Prosodie de la parole expressive : dimensionnalité d'énoncés méthodologiquement contrôlés authentiques et actés

Nicolas Audibert

INSTITUT POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque
|__|__|__|__|__|__|__|__|__|__|

T H E S E

pour obtenir le grade de

DOCTEUR DE L'Institut polytechnique de Grenoble

Spécialité : ***« Ingénierie de la Cognition, de la Création et des Apprentissages»***

préparée au laboratoire GIPSA-lab, Département Parole et Cognition
dans le cadre de l'Ecole Doctorale « Ingénierie pour le vivant : santé, cognition, environnement »

présentée et soutenue publiquement

par

Nicolas Audibert

le 2 décembre 2008

Prosodie de la parole expressive : dimensionnalité d'énoncés méthodologiquement contrôlés authentiques et actés

DIRECTEURS DE THESE
Véronique Aubergé et Jean-Luc Schwartz

JURY

M. Christian Boitet,	Président
M. Jean-François Bonastre,	Rapporteur
Mme. Jacqueline Vaissière,	Rapporteur
M. Jean-Luc Schwartz,	Directeur de thèse
Mme. Véronique Aubergé,	Co-encadrante
M. Nick Campbell,	Examinateur
M. Thierry Moudenc,	Examinateur
M. Laurent Besacier,	Examinateur

Remerciements

Les travaux de recherche présentés dans cette thèse doivent beaucoup aux personnes avec qui j'ai collaboré, ainsi qu'à celles qui m'ont permis de mener ces travaux dans des conditions favorables. Sans avoir la prétention de rédiger un texte digne de l'aide que ces personnes m'ont apporté, et tout en étant conscient de certainement en oublier certains, je tiens à les y associer à travers ces remerciements.

Tout d'abord, je remercie Véronique Aubergé qui, au-delà de l'encadrement scientifique de ma thèse, a toujours eu foi en mes capacités, et ce dès mon arrivée au Département Parole et Cognition de Gipsa-lab (qui s'appelait encore Institut de la Communication Parlée) pour mon stage de fin d'études d'ingénieur en informatique alors que je n'avais qu'une idée très vague de ce que pouvait être la recherche scientifique. Au cours des années que j'ai passé à Grenoble, elle n'a eu de cesse de m'encourager à poursuivre dans cette voie et à ne jamais céder au découragement ni à la facilité, y compris pendant les moments très difficiles qu'elle a pu traverser. Son immense culture scientifique, notamment dans les domaines des sciences du langage, de l'informatique, de la psychologie expérimentale mais aussi de la philosophie, et ses nombreuses idées à la fois novatrices et ambitieuses, ont été une source d'inspiration inépuisable pour mon travail et je suis fier d'avoir pu en mettre en musique quelques unes de ses idées dans cette thèse. Si sa préférence marquée pour les envolées lyriques autour d'un café au détriment des contigences pratiques m'a parfois angoissé, tout particulièrement les veilles de date limite de rendu d'articles, j'ai bien souvent réalisé plus tard la portée de ces discussions qui m'ont permis de progresser dans ma réflexion scientifique. De plus, c'est grâce à elle et à son rayonnement dans la communauté scientifique que j'ai pu nouer des contacts avec des acteurs majeurs de la recherche dans les domaines concernés par mes travaux de thèse et m'impliquer dans de nombreux projets collaboratifs. Merci Véro pour toutes ces idées que tu as bien voulu partager avec moi et pour ta bienveillance à mon égard avant, pendant et après ma thèse.

Merci à Jean-Luc Schwartz, directeur officiel de ma thèse, qui, s'il n'a pas été directement impliqué dans son encadrement scientifique, s'est plié de bonne grâce aux obligations administratatives liées à cette fonction quand bien même je m'y prenais tardivement, m'a régulièrement complimenté sur la qualité des travaux que nous menions, et a fait l'effort de relire une version préliminaire de ce manuscrit en apportant quelques commentaires constructifs.

J'adresse également mes remerciements aux chercheurs reconnus qui m'ont fait l'honneur d'être membres de mon jury de thèse malgré les nombreuses sollicitations dont ils sont l'objet, et dont les commentaires positifs sur mes travaux formulés avant et pendant ma soutenance m'ont beaucoup touché, à commencer par les deux rapporteurs : Jacqueline Vaissière qui a réalisé une relecture minutieuse et a commencé à m'informer au fur et à mesure des erreurs qu'elle relevait dès le lendemain de l'envoi de la version électronique, ce qui a facilité ma tâche de correction, et Jean-François Bonastre qui, malgré le choc que lui a causé la découverte du volume de cette thèse, inhabituel en informatique, ne m'en a pas tenu rigueur et l'a également relu et corrigé attentivement, allant ensuite jusqu'à me proposer d'intégrer son équipe pour un stage post-doctoral. Merci à Christian Boitet d'avoir accepté de présider ce jury et fait ensuite en sorte que le rapport de soutenance soit terminé pour la campagne de qualifications malgré les délais très restreints. Merci enfin aux examinateurs : Nick Campbell qui après avoir été à l'origine du projet *Expressive Speech Processing* mené en collaboration avec le centre de recherche japonais ATR qui a financé l'enregistrement du corpus Sound Teacher et permis l'évaluation du paramètre NAQ, a fait l'effort de lire un document rédigé en français et a effectué un long voyage depuis le Japon pour se rendre à ma soutenance, Thierry Moudenc sans qui la collaboration avec France Telecom R&D à travers le Contrat de Recherche Externalisé *Synthèse adaptative de la parole expressive* qui a abouti à une partie des résultats relatifs aux dimensions acoustiques de la prosodie expressive présentés dans cette thèse n'aurait pas été possible, et Laurent Besacier qui à la suite de ma thèse m'a intégré à son groupe de recherche en parole au LIG pour des discussions fructueuses.

Bien qu'il ne soit pas officiellement associé à cette thèse, Albert Rilliard mérite une place de choix dans ces remerciements. Sur place à Grenoble puis à distance depuis Orsay après sa mutation au LIMSI, il a assuré un rôle majeur dans l'encadrement de mon travail à travers son implication dans le développement du scénario Sound Teacher, le recueil et la numérisation du corpus, sa participation à l'élaboration des protocoles d'évaluation perceptive et des interfaces graphiques associées, ses talents de pédagogue pour me transmettre ses connaissances en statistiques, et enfin les corrections et les commentaires à la fois pertinents et drôles qu'il a apporté aux versions préliminaires de ce manuscrit. Merci à lui pour son humour qui m'a permis de relativiser les moments de stress, pour les nombreuses discussions autour de la gastronomie et de l'esprit de clocher, et pour sa gentillesse et tous les bons moments partagés au laboratoire et pendant les conférences. Toutes ces qualités compensent plus que largement ses moqueries répétées sur ma tenue vestimentaire (certes bien moins élégante que ses légendaires vestes de contrôleur), que je lui pardonne bien volontiers.

Merci aux personnes que je n'ai pas déjà cité, avec qui j'ai collaboré dans le cadre de différents projets, dans une ambiance à la fois studieuse et amicale : Alain Arnal et Christophe Savariaux du Gipsa-lab sur la mise en oeuvre du scénario Sound Teacher, la collecte et la numérisation et la synchronisation des signaux, Solange Rossato du Gipsa-lab et Parham Mokhtari d'ATR (Nara, Japon) sur l'évaluation de l'algorithme d'extraction du paramètre NAQ développé par ce dernier, Olivier Rosec et Damien Vincent de France Télécom R&D sur l'application à des expressions émotionnelles de leur méthode d'analyse/synthèse ARX-LF, Petri Laukka du département de psychologie de l'Université d'Uppsala (Suède) sur l'évaluation des déterminants perceptifs de la typicalité d'expressions émotionnelles, Janos Szabo, Mark Fék et Géza Németh du laboratoire SpeechLab de Budapest (Hongrie) sur l'application du scénario Sound Teacher pour la collecte d'expressions émotionnelles en hongrois, et Tamas Bohm du même laboratoire sur la tranformation de signaux de parole pour la synthèse de la glottalisation.

Merci à Nathalie Henrich et Solange Rossato qui ont contribué à cette thèse en me faisant bénéficier de leur expertise sur l'étude de la qualité de voix et l'acoustique de la parole. Merci aussi à Juliette Kahn qui a relu des passages de mon manuscrit et a contribué à la rendre plus lisible en traquant impitoyablement les phrases de plus de 10 lignes, à Isabelle Rousset qui m'a aidé à rectifier les incohérences bibliographiques, et enfin à Virginie Zampa et Auriane Faure qui m'ont permis de tenir les délais promis aux rapporteurs pour l'envoi de mon manuscrit en m'aidant à corriger les derniers problèmes de mise en page pendant que je m'escrimait sur la rédaction de la conclusion.

Un grand merci aux locuteurs piégés par Sound Teacher, qui ont eu l'amabilité de se prêter ensuite au protocole d'auto-évaluation. Merci aussi aux nombreux sujets, collègues, étudiants, amis ou amis d'amis qui ont participé aux évaluations perceptives malgré leur côté fastidieux, pour certains sans même avoir à les soudoyer avec des tablettes de chocolat, avec une mention particulière aux filles du laboratoire qui ont été le plus souvent sollicitées et qui se reconnaîtront.

Dans le cadre de ces projets et des nombreuses conférences auxquelles j'ai eu la chance de participer, j'ai été amené à solliciter fréquemment l'équipe administrative du laboratoire et je tiens aussi à remercier les personnes concernées. Merci aux secrétaires qui se sont succédées dans la partie Stendhal de l'ICP puis du DPC Gipsa-lab, et ont toujours su accéder à mes demandes avec le sourire en réalisant parfois des prouesses administratives : Dominique, Marie-Thé, Aline, Marie-Ange et Houria. Merci aussi à Jeanine, la femme de ménage du laboratoire, qui ne m'a pas tenu rigueur de mon approche très personnelle du rangement de mon bureau ni de mes promesses d'y mettre bon ordre qui n'ont pas toujours été suivies d'effet.

Parmi les personnes que je n'ai pas déjà cité nommément pour d'autres raisons mais qui devraient aussi figurer dans cette rubrique, merci aussi aux chercheurs, techniciens, doctorants ou stagiaires restés plus ou moins longtemps au laboratoire, qui par leur sourire, leurs blagues, leur gentillesse, et pour certaines et certains en partageant mon goût immodéré des calembours, ont rendu aussi agréable tout le temps que j'ai passé sur place. Je les citerai par ordre alphabétique pour ne pas avoir à les hiérarchiser, en m'excusant platement auprès de ceux que j'oublie : les deux Anne, Carole, Coriandre, Claire, Elisabetta, Emilie, Fanny, Hien, Jean-Pierre, Julien, Kévin, Krystyna, Lionel, Luca, Marc, Maria, Myrtille, Nadia, Nathalie, Omran, Sandra, Silvia, Taka, Tiphaine, Virginie et Yvon.

Merci aux amis de l'Université Stendhal que je n'ai pas déjà cité, avec qui j'ai eu plaisir à partager déjeuners et pauses café dans le patio : les gentils *geeks* du Département d'Informatique Pédagogique ou assimilés, Agnès, Aïcha, Claude, Lucy, Mathieu, Roseline et Thomas, et les copines du LIDILEM, Aurélie, Laurence et Myriam.

Merci aux groupes d'amis grenoblois avec qui j'ai pu me changer les idées dans une ambiance festive : les musiciens et assimilés Oxa, Greg, Claire, Mitch, Stouf, Aurel' et les jeunes EDF (j'aime beaucoup ce que vous faites, vous l'ai-je déjà dit ?) ; Pierre, Caro, Bibi et Cyprien ; Yannick, Guillaume, Milou, Kristel et le reste de la bande des « pépés ». Et merci aussi aux anciens rouennais et anciens belfortains que je n'ai pas pu voir aussi souvent, mais qui à chaque occasion de nous retrouver m'ont donné l'impression de les avoir quitté la veille.

Merci enfin à ma famille de m'avoir soutenu dans mes choix, et de ne pas trop m'avoir tenu rigueur des week-ends familiaux qui se métamorphosaient sans préavis en week-ends studieux. Un grand merci aussi à eux pour avoir organisé un magnifique pot de thèse qui m'a valu de nombreux compliments qui auraient dû leur revenir.

Table des matières

Conventions de typographie et de notation

Références bibliographiques

Les appels de références bibliographiques dans le texte sont notés sous la forme (Auteur, année) dans le cas d'un article ou ouvrage écrit par une seul auteur, sous la forme (Premier auteur & Second auteur, année) dans le cas de deux auteurs, et sous la forme (Premier auteur *et al.*, année) dans le cas d'un nombre d'auteurs supérieur à deux.

Dans certains cas cette notation est insuffisante pour identifier sans ambiguïté une référence. En conséquence, les références dont la notation est semblable sont distinguées à la fois dans la bibliographie et le texte par une lettre minuscule ajoutée à l'année, par exemple (Ekman, 1999a) et (Ekman, 1999b).

Par commodité, le nom du ou des auteurs sont fréquemment mentionné dans le texte, l'année étant indiquée entre parenthèses. Par exemple : « Suivant les propositions de Campbell (2000) ... ».

Traduction de termes anglais

Nous nous référons dans ce document à un nombre important de notions issues de la littérature scientifique anglophone, désignées pour certaines par un terme pour lequel il n'existe pas de traduction française consacrée par l'usage, ou pour lequel les traductions proposées divergent selon les auteurs. Dans la mesure du possible, nous nous sommes efforcé de proposer une traduction de ces termes, en mentionnant entre parenthèses et en italiques le terme anglais correspondant afin d'éviter toute ambigüité dans leur interprétation.

Les citations dans le texte d'extraits d'ouvrages ou articles publiés en langue anglaise ont été traduites par nos soins, la version originale de la citation étant indiquée en note de bas de page.

Nombres décimaux

La norme française de notation des nombres décimaux selon laquelle la partie entière est séparée de la partie décimale au moyen d'une virgule peut prêter à confusion, notamment dans le cadre de la présentation de résultats de tests statistiques d'hypothèses.

Prenons ainsi l'exemple de l'effet de l'émotion présentée sur leur reconnaissance par les sujets dans la première de nos évaluations perceptives, dont les résultats sont détaillés en section 8.4. La valeur prise par la fonction F après application du terme correctif de Huynh-Feldt (Huynh & Feldt, 1976) devrait selon cette convention être notée F(9,5, 236,4) = 23,8, d'où une possible confusion liée aux deux usages distincts de la virgule.

Dans un souci de lisibilité, nous adoptons donc la notation anglo-saxonne des nombres décimaux et séparerons donc les décimales de la partie entière au moyen d'un point plutôt que d'une virgule. Ainsi dans l'exemple ci-dessus devient F(9.5, 236.4) = 23.8.

Afin d'assurer la cohérence du document, tous les nombres décimaux sont notés selon cette convention anglo-saxonne.

Résumés des chapitres

Afin de faciliter la lecture de ce document, nous proposons en début de chapitre un résumé des différents points abordés, pour chacun des chapitres à l'exception de l'introduction étendue (section 1) et des sections finales dont la conclusion proprement dite (sections 1, 1 et 1). Ce résumé est placé en tête de chapitre sous le titre, et est encadré selon le format suivant :

Ceci est un résumé de chapitre.

1. Affects et cognition : enjeux théoriques et technologiques

1.1. Les émotions dans la communication

La communication des émotions est considérée comme un fait majeur du discours (Fonagy & Magdics, 1963 ; Rossi *et al.*, 1981 ; Ohala, 1996 ; Hirst & di Cristo, 1999). Modéliser les performances communicatives des humains durant leurs interactions verbales (Bottineau, 2007), rendre un robot physique ou virtuel à même de simuler ces performances (Cassel, 2000), comprendre les mécanismes cognitifs sous-jacents (Scherer *et al.*, 2001) sont des enjeux de recherches centraux dans ce domaine. Chacun de ces objectifs scientifiques nous amène, dans le courant d'un remarquable regain d'intérêt pour ces questions, à mettre avec cette thèse les émotions au centre des processus de la communication. Au-delà du cloisonnement classique entre cognition, affect et conation, des hypothèses fortes plus récentes en psychologie cognitive (voir Scherer (2000) pour un état de l'art) et neuropsychologie (Damasio, 1994) donnent aux émotions un statut central. A l'heure où les « sciences affectives » s'imposent comme un thème de recherche spécifique (un ouvrage de référence, le *Handbook on Affective Sciences*, leur a d'ailleurs été consacré en 2003, et un cycle de conférences internationales de grande ampleur dont le succès ne s'est pas démenti, *Affective Computing and Intelligent Interaction*, a vu le jour en 2005), les liens entre affects et cognition se retrouvent à la base des théories de l'émotion.

La perception des affects conditionne en amont le schéma attentionnel de l'humain (Paschler, 1997) ainsi que les mécanismes cognitifs de la prise de décisions. Frijda (1987) utilise la notion de préparation à l'action pour décrire les motivations liées à l'émotion. Pour Damasio (1994, 2003), émotion et raison fonctionnent comme les deux composantes d'une seule entité psychologique émergente. Il affirme ainsi que la capacité d'exprimer et ressentir des émotions serait indispensable à la mise en œuvre des comportements rationnels. Ainsi l'émotion serait à la base de l'évaluation des processus cognitifs qui fondent ce qui a classiquement été appelé la « raison » : pour déclencher les traitements cognitifs (et donc les actions) appropriés à la situation, le sujet doit se trouver dans un état émotionnel adéquat. Cette hypothèse appliquée aux actes de paroles (Searle, 1969) est déterminante. Non seulement la production d'énoncés langagiers permettrait de véhiculer les émotions ressenties par le locuteur, mais celles-ci seraient à l'origine même de la production d'un acte de parole plutôt qu'un autre, voire d'aucun acte de parole.

D'un point de vue applicatif, la production et la perception des indices émotionnels par un robot parlant en face-à-face avec un humain, peuvent sembler secondaires face aux casse-têtes théorique et technologique rencontrés. Pourtant, lorsque les systèmes artificiels quittent le cadre confortable des situations courantes du dialogue personne-

machine et s'inscrivent dans des situations d'interaction sensibles aux prises de décisions, les émotions n'ajoutent pas seulement du « naturel » mais ont un impact direct sur l'efficacité même de la communication. Par conséquent, il paraît légitime de supposer que la composante affective de la communication sera d'autant plus nécessaire que les systèmes artificiels s'inscriront dans des écologies proches du réel. Ainsi les concepteurs de tels systèmes devront s'interroger sur la façon d'éviter de perturber le mécanisme psychologique de l'attention, fortement lié aux émotions. En effet, Pashler (1997) souligne que l'accès à l'information est facilité et motivé par les variations émotionnelles. De plus, ils devront prendre en compte une autre question toute aussi centrale, celle de la réaction d'un utilisateur humain lorsque son mécanisme d'empathie (dont les bases neurales commencent à être décrites (Harris, 2003)) fait face à un artefact communicant dont il ne peut éviter de traiter les indices émotionnels, même si ceux-ci sont de simples biais non contrôlés par un système supposé ne produire que des énoncés déclaratifs. En effet, comme le souligne Daneš (1994), qui s'inscrit dans la tradition pragmatique de l'école de Prague, ne pas exprimer d'émotion ou d'attitude est en soi une attitude. Les énoncés présentés donnent donc, volontairement ou non, des informations sur la personnalité du sujet parlant humain ou virtuel, ce qui tendrait à réarticuler les enjeux de recherche autour du concept de «pertinence» du clone parlant, et plus particulièrement de la gestion de l'expressivité de sa parole. Cette dimension expressive est cependant, dans la plupart des systèmes de dialogue existants, considérée comme relevant des problèmes à résoudre lorsque la complexité strictement linguistique sera mieux maîtrisée.

Les avatars et autres *chatterbots*, en interaction communicative avec l'humain, commencent à être doués de croyances, de désirs et d'intentions (*Belief, Desire & Intention* (Rao & Georgeff, 1995)). Ainsi sont nés les *Affective Agents* au Medialab-MIT (Picard, 1997), les *Embodied Conversational Interface Agents* (Cassel, 2000) ou encore les *Believable Social and Emotional Agents* du projet Oz (Bryan Loyall, 1996). Que doit-on modéliser pour simuler les performances du vendeur qui « sait » vendre, avec des compétences langagières pourtant identiques au vendeur moins efficace ? Tant que la modalité est celle de l'écrit, les affects des agents prennent forme dans leurs stratégies interactionnelles et leur expressivité langagière. Mais ces technologies visent à terme la communication face-à-face : les clones parlants seront incarnés dans un corps et un visage. L'expression verbale ou non-verbale est en premier lieu concernée, avec toute la complexité de la multi-modalité faciale, gestuelle et parlée de ces expressions (voir par exemple Scherer & Ellgring (2007)).

Comme nous l'avons évoqué ci-dessus, et malgré les écueils à la fois techniques et théoriques restant à résoudre dans le domaine des Agents Conversationnels Animés,

faire passer la gestion des émotions exprimés par le clone virtuel à l'arrière plan serait sans doute contre productif. Ainsi, si l'on s'intéresse à la seule modalité de la parole transmise par le canal acoustique, cette problématique est d'ores et déjà centrale. En effet les récents progrès des technologies de la parole, et plus particulièrement la grande naturalité atteinte par la synthèse par sélection, ont redéfini les enjeux majeurs de la communication. L'intelligibilité du contenu n'est plus le seul vecteur d'information : l'opinion, les attitudes, les intentions, les émotions du sujet humain ou virtuel sont également des informations essentielles interprétées par l'auditeur, qui justifient les efforts de recherche croissants dans cette direction.

1.2. La prosodie expressive : questions de recherche

Les émotions, et de façon plus générale les affects, constituent un objet complexe difficile à définir. La question de leur interrelation avec la parole, et plus spécifiquement avec la prosodie (l'objet qui nous intéresse plus particulièrement dans cette thèse) peut sembler inextricable. En effet, les dimensions acoustiques de la prosodie qui permettent de véhiculer des informations d'ordre linguistique donnent également des informations sur le locuteur lui-même (genre, âge, etc.), ainsi que sur l'état affectif dans lequel il se trouve au moment auquel il parle, et sur les intentions qu'il exprime consciemment ou non. S'il ne fait aucun doute que le cerveau humain est capable de traiter l'essentiel des nombreuses informations transmises par la prosodie et la voix, aucune des modélisations proposées à l'heure actuelle n'est en mesure de représenter de façon fine le décodage des différents niveaux d'information véhiculés par la prosodie.

Outre l'intérêt que revêt l'étude en tant que telle des mécanismes cognitifs mis en œuvre dans ce décodage, ceci a également des implications d'un point de vue applicatif. En effet, il serait envisageable de construire un système de reconnaissance automatique des affects exprimés dans la parole en ne considérant que les caractéristiques acoustiques les plus saillantes des affects visés, et sans une connaissance fine de la morphologie prosodique de ces affects. Pourtant, il ne semble pas possible de se passer de telles connaissances si l'on poursuit l'objectif de construire un système de synthèse de la parole expressive. Considérer que les indices les plus saillants, potentiellement suffisants en reconnaissance, le sont également en synthèse pour générer des formes expressives écologiquement valides reviendrait à considérer que toute expression d'affect dans la parole pourrait être réduite à une forme normalisée de variation des paramètres acoustiques. Cette variation serait alors paramétrisée par un ensemble de valeurs globales qu'il suffirait de rééchelonner pour obtenir des expressions d'affects bien formées, de valeurs différentes. Or, un tel postulat semble hasardeux. Pour l'illustrer par

une métaphore triviale, ceci serait comparable au fait de considérer que souffler dans une souris permettrait d'obtenir un éléphant.

L'étude des expressions prosodiques des affects, et plus spécifiquement celles des expressions d'émotions, soulève de nombreuses questions, parmi lesquelles nous soulignons celles qui nous semblent cruciales.

- L'information affective véhiculée par la prosodie est-elle portée par des indices gradients (c'est-à-dire par des valeurs globales de paramétrisation d'une forme de base) ou par des contours prosodiques (c'est-à-dire par la dynamique de ses variations) ? Cette question implique les trois conséquences suivantes :
  (1) Est-il possible d'assurer la complétude de la modélisation dans l'optique de la synthèse ?
  (2) Comment intégrer dans les mêmes dimensions les prosodies émotionnelle et linguistique ?
  (3) Comment mesurer ces contours prosodiques alors que la mesure des phénomènes regroupés sous le terme de « qualité de voix » reste un problème non résolu ?
- La prosodie émotionnelle est-elle véhiculée prioritairement par certaines dimensions prosodiques, différentes de celles empruntées par la prosodie linguistique ? Cette question a principalement des conséquences sur la synthèse de la parole expressive. En effet, l'application serait plus résistante si ces dimensions pouvaient contrôlées indépendamment.
- Dans quelles mesures les expressions d'affects contrôlées volontairement sont-elles équivalentes à celles contrôlées involontairement ? Si elles ne le sont pas, selon quels critères morphologiques sont-elles discriminées ? Même si elles sont négligées dans la grande majorité des études du domaine, ces questions trouvent un certain écho depuis quelques années. Ce regain d'intérêt est dû à une remise en question plus large de l'idée, selon laquelle la parole émotionnelle produite par des acteurs ou simplement simulée par des locuteurs non entraînés fournirait une référence valide pour modéliser des phénomènes propres à l'expression spontanée d'émotions.
- Dans quelle mesure les expressions dans la parole de différents types d'affects sont-elles régies par des universaux ? Cette dernière question peut paraître plus simple de prime abord. Cependant, ces différents types d'affects ne sont pas toujours clairement distingués, ni catégorisés de manière similaire selon les langues et les cultures. De plus, étant donné que leurs expressions sont instanciées dans des langues différentes et que des normes socio-culturelles peuvent fortement influer sur

la tendance à inhiber ou masquer certaines émotions en fonction de la situation dans laquelle celle-ci apparaît (Scherer *et al.*, 2001), l'évaluation de l'existence ou non de tels universaux soulève d'importants défis méthodologiques.

1.3. Présentation des travaux réalisés

En section 2, après un rapide aperçu de l'historique de l'étude des émotions, nous passons en revue les principales théories des émotions proposées en psychologie. Nous développons plus particulièrement les postulats établis en neuropsychologie à la suite de Damasio, ainsi que les théories cognitives de l'*appraisal* et en particulier la théorie de processus composants de Scherer.

En section 3, nous présentons un bref état de l'art des principales technologies de la parole expressive que sont la synthèse expressive et la classification automatique des affects exprimés à partir d'un flux de parole.

Nous abordons en section 4 la problématique de la définition de la prosodie et adoptons une approche fonctionnaliste. Nous présentons de façon détaillée la fonction expressive de la prosodie, les propositions issues de la phonostylistique et leur lien avec l'étude des expressions vocales issue de la psychologie, puis l'hypothèse d'une prosodie en contours superpositionnels dirigée par ses fonctions que propose Aubergé.

En section 5, nous dressons un état de l'art des différentes méthodes de recueil de corpus de parole émotionnelle, en soulignant pour chacune leurs avantages et inconvénients. Nous présentons la plateforme logicielle E-Wiz, que nous avons développée afin d'induire et recueillir des expressions authentiques d'affects selon un scénario de Magicien d'Oz. Nous détaillons les deux scénarios successifs mis en place à l'aide de cette plateforme, et plus particulièrement le second, Sound Teacher. Ce scénario imite un logiciel d'aide à l'apprentissage des langues étrangères, dans le but de collecter des expressions émotionnelles produites sur des énoncés pour la plupart monosyllabiques, hors de toute interaction sociale. Puis nous présentons l'extension de ce scénario, proposée pour les sujets acteurs, afin de leur faire rejouer sur les mêmes énoncés les états émotionnels ressentis et exprimés au cour de l'expérience. Nous discutons enfin les différentes approches de l'annotation des affects exprimés dans les corpus spontanés, et proposons, en nous appuyant sur la littérature relative à la mémoire autobiographique, un protocole d'auto-annotation des affects ressentis et exprimés par les sujets de l'expérience.

En section 6, nous présentons la mise en œuvre du scénario Sound Teacher pour l'enregistrement d'un corpus multimodal de 17 locuteurs (10 femmes, 7 hommes) dont 7 acteurs, corpus sur lequel nous avons fondé les analyses présentées dans la suite de cette

thèse. Nous détaillons les signaux enregistrés ainsi que leur synchronisation et leur prétraitement.

Nous abordons en section 7 la notion de qualité de voix et commençons par présenter sa description phonétique dans la littérature avant de passer en revue les différentes approches proposées pour sa mesure et son étude. Nous présentons également les résultats d'une expérience visant à évaluer, sur les productions spontanées de deux locuteurs masculins, la validité pour des mesures fines de type suivi de contour d'un algorithme d'extraction du paramètre NAQ opérant par filtrage inverse et auparavant validé sur un corpus expressif spontané de très grande taille, en confrontant ces mesures à des mesures électroglottographiques.

En section 8, après avoir présenté les principaux outils d'analyse statistique auxquels nous avons recours dans nos analyses de résultats perceptifs, nous proposons une analyse perceptive et acoustique d'énoncés monosyllabiques sélectionnés à partir des productions de l'un des locuteurs masculins, dans laquelle nous nous efforçons d'évaluer séparément les différentes dimensions prosodiques. Nous étudions tout d'abord les formes de contours d'expressions actées et spontanées produites par ce locuteur, puis présentons les résultats d'une évaluation perceptive d'énoncés actés en conditions audio et audiovisuelle.

En section 9, ous présentons les résultats de deux études d'évaluation perceptive fondées sur une projection séparée de dimensions prosodiques (à l'aide de TD-PSOLA et de l'algorithme ARX-LF) d'énoncés actés sélectionnés parmi les productions du même locuteur. Enfin, nous détaillons les résultats d'une étude de dévoilement progressif d'énoncés actés.

Nous nous intéressons en section 10 aux résultats d'une étude perceptive portant sur les déterminants de la typicalité d'expressions émotionnelles actée et spontanées produites par 6 des 7 acteurs enregistrés. Ensuite, nous présentons les résultats d'une étude de discrimination d'expressions actées et spontanées produites par les mêmes locuteurs.

En section 11, nous revenons sur quelques limites de l'étude que nous avons réalisée, avant de conclure en section 12 en récapitulant les principaux résultats que nous obtenons et d'ouvrir en section 13 sur quelques perspectives.

2. Les théories des émotions

Dans cette section, après avoir introduit l'intérêt que nous accordons à la théorisation des émotions, nous dressons un bref historique de l'étude des émotions par la philosophie, et notamment les points de vue cartésien et spinoziste. Nous passons ensuite en revue les principales approches de l'émotion proposées par la psychologie : l'approche darwinienne qui postule l'existence d'un nombre limité d'émotions dites « de base » qui seraient universellement partagées en tant que produits de l'évolution, l'approche jamesienne qui accorde à l'activation physiologique un rôle central dans le déclenchement des processus émotionnels, le constructivisme social qui considère les émotions comme des constructions sociales fonction du contexte dans lesquelles elle apparaissent, et la perspective de l'*appraisal* selon laquelle l'émotion est déterminée par l'évaluation cognitive de la situation.

Nous développons plus avant la théorie des processus composants de Scherer, dans laquelle la survenue d'une émotion est considérée comme relevant d'une séquence d'adaptations à l'environnement de l'organisme au fur et à mesure des processus plus ou moins complexes d'évaluation cognitive des stimuli. Nous revenons également sur les théories avancées par Damasio à partir de l'étude de cas cliniques en neuropsychologie, qui mettent en avant la nécessité d'émotions bien formées pour une prise de décision adaptée, et donnent un éclairage sur les corrélats neuraux des émotions ressenties vs. simulées.

2.1. *Pourquoi théoriser les émotions ?*

Se poser le problème de l'expression des émotions à travers les événements de la parole, c'est forcément situer d'abord l'émotion dans son contexte théorique général. Il ne paraît pas possible de traiter le problème de la parole émotionnelle sans se placer dans un cadre théorique, car les technologies posent directement les problèmes résolus différemment, selon les courants théoriques, de la catégorisation, du niveau d'encodage, de l'universalité et de la pertinence indicielle des émotions. Construire un système automatique de reconnaissance ou de génération des expressions des émotions dépend fortement de la définition que l'on donne de l'émotion. Quels liens existent entre humeurs (par exemple le stress), émotions, attitudes, sentiments, passions ? Une émotion est-elle un état du cerveau, un état physiologique, psychologique, un processus cognitif, pourquoi et par quel processus est-elle exprimée ? Les émotions sont-elles catégorielles, finies, reliées aux étiquettes communes des lexiques des langues, ou bien réparties dans un espace continu multidimensionnel ? Comment les expressions sont-elles situées dans le processus émotionnel, et quels types de contrôle cognitif produisent ces expressions ? Peut-on utiliser comme référence les marqueurs et les signaux physiologiques, existe-t-il des aires neurales des émotions ?

2.2. *Bref historique*

Ces débats, avant de devenir scientifiques, que ce soit dans les domaines de la biologie, de la psychologie, de la neuropsychologie, de la linguistique ou encore de la sociologie, ont tout d'abord été ceux de la philosophie, le terme de *passion* étant fréquemment employé pour désigner les affects.

On peut trouver dès l'antiquité la trace d'un intérêt pour l'étude des émotions. Ainsi Platon considère-t-il les émotions comme des passions néfastes dont il faut tenter de se détourner car elles empêchent la raison de fonctionner. Aristote, quant à lui, réfute cette opposition et prête aux émotions une fonction informative ainsi qu'une tendance à l'action. Il est par ailleurs le premier à établir une catégorisation des émotions.

Par ailleurs, Damasio (1994) revient sur les conceptions cartésienne et spinozienne des émotions, en s'attachant principalement à l'articulation entre corps et esprit chez ces deux philosophes. Selon lui, à la lumière des récentes découvertes de la neurophysiologie, c'est la conception de Spinoza qu'il faudrait retenir, et non celle de Descartes.

Au 17e siècle, René Descartes développe la notion de dualité corps / esprit en considérant ces entités comme indépendante, la faculté de raisonner pouvant selon lui s'exercer séparément du corps. Dans le cadre de cette théorie, il rattache les émotions à l'activité

de l'esprit (Descartes, 1649) : c'est d'ailleurs pour lui ce qui fait des émotions une caractéristique spécifique à l'humain. Toutefois Descartes ne considère-t-il pas que les manifestations physiologiques sont totalement étrangères aux émotions : selon lui l'âme est en effet influencée par ces manifestations physiologiques, le lien se faisant par la glande pinéale (petite glande située en dessous du cerveau dont la fonction est en réalité de sécréter diverses hormones). De plus il reconnaît aux émotions une fonction informative, car elles nous renseignent sur les éléments auxquels notre esprit attache de l'importance. Enfin Descartes établit une liste de six états émotionnels fondamentaux, à savoir l'admiration, l'amour, la haine, le désir, la joie et la tristesse.

Spinoza développe au contraire une théorie moniste (c'est-à-dire dans laquelle corps et pensée sont vus comme une seule et même entité). Dans la troisième partie de son *Ethique* (Spinoza, 1677), intitulée *Des affects*, il tente de comprendre les passions et leur prête une fonction de régulation de la capacité d'agir. Ainsi, selon Spinoza, il existerait deux types de passions, bonnes ou mauvaises selon qu'elles augmentent ou diminuent notre capacité d'agir : les passions joyeuses et les passions tristes, joie et tristesse pouvant se décliner respectivement en chatouillement ou en hilarité et en douleur ou en mélancolie. Spinoza ne considère donc que deux émotions de base pouvant se décliner sous plusieurs formes suivant leur champ d'application. Il affirme ainsi, comme dernière proposition concernant la joie (qui au contraire de la tristesse concerne « l'âme agissante »), qu'« *Entre toutes les passions qui se rapportent à l'âme, en tant qu'elle agit, il n'en est aucune qui ne se rapporte à la joie ou au désir* ».

C'est à la fin du 19e siècle que l'étude des émotions est entrée dans le champ scientifique, avec l'intérêt qui lui a été porté par les psychologues. Cornelius (2000) propose de classer en quatre classes les différentes théories des émotions proposées jusqu'alors. Outre la perspective cognitive introduite par Arnold (1960), et particulièrement prégnante à l'heure actuelle, trois autres courants majeurs influencent encore les débats :

- La perspective évolutionniste ou darwinienne, étayée principalement par l'étude des expressions faciales.
- La perspective physiologique ou jamesienne, qui met en avant le rôle de l'activation physiologique dans le déclenchement des processus émotionnels.
- La perspective du constructivisme social qui considère les émotions comme le produit de constructions sociales.

Quoique ces différents courants théoriques ne s'excluent pas systématiquement, cette catégorisation permet de fournir une grille de lecture pour les théories de l'émotion. Nous présenterons ici plus en détails l'approche cognitive des émotions, en mettant

l'accent sur les approches de Scherer et de Damasio qui donnent des points d'ancrage aux propositions théoriques sur lesquelles cette étude se base plus particulièrement.

2.3. La perspective darwinienne

2.3.1 L'universalité des émotions

L'idée dominante dans le courant darwinien ou évolutionniste des théories des émotions est que les émotions constituent des phénomènes évolués qui sont apparus en adaptation à l'environnement car ils remplissent des fonctions de survie de l'espèce. La conséquence principale de ce postulat serait donc que les mêmes émotions devraient se retrouver chez l'ensemble des humains, voire chez les mammifères proches de l'humain en raison de leur passé commun. De nombreuses études ont étudié l'universalité des expressions faciales des émotions, dans le but d'étayer cette hypothèse.

Le premier auteur a avoir développé cette théorie est Charles Darwin lui-même, à travers son célèbre ouvrage *The expression of emotion in man and animal* (Darwin, 1872). Darwin a ainsi tenté de démontrer l'universalité des émotions, à partir de l'étude de leurs expressions faciales chez des peuples de culture variée effectuée par plusieurs observateurs basés en Afrique, en Amérique, en Australie, à Bornéo, en Chine, en Inde, en Malaisie et en Nouvelle Zélande. Toutefois, comme le souligne Paul Ekman (1999a), la méthodologie employée par Darwin est critiquable au regard des méthodes employées à l'heure actuelle en psychologie sur trois points fondamentaux :

- Le nombre de sujets interrogés dans chaque pays étudié était insuffisant pour pouvoir constituer une preuve statistique.
- Darwin s'est fondé sur les réponses données par les observateurs anglais sur leurs impressions par rapport à l'expression des émotions dans la population locale plutôt que de faire appel à des sujets natifs comme cela est la règle.
- Les réponses attendues aux questions posées étaient souvent implicitement données par la formulation même des questions, comme par exemple « La surprise est-elle communiquée par les yeux et la bouche grands ouverts, et par les sourcils levés ? ».

Bien que les résultats expérimentaux de Darwin ne soient pas directement exploitables, ses travaux ont eu une influence considérable sur les études ultérieures des émotions. Nombreuses sont en effet les théories qui, sans toujours se référer explicitement aux travaux de Darwin, considèrent les émotions comme des produits de l'évolution par sélection naturelle, et étudient les émotions à travers leurs fonctions.

Des expériences similaires à celle de Darwin ont par la suite été reproduites, notamment sous l'impulsion de Paul Ekman (Ekman & Friesen, 1978 ; Ekman, 1989, 1999a), de Carroll Izard (2000) ou encore de Sylvan Tomkins (1980), dans le but de montrer que les expressions faciales sont universellement partagées et surtout liées aux mêmes émotions. Dans la plupart des cas ces études sont parvenues à montrer qu'une majorité de sujets, bien que de cultures différentes, reconnaît bien la même émotion pour une expression donnée. Afin d'éliminer un éventuel biais dû à l'apprentissage par les médias de ces expressions, certaines expériences ont été effectuées dans des communautés n'ayant pas accès aux médias, notamment en Nouvelle-Guinée, avec des résultats similaires. La figure 2 présente des exemples de photographies utilisées dans le cadre d'études interculturelles sur les expressions faciales menées par Ekman.

Quoiqu'un certain nombre de critiques aient été formulées à l'encontre de ces études, Ekman (1999a) réfute l'ensemble de ces critiques. En particulier il affirme que pour pouvoir démontrer que les émotions ne sont pas universelles, il faudrait qu'une expression que la majorité des sujets d'une certaine culture associe à une émotion *X* soit associée majoritairement à une émotion *Y* par des sujets issus d'une autre culture, ce qui n'a jamais été le cas. Cependant Ekman reconnaît que, si les émotions sont universelles, leurs règles de déclenchement ou d'inhibition sont susceptibles de varier en fonction de facteurs culturels (Ekman, 1989, 1999a).

Figure 1 : Exemples de photographies d'expressions émotionnelles utilisées dans le cadre d'études interculturelles. Reproduit de (Ekman, 1999a)

Par ailleurs Ekman et Friesen (1978) ont proposé, à partir de l'étude des différents mouvements musculaires susceptibles de modifier les expressions faciales, un système

de codage de ces expressions en fonction de 46 paramètres musculaires appelé FACS (*Facial Action Coding System).*

2.3.2 Les émotions de base

L'universalité des émotions postulée par le courant évolutionniste va de pair avec la notion d'émotions de base ou émotions fondamentales, discrètes et en nombre restreint. Ekman (1999b), à partir de l'étude des expressions faciales mais aussi d'observations physiologiques, propose une définition générale de l'émotion de base. Il donne ainsi un ensemble de critères pour identifier ce qu'est une émotion de base dans l'ensemble des affects. Ainsi pour qu'une manifestation affective soit considérée comme une émotion de base elle doit notamment :

- Se déclencher rapidement et être brève.
- Se retrouver chez les primates non-humains.
- Etre universellement identifiable (par exemple à partir des expressions faciales).
- Etre associée à une physiologie propre (réponse nerveuse, rythme cardiaque, etc.).
- Se déclencher automatiquement, c'est-à-dire être contrôlée involontairement et non par un processus conscient.

Il identifie ainsi six émotions de base (Ekman, 1989) : joie, tristesse, peur, dégoût, colère et surprise. Ces six émotions de base, les plus fréquemment citées dans la littérature, sont appelées *Big Six* par Cornelius (2000). Cette catégorisation ne fait toutefois pas l'unanimité parmi les tenants de la perspective évolutionniste. Ainsi d'autres auteurs dénombrent entre 8 et 10 émotions de base, et Ekman lui-même ajoute parfois le contentement aux *Big Six* (Cornelius, 2000).

2.4. La perspective jamesienne

Les théories énoncées par James (1884) et Lange (1885) puis par Cannon (1927) accordent à l'activation physiologique un rôle central dans le déclenchement des processus émotionnels. Cette approche physiologique, qui a alimenté de nombreux débats en psychologie, s'inscrit également dans un cadre évolutionniste puisque l'émotion est également considérée comme une réponse adaptative à l'environnement. La différence majeure avec l'approche darwinienne des émotions est qu'ici les émotions elles-mêmes sont théorisées plutôt que leurs expressions.

2.4.1 Les théories de James et Lange

James (1884) associe les différents états émotionnels pouvant être ressentis à l'ensemble des changements corporels liés à la présentation d'un stimulus. Suivant cette définition, l'émotion se limite donc à la perception de ces changements corporels, qui constituent donc également la condition *sine qua non* pour ressentir une émotion. Ainsi James affirme-t-il que « *on se sent triste parce qu'on pleure, en colère parce qu'on frappe, apeuré parce qu'on tremble* ». D'après James les types de changements physiologiques impliqués dans l'émotion seraient donc extrêmement variés, pouvant inclure un contrôle moteur de haut niveau comme l'action de frapper. Il va d'ailleurs jusqu'à affirmer qu'on peut « *avoir peur parce qu'on court* ».

Les corrélats de cette théorie sont tout d'abord qu'on ne peut ressentir d'émotions en l'absence de perception, mais aussi que les réponses émotionnelles seraient déterminées directement par le système nerveux autonome et non par des aires neurales spécifiques. D'autre part, dans une telle optique, les changements corporels associés à différentes émotions sont nécessairement distincts.

La théorie de Lange (1885) quant à elle, se rapproche à de nombreux points de vue de celle de James. En effet pour Lange c'est également la perception des changements corporels périphériques qui constitue la nature même de l'émotion. De même que James, il propose d'ailleurs la séquence suivante pour modéliser le processus émotionnel :

- Apparition du stimulus
- Réponse physiologique spécifique
- Perception des changements corporels
- Emotion

La seule différence majeure que l'on peut relever entre les théories de James et Lange est leur désaccord quant au traitement cognitif des émotions. En effet Lange considère qu'un centre neural de traitement des émotions existe bien, et qu'il s'agirait du centre « vasomoteur » qui contrôle l'activité vasculaire. Les théories proposées par ces deux auteurs étant par ailleurs très proches, on les trouve fréquemment réunies dans la littérature sous le nom de théorie James-Lange.

2.4.2 La théorie centrale des émotions de Cannon

La théorie centrale des émotions proposée par Cannon (1927) s'inscrit également dans la perspective physiologique ou jamesienne en accordant à l'activation physiologique une place centrale dans le processus émotionnel. Cette théorie se présente toutefois comme une critique de la théorie de James, qu'elle prend à contre-pied : selon Cannon l'évaluation cognitive est un préalable nécessaire aux réactions physiologiques, qui ne sont donc que la conséquence des émotions.

Cannon énonce deux arguments principaux pour réfuter la théorie de James :

- Tout d'abord l'induction artificielle de changements physiologiques (changements viscéraux dans la terminologie employée par Cannon) ne permet pas seul de provoquer un état émotionnel défini
- Par ailleurs les mêmes changements viscéraux peuvent induire indifféremment des états émotionnels ou non-émotionnels.

2.5. Le constructivisme social

Averill (1980) marque le début de la prise en compte des interactions sociales et culturelles dans les émotions communiquées. Selon lui, les émotions seraient le produit de « constructions sociales », et dépendraient essentiellement du contexte social dans lequel elles apparaissent. Elles permettent d'établir une signification et une cohérence en fonction des règles sociales apprises.

Averill n'a pas l'apanage de la prise en compte des aspects socioculturels dans l'étude des émotions, en effet même des auteurs darwiniens convaincus comme Paul Ekman (1999a) leur accordent un rôle de régulation des expressions. Toutefois le courant constructiviste postule un rôle beaucoup plus central de ces aspects dans les processus émotionnels : selon eux l'émotion ne peut se réduire à une réponse adaptative et elle doit plutôt être vue comme une construction sociale, pour la compréhension de laquelle l'analyse doit nécessairement se faire au niveau de la société. Ainsi, pour les constructivistes, les émotions doivent leur sens et leur cohérence à des lois sociales apprises.

Le courant constructiviste s'oppose donc fortement à la notion darwinienne d'émotions de base primitives et universelles, et suggère au contraire des émotions variables selon les cultures et les codes sociaux partagés dans les diverses communautés. Selon Averill (1980), la colère, souvent considérée comme la plus « primitive » ou « animale » des émotions, constitue d'ailleurs un jugement moral très élaboré qu'on ne peut considérer comme une réponse automatique de l'organisme.

Ainsi une émotion serait un ensemble de réponses sociales, constituées de normes et d'attentes au sein d'un groupe qui définissent précisément qui doit éprouver quelle émotion à quel moment, et dont la signification serait seulement symbolisée par l'étiquette émotionnelle que le sujet lui-même et les autres membres du groupe appliquent à son comportement. Les réponses émotionnelles, y compris en termes de comportement, seraient donc entièrement régies par des *scripts* émotionnels fonctions de ces règles sociales.

2.6. La perspective cognitive

Dans les théories cognitives de l'évaluation, l'émotion est envisagée comme résultante de la perception et de l'évaluation d'un stimulus externe, et elle se traduit par des comportements expressifs. L'hypothèse de base des ces théories est que cognition et émotion sont indissociables, la nature même de l'émotion étant déterminée par l'évaluation cognitive des stimuli perçus. Plus spécifiquement, une notion centrale dans les théories cognitives est celle de l'*appraisal*, c'est-à-dire le processus par lequel les évènements de l'environnement sont jugés comme bons ou néfastes pour l'individu à partir de la perception qu'il en a. En cela, ces théories sont donc incompatibles avec des conceptions dualistes telles que celles de Platon ou de Descartes. Notons toutefois que, pour la plupart, ces théories, fondées sur des principes évolutionnistes et prenant en compte les phénomènes d'activation physiologique, ne se démarquent pas totalement des autres approches théoriques des émotions.

2.6.1 Les théories d'Arnold et de Frijda

Arnold (1960) est le premier auteur à avoir développé la notion de processus d'évaluation du caractère plaisant ou déplaisant d'une perception, ou *appraisal*. Selon elle ce processus est par nature inconscient et n'est pas contrôlé volontairement. Elle le définit ainsi comme direct, immédiat, non-réflexif, non-intellectuel et automatique.

Arnold reproche à James de ne pas suffisamment préciser dans sa théorie comment la perception du stimulus déclencheur de l'émotion est réalisée, lorsque ce dernier affirme que les changements corporels interviennent immédiatement après la perception de ce stimulus. Selon elle, le chaînon manquant est justement le processus d'*appraisal*, sans lequel il ne saurait y avoir d'émotion. Ainsi pour Arnold les émotions se différencient-elles chez un individu donné par divers schémas d'*appraisal*.

Ces schémas d'*appraisal* ne sont pas considérés comme immuables et universels, puisqu'ils seraient influencés par des facteurs liés au contexte social et environnemental de l'individu, les schémas étant définis comme dépendant à la fois des caractéristiques

propres de l'individu, de ses expériences passées, de son tempérament, de sa personnalité, de son état physiologique et enfin de la situation.

Arnold propose un modèle séquentiel pour décrire le processus émotionnel, dans lequel les changements physiologiques associés à l'émotion sont précédés d'une « tendance à l'action » positive ou négative (dirigée « vers » ou « contre ») en fonction du processus d'*appraisal*.

Frijda (1986) développe le concept de tendance à l'action introduit par Arnold. Selon lui, l'émotion est une évaluation de l'environnement, qui débouche sur une préparation à l'action. La préparation à l'action est ici définie comme le préalable à une action visant à établir, maintenir ou mettre fin à une relation de l'individu avec des aspects particuliers de l'environnement. Frijda établit une taxonomie de 18 tendances à l'action, parmi lesquelles figure l'apathie (qu'on peut définir comme tendance à ne pas agir) et d'autres tendances telles que l'approche (tendance à s'approcher pour posséder). Ces tendances à l'action sont regroupées en 6 catégories générales: vers, contre, activation, désactivation, évitement et inhibition, interruption.

2.6.2 La théorie de Scherer

Klaus Scherer (1984, 2001) s'inscrit également dans cette perspective cognitive. Il considère quant à lui les émotions comme des agents intermédiaires entre un environnement changeant et l'individu. Dans cette approche, les émotions remplissent donc essentiellement des fonctions adaptatives en étant un point d'intersection entre le milieu et l'organisme. Les émotions ne sont cependant pas assimilables à des réflexes innés car la réponse au stimulus nécessite une évaluation cognitive. On peut plutôt les considérer comme une série de changements de l'organisme qui s'adapte à l'environnement au fur et à mesure des processus d'évaluation des stimuli. Ces changements peuvent intervenir simultanément selon diverses modalités correspondant à des sous-systèmes de l'organisme. Ces sous-systèmes sont représentés chacun par un « composant des émotions » ayant une fonction bien définie (il s'agit ici essentiellement de fonctions adaptatives, au sens darwinien du terme). Parmi ces sous-systèmes, on peut citer la composante cognitive, dont la fonction est d'évaluer les stimuli provenant de l'environnement, la composante physiologique, qui assure la régulation, ou encore la composante de l'expression motrice destinée à communiquer les intentions.

Selon Scherer, le traitement du stimulus passe par cinq étapes successives (*SEC, Stimulus Evaluation Checks*). Cette évaluation, qui serait de plus accompagnée de changements affectant le conduit vocal (c'est en cela que le modèle de Scherer accorde une place prépondérante à la prosodie), intervient par ordre de complexité croissante :

- Evaluation cognitive de la nouveauté ou du caractère inattendu de l'événement : il s'agit ici d'évaluer s'il y a eu un changement dans les stimuli internes et externes.
- Evaluation du caractère intrinsèque plaisant ou déplaisant du stimulus (*appraisal*), à partir de quoi il découle un sentiment d'envie ou d'aversion, une tendance au rapprochement ou au recul de l'organisme.
- Evaluation des buts et intérêts favorables : il s'agit d'évaluer la manière dont le stimulus avantage ou entrave l'atteinte des buts et besoins.
- Evaluation de la capacité de maîtrise de l'individu face à une situation négative : Si la situation est évaluée comme ne présentant pas de danger pour l'organisme, l'émotion engendrée sera la colère. A l'inverse si la situation semble dangereuse, l'émotion sera la peur ou la dépression.
- Evaluation de compatibilité avec les normes et l'image de soi : Scherer pense que cette étape d'évaluation n'existe que chez l'humain. Si l'évaluation indique que le comportement de l'individu n'est pas conforme aux normes, il en résulte une émotion telle que la honte.

Le modèle proposé par Scherer ne s'intéresse pas qu'aux mécanismes cognitifs mis en œuvre par l'évaluation émotionnelle, mais également à l'expression des émotions. Il prend ainsi en compte les expressions vocales des émotions, aussi bien produites que perçues, et s'attache à mesurer les signaux véhiculés afin de les relier aux émotions qui sont exprimées. Dans les évolutions récentes du modèle, ce lien entre émotions et signaux est établi à l'aide d'un modèle corrélationniste en lentille de Brunswick, afin de pouvoir prédire les évolutions des signaux en fonction de l'émotion exprimée. Dans cette optique de nombreux paramètres acoustiques relatifs à divers aspects de la parole ont été mesurés, dont les tendances générales en fonction des expressions, relevées à partir d'un grand nombre d'études, sont synthétisées dans la Table 1. Il faut noter que certains des paramètres présentés dans ce tableau ne sont pas mesurables directement mais constituent une catégorie générale établie à partir d'un ensemble de mesures comparables, et que tous les paramètres présentés n'ont pas été systématiquement mesurés. Les différentes tendances sont notées >, >=, =, <= ou < selon la répartition des valeurs par rapport à la moyenne ou <> si aucune tendance claire n'a pu être établie (résultats contradictoires selon les études).

Paramètres acoustiques	stress	joie/	colère/ rage	tristesse	peur/ panique	ennui
Paramètres de durée						
Nombre de syllabes par seconde	>	>=	<>	<	>	<
Durée syllabique	<	<=	<>	>	<	>
Durée des voyelles accentuées	>=	>=	>	>=	<	>=
Nombre et durée des pauses	<	<	<	>	<>	>
Durée relative des portions voisées			>		<>	
Durée relative des portions non-voisées			<		<>	
Fréquence fondamentale						
Moyenne	>	>	>	<	>	<=
5e percentile	>	>	=	<=	>	<=
Variabilité	>	>	>	<	>	<
Etendue	>	>	>	<	<>	<=
Hauteur des syllabes accentuées	>	>=	>	<		
Gradient des chutes et montées de F0	>	>	>	<	>	<=
Etendue et gradient de la descente finale	>	>	>	<	>	<=
Intensité et type de phonation						
Intensité moyenne	>	>=	>	<=		<=
Variabilité de l'intensité	>	>	>	<		<
Gradient des chutes et montées d'int.	>	>=	>	<		<=
Energie spectrale en hautes fréquences	>	>	>	<	<>	<=
Pente spectrale	<	<	<	>	<>	>
Laryngalisation		=	=	>	>	=
Jitter		>=	>=		>	=
Shimmer		>=	>=		>	=
Ratio harmoniques / bruit		>	>	<	<	<=
Vitesse et précision de l'articulation						
Précision des formants		=	>	<	<=	<=
Bande passante des formants	<		<	>		>=

Table 1 : Synthèse des résultats empiriques relatifs aux effets de l'émotion sur les expressions vocales. D'après (Scherer, 2003).

2.6.3 L'approche de Damasio

2.6.3.1 La raison des émotions

Antonio Damasio, neurophysiologiste américain, a étudié avec Hannah Damasio les émotions à partir de cas de patients ayant subi une lésion du cerveau (souvent localisée au niveau du cortex préfrontal ventro-médian), constatant qu'ils perdaient à la fois la faculté d'éprouver des émotions et celle de prendre des décisions. La conséquence

systématique de cette pathologie est une dégradation spectaculaire des rapports sociaux de ces patients, tandis qu'ils conservent leurs facultés physiques et mentales. De cette approche neurophysiologique et cognitive a découlé l'hypothèse selon laquelle émotion et raison sont, bien que distinctes, deux parties d'une même entité psychologique irréductible, contrairement à l'idée de dualité « autonome » avancée par Descartes et largement répandue (Damasio, 1994).

A partir de cette hypothèse, Damasio a également mis en évidence la superposition de deux types d'émotions : les émotions primaires, qui seraient innées et dédiées à des fonctions de survie, et les émotions secondaires, acquises à partir des émotions primaires. Certains de ses patients pouvaient en effet toujours ressentir le premier type d'émotions, mais pas le deuxième.

Les émotions secondaires, quant à elles, n'apparaissent que lorsqu'on peut établir un lien entre la perception de l'émotion et la situation qui l'a produite. Elles se forment en modulant les émotions primaires en fonction de l'expérience personnelle et des interactions avec l'environnement. Selon Damasio, la prise de conscience de cette relation permet d'améliorer les stratégies de défense, par exemple en évitant préventivement les situations à risques, alors que les émotions primaires ne sont déclenchées que de manière automatique dans des situations d'urgence immédiate. D'autre part, les deux types d'émotion se différencient également par la possibilité de déclencher les émotions secondaires par un acte purement mental, tandis que les émotions primaires ne peuvent être déclenchées que par les stimuli en provenance du corps.

2.6.3.2 La perception des émotions

Le premier ouvrage de Damasio s'intéresse également à la façon dont les émotions sont perçues. Selon lui, l'émotion elle-même est la résultante de la juxtaposition cognitive de l'image corporelle (la carte des stimuli externes dans le cerveau ou encore la perception des changements d'état du corps) et de l'image de la situation (la perception de la situation, qui peut être imaginaire dans le cas d'émotions secondaires).

Ainsi, pour Damasio, la perception des émotions passe par la perception des changements de ses propres états somatiques par le sujet qui ressent l'émotion. C'est sur la base de ce postulat que Damasio établit la différence entre les émotions réellement ressenties (ceci peut inclure des émotions secondaires déclenchées par l'imagination) et les émotions simulées.

La figure 2 schématise la perception des émotions selon Damasio lorsqu'il s'agit d'émotions réelles ou simulées. Dans les deux images illustrant les deux sortes de

boucles, le cerveau est représenté par la ligne fermée du haut, et le reste du corps par celle du bas. Selon l'hypothèse de Damasio, les processus se déroulant dans la boucle de simulation restent circonscrits au cerveau, court-circuitant complètement la perception de l'état du corps. Dans ce cas, la perception des changements d'état physiologique est remplacée par le souvenir d'états somatiques associés à l'émotion que le locuteur veut simuler, qui servent de base pour reproduire l'expression de cette émotion.

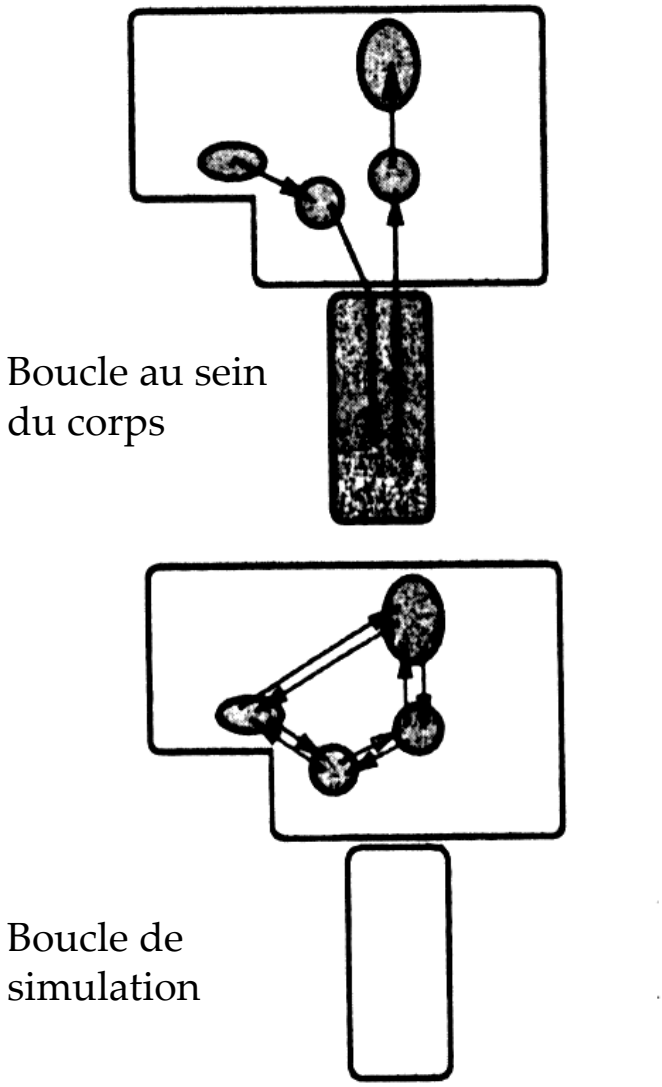


Figure 2 : Les mécanismes de perception des émotions, en boucle au sein du corps ou par le biais d'une boucle de simulation, d'après Damasio (1994)

2.7. Emotions et expressions

Les émotions sont exprimées, et c'est bien la raison de leur existence dans la communication entre humains. Les approches physiologiques ont montré que les changements bio-physiologiques liés aux variations d'état émotionnels sont « récupérés » selon diverses modalités (rosissement des joues, voix qui tremble, odeur de

sudation, etc.), mais les approches plus radicalement évolutionnistes ont eu l'intérêt de montrer que l'expression est aussi le produit d'un contrôle complexe. C'est ainsi que les expressions faciales, sous l'impulsion d'Ekman en particulier (Ekman & Friesen, 1978 ; Ekman, 1989, 1999a), ont été très largement étudiées (Tcherkassof (1999) a établi un état de l'art de ces études), modélisées et même simulées (Massaro, 2000). Par ailleurs la pragmatique s'est très tôt emparée, en s'ancrant également dans une perspective de psychologie cognitive, du problème de la modélisation de la gestualité associée aux expressions émotionnelles (MacNeil, 1992).

La modalité visuelle est le médium privilégié des contrôles d'expressions (de la même façon que l'acoustique est le médium privilégié de la parole « linguistique »). Pourtant même pour des variations d'états émotionnels complexes, pour lesquels la modalité visuelle seule apporte déjà des informations robustes (par exemple l'amusement, du sourire au rire), la parole acoustique est également porteuse d'expressions émotionnelles : Aubergé et Cathiard (2003) ont montré que l'on entend plus que la conséquence acoustique de la modification du conduit vocal (intégration audio-visuelle), c'est-à-dire que la parole aurait ses expressions spécifiques. Au-delà des propositions de Scherer, Aubergé et Cathiard proposent que les expressions des émotions ne sont pas seulement une intégration multi-modale d'un « geste produit pour être d'abord visible », mais qu'elles sont contrôlées spécifiquement pour les différentes modalités.

3. Technologies de la parole expressive

Nous donnons dans cette section un aperçu des principales applications technologiques de l'étude des affects dans le domaine du traitement automatique de la parole, à travers les deux objectifs de recherche majeurs que sont la synthèse de parole expressive et la classification automatique des affects.

En séparant les applications existantes en fonction du type de système de synthèse utilisé, nous présentons tout d'abord les principales approches existantes de la synthèse de la parole expressive et, en particulier, les différentes propositions d'intégration de la dimension expressive à la technique de synthèse la plus aboutie à l'heure actuelle qu'est la synthèse par sélection.

Nous présentons ensuite les principes généraux qui régissent les systèmes existants de classification automatique d'affects, puis les contributions qui nous semblent majeures dans ce domaine, et qui mettent l'accent sur la nécessité de considérer des expressions spontanées d'affects au-delà de la simulation des émotions dites « de base » sur laquelle se sont appuyés la majorité des systèmes proposés.

3.1. Synthèse expressive

3.1.1 Généralités

Nous n'abordons pas dans cette section l'intégralité des systèmes de synthèse expressive proposés dans la littérature, et nous contentons de mentionner quelques exemples représentatifs des différentes approches de cette problématique complexe, sans entrer dans les détails des résultats obtenus par ces études. La grande majorité des systèmes proposés s'est appuyée sur le principe de la synthèse à partir du texte (voir notamment Dutoit (1997)), dont le principe général est de procéder à une phonétisation et, dans la plupart des cas, à une analyse morpho-syntaxique d'un texte inconnu fourni en entrée du système sous forme écrite, avant de transformer ce texte en parole synthétique. La phonétisation d'un texte écrit est en elle-même un problème complexe, que nous n'aborderons pas ici. Les résultats de l'analyse morpho-syntaxique réalisée sont généralement utilisés pour appliquer des règles de transformation prosodiques et notamment des contours de fréquence fondamentale, afin de réaliser les fonctions linguistiques de la prosodie que sont la segmentation/hiérarchisation et la modalisation ou, plus rarement, la focalisation. L'objectif visé par la plupart des systèmes de synthèse vocale qui ont été développés, et atteint en grande partie par ceux développés récemment, est de parvenir à générer des énoncés assimilables à de la parole naturelle lue parfaitement intelligible. Comme nous l'avons mentionné en introduction, avec les progrès récents des systèmes de synthèse par corpus et la réalisation de cet objectif majeur des systèmes de synthèse à partir du texte, l'enjeu de la synthèse de la parole s'est déplacé de l'intelligibilité et de la naturalité vers la pertinence de l'énoncé généré en fonction du contexte d'énonciation. En conséquence, l'intégration d'une dimension expressive à ces systèmes est devenu un objectif central, d'où les efforts de recherche croissants dans cette direction depuis les premiers jalons posés notamment par Janet Cahn au début des années 1990.

Schröder (2001) dresse un état de l'art des systèmes de synthèse de la parole expressive, et propose une taxonomie des approches proposées dans la littérature en fonction de la méthode de synthèse utilisée. Nous nous appuyons sur cette taxonomie pour présenter des exemples que nous jugeons représentatifs, en complétant d'exemples plus récents lorsque cela nous semble utile.

3.1.2 Synthèse par formants

La synthèse par formants, parfois appelée synthèse par règles, est la méthode dont le résultat est le moins abouti en termes d'intelligibilité et plus généralement de naturalité

dans le cadre d'une approche entièrement automatisée. Cependant, cette méthode est également la plus flexible, et permet une synthèse de qualité lorsqu'elle est contrôlée manuellement de façon fine. En effet, son principe est de générer dans leur intégralité les sons de parole produits à partir de modèles acoustiques, sans réutiliser des enregistrements préexistants. L'exemple le plus répandu de synthèse par formants est le système DECtalk (Hallahan, 1996). Dans le domaine de la synthèse de parole expressive, le système pionnier Affects Editor de Cahn (1990) et le système HAMLET de Murray & Arnott (1995) peuvent notamment être mentionnés. Dans les deux cas, les règles de modification de la sortie du synthétiseur DECtalk ont été établies à partir d'une revue de littérature sur l'encodage des expressions émotionnelles dans la parole. En dépit du manque de naturalité généralement lié à la synthèse par formant, ces systèmes ont permis de générer des énoncés de parole synthétique de qualité suffisante pour être correctement classifiés à un niveau supérieur au hasard dans une tâche de catégorisation perceptive.

3.1.3 Synthèse par concaténation

La synthèse par concaténation, dont l'exemple le plus répandu est le système multilingue distribué librement MBROLA[1] (Dutoit *et al.*, 1996), permet d'atteindre un degré d'intelligibilité et de naturalité supérieur, sans pour autant pouvoir passer pour une voix naturelle. Le principe de ce type de système de synthèse est de s'appuyer sur une base de diphones, enregistrés par un même locuteur pour une langue cible donnée, et permettant de couvrir l'ensemble des combinaisons de deux phonèmes consécutifs de cette langue. Les diphones sont constitués de portions de signal s'étendant du milieu d'un phone au milieu du suivant. Ils sont généralement enregistrés avec une prosodie neutre. La synthèse de la chaîne sonore est alors fondée sur une concaténation de la suite de diphones correspondant au texte phonétisé, suivie de l'application de règles de transformation prosodiques fondées sur un modèle de durée et de variations de fréquence fondamentale. Ces modifications prosodiques sont le plus souvent réalisées par l'algorithme TD-PSOLA (Moulines & Laroche, 1995).

La technique de la synthèse par diphones a été utilisée pour construire des systèmes de synthèse de parole expressive, avec deux types distincts de transformations prosodiques. La plus répandue a été celle de l'analyse/synthèse, dans laquelle les variations prosodiques d'un énoncé naturel véhiculant l'affect visé sont extraites et appliquées à

[1] Le système MBROLA ainsi que des bases de diphones pour un ensemble étendu de langues est téléchargeable librement à l'URL http://tcts.fpms.ac.be/synthesis/mbrola.html

l'énoncé synthétisé. Cette méthode a, par exemple, été adoptée par Montero *et al.* (1999). L'approche plus ambitieuse de la modification de la prosodie par règles n'a à notre connaissance pas été appliquée dans le cadre d'un système complet de synthèse à partir du texte, mais l'a été avec l'objectif de confronter les prédictions d'un modèle à la perception que peuvent avoir des auditeurs naïfs des énoncés synthétiques correspondants. Cette approche a, par exemple, été retenue par Mozziconnacci (1999), qui a proposé une modélisation de contours de fréquence fondamentale supposés véhiculer un ensemble d'affects.

Comme le souligne Schröder (2001), la synthèse par concaténation ne permet pas de manipuler directement la qualité de voix, et la question de la possibilité de générer des expressions vocales d'affects sans prendre en compte la qualité de voix s'est donc posée de façon récurrente dans les études ayant eu recours à cette méthode de synthèse. La revue de littérature réalisée par Schröder (2001) aboutit à des résultats divergents quant à cette possibilité, certaines émotions produites par certains locuteurs semblant être véhiculées efficacement par de simples modifications de fréquence fondamentale et de durée. Il conclut toutefois que la possibilité d'étendre ces résultats à l'ensemble des types d'émotions existants n'est pas avérée.

3.1.4 Synthèse par sélection

Le principe de la synthèse par sélection, ou synthèse par corpus, est à la base de la plupart des systèmes de synthèse actuels de qualité professionnelle. Le plus abouti des systèmes de synthèse du français existants à l'heure actuelle est de notre point de vue celui développé par Orange-labs[2]. Un tel système de synthèse repose sur un corpus de taille importante produit par un même locuteur, étiqueté avec précision, comprenant des phrases phonétiquement équilibrées, et dépassant très largement la couverture minimale de la langue visée telle qu'elle est réalisée par les bases de diphones des systèmes de synthèse par concaténation. De même que dans le cas de la synthèse par concaténation, la génération de la chaîne sonore est réalisée en concaténant des unités extraites du corpus de parole naturelle sur lequel s'appuie le système de synthèse. Ces unités sont en revanche choisies comme les plus longues possibles répondant aux différents critères de sélection, qui ne se réduisent pas à l'adéquation de la suite de phonème à celle du message à synthétiser, incluant également des critères prosodiques. Ainsi, les unités sélectionnées préalablement à l'étape de concaténation consistent en des extraits des

[2] Une démonstration en ligne de ce système de synthèse, qui équipe par ailleurs un grand nombre de services téléphoniques et de transports en commun, est accessible à l'URL http://tts.elibel.tm.fr/tts

phrases constituant le corpus, en adéquation avec la suite de phonèmes à produire et porteurs d'une prosodie la plus proche possible de celle du message à synthétiser telle que prédite par le modèle utilisé à la suite de l'analyse morpho-syntaxique. Un lissage des unités concaténées est alors réalisé, complété par l'application de transformations prosodiques si nécessaire. Cependant ces modifications du signal, à l'origine du manque de naturalité des systèmes de synthèse par concaténation de diphones, sont réduites au minimum. Le cas idéal de la synthèse par sélection consiste ainsi en la synthèse d'un énoncé identique, selon l'ensemble des critères retenus, à l'un des énoncés présents dans le corpus de base qui est alors restitué sans qu'aucune modification du signal ne soit nécessaire. Le cas d'une erreur de prétraitement (lors de l'étape de phonétisation ou d'analyse morpho-syntaxique) mis à part, les systèmes les plus évolués de synthèse vocale par sélection permettent de transmettre le message fourni par écrit en entrée avec, dans la grande majorité des cas, une intelligibilité supérieure à celle atteinte par les systèmes de synthèse par concaténation. De plus, la synthèse par sélection atteint un degré de naturalité très élevé, à tel point que la plupart des énoncés synthétisés par ces systèmes peuvent aisément passer pour de la parole lue naturelle. Cependant dans les cas marginaux, d'autant plus rares que la base de données de parole étiquetée utilisée est étendue, pour lesquels des unités inappropriées doivent être choisies, les résultats de ces systèmes de synthèse peuvent être localement de piètre qualité.

Parmi les systèmes de synthèse ayant directement tiré parti de la technique de la synthèse par corpus, le système développé par Iida *et al.* (2003) s'appuie sur une base de données de parole étiquetée de grande taille, dans laquelle chacune des 525 phrases constituant le corpus est produite en triple exemplaire, en simulant une expression de joie, de colère ou de tristesse. Le système de synthèse proprement dit est le système CHATR (Campbell & Black, 1997), dans lequel les critères de sélection des unités à concaténer sont étendus au choix de l'émotion à produire parmi les trois cités ci-dessus. Bien que des confusions puissent être observées, les énoncés générés par ce système ont été jugés majoritairement comme exprimant l'émotion visée.

Marumoto & Campbell (2000), cité par Schröder (2001), plutôt que de dupliquer le corpus utilisé pour chacune des émotions visées identiques à celles adoptées par Iida *et al.* (2003), ont utilisé un corpus commun en sélectionnant les unités en fonction de critères prosodiques et de qualité de voix. L'évaluation des énoncés synthétisés a montré que les expressions de colère et de tristesse étaient reconnues mais pas celles de joie.

D'autres approches de la synthèse expressive par corpus s'écartent de l'objectif de générer des émotions dites « de base » et visent d'autres aspects de l'expressivité prosodique. Syrdal & Kim (2008) se sont ainsi appuyés sur une base étiquetée en

fonction, non seulement de critères phonétiques et prosodiques, mais également d'actes de parole. Bien que les auteurs n'y fassent pas explicitement référence, la taxonomie des actes de parole qu'ils retiennent est similaire à celle introduite par Searle (1969, 1975), notamment en ce qui concerne les actes de parole dits « affectifs ». L'enrichissement du système de synthèse par sélection d'AT&T qu'ils proposent permet ainsi de favoriser dans le processus de sélection les unités correspondant à l'acte de parole visé pour l'énoncé synthétisé.

Enfin, l'approche la plus novatrice de la synthèse de la parole expressive est probablement celle proposée par Suciu *et al.* (2006a, 2006b) et Kanellos *et al.* (2007). Celle-ci repose sur un corpus de textes de différents types, lus en faisant varier la situation d'élocution afin que celle-ci balaie des situations socialement reconnues, telles qu'un discours politique, un commentaire sportif ou encore un sermon. Les « formes discursives expressives » sont annotées à divers niveaux linguistiques, en partant de celui de la syllabe, et leur tolérance aux déformations prosodiques tout en conservant une même valeur d'expressivité est évaluée. Afin de générer des textes synthétiques semblables à d'authentiques discours, les unités ainsi annotées sont sélectionnées par un système de synthèse par corpus, avec comme critères de haut niveau le type de texte et la situation discursive.

3.1.5 Transformation de voix

Outre les systèmes complets de synthèse à partir du texte passés en revue par Schröder (2001), des algorithmes permettant de modifier la qualité de voix des signaux générés (cf. section 7 pour un développement de la notion de qualité de voix) ont été proposés. En particulier, d'Alessandro & Doval (2003) ont proposé un ensemble d'algorithmes spécifiquement dédiés à la modification de la qualité de voix avec l'objectif d'une application en synthèse de la parole expressive. Ces algorithmes permettent de modifier la qualité de voix d'un signal de parole, en traitant séparément ses composantes périodiques et apériodiques. Rosec & Vincent (2008) proposent quant à eux une méthode de modification de voix fondée sur une modélisation de la qualité de voix, qui constitue une extension de leur méthode d'analyse/synthèse que nous présentons en section 9.2.5.

D'Alessandro & Doval (2003) proposent également un algorithme permettant de caractériser les unités sélectionnées en fonctions de leurs propriétés de qualité de voix, dans une optique de synthèse par corpus.

3.2. Reconnaissance automatique des affects

3.2.1 Principes généraux

Parallèlement à l'accroissement des efforts de recherche visant la synthèse de parole expressive, l'autre grand domaine d'application du traitement automatique de la parole qu'est la reconnaissance automatique s'est également intéressé à la reconnaissance automatique des affects exprimés par un locuteur.

Deux types principaux d'application coexistent dans le domaine de la reconnaissance automatique appliquée à des signaux de parole. D'une part, la reconnaissance de la parole a pour objectif la transcription automatique du message prononcé par le locuteur. D'autre part, la reconnaissance automatique du locuteur s'intéresse à l'auteur du message et cherche à repérer les informations permettant son identification.

Autant dans le cas de la reconnaissance automatique de la parole que de la reconnaissance automatique du locuteur, l'extraction de l'information acoustique se fait au moyen d'un nombre souvent élevé de paramètres acoustiques. L'appariement des informations acoustiques et des phonèmes, mots ou locuteurs visés, se fait ensuite au moyen de modèles probabilistes.

La reconnaissance des affects consiste en pratique en une catégorisation des affects parmi un ensemble prédéfini, généralement d'étendue limitée, d'étiquettes ou de classes considérées (voir section 8 pour une discussion de la problématique de la reconnaissance vs. la discrimination d'affects dans le cadre d'évaluations perceptives d'affects). Le nombre de paramètres acoustiques utilisé dans ce cadre est très élevé, ces paramètres étant en grande partie redondants. En effet, dans le cas général, la quantité la plus importante possible de paramètres supposés différents et susceptibles de mesurer une partie de l'information acoustique est retenue. Le jeu de paramètres est, par la suite, restreint par des méthodes statistiques de réduction de données à un nombre fixé de « meilleurs paramètres » permettant d'expliquer la part la plus importante de la variance mesurée (par exemple 16 paramètres dans le cas de l'étude menée par Ververidis *et al.* (2004)).

Le jeu de paramètres initial inclut généralement des mesures diverses de durée, des caractéristiques des distributions de la fréquence fondamentale et de l'énergie, divers paramètres spectraux et cepstraux (notamment les coefficients MFCC qui ont fait la preuve de leur efficacité dans les champs d'application plus classiques de la reconnaissance automatique). Il peut comprendre également des paramètres supposés capturer plus explicitement des phénomènes liés à la qualité de voix. Lorsque le contenu

du message est inclus dans les données traitées dans le but de déterminer l'affect exprimé par le locuteur, des paramètres relatifs aux informations lexicales, aux actes de langage ou encore à des interprétations sémantiques de plus haut niveau peuvent également être pris en compte. Une description des 4244 paramètres utilisés dans la plus vaste étude de reconnaissance automatique des affects réalisée à ce jour est présentée par Schuller *et al.* (2007).

3.2.2 Bref état de l'art

La plupart des études portant sur la classification automatique d'émotions ou autres affects exprimés dans la parole se sont appuyées sur des corpus actés construits de façon à faire varier systématiquement les affects sur des énoncés constants (cf. section 5 pour une discussion des différents types de corpus de parole expressive et de leurs implications). En particulier, deux corpus de ce type, accessibles librement pour des utilisation non-commerciales, ont été utilisés dans ce cadre. Il s'agit du corpus allemand *Berlin emotional speech database* (Burkhardt *et al.*, 2005) et dans une moindre mesure du corpus danois *Danish Emotional Speech* (Engberg & Hansen, 1996). Quelques études ont pris en compte de façon conjointe les informations données par la face et la parole pour améliorer les performances en classification automatique des affects exprimés (voir par exemple Busso *et al.* (2004)).

Plus proche des données naturelles, l'application de la classification automatique d'états affectifs à des données tirées de films peut également être mentionnée. Ainsi Clavel (2007) a proposé une méthode de détection d'états affectifs proches de la peur, étalonnée sur des données extraites de films d'épouvante.

Dans le champ de la reconnaissance automatique des affects, de façon plus marquée que dans celui de la synthèse vocale, l'idée s'est imposée que les corpus d'émotions actées, et tout particulièrement ceux constitués d'expressions prototypiques, ne seraient pas représentatifs des affects exprimés naturellement et ne permettraient pas d'élaborer des systèmes de reconnaissance automatiques efficaces sur des données réelles (voir notamment Batliner (2003a)). Au cours des dernières années, de nombreuses études ont donc cherché à caractériser les affects exprimés dans des données spontanées, ce qui a impulsé un effort de recherche particulier en direction de l'annotation des affects exprimés dans ce type de données, afin de fournir une référence fiable pour l'entraînement des systèmes automatiques de classification comme pour la validation des résultats obtenus (cf. section 5 pour une discussion sur la problématique de l'annotation de données spontanées).

Ainsi, Vidrascu & Devillers (2005a) ont mis au point un système de catégorisation automatique des affects exprimés en français par des usagers d'un centre d'appel. Constatant que les expressions d'affects exprimés naturellement consistent souvent en des mélanges de différents affects, elles se sont également efforcées de faire en sorte que cette catégorisation automatique ne consiste pas obligatoirement en un choix forcé mais puisse également rendre compte de tels mélanges (Vidrascu & Devillers, 2005b). Batliner *et al.* (2005) ont, quant à eux, élaboré un système de classification automatique d'affects exprimés par des enfants germanophones en interaction avec un robot-jouet (Sony AIBO®). Cette étude a été récemment prolongée dans un projet coopératif de grande ampleur, dans lequel non seulement la classe d'affect exprimée mais également le degré de prototypicalité ont été étudiés (Seppi *et al.*, 2008). Les auteurs ont conclu que les expressions les plus prototypiques parmi celles étudiées n'étaient pas en mesure d'expliquer l'ensemble de la variance des expressions naturelles d'affects présentes dans le corpus étudié, tendant à confirmer que les corpus d'expressions actées ne seraient pas adaptés à la modélisation d'expressions spontanées d'affects dans la parole.

4. Prosodie et émotion

Dans cette section, nous commençons par revenir sur la définition la plus classique de la prosodie, qui considère les paramètres les plus couramment considérés comme constituants de la prosodie que sont la fréquence fondamentale, l'intensité et la durée.

Nous nous plaçons ensuite dans une perspective fonctionnaliste, et revenons sur quelques-unes des contributions majeures à l'étude de la fonction expressive de la prosodie dans le champ de la phonostylistique.

Nous discutons alors brièvement de la pertinence d'une prosodie non plus limitée à ses trois paramètres classiques, mais étendue à la qualité de voix.

Enfin, nous présentons de façon plus détaillée le modèle de superposition de contours gradients dirigée par les fonctions proposé par Aubergé.

4.1. La prosodie : une définition controversée

La prosodie est un objet complexe de la parole qui regroupe divers phénomènes tels que l'accent, le ton, la mélodie, le rythme ou encore le débit, mais qu'il est impossible de définir dans l'absolu, hors de la théorie ou de l'approche expérimentale dans le cadre de laquelle elle est abordée. Dès 1975, Di Cristo notait déjà plus de 4000 références bibliographiques relatives à la prosodie. Pourtant, le rôle de la prosodie autant que sa structure, son fonctionnement cognitif et maintenant ses structures neurales, sont toujours fortement discutés, sans que puisse se dégager de paradigme réellement objectif. Elle intègre la réalisation de multiples fonctions qui opèrent dans les niveaux linguistique, pragmatique (intentions et attitudes), émotionnel et idiolectal (âge, sexe, identité). Au niveau linguistique, qui a été de loin le plus étudié jusqu'alors, on s'accorde à prêter à la prosodie les fonctions de structuration de l'énonciation (également appelée fonction de démarcation ou de segmentation / hiérarchisation et fréquemment assimilée à l'intonation), de modalisation (expression de la déclaration, de l'interrogation…), et de focalisation (mise en valeur d'une partie de l'énoncé).

4.2. Les paramètres classiques

La réalisation phonétique de la prosodie ne peut pas être objectivement et strictement extraite du signal comme étant « tout ce qui ne fait pas la chaîne sonore », et reste en soi objet de discussions. A cette dichotomie entre segmental et suprasegmental, pour reprendre la terminologie anglo-saxonne, Rossi (1993) répond d'ailleurs que la prosodie constituerait une troisième articulation du langage.

Bien qu'il ne soit pas possible de s'appuyer sur une définition consensuelle de la prosodie, un accord est toutefois établi au niveau phonétique sur trois paramètres acoustiques la décrivant : la fréquence fondamentale, l'intensité et la durée (ce dernier paramètre impliquant de définir des événements temporels de segments, supposant la définition claire de ce qu'est un segment de parole).

Lacheret et Beaugendre (1999) les définissent comme suit :

- La fréquence fondamentale (notée F0) est définie comme la fréquence de vibration des plis vocaux à un instant donné. On appelle voisement la présence d'une fréquence fondamentale, c'est-à-dire la vibration périodique des plis vocaux.
- La durée est vue comme l'intervalle de temps nécessaire pour émettre un signal. En d'autres termes, c'est l'organisation temporelle du message qui comprend le débit de parole (nombre de syllabes réalisées par seconde), le tempo (accélération ou

ralentissement du débit dans le groupe prosodique), les pauses et la durée phonémique.

- L'intensité correspond à l'énergie présente dans le signal, soit la force avec laquelle un son est émis. Cette intensité résulte de l'amplitude des plis vocaux, de la force du flot d'air expiré et des variations de fréquence.

4.3. La fonction expressive de la prosodie

Au-delà de sa définition par ses paramètres, la prosodie peut être définie par les fonctions qu'elle réalise. Celles-ci incluent les fonctions linguistiques les plus couramment étudiées, ainsi que la fonction expressive de la prosodie, qui nous intéresse plus particulièrement ici.

La fonction expressive a été plus spécifiquement étudiée dans le domaine de la phonostylistique, développé par des auteurs comme Léon (1971, 1993) ou Fonagy (1983). Léon (1971) introduit ainsi la notion de « phonostylème » comme unité minimale d'étude de l'expressivité transmise, définit comme l'ensemble des traits permettant d'identifier les caractéristiques du message expressif de façon pertinente. Selon Léon, ces unités expressives incluent à la fois des informations relatives aux états psychologiques (dont ses émotions) et à la situation sociologiques du locuteur et ses intentions ou ses attitudes vis-à-vis de son interlocuteur. Léon (1993) établit notamment une distinction entre signaux volontaires conventionnalisés dans la langue et la culture considérée (parmi lesquels il classe l'accent de focalisation), et des signaux involontaires qui consisteraient en des manifestations biologiques (rire, toux, etc.). Ces signaux volontaires et involontaires sont considérés par Léon (1993) plutôt comme des facettes distinctes co-occurrentes de l'expressivité que comme deux pôles antagonistes entre lesquels un intermédiaire doit être trouvé. Bien que ne se recouvrant donc pas entièrement, cette distinction entre contrôle cognitif volontaire et involontaire pouvant apparaître au sein d'un même message est à mettre en relation avec les notions d'effets *pull* et d'effets *push* introduits en psychologie cognitive par Scherer (2001) pour désigner respectivement la conventionnalisation de l'expression et la transmission directe dans la parole de changements physiologiques liés à l'émotion.

Fonagy (1983) propose un modèle dans lequel tout message transmis dans la parole subirait un double encodage. Selon lui, le premier niveau consiste en un encodage linguistique qui transforme le concept abstrait correspondant au message à transmettre en une suite de phonèmes, tandis que le second niveau consisterait en une modulation expressive de ce message, modifiant l'expression stylistique et émotionnelle de ce message à un niveau de contrôle cognitif différent. Selon lui, toute instanciation concrète

d'un phonème est en elle-même expressive, de même que l'ensemble de la prosodie, fonctions linguistiques incluses. Ceci a pour conséquence que la prise en considération de ce second niveau d'encodage devient indispensable, quand bien même une expression « neutre » serait considérée.

Fonagy (1983) considère le mécanisme de décodage comme le processus miroir du processus d'encodage. Ainsi le décodage de tout message parlé relèverait en premier lieu d'un décodage de l'expressivité relevant du même code paralinguistique que celui utilisé pour le second niveau d'encodage, puis d'un décodage linguistique de la suite de phonèmes constituant le message transmis, relevant du même code linguistique que celui utilisé dans le processus d'encodage. De même que dans le cas de l'encodage, le décodage des niveaux paralinguistique et linguistique du message relèverait de niveaux cognitifs distincts.

Gussenhoven (2002) propose quant à lui, à la suite des propositions d'Ohala (1983, 1996) d'une explication éthologique de l'utilisation des registres de fréquence fondamentale, d'expliquer l'ensemble des utilisations paralinguistiques des variations de fréquence fondamentale en termes de Code de Fréquence, Code d'Effort et Code de Production. Le principe fondateur du Code de Fréquence (Ohala, 1983) est que, tant chez l'animal que l'humain, les larynx de dimensions réduites tendent à produire des fréquences plus élevées, tandis que ceux de plus grandes dimensions produisent des fréquences plus basses. Ohala (1996) a proposé une explication évolutionniste du Code de Fréquence, selon laquelle le fait de produire des vocalisations avec une fréquence fondamentale plus basse afin d'être considéré comme massif et menaçant par les congénères et les autres espèces constituerait un avantage adaptatif. En effet un gros larynx est généralement lié à une taille importante de l'animal, ce qui expliquerait que les cris menaçants, tant dans la parole que dans l'éthologie animale, soient généralement produits avec une fréquence plus basse que les cris de soumission. Le Code d'Effort et le Code de Production, proposés par Gussenhoven (2002) en complément du Code de Fréquence, sont directement liés au mécanisme de production de la parole : ils font référence respectivement au fait que la production de sons de parole demande un certain niveau d'énergie dont les variations sont reflétées dans le signal de parole, et que cette énergie n'est disponible que dans certaines phases de la production de parole, en lien avec les processus respiratoires.

4.4. Quels paramètres retenir ?

Si la fonction idiolectale de transmission des informations relatives à l'âge et au sexe du locuteur, et surtout la fonction émotionnelle, sont considérées comme faisant partie des

fonctions réalisées par la prosodie, il est nécessaire d'ajouter aux paramètres la décrivant les paramètres spectraux relatifs au timbre qui renvoient à la notion de qualité de voix. En effet, diverses études, parmi lesquelles (Roach, 2000 ; Scherer, 1986 ; Johnstone & Scherer, 1999 ; Gobl & Ní Chasaide, 2000), ont montré que la qualité de voix était porteuse d'informations sur l'état émotionnel du locuteur. En conséquence les trois paramètres classiques définissant la prosodie ne sont pas suffisants pour décrire des phénomènes paralinguistiques tels que l'expression des émotions. Nous développons en section 7 la problématique de la qualité de voix.

Si la prosodie est définie par ses trois paramètres classiques organisés dans un modèle, on se trouve confronté à une fonction émotionnelle partagée entre prosodie et qualité de voix. Si en revanche la prosodie est définie par ses fonctions, son « matériel » acoustique se caractérise par l'évolution organisée de tous les paramètres pertinents pour la réalisation de la fonction visée.

De nombreux modèles prosodiques ont été proposés dans la littérature, qui pour la plupart rendent compte principalement de la prosodie linguistique ou intonation. Notre objet ici n'est pas de passer en revue de tels modèles. Pour un tel état de l'art, dans lequel les implications des différentes classes de modèles de l'intonation ont discutées, le lecteur pourra notamment se reporter à Rossi (1999).

L'une des classes de modèles prosodiques abordée par Rossi (1999) propose une approche par superposition de contours globaux, définis sur l'ensemble de l'énoncé. Parmi ces modèles de la prosodie, celui qui a été le plus largement utilisé est probablement le modèle proposé par Fujisaki (1988), qui se fonde essentiellement sur la physiologie du processus de phonation en développant notamment la notion de ligne de déclinaison (mouvement général descendant du contour de F0, qui correspondrait à la fois à une phonation sans effort particulier et à un énoncé déclaratif (Maeda, 1976 ; Vaissière, 1997). Le modèle prosodique développé par Aubergé (2002), que nous présentons ci-dessous de manière plus détaillée et sur lequel nos analyses s'appuient, appartient également à la classe des modèles par superposition de contours globaux.

4.5. Le modèle prosodique d'Aubergé

4.5.1 Un modèle fonctionnel de la prosodie

Le modèle d'Aubergé (2002) est un modèle fonctionnel, qui étudie les phénomènes langagiers à travers les fonctions du système de communication implémentées via ses

différents agents. Ce modèle, qui décrit plus particulièrement l'agent prosodique lui-même, prend en considération trois types de fonctions réalisées par la prosodie :

- Les fonctions linguistiques, soient les fonctions de segmentation / hiérarchisation, de modalisation et de focalisation,
- Les fonctions attitudinales, expressions des intentions du locuteur (attitudes et émotions simulées),
- Les fonctions émotionnelles, expression des états émotionnels du locuteur.

La distinction essentielle entre fonctions émotionnelles et attitudinales réside dans les domaines temporels par lesquels ces expressions sont régies, correspondants aux différents types de contrôle cognitif exercés. Tandis que les expressions attitudinales, contrôlées volontairement, sont régies par le « temps linguistique », les expressions émotionnelles sont régies par un temps contraint par les évènements qui provoquent l'état émotionnel du locuteur (et qui peuvent être externes au système communicatif). La Figure 3 est une représentation schématique des fonctions prises en compte par le modèle.

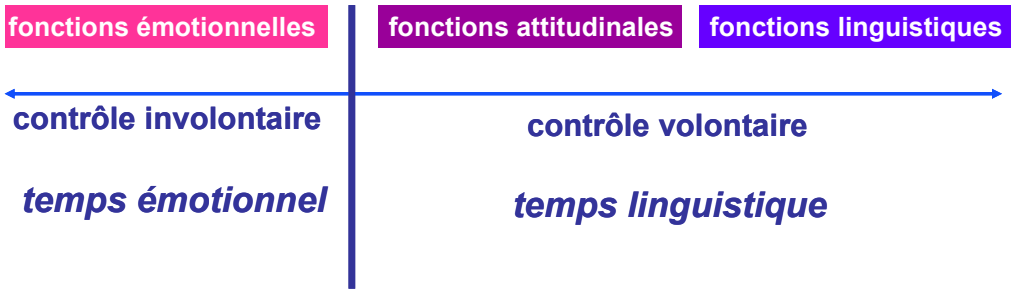


Figure 3 : Types de contrôle et de timing des fonctions de la prosodie. D'après Aubergé (2002)

4.5.2 Les principes du modèle

Le modèle prosodique d'Aubergé est fondé sur cinq grands principes morphologiques, que l'on peut séparer selon le type d'encodage (direct *vs.* indirect) et le domaine temporel auquel les fonctions concernées se rapportent. Les trois premiers principes du modèle concernent les fonctions linguistiques et pragmatiques. Celles-ci sont considérées comme encodées directement, contrôlées volontairement et ancrées dans le domaine temporel linguistique.

- **Principe cognitif :**

  Le traitement cognitif de la morphologie prosodique s'appuie sur les mouvements prosodiques globaux, sur un *domaine* linguistique donné (autrement dit le segment porteur des valeurs des fonctions linguistiques), aussi bien en perception qu'en

production. L'unité phonologique de la prosodie ou prosodème est donc, dans ce modèle, un contour global émergent, qui permet d'accéder à la valeur fonctionnelle associée.

- **Principe de morphologie syntagmatique :**

  Chaque niveau délimité par la fonction de segmentation / hiérarchisation et recevant des valeurs fonctionnelles telles que démarcation, modalisation, focalisation ou encore attitudes, a sa propre morphologie, indépendamment des autres niveaux. Le contour global résulte de la superposition de ces contours indépendants.

- **Principe de morphologie paradigmatique :**

  Plusieurs valeurs fonctionnelles (par exemple démarcation et focalisation) peuvent être attribuées à un même segment, la morphologie associée à la combinaison de ces fonctions étant la superposition des différents contours (dirigés par les valeurs des fonctions) pour le même segment, au même niveau.

- **Principe d'organisation coopérative :**

  Ce principe se rapporte à la fonction expressive, également ancrée dans le domaine temporel linguistique, mais dont l'encodage est indirect. Les fonctions linguistique et pragmatique guident le système communicatif. Les agents, y compris l'agent prosodique, coopèrent en fonction de stratégies définies pour construire le sens et véhiculer les valeurs fonctionnelles induites par le système communicatif. Cette coopération s'organise à travers des rendez-vous structurels entre les agents du système. La fonction expressive, quant à elle, est implémentée par la stratégie de contrôle du choix de ces rendez-vous structurels.

- **Principe d'intégration des évènements extra-linguistiques :**

  Ce dernier principe concerne la fonction émotionnelle, directement encodée et ancrée dans le domaine temporel des évènements qui causent l'apparition de l'émotion. La fonction émotionnelle est partagée entre l'agent prosodique et d'autres agents tels que celui des expressions faciales. Ces valeurs fonctionnelles sont associées à des schémas prosodiques qui ne sont pas contrôlés sur le domaine d'un segment linguistique (défini par la fonction de segmentation / hiérarchisation), mais contraints par ces segments pour ne pas détruire les valeurs linguistiques portées par la prosodie.

4.5.3 C-Clone : une modélisation de la communication par système coopératif

Le modèle prosodique d'Aubergé s'inscrit dans une modélisation cognitive globale du système communicatif, le modèle C-Clone (Aubergé *et al.*, 2006). Ce modèle, illustré par la Figure 4, est une architecture cognitive, de type système multi-agents, proposée pour décrire le processus de communication d'un humain, mettant en synergie cooperative auto-organisée les structures organisationnelles de tous les matériaux impliquées dans la communication sur le principe d'un contrôle fonctionnaliste. Ces fonctions communicatives sont organisées en priorité selon les buts de l'interaction, et sont globalement émergentes du système C-Clone. C-Clone, contrairement aux architectures cognitives fodoriennes ou de type Levelt, n'est ni une structure à gouvernement centralisé, ni une organisation de type « pipe-line ». Chaque module (phonologie, lexique, morphologie, syntaxe, prosodie, gestualité) est construit dans une morphologie autonome, selon des degrés de liberté et des contraintes qui lui sont spécifiques, et dépend des autres modules à travers le contrôle des fonctions globales, répartie en auto-organisation dans les modules. Les traces saillantes de cette répartition sont par exemple les rendez-vous structurels décrits par Aubergé (1991) pour les fonctions de segmentation/hiérarchisation de l'énonciation (segmentation syntagmatique et ordonnancement paradigmatique), et de modalisation, et sont prises en charge au moins par la syntaxe, la prosodie, la gestualité. La fonction de focalisation est réalisée également par la syntaxe pour certaines langues (processus d'extraction), la morphologie et le lexique (adverbes intensifieurs, choix lexicaux), la prosodie (Aubergé & Rilliard, 2006) et la gestualité. La fonction d'expression des attitudes relève également de la coopération de ces mêmes modules.

Le système communicatif est donc considéré comme l'interaction de l'ensemble des agents linguistiques autonomes, y compris l'agent prosodique, sans qu'une quelconque hiérarchisation entre ces agents ne soit établie. Ce modèle va ainsi à l'encontre de la conception classique d'une prosodie guidée par les fonctions syntaxiques, en considérant au contraire que les fonctions du système communicatif sont globales à ce système et réparties de manière interactive entre les agents du système. La prosodie est donc vue comme un agent autonome en interaction coopérative avec les autres agents du système communicatif. Cette hypothèse d'une autonomie partielle de la prosodie est notamment étayée par une étude développementale menée par Morgan et Demuth (1996), qui ont mis en évidence un amorçage (*bootstrapping*) de la prosodie sur la syntaxe : chez l'enfant,

les fonctions linguistiques de la prosodie sont en effet maîtrisées avant la réalisation syntaxique des mêmes fonctions, et agissent comme « déclencheur » de ces fonctions.

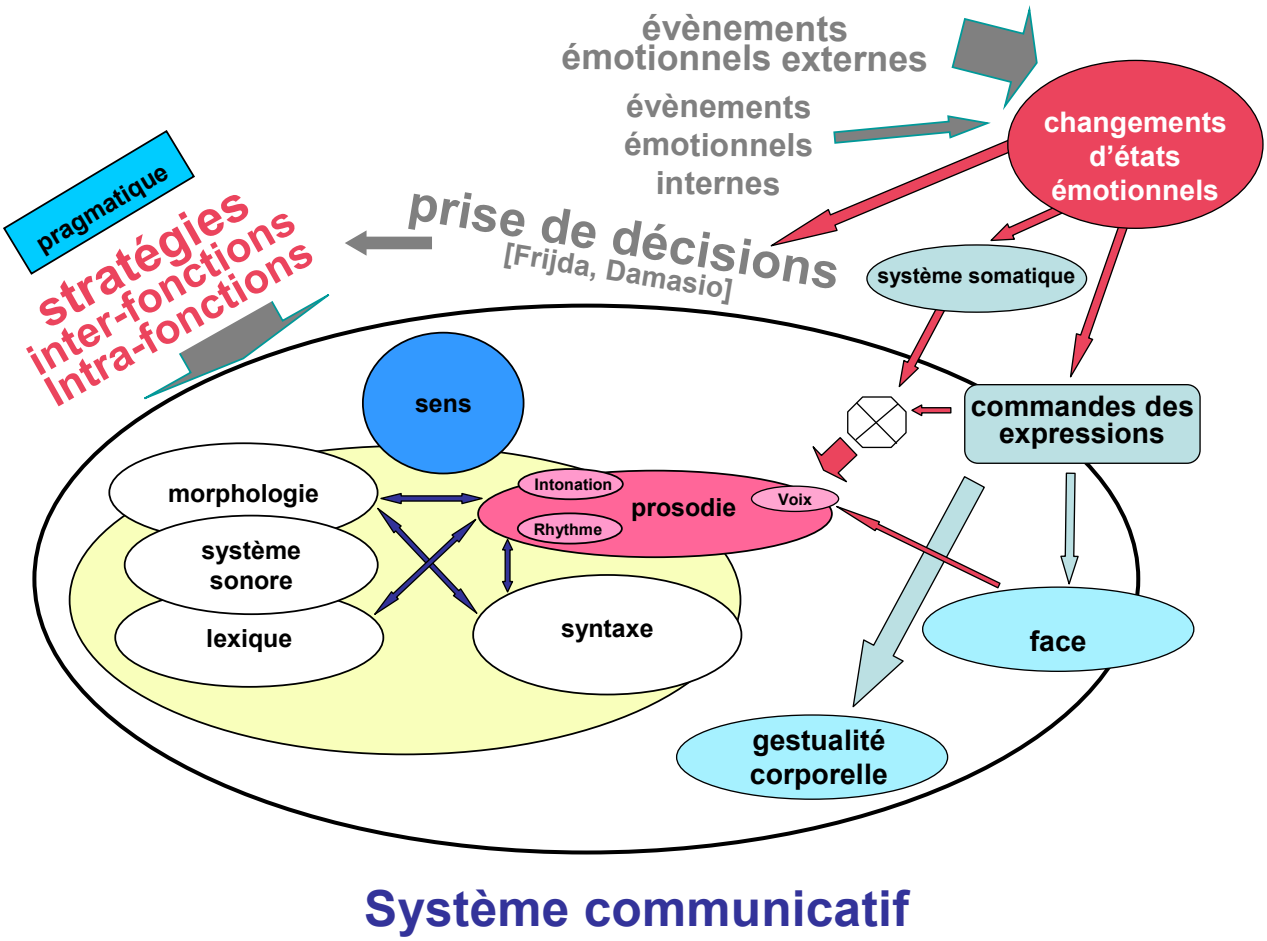


Figure 4 : Représentation des modules impliqués dans le modèle cognitif C-Clone et de leurs interactions, d'après Aubergé *et al.* (2006)

5. Méthodologie de recueil de données

Dans cette section, nous introduisons tout d'abord les différentes méthodologies permettant l'étude expérimentale de la prosodie expressive, ainsi que notre choix de nous appuyer sur des corpus supposés denses en affects du niveau visé. Nous dressons un état de l'art des différentes méthodes existantes de recueil de corpus de parole émotionnelle, en soulignant leurs avantages et inconvénients. Nous justifions le choix que nous avons fait de nous orienter vers l'induction d'affects en laboratoire afin de concilier spontanéité de l'expression et contrôle expérimental.

Nous présentons ensuite la plateforme logicielle E-Wiz que nous avons développée afin d'induire et recueillir des expressions authentiques d'affects, selon des scénarios de Magicien d'Oz dans lesquels la tâche prétexte imite le fonctionnement d'un système fondé sur la reconnaissance de commandes vocales. Nous détaillons les deux scénarios successifs mis en place à l'aide de cette plateforme, et plus particulièrement le scénario Sound Teacher. Ce scénario imite un logiciel d'aide à l'apprentissage de la prononciation de voyelles de langues étrangères. Il a pour but la collecte d'expressions d'affects produites sur des énoncés pour la plupart monosyllabiques, afin de geler les variations de la prosodie linguistique, hors de toute interaction sociale afin de favoriser les expressions d'émotions contrôlées involontairement.

Nous présentons alors l'extension de ce scénario, proposée pour les sujets acteurs, afin de leur faire rejouer sur les mêmes énoncés les états émotionnels ressentis et exprimés au cours de l'expérience.

Enfin, nous discutons les différentes approches proposées dans la littérature pour l'annotation des affects exprimés dans les corpus spontanés, et proposons, en nous appuyant sur la littérature relative à la mémoire autobiographique, un protocole d'auto-annotation des affects ressentis et exprimés par les sujets de l'expérience.

5.1. Une méthodologie hypothético-inductive

Le modèle prosodique d'Aubergé (1991, 2002) sur lequel nous nous appuyons (cf. section 4), considère que la morphologie prosodique résulte de la superposition de contours multiparamétriques (Gestalts). Ces contours correspondent aux différentes fonctions réalisées par la prosodie.

Dans l'optique d'une étude des expressions prosodiques des affects, il nous faut décider de la démarche méthodologique à adopter :

- Une méthodologie déductive, fondée sur la construction et l'évaluation d'énoncés construits à partir du modèle choisi. Dans le cadre de l'étude de la parole expressive, cette démarche a été adoptée par divers auteurs. Par exemple, Ní Chasaide & Gobl (2003) ont généré des stimuli synthétiques porteurs de différentes qualités de voix, avant l'évaluation perceptive de leur contenu affectif.
- Une méthodologie hypothético-inductive, qui repose sur l'analyse de corpus naturels supposés « denses » en phénomènes prédits par le modèle. C'est par exemple la démarche adoptée par Scherer *et al.* (1984) qui ont enregistré des interactions dans un contexte propice à l'expression d'affects.

Deux arguments majeurs nous ont conduit à trancher en faveur d'une méthodologie hypothético-inductive plutôt que déductive.

En premier lieu, l'intégration des domaines temporels émotionnel et linguistique n'est pas résolue dans le modèle. Ceci ne nous permet pas de décider de règles de génération d'énoncés porteurs d'expressions émotionnelles et en accord avec le modèle prosodique que nous utilisons. En second lieu, cette démarche permet d'éviter de produire et analyser des énoncés qui ne correspondent pas à une écologie naturelle, mais qui sont pourtant susceptibles d'être très bien acceptés par les auditeurs. Ce phénomène peut être défini comme le problème de « l'artefact acceptable ». Dans le cadre d'expériences d'évaluation perceptive dans lesquelles la tâche n'est pas définie de manière extrêmement stricte, Morlec *et al.* (1998) ont ainsi montré que les auditeurs acceptent facilement des stimuli synthétisés à partir des variations mélodiques d'un énoncé, appliquées à un autre énoncé ne partageant pas la même structure syntaxique, bien que ces stimuli constituent de véritables « chimères prosodiques ».

Le problème auquel nous sommes confronté dans le cadre de cette méthodologie hypothético-inductive est d'adopter une démarche qui nous permette d'évaluer sans risque de confusion les morphologies prosodiques des trois niveaux d'affects considérés. Nous optons pour une démarche, classique en phonétique expérimentale, qui est de construire des corpus dans lesquels les phénomènes à observer sont concentrés et, surtout, isolés des autres phénomènes qui pourraient interférer. Il s'agit donc de ne faire varier dans les corpus collectés que les phénomènes que nous cherchons à étudier, suivant le principe éprouvé en sciences expérimentales du *ceteris paribus* (toutes choses égales par ailleurs).

Étant donné que nous considérons, dans le cadre des hypothèses formulées par Aubergé (2002, voir également Aubergé *et al.* (2006)) les différents niveaux d'affects comme encapsulés (autrement dit le niveau le plus élevé d'expressions des affects, celui de l'expressivité langagière, inclut les expressions directes volontaires et involontaires), la construction des corpus se fait en suivant cette encapsulation ascendante. La première étape est donc de recueillir des corpus d'expressions directes des émotions en gelant autant que possible la variation des expressions linguistiques et métalinguistiques. Plus spécifiquement, notre objectif sera ici de recueillir en priorité des expressions involontaires d'émotions dans la parole, exprimées dans la mesure du possible sur des énoncés dont le contenu phonétique est comparable.

5.2. Capture de corpus d'expressions émotionnelles

5.2.1 Les méthodes de recueil

Les méthodes de recueil de corpus en phonétique sont séparables en deux grandes classes, qui peuvent être comparées aux différentes méthodes de prélèvement adoptées par les biologistes. Le prélèvement peut en effet être effectué des deux manières suivantes :

- *In vivo*, sur des organismes plongés dans leur écologie naturelle, sélectionnés en fonction de la théorie qui est évaluée. Cette première méthode correspond en phonétique à un enregistrement de parole spontanée, en allant enregistrer sur le terrain les phénomènes à étudier.
- *In vitro*, en cultivant les organismes en laboratoire, afin de contrôler au maximum les conditions expérimentales. Cette méthode, qui peut généralement se rattacher à la méthodologie déductive, correspond à l'enregistrement contrôlé de corpus en laboratoire.

Pour ce qui est de l'étude de phénomènes prosodiques, qui constituent notre objet d'étude, la méthode la plus couramment utilisée est la fabrication de corpus *in vitro.* Le locuteur reçoit, dans ce cas, des consignes précises sur les caractéristiques de la parole à produire. Ceci consiste typiquement en des tâches de lecture de phrases sélectionnées en fonction des phénomènes étudiés, comme des critères syntaxiques (Morlec, 1997), ou encore en des tâches de focalisation sur certaines parties d'un énoncé lu (voir par exemple Dohen *et al.* (2004)).

Quelques travaux étudient également la parole dite « spontanée », avec des données recueillies sans que le locuteur soit informé du véritable objectif de l'enregistrement. Toutefois, dans ce cas, les auteurs sont confrontés à une très grande variabilité des phénomènes observés. En effet, ces variations, attachées à certaines fonctions du langage qui ne constituent pas l'objet d'étude et peuvent ainsi dans un tel contexte être assimilées à du bruit de mesure, ne sont pas contrôlées dans un contexte de recueil *in vivo*. Ainsi, les variations de ces fonctions « indésirables » qui peuvent, pour établir une analogie avec le domaine du traitement du signal, être considérées comme du bruit dans la fonction observée, diminuent la densité des phénomènes étudiés. La probabilité est donc élevée, en ne voulant collecter que des expressions émotionnelles directes, de récupérer également des valeurs attitudinales et d'expressivité indirecte.

Il était donc naturel que la plupart des travaux sur la parole émotionnelle (parmi lesquels peuvent notamment être cités ceux de Banse & Scherer (1996)) s'appuient sur des corpus construits avec une tâche explicite de production d'émotions, en faisant appel aux compétences développées par le groupe socio-culturel des acteurs. Juslin & Laukka (2003) ont ainsi effectué une revue de 104 études relatives aux expressions vocales des émotions, et relatent que 87% de ces études se sont fondées sur des expressions émotionnelles simulées par des acteurs.

Aubergé & Cathiard (2003) ont cependant montré que les expressions d'amusement actées *vs.* spontanées pouvaient être perceptivement discriminées des expression non-actées par des auditeurs et, surtout, que cette distinction se faisait avec un effet inter-juges. Elles ont pour cela prolongé les travaux de Schröder (1998) qui a construit un corpus composé de productions spontanées et actée de l'amusement réalisées par les mêmes locuteurs (piégés lors de l'enregistrement d'un corpus pour obtenir des expressions spontanées d'amusement). En d'autres termes, certains juges parviennent systématiquement à discriminer les productions actées des productions non-actées tandis que d'autres n'y parviennent jamais, et ceci quelles que soient les capacités des locuteurs à simuler des expressions authentiques. La conséquence principale de cet effet inter-juges, essentielle dans l'optique d'un recueil de données émotionnelles

représentatives, est que les expressions émotionnelles simulées peuvent toujours différer des expressions authentiques sans qu'il soit possible d'évaluer avec certitude cette différence par des évaluations perceptives.

Par ailleurs, Damasio propose une hypothèse de boucle de simulation (cf. section 2) qui impliquerait une différence d'état intérieur du locuteur entre les émotions réellement ressenties et les émotions simulées (Damasio, 1994). Il s'agit donc de savoir si la participation, ou non, de l'ensemble du corps du locuteur à l'expression d'un état affectif entraîne aussi des différences sur l'expression vocale de l'amusement. Il est possible de se demander si l'expression de l'amusement spontané qui se traduit par un « sourire de Duchenne » (Ekman *et al.*, 1990)) induit d'autres changement sur la qualité vocale de la parole, au-delà des modifications liées à une déformation mécanique de la géométrie des lèvres relevées par Tartter (1980). Cette dernière étude relève en effet que les déformations du conduit vocal induites par le simple geste de sourire sont perceptibles dans la voix, et a pu mettre en relation au niveau acoustique la présence du sourire avec une élévation des valeurs du second formant (Tartter & Braun, 1994).

La question de savoir si les acteurs utilisent cette compétence de simulation décrite par Damasio ou s'ils sont capables de reproduire pendant leur jeu le processus physiologique mis en action lors de la survenue d'une émotion authentique reste ouverte. Il reste néanmoins vraisemblable que ces acteurs, enfermés dans une chambre sourde et dans un contexte inhabituel, ne soient pas en mesure de reproduire une performance en tous points identique à ce que serait une production spontanée de la même émotion ou de façon plus générale du même affect. De plus, nous pouvons supposer que les processus mis en œuvre dans le jeu d'acteur varient en fonction de l'apprentissage, et, bien entendu, en fonction des individus. Plus généralement, il semble impossible de tirer des conclusions définitives quant à la parole expressive actée sans étudier les différentes techniques d'acteurs selon les époques les cultures, ainsi que selon les contextes de jeu (théâtre, cinéma « vérité », ligue d'improvisation etc.). Il convient simplement de noter que, à notre connaissance, aucune étude n'a pu proposer d'évaluation fiable de la similitude de la performance de l'acteur avec des productions involontaires.

Afin d'assurer la représentativité des expressions émotionnelles recueillies, il est donc nécessaire de collecter des données spontanées. Cependant, comme le souligne Campbell (2000), le recueil de données spontanées ne doit pas se faire au détriment de la qualité des conditions d'enregistrement, qui doivent permettre d'effectuer des analyses acoustiques fines. Il nous faut en conséquence développer des méthodes permettant de concilier ces deux contraintes.

Nous dressons dans cette partie un état de l'art des méthodes de recueil de corpus de parole émotionnelle, en tentant de mettre en lumière les avantages et les limites de chacune et en présentant quelques exemples représentatifs. En raison du grand nombre d'études dans ce domaine, ce tour d'horizon n'a cependant pas la prétention de l'exhaustivité. Pour une revue plus large, il est possible de se référer à Douglas-Cowie *et al.* (2003) et Scherer (2003).

5.2.2 Les émotions *in vivo*

Les travaux expérimentaux qui s'appuient sur des corpus collectés en milieu écologique sont assez peu nombreux. En effet, comme le souligne Campbell (2000), ce type de protocole présente des limites conséquentes. D'une part, il est difficile de recueillir dans ces conditions des échantillons de qualité suffisante pour une analyse acoustique fine. D'autre part, un tel protocole ne permet pas de contraindre l'accès lexical du locuteur. Il est alors moins évident d'effectuer des comparaisons entre locuteurs et entre émotions. Quelques études se sont néanmoins appuyées sur des expressions émotionnelles spontanées recueillies à l'occasion d'événements comme des catastrophes. Williams & Stevens (1972) ont ainsi utilisé l'enregistrement radiophonique du commentaire en direct d'un accident de dirigeable, et ont comparé la version originale de cet enregistrement à celle produite par un acteur. Protopapas & Lieberman (1997) ont pour leur part utilisé l'enregistrement d'un pilote d'hélicoptère sur le point de s'écraser afin d'étudier l'influence du stress sur la fréquence fondamentale.

En dépit de la perte de qualité induite par ce média, plusieurs études se sont appuyées sur des extraits de parole expressive issus de conversations téléphoniques pour des applications de reconnaissance automatique des émotions. Une part importante des enregistrements étudiés étaient notamment constitués d'appels d'usagers vers des centres d'aide et de support. Ainsi Ang *et al.* (2002) ont utilisé des expressions d'ennui et de frustration obtenues à partir d'interactions d'usagers appelant un système de planification de voyages. Vidrascu & Devillers (2005) ont quant à elles travaillé sur un corpus de 10 heures de parole représentant 404 conversations téléphoniques entre usagers et agents d'un centre d'appel spécialisé dans le conseil médical, dont une part importante a été étiquetée comme des expressions d'anxiété.

Des extraits de parole émotionnelle radiophonique ou télévisuelle ont également été utilisés. Trouvain & Barry (2000) ont ainsi analysé des enregistrements radiophoniques de commentaires de courses de chevaux. Douglas-Cowie *et al.* (2000) ont enregistré des conversations informelles et également extrait des séquences télévisuelles au cours desquelles des membres du public interagissent de façon spontanée.

Chung (2000) a pour sa part utilisé des enregistrements d'entretiens télévisés au cours desquels des locutrices coréennes ainsi qu'une locutrice anglaise, s'exprimant initialement sur un ton enjoué, étaient ensuite invitées à parler de leurs problèmes personnels et exprimaient de la tristesse. En raison du caractère inattendu pour les locutrices des questions posées par le présentateur, l'auteur a considéré les émotions exprimées comme authentiques. Des entrevues radiophoniques et télévisuelles similaires ont été également été recueillies par Greasley *et al.* (1995).

Une réserve peut toutefois être émise quant à l'authenticité de la parole expressive recueillie dans un type de contexte médiatisé. En effet, qu'il y ait eu ou non une répétition préalable, il est peu vraisemblable que le discours d'un locuteur placé dans ces conditions soit dénué d'intentions pragmatiques. Ce problème peut s'accentuer encore selon le thème du discours qu'il doit produire. Se préparer à parler de ses problèmes personnels, revient indirectement à se préparer à transmettre des émotions négatives (tristesse, colère, etc.). Dans de telles conditions, la question d'une tendance du locuteur à devenir acteur amateur peut se poser.

D'une manière plus générale, l'influence de l'observateur sur la situation observée a été mise en avant en sociolinguistique, sous le nom de « paradoxe de l'observateur » par Labov (1970). Lors des campagnes d'enregistrement qu'il a effectuées, il a remarqué que les locuteurs étudiés, se sachant observés, adoptaient une façon de parler plus formelle. Labov a en conséquence énoncé le paradoxe de l'observateur comme la contradiction entre le fait que l'objet d'étude de la linguistique devrait être la façon dont les gens parlent lorsqu'ils ne se savent pas systématiquement observés, alors même que la seule façon d'obtenir ces données soit l'observation systématique.

Batliner *et al.* (2003a, p. 10) proposent une relecture de ce paradoxe de l'observateur adaptée à l'étude des expressions émotionnelles, aux conclusions plutôt pessimistes pour ce champ de recherche. Les auteurs mettent ainsi en avant trois « paradoxes émotionnels » :

- « Plus vous pouvez observer d'émotions, moins il est probable que vous ayez le droit de le faire *(problème des données clairsemées)* »[3]. Le problème de la spontanéité qui se pose en sociolinguistique se superpose dans l'étude des émotions à un problème éthique d'intrusion de l'observateur dans l'intimité des sujets. Les auteurs

[3] Traduction personnelle de la citation originale : *"The more emotions you could observe, the less likely it is that you are allowed to do so (sparse data problem)."*. Les auteurs n'explicitent pas le choix, surprenant puisque le problème ici se situe plus au niveau de l'accès aux données que de leur densité, de désigner ce premier paradoxe par le terme de "*sparse data*".

considèrent néanmoins que, bien que cela revienne à renoncer à recueillir des corpus d'expressions émotionnelles spontanées de grande taille, ce problème peut être contourné en se concentrant sur des situations particulières susceptibles de favoriser l'apparition d'émotions et leur expression sans violer l'intimité des locuteurs.

- « Plus vous êtes autorisé à observer des émotions, moins il est probable qu'elles soient exprimées de façon claire et simple. *(problème de la référence vague)* »[4]. Les émotions induites dans ces situations particulières seraient, selon eux, peu susceptibles d'être exprimées de façon aussi manifeste qu'elles le seraient dans la sphère privée. Les auteurs arrivent alors à différentes conclusions. D'une part, il est nécessaire de prendre en compte la complexité des expressions émotionnelles recueillies dans ces conditions, plutôt que de tenter de contourner ce problème. D'autre part, le degré d'accord inter-annotateurs quant aux émotions exprimées dans les données recueillies ne peut en conséquence demeurer l'unique critère d'évaluation de ces expressions.
- « Plus vous aurez pu modéliser d'émotions pures, moins il est probable que celles-ci soient pertinentes dans le cadre d'applications réalistes. (*problème de l'acceptabilité)* »[5]. Au-delà de la difficulté pour collecter des expressions spontanées de telles émotions et les modéliser, les auteurs considèrent qu'il n'existe pas de projet convainquant d'application pour laquelle une telle modélisation serait utile car correspondant à un besoin ou un désir exprimé par des utilisateurs.

Ces positions méritent toutefois d'être nuancées. Nous pouvons effectivement considérer comme peu probable le fait de collecter des expressions intenses de joie, de tristesse ou encore de colère sans s'immiscer dans l'intimité du sujet, avec les problèmes d'ordre éthique que cela implique. Il est cependant peu vraisemblable qu'une quantité importante d'expressions de ce que Batliner *et al.* (2003a) nomment « émotions pures » et qui sont à rapprocher des *Big Six* puisse être recueillie en observant les interactions de la vie de tous les jours, indépendamment du débat quant à la pertinence de telles émotions « de base » (voir notamment Sander (2008) pour une brève mais très informative discussion sur cette thématique). Ce type d'expression est en effet rare dans les interactions courantes : c'est notamment l'une des conclusions majeures de Campbell (2005), qui souligne que ses corpus d'expressions spontanées recueillis à très large

[4] Traduction personnelle de la citation originale : *"The more you are allowed to observe emotions, the less likely it is that they are indicated in a clear and simple way (vague reference problem)."*

[5] Traduction personnelle de la citation originale : "*The more pure emotions you eventually could model, the less likely it is that these are relevant for realistic applications (acceptability problem)."*

échelle, bien que riches en expressions variées d'affects, comportent remarquablement peu d'expressions claires des émotions couramment désignées sous le terme de *Big Six*.

La rareté de telles expressions dans l'interaction ne remet toutefois pas en question leur pertinence, et nous pouvons même affirmer que c'est précisément de cette rareté que découle la valeur informative de ces expressions. De plus, si l'intérêt d'une modélisation acoustique d'expressions prosodiques des seules *Big Six* peut être mis en cause pour des buts applicatifs tels qu'un système de synthèse restreint à un domaine dans lequel le sujet virtuel parlant n'est pas supposé exprimer de telles émotions (comme ce peut être le cas par exemple de la vente en ligne), d'autres applications qui bénéficieraient de cette modélisation peuvent aisément être imaginées sans apparaître comme irrémédiablement inutiles ou irréalistes. Ainsi un système de reconnaissance des émotions destiné à déclencher des alertes, tel que celui élaboré par Clavel (2007) pour la détection d'expressions intenses de peur, pourrait directement tirer parti d'une telle modélisation.

Campbell (2005) revient lui aussi sur le paradoxe de l'observateur de Labov en considérant que si la simple présence d'un dispositif d'enregistrement modifie la façon de parler des locuteurs, ce paradoxe peut néanmoins être dépassé. Campbell estime ainsi qu'il est possible d'imaginer des protocoles pour recueillir des corpus de parole émotionnelle spontanée telle qu'elle peut être observée dans la vie de tous les jours, et l'illustre en mettant en œuvre à grande échelle de tels protocoles.

Une large collecte de parole expressive spontanée a ainsi été orchestrée dans le cadre du projet Expressive Speech Processing (ESP). Parmi les deux corpus de parole expressive spontanée recueillis (Campbell, 2003, 2004), le plus court consiste en des conversations téléphoniques (mais enregistrées simultanément en haute qualité) hebdomadaires sur une période de trois mois entre des locuteurs japonais et des proches, des inconnus de langue japonaise et des locuteurs chinois et anglais. Le second corpus du projet ESP consiste en l'enregistrement sur une période de plusieurs années des conversations à la fois d'ordre privé et professionnel de 12 locuteurs japonais (10 femmes et 2 hommes). Ces enregistrements ont été réalisés à l'aide d'un dispositif miniaturisé porté en permanence par les locuteurs et permettant d'obtenir une haute qualité d'enregistrement.

Les travaux de Scherer *et al.* (1984) peuvent également être mentionnés pour leurs études où ils se sont attachés à collecter des échantillons de parole émotionnelle en milieu écologique tout en assurant un certain contrôle de la qualité d'enregistrement. Ils ont enregistré dans un studio d'enregistrement aménagé en bureau les interactions verbales d'employés de l'assistance sociale en entretien avec des acteurs non-professionnels.

Scherer et Ceschi (2000) ont également filmé à leur insu 110 passagers déposant une réclamation à l'aéroport quant à la perte de leur bagage.

Bien que n'étant pas systématiquement dédiés à l'étude des expressions émotionnelles, des corpus audiovisuels d'interaction en contexte de réunions professionnelles ont été recueillis. Pour cela, des salles de réunions spécialement équipées pour permettre l'enregistrement synchrone des productions vocales, des expressions faciales et de la gestualité de l'ensemble des participants, ont été utilisées, notamment par Campbell *et al.* (2006) et par Carletta (2006) avec le corpus AMI.

5.2.3 Les expressions actées des émotions

En dépit des réserves pouvant être soulevées quant à la validité de la parole actée pour la modélisation de phénomènes relatifs à la parole expressive spontanée, le recours à des locuteurs simulant l'expression vocale d'émotions en fonction des instructions données par les expérimentateurs a été de loin la méthode la plus utilisée, en raison de sa mise en œuvre plus aisée. En effet, si le contrôle de la qualité d'enregistrement et du contenu phonétique des énoncés dans le cas du recueil de productions spontanées d'expression émotionnelles peuvent s'avérer extrêmement problématique (Campbell, 2000), le recours à des acteurs permet de mettre en œuvre de façon beaucoup plus aisée des enregistrements en chambre sourde, avec du matériel d'enregistrement de haute qualité. De plus l'enregistrement de parole actée permet de contrôler le contenu phonétique des énoncés en vue d'analyses acoustiques et perceptives ultérieures. Ce type de corpus permet ainsi de faire varier l'émotion indépendamment de l'énoncé, selon le paradigme du contenu standard.

Les corpus de parole émotionnelle actée se sont pour la plupart appuyés sur un protocole d'enregistrement simple. Le ou les locuteurs avaient pour instruction de lire la ou les phrases présentées en « mettant le ton » d'un certain nombre d'émotions, les laissant libre du choix de la méthode d'acteur. Des exemples représentatifs de tels protocoles d'enregistrement peuvent être trouvés dans Dellaert *et al.* (1996) ou encore Petrushin (2000). La capacité des locuteurs à catégoriser correctement ce type de corpus a généralement été retenue comme un critère suffisant pour que les énoncés recueillis soient considérés comme des expressions émotionnelles valides. Or, la parole expressive actée est réputée stéréotypique voire caricaturale (Campbell, 2003), ce qui explique qu'elle soit bien reconnue, généralement bien au-delà des taux de reconnaissance qui peuvent être observés pour la parole expressive spontanée.

L'hypothèse d'une tendance moindre à l'exagération des expressions produites dans le cas de corpus produits par des locuteurs non-acteurs pourrait être émise, comme le

suggèrent par exemple Burkhardt *et al*. (2005). Néanmoins, dans leurs travaux, Wilting *et al.* (2006) puis Shahid *et al.* (2008), ont comparé en condition visuelle l'intensité émotionnelle véhiculée par les productions spontanées et actées de locuteurs (respectivement néerlandais et pakistanais) ne possédant pas d'expérience particulière du jeu d'acteur. Leurs conclusions ne vont pas dans le sens de productions actées moins stéréotypées lorsque les locuteurs n'ont pas de formation d'acteur, ceux ci auraient en effet, à l'inverse, tendance à surjouer.

Campbell (2005, p. 115) se montre particulièrement sceptique quant à l'authenticité de tels corpus de parole émotionnelle :

« Lorsque [une prosodie « émotionnelle » de ce type] est actée ou produite à la demande, elle n'est pas exprimée comme un énoncé contextualisé ou situé, mais simplement générée en tant qu'échantillon. Cela peut fournir de bonnes données, mais il ne s'agit pas d'un corpus dont on peut tirer des connaissances. [Ces données] ne sont pas authentiques, n'adviennent pas naturellement, ne sont probablement même pas représentatives d'une parole située normale, et ne nous aident pas à étudier le « langage dans son usage » puisqu'elles n'ont jamais été « utilisées ». »[6]

Néanmoins, un nombre très élevé d'études dédiées aux expressions d'émotions dans la parole continuent à se fonder sur des productions actées. Les auteurs s'efforcent alors de contourner les limites inhérentes aux émotions actées, afin de s'approcher de l'écologie de l'acteur, en ayant recours de façon plus ou moins explicite à des méthodes d'acteurs « réalistes » comme celle développée par l'Actor's Studio (voir également Stanislavski (1963, 1966)).

Bänziger & Scherer (2007) reviennent ainsi sur les principales critiques émises à l'encontre des expressions actées des émotions. La plus fréquemment émise de ces critiques est que ces expressions reflèteraient des stéréotypes plutôt des émotions authentiques. De plus, les acteurs ne reproduiraient que les indices les plus saillants associés à l'expression d'une émotion donnée sans parvenir à en reproduire toute la subtilité.

---

[6] Traduction personnelle de la citation originale : *"When [this type of 'emotional' prosody] is acted or produced at a prompt, it is not expressed as a contextualised or situated utterance, but simply generated as a sample. It may be good data, but it is not part of a corpus that we can learn from. It is not authentic, not naturally occurring, probably not even representative of normal situated speech, and does not help us to study 'language in use' since it has never been 'used' […]."*

Selon Bänziger & Scherer (2007), le fait d'encourager les acteurs à réactiver des expériences émotionnelles passées leur permettrait plus facilement de produire des expressions similaires à des expressions spontanées. Les auteurs préconisent une évaluation de la crédibilité des expressions produites plutôt qu'une validation fondée sur la catégorisation correcte des émotions exprimées.

Les auteurs insistent en outre sur le fait que, bien que les expressions actées soient tenues pour hypercontrôlées et insuffisamment spontanées par leurs détracteurs, les expressions émotionnelles observables dans la vie de tous les jours sont également fortement contrôlées, notamment par des stratégies de présentation de soi. Ainsi Banse & Scherer (1996) considèrent que seuls les « éclats d'affects » (*affect bursts*) relèvent de façon exclusive d'une externalisation d'états internes (effets *push*). Ces éclats d'affects sont définis par Scherer (1994, p. 170) comme « des expressions non verbales d'affects à la fois dans la face et la voix, déclenchées par des événements clairement identifiables »[7].

Toute vocalisation d'affect reflétant une émotion authentique résulterait alors, conjointement, de ces effets *push* et de contraintes normatives sur l'expression des affects spécifiques à la culture du sujet parlant (effets *pull*). Banse & Scherer (1996) considèrent en conséquence que la plupart des expressions d'affects réputées naturelles sont aussi en partie actées, et que les expressions actées sont en partie naturelles, le degré de naturalité dépendant des stratégies de production du locuteur.

Bänziger & Scherer (2007) estiment également que la critique portant sur le supposé manque de représentativité des expressions émotionnelles actées peut être contournée en ne se focalisant pas exclusivement sur le sous-ensemble que constitue les émotions supposées basiques (et notamment les *Big Six*), exprimées de plus uniquement sous leur forme la plus intense. Ils suggèrent ainsi d'étendre les corpus d'expressions émotionnelles actées à des émotions moins intenses et non limitées aux émotions « basiques », choisies notamment en fonction du contexte de recherche. Enfin, ils insistent sur la nécessité de placer l'acteur dans un contexte propice à l'apparition des émotions visées à l'aide d'un scénario crédible, incluant, dans la mesure du possible, une interaction de l'acteur avec d'autres personnages.

La technique de l'élicitation, issue des méthodes d'acteurs, consiste à activer chez le locuteur la mémoire d'une expérience émotionnelle passée proche de celle à reproduire, afin qu'il puisse la simuler de la manière la plus naturelle possible. Parmi les techniques d'élicitation, une distinction peut être opérée entre l'élicitation grâce au corps et

[7] Traduction personnelle de la citation originale *''very brief, discrete, nonverbal expressions of affect in both face and voice as triggered by clearly identifiable events''*.

l'élicitation grâce à la mémoire émotive. Dans la première, fondée sur les travaux de Stanislavsky (1966), l'acteur adopte des postures destinées à l'aider à ressentir l'émotion souhaitée. La seconde technique se fonde sur un processus de remémorisation par les acteurs d'un événement ayant suscité chez eux une émotion très proche de celle qu'ils doivent exprimée.

Parmi les premiers corpus à se réclamer du recours à l'élicitation, citons tout d'abord celui enregistré par Mozziconacci (1998). Cette dernière a demandé à 3 acteurs néerlandais d'exprimer 13 émotions différentes sur 8 phrases sémantiquement neutres, en prononçant au préalable des phrases dont le contenu verbal émotionnellement chargé était supposé induire chez les locuteurs l'état émotionnel recherché. Par la suite, Burkhardt *et al.* (2005) sont allés légèrement plus loin dans l'utilisation de techniques d'élicitation. Ils ont donné aux acteurs quelques indications quant au contexte de l'émotion à exprimer, ainsi que l'instruction explicite de s'appuyer sur le souvenir d'épisodes émotionnels passés avec la possibilité de recourir à des gestes et postures à même de faciliter le ressenti de l'émotion à exprimer.

Enos & Hirschberg (2006) dressent un état de l'art des méthodes utilisées par les acteurs professionnels. Ils proposent un cadre méthodologique pour l'élicitation des émotions et le recueil de leurs expressions vocales. Ils suggérent deux approches distinctes en accordant une place centrale aux intentions et aux buts du personnage joué par l'acteur, l'approche par script et l'approche par scénario.

La première consiste à s'appuyer sur des pièces de théâtre éprouvées et bien connues des acteurs, afin de tirer parti de la familiarité des acteurs avec le contexte et les buts poursuivis par les personnages impliqués.

La seconde revisite des méthodes déjà utilisées auparavant pour la collecte de corpus de parole émotionnelle. Elle met en œuvre des descriptions détaillées de la situation, qui incluent les possibles obstacles à la réalisation du but poursuivi par le personnage incarné par l'acteur, ainsi que de l'ensemble des personnages impliqués. Dans une telle approche, le personnage joué par l'acteur doit tendre vers un but particulier spécifié de façon précise par le scénario. Le texte lui permettant d'atteindre ce but doit également être prédéfini. Ces différents éléments sont ici vus comme les composantes d'une équation qui doivent être réglées méticuleusement par le chercheur assisté d'un metteur en scène compétent, pour aboutir à l'expression par l'acteur de l'émotion souhaitée. Cependant les auteurs insistent sur le fait que dans cette approche, l'émotion souhaitée ne doit pas être indiquée à l'acteur. Ce dernier doit se concentrer sur l'action à effectuer pour accomplir le but défini par le scénario et non sur une émotion à exprimer. Enos & Hirschberg (2006) suggèrent deux méthodes de validation *a priori* des émotions élicitées

au moyen d'une approche par scénario. Il s'agit d'une part de l'évaluation par un panel de juges de l'émotion qu'ils éprouveraient à la place du personnage dans une telle situation, ou de la construction du scénario à partir de véritables expériences, rapportées de façon convergentes avec leur contexte, par un groupe de sujets.

Des méthodes semblables aux approches proposées par Enos & Hirschberg (2006) ont été adoptées par certains auteurs. Bien que les auteurs ne donnent que peu de détails sur le protocole retenu, on peut considérer Williams et Stevens (1972) comme des pionniers de l'approche par script. Ces derniers ont en effet eu recours à un metteur en scène et à trois acteurs, anciens membres de l'Actor's Studio, dans de courtes pièces mettant en scène trois personnages masculins. Ces pièces étaient spécifiquement écrites pour permettre l'expression de peur, de colère et de tristesse par les acteurs, en incluant des répliques de contrôle commune aux différentes pièces.

Banse & Scherer (1996) ont élaboré, à partir de rapports individuels d'expériences émotionnelles récoltés au préalable, 14 scénarios détaillés destinés chacun à éliciter une émotion particulière. Ces scénarios étaient communiqués aux acteurs germanophones avec le texte-cible composé d'énoncés pseudo-linguistiques trois à sept jours avant la session d'enregistrement. Dans cette expérience, la durée d'enregistrement n'était pas contrainte, les acteurs étant libres de prendre le temps nécessaire pour parvenir à ressentir l'émotion visée. La différence majeure avec la méthodologie proposée par Enos & Hirschberg (2006) est que les scénarios proposés aux acteurs étaient, dans ce cas, explicitement liés aux émotions visées.

Bänziger & Scherer (2007) ont affiné, pour la conception et l'enregistrement du corpus audiovisuel GEMEP, la méthodologie mise en œuvre par Banse & Scherer (1996). Ce corpus, dans lequel les phrases-cibles pseudo-linguistiques ont été complétées par des séquences improvisées, a ainsi été élaboré à partir de scénarios faisant autant que possible référence à des partenaires en interaction avec l'acteur. Les acteurs professionnels francophones recrutés pour l'enregistrement de ce corpus ont, de plus, été encadrés par un metteur en scène expérimenté.

Busso & Naranayan (2008) ont enregistré 10 acteurs professionnels pour constituer le corpus audiovisuel USC IEMOCAP, en mettant l'accent sur la contextualisation des expressions produites. Les acteurs enregistrés se trouvaient toujours en situation d'interaction avec un autre acteur. Pour chacun des acteurs, deux approches ont été adoptées, semblables aux approches par scénario et par script proposées par Enos & Hirschberg (2006). Néanmoins dans les enregistrements correspondant à l'approche par scénario, que les auteurs ont qualifié de partie « spontanée » du corpus, le choix des

auteurs a été de définir des contraintes scénaristiques souples afin de permettre aux acteurs d'improviser.

La quasi-totalité des corpus d'expressions émotionnelles actées ont été évalués sur la base d'une catégorisation supérieure au niveau du hasard par des juges naïfs, ou alternativement, à partir de jugements experts de la capacité des acteurs à exprimer l'émotion visée, comme par exemple dans le cas du corpus GEMEP (Bänziger & Scherer, 2007). Or, la tendance des expressions émotionnelles actées à être plus intenses et stéréotypiques que les expressions spontanées (Campbell, 2003) a pour conséquence de produire des scores plus élevés dans les tâches perceptives de catégorisation ou d'identification des émotions exprimées comparativement à ceux obtenus dans le cas d'expressions spontanées. Ceci explique d'ailleurs, pour partie, l'utilisation fréquente qui est faite de corpus d'expressions émotionnelles actées. Des doutes quant à l'authenticité de ces émotions peuvent toutefois être émis, quand bien même le cas d'un hypothétique « acteur parfait » placé en situation optimale à l'aide d'un scénario approprié et préparé, d'une mise en contexte adéquate et d'une élicitation réussie, serait considéré. En effet, rien ne nous permet de nous assurer avec certitude que l'acteur simule l'émotion et non certaines caractéristiques de son expression, et encore moins que le même locuteur en situation de ressentir et exprimer la même émotion l'aurait exprimée de manière en tous points similaire.

5.2.4 L'induction d'émotions en laboratoire

Nous avons vu que ni la collecte d'émotions *in vivo*, ni les méthodes d'acteurs, n'apportent de solution pleinement satisfaisante. Afin de concilier authenticité des émotions et conditions d'enregistrement de laboratoire, il convient donc de pouvoir susciter ces émotions en laboratoire, mais dans des conditions semblables à celles du *in vivo*. Des techniques d'induction variées ont été utilisées à cet effet. Helfrich *et al.* (1984) ont ainsi administré des antidépresseurs à des sujets sains afin d'induire des états émotionnels positifs et d'en étudier les effets sur leur prosodie. Tolkmit & Scherer (1986) ont demandé à des locuteurs allemands de lire à voix haute des codes sans signification, tandis que leur étaient présentées des tâches complexes de résolution de problèmes logiques ainsi que des images de maladies de peau et d'accidents. Wilting *et al.* (2006) ont, quant à eux, adapté la méthode d'induction de Velten (1968) pour enregistrer des expressions spontanées d'émotions chez des locuteurs néerlandais naïfs. Le principe de cette méthode est d'induire chez les sujets des émotions positives (respectivement négatives) en leur faisant lire des phrases véhiculant une valence positive (respectivement négative) croissante, Un état de l'art plus complet des techniques

d'inductions émotionnelle appliquées au recueil d'expressions vocales est présenté par Scherer (2003).

Néanmoins, la majorité des études ayant eu recours à de tels protocoles d'induction se situent hors du domaine de l'étude de la parole, une part importante ayant eu recours à des vidéos comme stimuli d'induction. Ainsi Ekman *et al*. (1990), de même que Rosenberg & Ekman (1994), ont fait visionner des films plaisants ou déplaisants à des sujets, tout en filmant l'expression de leurs visages au moyen d'une caméra cachée. Une méthode similaire a été utilisée par Sinha (1996) pour l'étude des réactions somatiques des sujets à une induction émotionnelle. Une revue des techniques d'induction émotionnelle en laboratoire est donnée par Gerrards-Hesse *et al.* (1994).

D'autres techniques se fondent sur l'utilisation de la mémoire émotive, telle la méthode *imagination* utilisée par Johnstone & Scherer (1999) qui ont demandé aux sujets d'imaginer le ressenti d'une émotion donnée. Amir & Ron (1998) ont pour leur part demandé à des étudiants qui devaient garder les yeux fermés de raconter un événement émotionnel passé. Cependant, ce type de corpus présente l'inconvénient majeur de placer les sujets dans une situation de passivité : il leur est demandé de se concentrer uniquement sur eux même, sur leur vécu ou sur leur imagination, mais l'interactivité et la spontanéité ne sont pas présentes. Dans le but d'obtenir des émotions spontanées, nous préfèrerons en effet les situations interactives durant lesquelles le sujet participe à l'action.

De fait, ces méthodes où le sujet est soit passif, soit centré sur un monologue chargé émotionnellement, tendent à être supplantées par d'autres méthodes mettant l'accent sur l'interactivité. Dans ce but, les scénarios fondés sur le paradigme du Magicien d'Oz ont connu un certain succès. En effet, ces techniques, que nous présentons plus en détails dans la partie suivante, permettent de placer le sujet en situation de ressentir et d'exprimer des émotions sans qu'il soit conscient que cela constitue le but de l'expérience, par exemple un jeu vidéo ou des simulateurs divers (simulateur de conduite, de vol, etc.).

5.3. Le paradigme du Magicien d'Oz « perturbé »

5.3.1 Définition

Le paradigme du Magicien d'Oz, dans le cadre des recherches en Communication Homme-Machine, consiste à faire croire à un utilisateur qu'il utilise un système « intelligent », alors qu'il se trouve face à une simulation contrôlée par un compère, désigné par le terme de magicien. Dans un autre cadre, il est possible de s'appuyer sur ce

paradigme pour mettre en place des expériences d'induction. Il s'agit alors de placer un utilisateur face à une tâche cognitive lourde, afin que celle-ci accapare son attention, et d'introduire à son insu des perturbations dans cette tâche, afin de provoquer chez lui certaines émotions.

Ce paradigme tient son appellation du personnage du roman de L. Franck Baum dans lequel le magicien se fait passer pour une imposante statue, tandis qu'il ne s'agit que d'un petit homme dissimulé derrière un rideau pour contrôler cette statue. Ce paradigme a été baptisé ainsi par J.F. Kelley, qui l'a proposé pour la mise au point d'interfaces personne-machine en langage naturel (Kelley, 1983). Le paradigme du Magicien d'Oz est parfois également désigné par l'acronyme PNAMBiC[8] comme relevé par Fraser & Gilbert (1991).

5.3.2 Des magiciens d'Oz classiques …

Le paradigme du Magicien d'Oz, s'il est très largement utilisé à l'heure actuelle, a diverses applications. La plus répandue est l'évaluation d'interface, et les exemples d'une telle utilisation du paradigme du Magicien d'Oz sont légion. Citons en particulier le système Neimo qui consiste en une plate-forme générique pour l'analyse d'interfaces multimodales, grâce à la présence de plusieurs magiciens, et propose un cadre méthodologique détaillé (Salber & Coutaz, 1993). Fraser & Gilbert (1991) ont quant à eux étudié en détail le cas d'applications dans lesquelles un système de reconnaissance vocale doit être simulé.

Le Magicien d'Oz a également été étendu à l'étude de la multimodalité. Dans ce cas, ce sont les stratégies élaborées par le sujet pour exprimer un but communicatif lorsqu'il est placé dans des situations particulières qui sont étudiées. C'est le cas par exemple d'un corpus en langue suédoise (Lewin *et al*., 1999), portant sur des réservations de trains, hôtels, avions et taxis. Ce corpus a été enregistré afin d'évaluer le gain sémantique pouvant être apporté par des systèmes de reconnaissance incluant une analyse linguistique plus fine, au-delà des langages de commandes. Ce corpus a été réalisé par des demandes de reformulation successives, ce qui d'après les auteurs a permis d'obtenir des requêtes « naturelles ».

[8] Pay No Attention to the Man BehInd the Curtain (*Ne fais pas attention à l'homme derrière le rideau*)

5.3.3 … et perturbés

Le principe du Magicien d'Oz peut aussi être utilisé pour l'induction d'émotions. En particulier, il est possible de provoquer l'irritation du sujet en simulant un mauvais fonctionnement du système, sous la forme de ralentissements ou de réponses inappropriées. Toutefois, l'utilisation d'un Magicien d'Oz à cette fin suppose de disposer de sujets suffisamment « naïfs » pour être persuadés qu'ils se trouvent face à un véritable système, ou à l'inverse d'un système suffisamment réaliste pour que les sujets ne soupçonnent pas le caractère factice de la tâche proposée. De ce fait, il est nécessaire d'élaborer un scénario qui constitue un bon compromis entre les contraintes imposées par les conditions d'expérience à respecter et un comportement réaliste de l'application. De plus, et bien que les techniques de Magicien d'Oz soient considérées comme comptant parmi les plus adaptées à l'induction d'une large palette d'émotions spontanées, il peut être difficile d'induire chez le sujet des émotions telles que la colère. En effet, le sujet pense être en présence d'un véritable ordinateur, et l'expérience montre que les sujets sont plus coopératifs et, surtout, plus tolérants à des attentes répétées, dans ce type de situation que dans une situation de communication humain-humain (Batliner, 2003a).

Dans ce domaine de l'induction d'émotions, les exemples d'utilisation de protocoles de Magicien d'Oz sont moins nombreux et tous ne s'inscrivent pas strictement dans le cadre des simulations par ordinateur. L'expérience pionnière en ce domaine semble être celle menée par Hecker *et al.* (1968), qui ont mis en place un protocole afin de provoquer le stress chez des sujets enregistrés en chambre sourde, en leur donnant une tâche complexe de surveillance de voyants à effectuer, tout en jouant sur les délais de réponse accordés. Schröder (1998) a proposé à des locuteurs professionnels des tâches de lecture classiques pour eux. En les distrayant par des événements « amusants » inattendus, il a induit des expressions d'amusement spontané sur une tâche contrôlée de lecture. Les travaux de Fernandez & Picard (2000), qui ont mis en place un protocole permettant de collecter de la parole émotionnelle dans le cas du stress au volant, peuvent également être cités. Pour cette expérience, il était demandé aux sujets de résoudre par téléphone des questions de calcul mental tout en conduisant, la vitesse de conduite et la fréquence des questions étant fixées. De par le caractère extraordinaire de la situation dans laquelle le sujet est placé, il est néanmoins possible de s'interroger sur la mesure dans laquelle la parole collectée dans ce type de situation reflète les conditions normales de production (Campbell, 2000).

Klasmeyer *et al.* (2000) ont, quant à eux, utilisé des tests de logique dans lesquels l'état émotionnel du locuteur, à qui il était demandé de lire des phrases sémantiquement

neutres, était manipulé en simulant des dysfonctionnements ayant un impact négatif sur son score ou au contraire en le surévaluant. Une expérience de ce type a également été réalisée par Loyau (2003), afin d'étudier la parole émotionnelle chez les pilotes de chasse. Pour cela, le simulateur de vol a été modifié afin de pouvoir perturber à distance les commandes du sujet et induire ainsi émotions et attitudes, tout en restant dans le cadre du langage de commandes dédié à la communication entre pilotes. Un Magicien d'Oz a également été utilisé par Kehrein (2002) pour induire et collecter des expressions émotionnelles authentiques. Dans cette expérience, plusieurs participants, qui devaient collaborer pour mener à bien un jeu de construction, ont été enregistrés au cours de leurs interactions. Egalement dans un contexte ludique, Johnstone *et al.* (2005) ont induit des expressions vocales d'affects au moyen d'un jeu vidéo.

Mishra *et al.* (2004) ont enregistré dans un simulateur de conduite 12 pilotes en interaction avec un soi-disant système de reconnaissance de la langue naturelle leur permettant de contrôler l'appareillage électronique de l'habitacle, afin de recueillir des échantillons de parole expressive authentique en situation de stress. Batliner *et al.* (2004) ont enregistré les interactions d'enfants anglais et allemands avec un robot-chien AIBO®, en incluant une condition dans laquelle le robot suivait un scénario prédéfini plutôt que d'obéir aux ordres de l'enfant.

Plus récemment, Février *et al.* (2006) ont élaboré un protocole de Magicien d'Oz mettant en jeu des sujets placés en situation de vidéo-conversation avec un vendeur complice de l'expérimentateur. Les scénarios suivis par les complices ont permis d'induire chez les sujets des réactions de satisfaction, d'amusement, d'embarras, d'incompréhension et de surprise.

5.3.4 Quelle tâche prétexte, quelles perturbations pour quelles émotions?

La forme générale que nous avons retenue pour le développement de scénarios d'induction de variations émotionnelles est un Magicien d'Oz imitant le fonctionnement d'une application à commandes vocales (Audibert *et al.*, 2003, 2004). Toutefois l'usage que nous faisons de la technique du Magicien d'Oz s'écarte du rôle qui lui est classiquement dévolu : en effet notre objectif ici n'est pas de collecter des données directement liées à la tâche, mais relatives aux effets induits par la tâche proposée sur les locuteurs

D'autre part, le choix de la tâche proposée est primordial pour assurer le succès des scénarios d'induction. Cette tâche doit être en adéquation avec le profil psychologique des sujets, afin de faire en sorte que leur motivation pour passer l'expérience prétexte

provienne de la nature de la tâche elle-même, dans laquelle les sujets se sentent réellement impliqués. Cela exclut donc les scénarios dans lesquelles la motivation est suscitée par l'aspect ludique de la tâche, ou par une promesse de rémunération comme cela a été appliqué dans d'autres études telles que (Kaiser & Wehrle, 1994) ou (Kehrein, 2002). Un effort particulier doit en outre être consacré à la crédibilité de la tâche.

Cet objectif de motivation par la tâche est toutefois compliqué par les conditions expérimentales, susceptibles de perturber le sujet et par là même de le détourner de cette motivation. En effet, les mesures de données diverses au cours des enregistrements, quoique indispensables à toute analyse ultérieure, peuvent également s'avérer être un facteur perturbant (Campbell, 2000). Il convient donc de trouver un équilibre entre les contraintes imposées par l'ensemble des appareils de mesure et la naturalité de la tâche, c'est-à-dire sa crédibilité par rapport à l'écologie quotidienne du locuteur, afin d'éviter l'écueil de retirer de l'information (les variations émotionnelles du locuteur, qui sont le but recherché) en cherchant à en ajouter par des mesures plus complètes. Outre l'enregistrement acoustique, pour lequel l'utilisation d'un microphone se justifie aisément par le prétexte de l'utilisation d'une application à commandes vocales, et sauf à induire des émotions très intenses, le choix des mesures à effectuer doit en conséquence se limiter à des instrumentations peu invasives.

5.4. Une plate-forme dédiée : E-Wiz

Afin de permettre l'implémentation de scénarios de type « Magicien d'Oz » appliqués à la collecte de corpus de parole émotionnelle authentique et répondants aux principes énoncés ci-dessus, une plate-forme d'expérimentation baptisée E-Wiz (*Emotional Wizard of Oz*) a été développée. Le cahier des charges ainsi qu'une version préliminaire de l'application ont été développés en collaboration avec Julien Rebreyend, stagiaire en DUT Informatique (Audibert, 2002 ; Rebreyend, 2002). La plate-forme a, par la suite, été documentée et enrichie de nombreuses fonctionnalités (Audibert, 2003). Le logiciel E-Wiz, développé en langage Java à des fins de portabilité et de réutilisabilité, s'inscrit dans le cadre du projet *Expressive Speech Processing* du CREST/JSP (ministère japonais de la recherche) et a été conçu comme un outil permettant non seulement de recueillir des corpus de parole émotionnelle spontanée en français, mais également d'être adapté à d'autres langues et cultures.

Cette plate-forme, dotée d'une interface ergonomique et multilingue (français et anglais), permet de concevoir et mettre en place des scénarios d'induction d'états émotionnels, sans recourir à des compétences particulières en informatique de la part de l'utilisateur. Les scripts de configuration, qui décrivent les scénarios proprement dits, sont en effet générés automatiquement par l'application. Le cadre commun de ces scénarios est

d'imiter et de perturber le comportement d'une interface personne/machine à commandes vocales, afin de recueillir des expressions émotionnelles directes dans la production vocale du sujet. En effet, l'expérimentateur, ou « magicien », a la possibilité de contrôler à distance le comportement de l'application, en fonction des soi-disant commandes vocales émises par le sujet. De cette façon, le magicien peut d'une part faire en sorte que le sujet soit convaincu qu'il a réellement affaire à une application à commandes vocales, et, d'autre part, provoquer à dessein des perturbations dans le comportement attendu de l'application, afin d'induire chez le sujet certains états émotionnels.

La plate-forme E-Wiz se décompose en trois applications distinctes :

- Un éditeur spécifique (dont un exemple est présenté Figure 5) dédié à la conception des scénarios, qui permet de définir graphiquement ces derniers, en générant automatiquement les scripts correspondants.

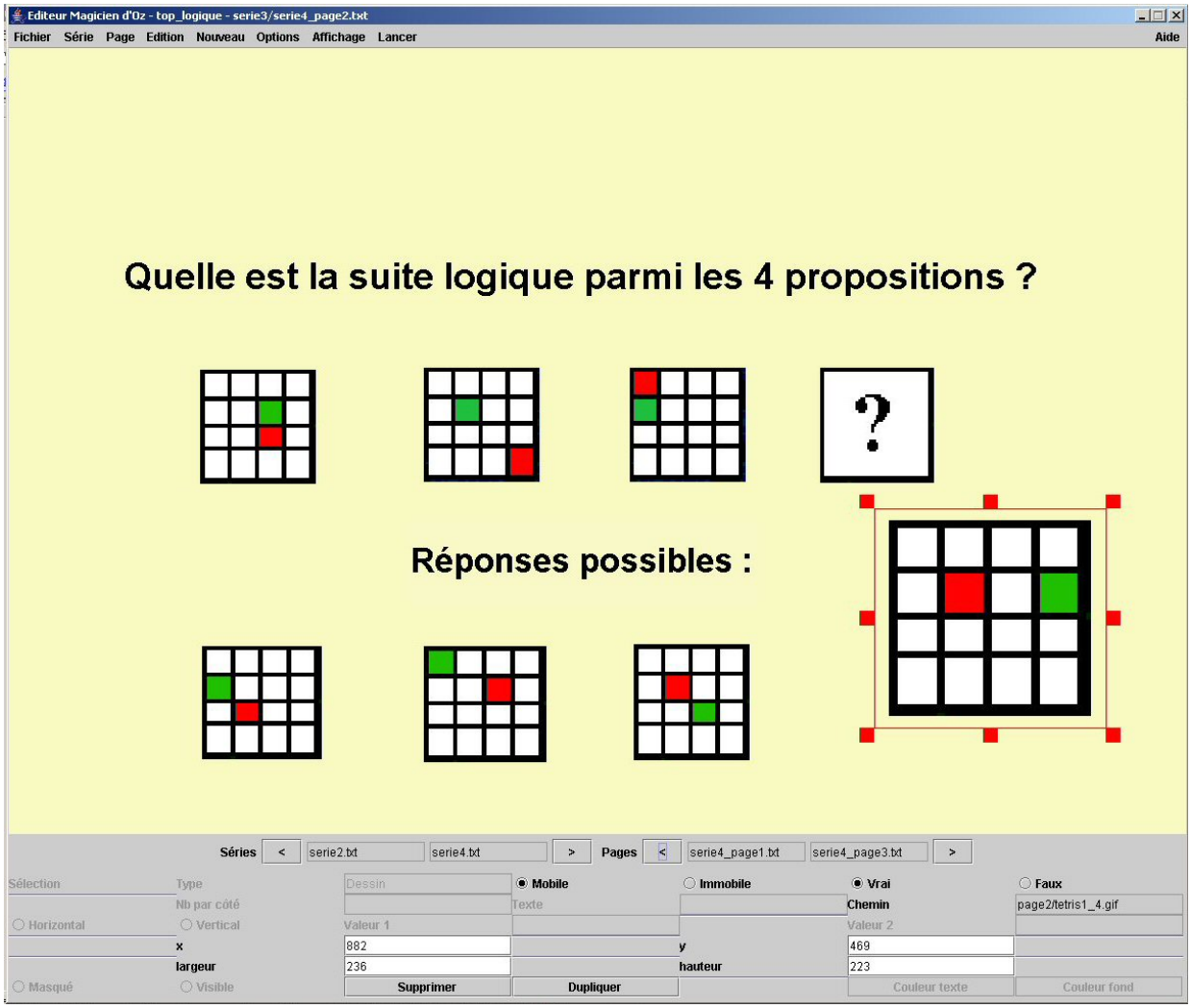


Figure 5 : L'application Editeur de la plateforme E-Wiz

- Des applications client et serveur qui utilisent les scripts générés par l'éditeur pour la phase d'enregistrement effectif des corpus. Lors de cette phase d'enregistrement, le serveur, placé entre les mains du magicien, permet d'adapter le comportement de l'application client, présentée au sujet, en fonction des réactions de ce dernier. Un exemple de l'application serveur est présenté Figure 6.

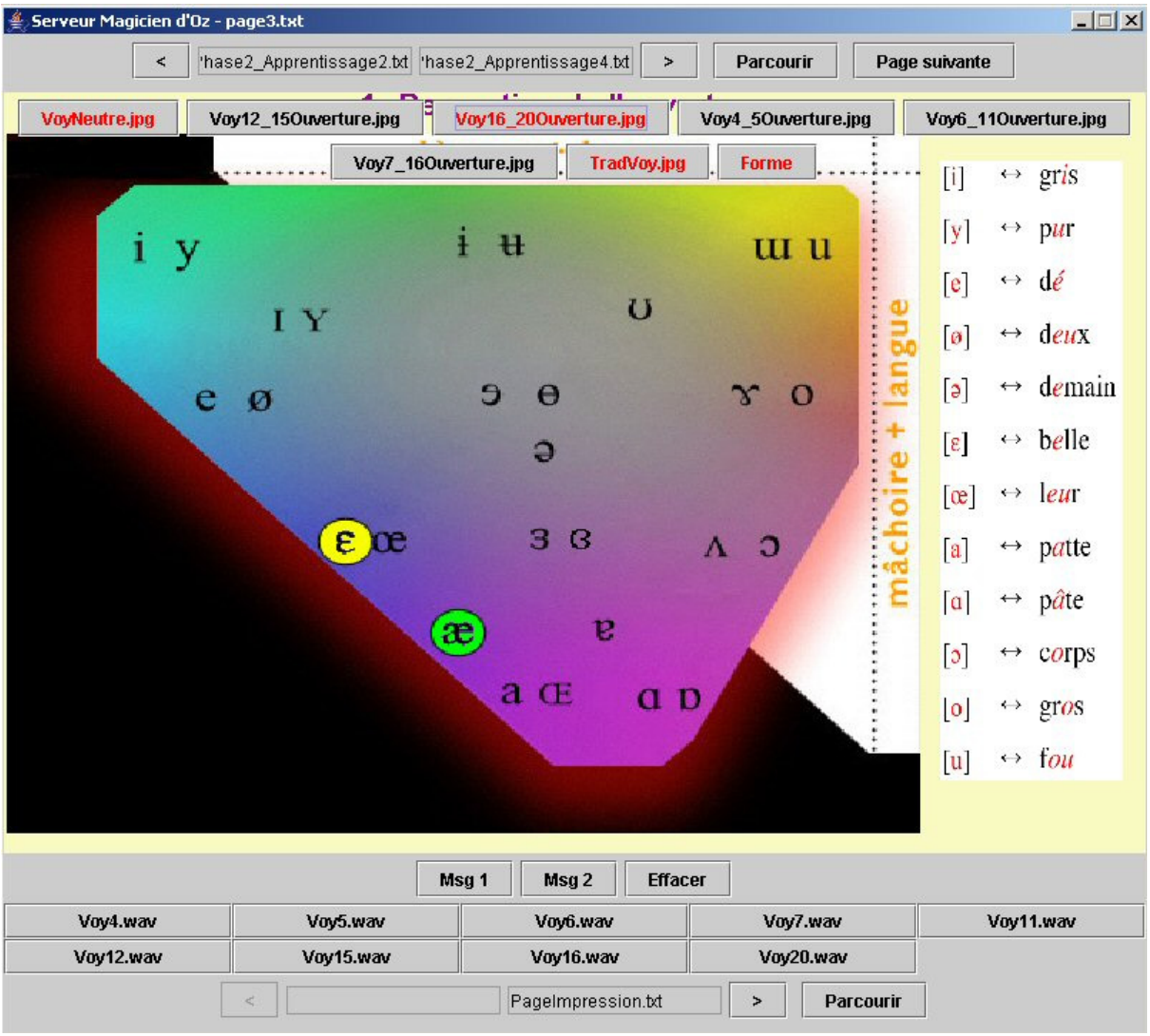


Figure 6 : Exemple de présentation de l'application Serveur de la plateforme E-Wiz

Les scénarios implémentés à l'aide de ce logiciel peuvent mettre en œuvre différents types de données multimédia, telles que du texte, des sons et des images. Les zones de texte ainsi que les images peuvent être masquées ou déplacées par le magicien afin de produire une sorte de diaporama présenté au sujet sur l'application client. Afin de faciliter le processus de mise en place des scénarios, un effort particulier a été consacré à la réalisation d'une interface conviviale et intuitive. Ainsi, des fonctionnalités de traitement de texte et d'édition telles que le copier / coller, ou encore des raccourcis

clavier, ont été implémentés. De plus, une aide à l'utilisateur permettant une prise en main rapide du logiciel est proposée.

Les scénarios ainsi définis se composent d'un ensemble de pages successives, parmi lesquelles des séries de pages dotées d'un comportement similaire peuvent être définies. Il est possible grâce à l'application éditeur de simplifier la tâche du magicien au cours de l'expérience, en automatisant le comportement de certains objets dans une série de pages donnée. La lecture des sons, plutôt que d'être déclenchée manuellement par le magicien, peut ainsi être liée à l'ouverture de certaines pages. On peut de plus intégrer aux pages du scénario des compteurs dont le comportement peut être prédéfini. D'autre part, les déplacements d'objets d'un point vers un autre sont calculés automatiquement afin que l'intervention du magicien ne soit pas décelable par le sujet.

Dans un souci d'ouverture à la communauté des utilisateurs potentiels, E-Wiz est distribué librement avec le code source de l'application sous licence GPL (GNU Public License), permettant aux utilisateurs d'y apporter toute modification jugée utile à condition que ces modifications soient distribuées selon les mêmes termes de licence[9].

5.4.1 Un échec riche en enseignements : le scénario Top Logic

Un premier scénario, « Top Logic », a été conçu et implémenté en collaboration avec Amandine Fouard (2002) à l'aide d'une version préliminaire de la plate-forme E-Wiz. Ce scénario s'inspire de l'expérience de Klasmeyer *et al.* (2000), qui ont imaginé un scénario de Magicien d'Oz dans le but d'induire de l'irritation et de la satisfaction. Dans le cadre du scénario imaginé par Klasmeyer *et al.* (2000), les locuteurs devaient accomplir des tests simples de logique pour lesquels le temps de réponse indiquait leurs performances. Les locuteurs étaient invités à compléter des séquences de nombres en choisissant le nombre approprié dans une liste. Pour obtenir des enregistrements de parole émotionnelle, le scénario prévoyait d'une part que les sujets lisent des phrases sémantiquement neutres affichées à l'écran tout en réalisant la tâche de résolution des suites logiques (l'expression de ces phrases devant être influencée par l'état émotionnel du locuteur), et, d'autre part, que la parole spontanée éventuelle soit enregistrée. Les états émotionnels recherchés devaient être induits par l'apparition de soi-disant bogues diminuant injustement les performances du sujet, ou au contraire par une évaluation exagérément positive de leurs résultats. Aucun résultat de l'expérience de Klasmeyer *et al.* (2000) n'ayant été communiqué, il paraît délicat d'évaluer l'efficacité de ce scénario. Il semble cependant problématique de demander aux locuteurs d'effectuer deux tâches

[9] E-Wiz peut être téléchargé à l'adresse http://www.icp.inpg.fr/EMOTION/EWiz

simultanément, à savoir la résolution de suites logiques et la lecture à l'écran de phrases, dans le but de récupérer l'expression vocale des émotions. Cela paraît en effet représenter une charge cognitive très lourde pour les locuteurs.

Le scénario développé par Fouard (2002) s'appuie sur un enchaînement de tests logiques et se compose de cinq séries de dix questions. Pour chacune de ces questions, le sujet doit compléter la suite logique présentée (composée de chiffres, dominos, et autres motifs divers) en désignant par sa position spatiale l'une des réponses parmi les quatre présentées (Figure 7). La sélection de la réponse se fait grâce à la phrase de commande « *le premier / deuxième / troisième / quatrième en partant de la gauche* ». La tâche est présentée comme une évaluation des facultés intellectuelles des sujets, à la manière des tests de QI. L'évaluation fictive donnée au sujet se fait comparativement aux performances moyennes des nombreux sujets supposés avoir déjà passé ce test.

Dans ce test, les variations émotionnelles sont induites de deux façons :

- En manipulant les performances du sujet, dont le score est comparé à un soi-disant score moyen des sujets précédents, beaucoup plus élevé ou plus faible, afin d'induire des émotions positives ou négatives. Ses performances sont également manipulées par l'utilisation de tests logiques de difficulté variable, ainsi que par le choix du temps maximum alloué pour la réponse.
- Par la manipulation du comportement attendu de l'application en simulant ralentissements et bogues alors que le compteur continue de défiler, afin d'induire chez les sujets de l'irritation ou de la frustration.

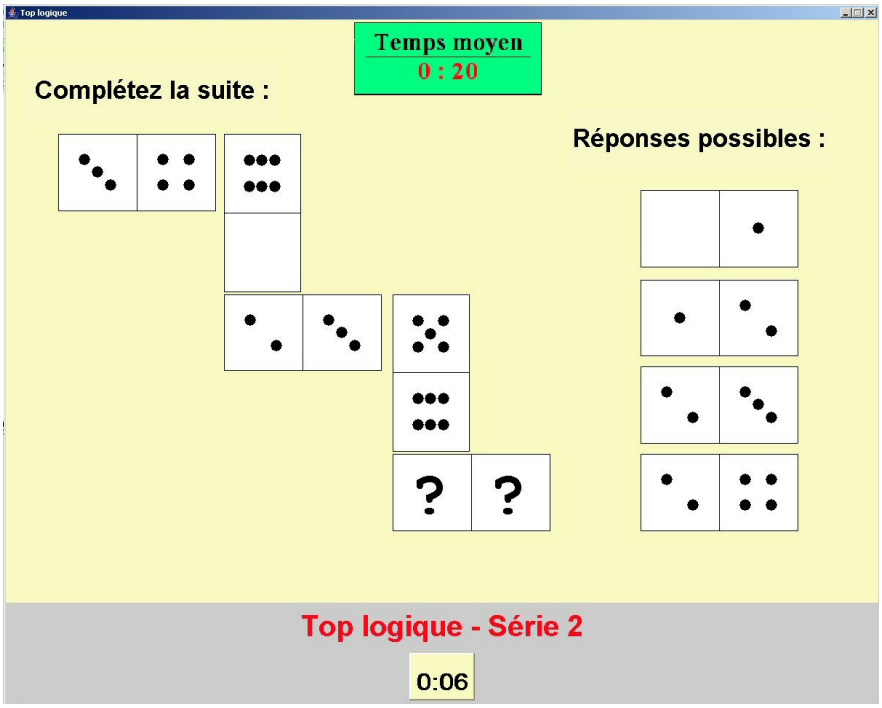


Figure 7 : Exemple de page extraite d'une tâche du scénario d'induction Top Logic

Le scénario « Top Logic » a été utilisé comme un premier test de la plate-forme E-Wiz. Deux locuteurs masculins, tous deux étudiants en sciences, ont été enregistrés. Bien que pour l'un de ces sujets l'objectif d'induction d'irritation paraisse avoir été atteint, ce scénario a montré des limites quant à la question de la motivation des sujets pour cette tâche. En effet, les locuteurs testés n'ont pas semblé suffisamment impliqués par la tâche elle-même, en dehors de son aspect ludique.

Or, la motivation et l'implication des sujets revêtent une importance primordiale dans une telle expérimentation. En effet, outre la difficulté qu'il peut y avoir à tenter d'induire des états émotionnels chez un sujet peu impliqué, une motivation comparable chez les différents sujets placés face à un même scénario d'induction permet de limiter la variabilité des émotions induites aux profils psychologiques des sujets et à leur état émotionnel au moment de l'expérience.

Cette première expérience nous a donc permis de mettre l'accent sur l'aspect primordial de la motivation des sujets pour la tâche proposée, afin de s'assurer de leur implication. La solution d'une rétribution de la participation des sujets pourrait être envisagée, ce qui permettrait d'en faciliter le recrutement et, selon la procédure retenue, de s'assurer d'un minimum de sérieux. Cependant nous considérons que lorsque cette rétribution constitue la motivation principale des participants, une implication dans la tâche proposée comparable à celle de sujets directement motivés par cette tâche n'est pas assurée.

5.4.2 Le scénario Sound Teacher

5.4.2.1 Description du scénario

Nous avons ensuite développé à l'aide de la plate-forme E-Wiz un second scénario, « Sound Teacher » (Rilliard, Aubergé & Audibert, 2004). Ce scénario imite un logiciel d'apprentissage des sons de diverses langues du monde tiré des travaux de Boë *et al.* (1995a, 1995b), prétendu permettre un apprentissage intuitif des langues étrangères. Ce logiciel est présenté comme l'application à l'apprentissage des langues étrangères de découvertes neurophysiologiques récentes liées à la théorie de la perception-action et des neurones miroirs. L'argument factice principal, apporté aux sujets comme introduction à la tâche proposée, est que le cerveau humain crée des cartes corticales lors de l'observation d'un mouvement biologique, également utilisées pour la production du même mouvement. L'apprentissage proposé aux sujets résulte donc de l'application de cette théorie aux mouvements articulatoires. L'application est présentée aux sujets comme en cours de test, ce qui permet de justifier l'utilisation de certains capteurs ainsi que les demandes de commentaires. Diverses raisons nous ont conduites au choix de ce scénario, que nous avons tenté de rendre le plus motivant possible pour les sujets, à commencer par l'engouement pour les études relatives au fonctionnement du cerveau, qui trouvait déjà au moment de la mise en œuvre de ce scénario un certain écho auprès du grand public. De plus, le laboratoire dans lequel les enregistrements prenaient place étant situé dans l'enceinte d'une université dans laquelle sont notamment enseignées les sciences du langage et les langues étrangères, il était relativement aisé de recruter des étudiants grandement motivés pour comprendre le fonctionnement de leur cerveau lors de l'apprentissage de langues étrangères.

L'apprentissage est fondé sur la présentation sonore de prototypes de voyelles issues de diverses langues du monde, réalisés par un synthétiseur vocal, couplée à la présentation visuelle de quatre traits articulatoires : ouverture, avant/arrière, centralisation, arrondissement des lèvres. De plus, les positions des voyelles dans le triangle vocalique (projection en 2 dimensions de l'espace acoustique décrit par les deux premiers formants) sont représentées, associées à leur symbole dans l'alphabet phonétique international. L'interface est basée sur une présentation multimédia des voyelles des langues du monde développé au laboratoire par Boë *et al.* (1995a, 1995b) pour la cité des sciences de la Villette (Vallas *et al.*, 2000). Les sujets sont entraînés à reconnaître les valeurs de ces paramètres, c'est-à-dire le mouvement articulatoire, lors d'une phase d'écoute de ces voyelles, puis à appliquer lors d'une phase de production (Figure 8) les nouvelles compétences qu'ils sont censé avoir ainsi acquises.

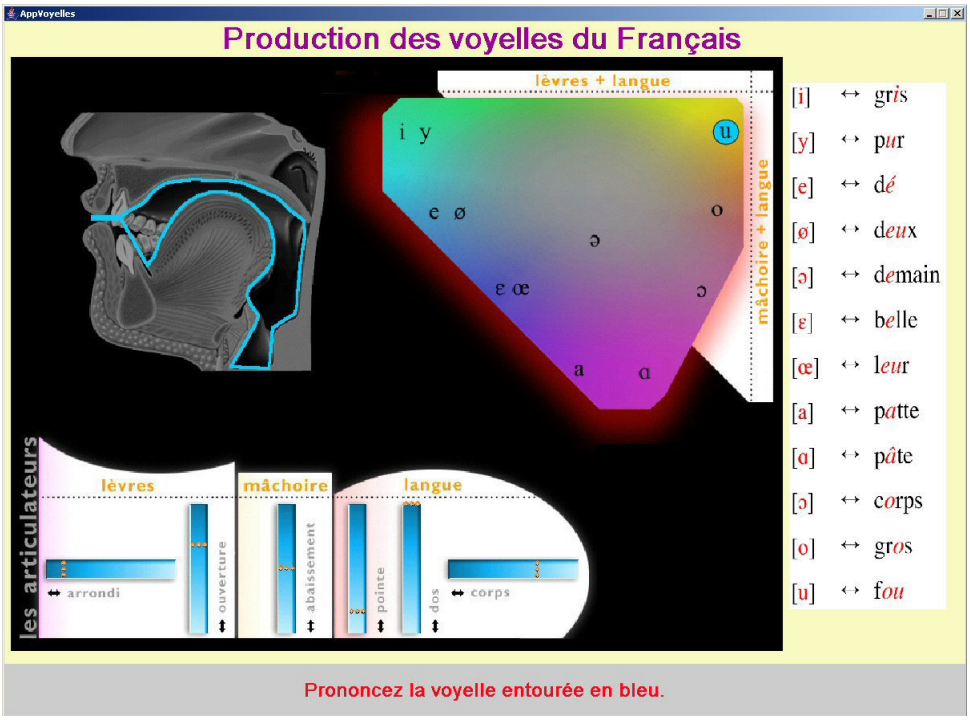


Figure 8 : Exemple de tâche de production proposée par le scénario Sound Teacher

Le scénario, après une phase de présentation, est organisé en quatre étapes, de difficulté croissante du point de vue de la tâche d'apprentissage. Du point de vue de l'expérimentateur, ces quatre phases sont destinées à induire des variations émotionnelles, tout d'abord positives (phases 1 et 2), puis négatives (phases 3 et 4). Chaque phase se décompose en une tâche de production et 4 tâches de perception (Figure 9), chacune centrée sur l'un des traits articulatoires. Lors des tâches de perception, le sujet écoute un stimulus sonore (une voyelle synthétisée) et doit prononcer la couleur correspondant à celle qu'il pense avoir reconnue parmi les deux voyelles entourées dans le triangle vocalique. Toutes les tâches proposées sont séparées par des pages de présentation au sujet des résultats qui lui sont attribués, comparés aux résultats obtenus par les sujets précédents. L'ensemble de ces résultats est prédéterminé au sein du scénario en fonction des objectifs d'induction de variations émotionnelles. Outre les tâches de production, les sujets doivent prononcer la commande vocale « *page suivante* » pour avancer dans le déroulement du scénario, et des noms de couleurs monosyllabiques pour fournir les réponses lors des tâches de perception, en fonction des instructions présentées à l'écran. Des commentaires lui sont également demandés régulièrement.

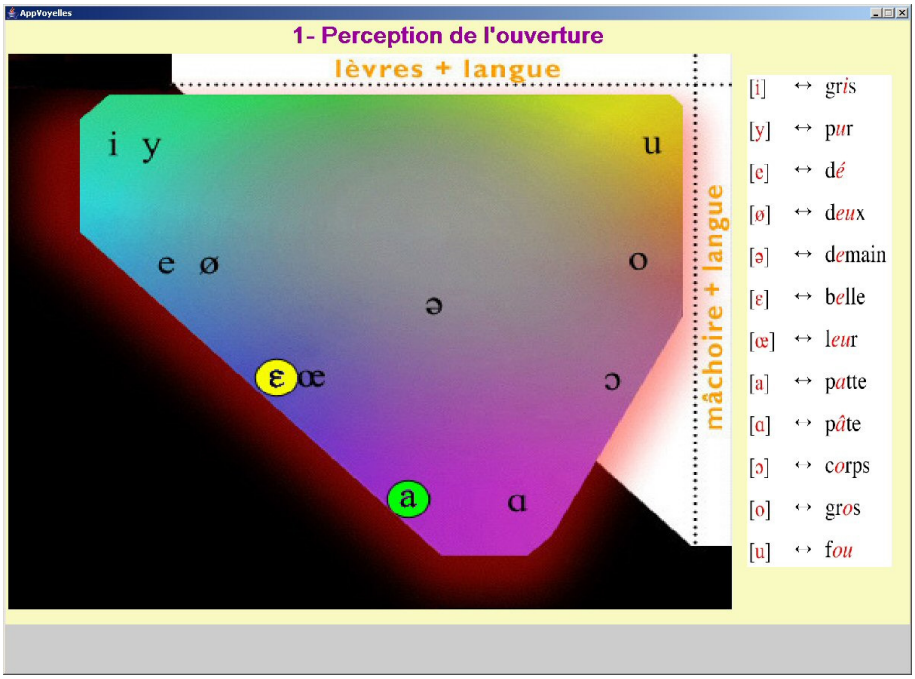


Figure 9 : Exemple de tâche de perception proposée par le scénario Sound Teacher

La première phase consiste en une vérification des compétences des sujets (de langue maternelle française) pour la production et la perception de voyelles du français. Un retour d'information artificiellement positif est donné aux sujets, avec un score significativement plus haut que le prétendu score moyen des apprenants précédents. Le sujet doit, lors de la deuxième phase, apprendre des voyelles proches du système phonologique du français. Le choix des voyelles perceptivement proches ou éloignées de celles du système phonologique du français a été fondé sur les résultats d'une étude perceptive menée par Chauny (1996) à partir du même modèle de synthèse articulatoire des voyelles. Durant cette phase, le sujet reçoit toujours un score élevé après chaque tâche. Il est averti, à l'issue de cette phase d'apprentissage, que ses performances le classent parmi les 5 meilleurs apprenants évalués jusqu'alors, ce qui lui permet de passer directement à une phase de généralisation à des voyelles complexes.

Lors de cette troisième phase, le sujet est confronté à des voyelles perceptivement éloignées de celles du français. Son score est évalué comme très nettement inférieur à la moyenne des autres sujets. Il est alors mis en garde de l'anormalité de ses résultats, et il lui est signifié que ses compétences pour la perception et la production des voyelles du français doivent être vérifiées, car elles pourraient avoir été dégradées par Sound Teacher. La quatrième et dernière phase est donc similaire à la première, à ceci près que les stimuli sonores présentés ont été modifiés afin de réduire fortement le contraste perceptif et d'obliger ainsi les sujets à répondre au hasard. A nouveau, des scores très faibles sont attribués au sujet, et les commentaires qui lui sont présentés vont dans le sens d'une dégradation de ses capacités perceptives. Néanmoins afin d'éviter de

provoquer chez lui une peur trop intense, ce qui serait problématique d'un point de vue éthique, il lui est signifié que ses capacités perceptives devraient revenir rapidement à la normale.

A la fin du déroulement du scénario, le sujet est informé des buts réels de l'expérience. Une brève interview prend alors place, dans laquelle il lui est demandé de décrire brièvement les différents états émotionnels ressentis pendant l'enregistrement.

5.4.2.2 Spécificité du contrôle des données

Les données collectées pour constituer les corpus sont les pseudo-commandes vocales, à savoir les occurrences de « *page suivante* » et des cinq noms de couleurs monosyllabiques « *rouge* », « *jaune* », « *sable* », « *vert* », « *brique* », ainsi que les commentaires libres. Le choix de retenir essentiellement des stimuli monosyllabiques dans le langage de commande imposé au sujet a été dicté par la volonté de collecter des énoncés dans lesquels les seules variations prosodiques soient celles dues aux affects exprimés par les sujets. En effet, dans le cadre d'un modèle prosodique de superposition de contours gradients (Aubergé, 2002), un énoncé d'une longueur supérieure à la syllabe (et plus particulièrement d'une longueur supérieure à trois syllabes si la notion de mot prosodique est retenue) verra se superposer aux expressions d'affects les variations prosodiques liées aux fonctions linguistiques de la prosodie. En l'absence d'une modélisation suffisamment fine de la prosodie linguistique, qui permettrait de simplement retrancher les contours correspondants pour n'étudier que les variations prosodiques liées aux affects, l'utilisation d'énoncés minimaux est destinée à geler ces fonctions, et tout particulièrement la fonction de segmentation/hiérarchisation de l'énoncé.

Le choix des noms de couleurs a été effectué afin d'obtenir des voyelles différentes, distribuées dans l'espace vocalique du français : [i], [u], [o], [a] et [ɛ]. Cette distribution a été choisie afin de pouvoir utiliser les données recueillies pour tester des algorithmes de mesure, notamment pour l'extraction de paramètres acoustiques de qualité de voix, en évaluant d'éventuels artefacts liés aux caractéristiques acoustiques des différentes voyelles.

Pour l'élaboration du scénario Sound Teacher, le choix que nous avons fait a été de placer le sujet dans une situation dans laquelle il est persuadé de n'interagir qu'avec une machine, par le biais d'un langage de commandes, afin d'éliminer les expressions de l'intentionnalité du sujet, c'est-à-dire l'élaboration de stratégies communicatives dans le but d'influer son interlocuteur. Ces stratégies communicatives peuvent s'exprimer selon diverses modalités, notamment via le choix lexical ou l'utilisation d'attitudes prosodiques (ou encore par l'expression dans la parole d'une « émotion intentionnelle »,

c'est-à-dire simulée). Or, il convient d'éliminer *a priori* de nos corpus l'expression de ces attitudes afin d'éviter autant que possible qu'elles n'interfèrent avec la prosodie des émotions que nous voulons collecter. En conséquence, l'hypothèse que nous formulons est que cette intentionnalité n'apparaît pas dans le cadre d'une communication personne-machine, et donc que le deuxième niveau prosodique (celui des attitudes émotionnelles) n'est pas exprimé. En fixant le troisième niveau (celui des fonctions linguistiques de la prosodie) ainsi que l'accès lexical grâce à un langage de commandes imposé au sujet, il devient alors possible d'obtenir un corpus dans lequel seul le niveau émotionnel de la prosodie varie.

5.4.2.3 Un scénario particulier pour les acteurs

Au-delà du protocole décrit ci-dessus et destiné à capturer des expressions vocales d'émotions induites chez les sujets par la manipulation de leurs performances, le scénario Sound Teacher a également été imaginé pour permettre de motiver des acteurs expérimentés ayant une motivation pour l'apprentissage des langues étrangères.

En effet, ce scénario prévoit un protocole particulier pour le cas des acteurs recrutés pour l'enregistrement d'un corpus d'émotions actées, élaboré par Noiray (2003). Les acteurs sont dans un premier temps contactés uniquement pour l'enregistrement de ce corpus acté, qui leur est présenté à l'avance, dans lequel il est prévu de leur faire produire des expressions de joie, colère, peur, surprise, tristesse et dégoût (les *Big Six* de Cornelius (1996)) sur 10 phrases sémantiquement neutres, dont la longueur varie de 3 à 7 syllabes :

1) Le chat dort
2) Le bus passe
3) Les gens parlent
4) Le cheval marche
5) La lampe est éteinte
6) Il est dans la chambre
7) Jean met son pantalon
8) J'entre dans le bâtiment
9) Demain je mange à midi
10) Le papier est sur la table

Afin de permettre des comparaisons ultérieures d'expressions émotionnelles actées et spontanées produites par les mêmes locuteurs sur les mêmes énoncés, une extension du scénario Sound Teacher prévoit de combiner ces deux approches en privilégiant les acteurs se déclarant intéressés par l'apprentissage des langues étrangères et pratiquant le théâtre d'improvisation ou de rue. Ainsi, les acteurs recrutés pour l'enregistrement d'un corpus acté et correspondant à ces critères d'après leurs déclarations (Noiray, 2003) sont mis à contribution pour participer au test de Sound Teacher : après que les modalités de l'enregistrement du corpus acté leur aient été détaillées par l'expérimentateur avec lequel ils ont été en contact et qu'ils soient installés pour l'enregistrement de ce corpus, un expérimentateur complice survient et entame une négociation simulée avec le premier expérimentateur pour obtenir la priorité dans l'utilisation de la chambre sourde. Cette négociation aboutit à la demande inattendue pour le sujet acteur de participer, avant l'enregistrement du corpus acté pour lequel il a été recruté, au test du logiciel Sound Teacher, en insistant sur l'intérêt de la présence d'acteurs dans le panel de test.

Le déroulement du scénario Sound Teacher est le même pour les acteurs que pour les autres sujets. Cependant, dans le cas des acteurs, les états émotionnels ressentis au cours de l'expérience et décrits par le sujet à la fin de l'enregistrement sont directement réutilisés : en effet ces émotions sont incluses dans l'ensemble d'émotions à reproduire par l'acteur en complément des *Big Six*, avec l'instruction de les produire de la façon la plus similaire possible à leur ressenti pendant l'expérience. Afin de permettre une comparaison ultérieure plus aisée des expressions actées et spontanées, les énoncés constituant les commandes vocales dans le scénario Sound Teacher (noms de couleurs monosyllabiques et « page suivante ») sont également inclus dans l'ensemble d'énoncés sur lesquels doivent être produites les expressions émotionnelles actées.

5.5. La problématique de l'annotation des expressions spontanées

5.5.1 Quelle méthode pour annoter les affects ?

Le recueil de corpus expressifs spontanés pose le problème de l'annotation des expressions d'affects recueillies. En effet, en l'absence d'indices mesurables et permettant de lier de façon non ambigüe une valeur ou un ensemble de valeurs à un affect ou une classe d'affects, l'annotation doit nécessairement reposer sur des jugements subjectifs.

La première question qui se pose est celle de l'ensemble de descripteurs qui doivent être utilisés pour étiqueter les expressions d'affects présentes dans le corpus recueilli. En effet, si le choix de ces descripteurs se porte sur des étiquettes, ce qui peut sembler être la solution la plus intuitive, la question du choix du jeu d'étiquettes à utiliser et de la

possibilité ou non d'en combiner plusieurs se pose rapidement. Une possibilité pour éluder cette question est le recours à une méthodologie hypothético-inductive comme celle proposée par Douglas-Cowie *et al.* (2000) avec le système Feeltrace, qui s'appuie sur des hypothèses fortes quant à la nature des émotions, à savoir une représentation continue selon les axes de valence et d'activation. Une telle méthodologie permet de définir *a priori* non seulement l'ensemble d'étiquettes à utiliser mais également le paradigme d'annotation. L'expert doit alors posséder une connaissance fine à la fois des concepts scientifiques par lesquels sont désignés les objets émotionnels et des outils permettant de spécifier la place d'un objet émotionnel particulier dans l'espace théorique.

5.5.2 L'annotation experte

Quel que soit l'ensemble d'étiquettes ou de dimensions utilisé pour l'annotation, il est nécessaire de définir la procédure selon laquelle ces étiquettes doivent être associées aux signaux audio, vidéo ou audiovisuels enregistrés. La méthode la plus répandue pour mener à bien cette tâche est de faire appel à des experts humains. Plutôt que de se fier aux annotations réalisées par un seul et même expert, des méthodes plus élaborées mais également plus lourdes à mettre en œuvre sur un volume important de données ont été proposées, dans lesquelles l'étiquetage est réalisé par différents experts en parallèle (Douglas-Cowie *et al.*, 2000, Abrilian *et al.*, 2005, 2006).

Dans ces méthodes, la cohérence inter-annotateur de l'étiquetage est ensuite validée au moyen d'indices statistiques tels que le kappa de Cohen (Cohen, 1960) dans le cas de variables catégorielles ou l'alpha de Cronbach (Cronbach, 1931) dans le cas de variables continues. Ces deux indices prennent une valeur d'autant plus proche de 1 que les annotations réalisées par les différents experts sont cohérentes entre elles. En l'absence de fondations théoriques pour l'interprétation des valeurs prises par ces indices, celle-ci ne peut se faire que par comparaison avec des valeurs repères établies de façon empirique dans la littérature et communément acceptées. Ainsi, Landis et Koch (1977) proposent une grille d'interprétation des valeurs de kappa, selon laquelle la valeur de kappa doit être supérieure à 0.20 pour que le degré d'accord soit considéré comme acceptable, tandis qu'un kappa compris entre 0.41 et 0.60 correspond à un accord modéré, une valeur entre 0.61 et 0.80 à un degré d'accord substantiel et une valeur supérieure à 0.80 à un accord presque parfait. Comme relevé par Peterson (1994) dans sa méta-analyse des utilisations de l'alpha de Cronbach, les valeurs recommandées sont fixées arbitrairement et varient selon les auteurs, mais il y a un certain consensus à définir 0.70 comme seuil en deçà duquel la cohérence des données doit être considérée comme trop faible. Néanmoins il faut souligner que cet indice a été plus fréquemment

utilisé comme une mesure de cohérence intra-sujet (notamment entre différentes échelles de type Likert supposées capturer le même concept dans un questionnaire, cas dans lequel l'interprétation des valeurs prises par l'alpha de Cronbach est plus controversée) que pour évaluer la cohérence inter-sujets comme dans le cas de l'annotation experte.

Les méthodes d'annotation de corpus d'expressions émotionnelles reposant sur le recours à plusieurs experts ont connu un important essor dans le cadre du réseau européen d'excellence HUMAINE. Ainsi, le corpus EmoTV d'expressions émotionnelles dans le cadre d'entrevues télévisées (Abrilian *et al.*, 2005) a été annoté dans un premier temps par 2 experts, à la fois sous forme de texte libre et selon les dimensions de valence et d'intensité. Les annotations données sous forme de texte libre ont ensuite été réinterprétées par les auteurs comme appartenant à une catégorie émotionnelle parmi quatorze. Cette annotation a mis en évidence un faible accord entre locuteurs, avec une valeur de kappa de 0.37 pour les catégories émotionnelles (dénotant selon Landis et Koch (1977) un niveau médiocre d'accord inter-annotateurs). Cette constatation a ensuite conduit les auteurs à expérimenter, sur un sous-ensemble plus restreint du corpus EmoTV, une représentation sous forme de mélange pondéré d'émotions afin de tenir compte de la variabilité entre annotateurs, en faisant appel à 40 annotateurs (Abrilian *et al.*, 2006).

De même, Reidsma *et al.* (2006) ont fait annoter par 3 experts le corpus multimodal AMI, composé d'interactions lors de réunions professionnelles, en leur demandant d'indiquer à la fois une estimation de la valence et de l'activation et une catégorisation à l'aide d'une liste d'étiquettes pouvant être étendue par les annotateurs, selon un protocole dans lequel les annotateurs devaient définir les segments à annoter leur semblant cohérents du point de vue de l'expression émotionnelle. Les auteurs n'indiquent pas de valeur globale d'accord, mais une valeur d'alpha de Cronbach[10] pour chaque paire d'annotateurs comprise entre 0.061 et 0.443, soit un niveau nettement inférieur aux différentes recommandations relevées dans la littérature par Peterson (1994). L'alpha de Cronbach mesurant l'accord entre annotateurs sur l'alignement des segments atteint néanmoins une valeur comprise entre 0.672 et 0.781 après regroupement des étiquettes selon leur valence.

Une méthodologie similaire a été appliquée à l'annotation de productions actées dans quelques études. Ainsi Clavel (2007) a fait appel à 3 experts pour annoter le corpus SAFE d'expressions de type peur extraites de films, obtenant une valeur de kappa de 0.49.

[10] Les auteurs présentent également des valeurs de kappa mais il s'agit d'un indice calculé sur des variables continues, qui n'est pas équivalent au kappa de Cohen (1960).

Cette valeur, quoique plus élevée que celles généralement relevées pour l'annotation des affects présents dans des productions spontanées, correspond selon Landis et Koch (1977) à un niveau d'accord modéré entre les annotateurs. De même Busso et Naranayan (2008) ont fait appel à 3 annotateurs pour l'étiquetage du corpus acté USC IEMOCAP. Les valeurs de kappa obtenues peuvent être interprétées comme correspondant à un niveau d'accord correct à modéré : 0.43 dans le cas des énoncés produits en improvisant à partir d'un scénario choisi comme susceptible d'éliciter la joie, la colère, la tristesse, la frustration ou un état neutre, 0.36 dans le cas de scènes théâtrales apprises et répétées au préalable par les acteurs.

Dans le cas de l'alpha de Cronbach, l'interprétation largement répandue selon laquelle la valeur prise par cet indice dépend directement du nombre d'éléments dans l'échelle (ce qui, dans le cas de l'annotation experte de corpus d'expressions émotionnelles, correspondrait au nombre d'étiquettes utilisées) a été battue en brèche par l'étude de Peterson (1994). Cependant, le niveau d'accord inter-annotateur obtenu est fortement dépendant du jeu d'étiquettes ou de dimensions proposées, ainsi que de la possibilité ou non de combiner différentes valeurs affectives. Il est donc délicat de prétendre comparer des degrés d'accord entre annotateurs pour différents corpus lorsque les protocoles d'annotation diffèrent. De plus, ce degré d'accord reste généralement relativement faible par rapport aux valeurs de références communément acceptées pour ces mesures statistiques.

Les méthodes d'annotation experte ont néanmoins fait la preuve de leur efficacité dans le cadre de l'annotation de données émotionnelles complexes extraites de situations de la vie courante, permettant d'isoler les portions du corpus pour lesquelles les annotateurs s'accordent le mieux sur une ou plusieurs valeurs d'affects. Cependant de telles méthodes présupposent une compétence de l'expert développée au-delà de ses capacités communicatives normales. Il doit en effet être capable de se placer dans une situation méta-communicative d'observation afin de pouvoir légitimement revendiquer la qualité d'expert. Une telle approche est classique lorsqu'il s'agit d'annoter des productions humaines et, plus particulièrement, des productions langagières comme par exemple dans des tâches de transcription ou d'étiquetage phonétique. Il semble en conséquence naturel de l'étendre à l'annotation d'expressions d'affects volontaires, tels que les attitudes dont les expressions sont normalisées par les langues et les cultures. En outre, cette approche se justifie dans le cadre de méthodes d'annotations guidées par des principes théoriques, qui fournissent à l'annotateur à la fois les règles et les outils nécessaires à l'interprétation des affects exprimés dans le corpus, au-delà de ses compétences d'humain.

5.5.3 Des limites de l'annotation experte

Cependant, une première objection quant à la validité d'une telle approche d'annotation peut être soulevée. En effet il convient de garder à l'esprit la grande variabilité inter-annotateur inhérente à l'annotation experte de productions langagières. En effet, même dans le cas de transcriptions phonologiques qui suivent un code consensuel et sont réalisées par des experts spécifiquement entraînés, les transcriptions réalisées par différents experts sont sujettes à d'importantes variations en fonction de leur expérience phonologique. Ainsi, Cole *et al* (1994) ont montré que pour l'étiquetage phonétique d'environ 30 minutes de parole continue en 4 langues (anglais, allemand, mandarin et espagnol), 2 annotateurs natifs formés et entrainés à cette tâche étaient généralement en désaccord pour plus de 30% des segments étiquetés. Cette variation est d'autant plus forte que la définition de l'objet à étiqueter n'est pas consensuelle : ainsi dans une étude réalisée dans le cadre du projet Phonologie du Français Contemporain et rapportée par Morel *et al.* (2006), 7 spécialistes de l'étude de la prosodie ont étiqueté pour 162 syllabes la présence ou l'absence de proéminence, avec un taux de syllabes jugées proéminentes compris entre 19% et 49% selon les experts. Sans que cela n'implique que tout protocole d'annotation experte doive être radicalement rejeté, il peut être problématique de se fier aux étiquettes ou valeurs dimensionnelles attribuées par un nombre restreint d'annotateurs, tandis que la difficulté à mettre en œuvre un protocole impliquant un nombre d'annotateurs très élevé est susceptible d'être rédhibitoire.

Au-delà de cette variabilité des jugements d'experts, il faut souligner que quel que soit le niveau d'expertise de l'annotateur par rapport à la théorie à laquelle le protocole d'annotation utilisé se réfère, ainsi qu'aux outils utilisés, il serait hasardeux de considérer comme relevant d'une quelconque expertise l'association proprement dite de ces connaissances expertes à une réalité capturée. En effet, cela reviendrait soit à considérer que cette association pourrait se faire explicitement à partir de la morphologie des données capturées, ce qui impliquerait que la problématique de la caractérisation des expressions d'affects dans la parole serait alors résolue et que le recours à des experts ne serait plus nécessaire, soit qu'il pourrait résulter d'un apprentissage, même empirique. Si cette seconde possibilité apparaît comme plus réaliste, cela implique que c'est alors la compétence « naïve » de l'humain pour l'identification d'expressions émotionnelles qui est utilisée dans toute tâche d'association entre descripteurs émotionnels et données capturées, quand bien même l'annotateur est par ailleurs expert et quelle que soit la complexité de ces descripteurs. Bien que nous ne disposions d'aucun moyen autre que les mesures de fiabilité d'échelles pour évaluer cela, il est vraisemblable que les

jugements réalisés par des experts soient, *in fine*, de qualité équivalente à ceux de sujets naïfs.

Une telle méthodologie appliquée à l'annotation des expressions émotionnelles dans un corpus spontané peut souffrir d'un artefact dû à l'implication de l'annotateur dans la situation observée. Par exemple, l'expert peut, sur la base d'une ressemblance physique ou comportementale, établir une association entre un sujet observé en interaction et une personne de sa connaissance et attribuer au sujet observé, parfois à tort, des réactions fréquentes chez la personne connue présentant des traits similaires. En effet, l'observation d'une interaction entre humains est en soi une situation écologique : l'interprétation de la situation par l'expert et donc sa compétence d'étiquetage résulteraient alors de ses compétences de sujet naïf pour l'observation d'autres humains éprouvant et exprimant des émotions, en partie construites par ses propres expériences antérieures et sa personnalité, ce qui remettrait en cause sa qualité d'expert (voir notamment Campbell *et al.* (2006) pour une discussion sur certains de ces points).

Le décodage des affects se trouverait ainsi limité par le processus d'empathie, ce qui est illustré par les travaux de Frijda (1987), pour qui l'un des critères fondamentaux de l'*appraisal* est la compatibilité d'un évènement émotionnel avec les normes et valeurs du sujet, qu'elles soient sociales ou personnelles. Ce point de vue est renforcé par les travaux de Garcia-Prieto & Scherer (2006), dont les résultats suggèrent que certaines dimensions de l'*appraisal* impliquées dans le processus émotionnel seraient influencées par l'identité sociale du sujet, ce qui pourrait être le cas de l'expert placé en situation d'annotation émotionnelle.

5.5.4 Une méthode alternative : l'auto-annotation des affects

Afin de contourner le risque d'une implication de l'expert dans la situation à annoter, et d'éviter de devoir poser *a priori* des hypothèses théoriques quant à l'ensemble d'étiquettes à utiliser, le choix effectué pour l'annotation des expressions d'affects recueillies dans le corpus E-Wiz a été d'expérimenter une méthode alternative d'étiquetage. Nous nous sommes donc orienté vers une méthode qui n'est certes pas exempte de subjectivité, mais dans laquelle cette subjectivité est celle du sujet lui-même, en demandant aux sujets d'annoter eux-mêmes leurs propres productions.

5.5.4.1 Le rôle de l'*appraisal* dans la conscience de soi

Une telle méthode d'étiquetage s'appuie sur la capacité des sujets à être conscients de leurs propres expériences émotionnelles, et sur la fiabilité de leur mémoire pour leurs émotions propres. Cette capacité des sujets à être conscients de leurs expériences

émotionnelles a été mise en avant dans les théories de l'*appraisal*. Ainsi Lazarus (1991) considère que l'*appraisal*, c'est-à-dire l'évaluation cognitive de la significativité d'une situation qui peut être à l'origine d'une réponse émotionnelle, influence la tendance à l'action et l'expérience émotionnelle, de même que les comportements individuels et les stratégies de *coping* (stratégies adoptées par le sujet pour faire face à la situation). Selon ce point de vue, les *appraisals* seraient responsables de la prise de conscience de l'expérience émotionnelle. Lazarus (1991) distingue l'*appraisal* automatique, non-réflexif et non-conscient, d'un *appraisal* étendu qui serait délibéré et conscient. Ekman (1999b, p. 52) souligne ainsi que :

« Souvent dans la vie civilisée, nos émotions apparaissent en réponse à des mots plutôt qu'à des actions, à des évènements qui sont complexes et indirects, et c'est ce processus d'appraisal étendu qui opère consciemment et délibérément. La personne est alors bien consciente de ce que Lazarus nomme « l'analyse du sens ». »[11]

Lewis (2005) considère quant à lui les évènements émotionnels comme contribuant à l'évolution et au renforcement des *patterns* d'*appraisal*. Ce qui évolue ne serait donc pas l'*appraisal* seul, mais l'ensemble formé par l'*appraisal* et l'émotion, c'est-à-dire en d'autres termes une « interprétation émotionnelle ».

5.5.4.2 Nous souvenons-nous correctement de nos propres émotions ?

La mémoire autobiographique des émotions a été considérée par de nombreux auteurs, parmi lesquels LeDoux (1996), comme détaillée, précise et à long terme, voire indélébile. Levine & Pizzaro (2004) nuancent ces affirmations en s'appuyant sur des études ayant montré que, bien qu'effectivement intenses et de longue durée, les souvenirs détaillés d'événements inattendus aux conséquences importantes ou à l'origine d'émotions intenses appelés « mémoires flash »[12] (Brown & Kulik, 1977), étaient susceptibles d'être sensiblement affectés par des événements ultérieurs. Levine & Pizzaro (2004) estiment que le point de vue selon lequel la mémoire des émotions serait indélébile relève d'une généralisation excessive de résultats tirés de l'étude de ces « mémoires flash », dont l'exemple le plus classique est celui utilisé par Brown & Kulik (1977) de l'effet sur la population américaine de l'annonce de l'assassinat du président Kennedy, ou d'épisodes

[11] Traduction personnelle de la citation originale : *"Often in civilized life, our emotions occur in response to words, not actions, to events which are complex and indirect, and it is an extended appraisal process which operates with consciousness and deliberation. Then the person is quite aware of what Lazarus calls the 'meaning analysis'."*.

[12] Traduction du terme anglais « *flashbulb memories* »

traumatiques ayant induit une peur très intense. Ils concluent que les souvenirs des épisodes émotionnels doivent être considérés comme partiellement reconstruits en fonction d'*appraisals* plus récents, de la même manière que les épisodes dénués de charge affective particulière. Néanmoins, plusieurs études ont mis en évidence des biais dans le souvenir de ces « mémoires flash », en dépit de la confiance très élevée des sujets envers ces souvenirs. Ainsi Talarico & Rubin (2003) ont montré que la précision de ces souvenirs n'était pas meilleure que celle du souvenir d'événements de la vie de tous les jours et déclinait avec le temps selon un schéma similaire. Dans un état de l'art des études portant sur l'impact du temps écoulé depuis un épisode passé sur la capacité d'un sujet à se souvenir de cet épisode, Thompson *et al.* (1996, p.40-42) notent que cette capacité décroît avec le temps suivant une fonction linéaire ou curvilinéaire. Dans l'optique de la construction d'un protocole d'auto-annotation de corpus d'expressions émotionnelles spontanées, il semble donc préférable de minimiser le délai entre l'enregistrement du corpus et la tâche d'annotation.

Parmi les expérimentations relatives au souvenir d'épisodes émotionnels et prenant en compte des expressions de moindre intensité, il est à noter que Thomas & Diener (1990) ont évalué la capacité de sujets sains à se souvenir après 3 ou 6 semaines d'émotions ressenties dans la vie de tous les jours ainsi que de leur intensité, évaluées quotidiennement ou 3 fois par jour par les sujets. Les auteurs ont conclu à une surestimation de l'intensité attribuée aux souvenirs des épisodes émotionnels comparativement aux rapports quotidiens établis par les sujets, ainsi qu'à une sous-estimation de la fréquence des épisodes émotionnels à valence positive. D'autre part, Talarico *et al*. (2004) ont mis en évidence l'intensité des émotions passées comme le composant principal de la précision et de la persistance des souvenirs des expériences émotionnelles associées, au-delà des variations liées à la valence des épisodes émotionnels.

Par ailleurs, des études centrées sur des aspects plus spécifiques de la récupération d'épisodes émotionnels passés (voir Healy & Williams (1999) et Levine & Safer (2002) pour des revues de ces études) ont pu montrer des biais systématiques dans le souvenir d'épisodes émotionnels passés, tels que l'influence de l'état émotionnel du sujet au moment de la récupération de ces émotions, plus spécifiquement dans le cas de troubles ou pathologies comme la dépression ou la schizophrénie. En conséquence, Levine & Safer (2002, p. 4) concluent que :

« […] les tests expérimentaux fondés sur des comptes-rendus personnels d'émotions passées, ainsi que les témoignages concernant l'impact émotionnel d'événements passés, doivent être interprétés

avec circonspection, particulièrement lorsque le compte-rendu donné par un individu fait suite à des changements majeurs de ses buts et croyances »[13]

Nous considérons toutefois que demander à des sujets ne souffrant pas de troubles psychologiques particuliers d'étiqueter leurs productions très peu de temps après la session d'enregistrement permet d'éviter de biaiser de la sorte le souvenir que les sujets ont de leur propre expérience émotionnelle, et ce d'autant plus que le scénario Sound Teacher n'est pas conçu pour induire des états émotionnels très intenses.

5.5.4.3 Timing

Rosenberg & Ekman (1994) ont par ailleurs mis en évidence une grande cohérence, non seulement catégorielle mais également temporelle, entre des rapports d'émotions propres datés très précisément et les expressions faciales des sujets suite à l'induction d'émotions négatives intenses au moyen de courtes vidéos, les rapports donnés par les sujets étant déclenchés par la visualisation des stimuli d'induction. Ce résultat revêt un intérêt tout particulier dans l'optique de l'annotation par les locuteurs de leurs productions dans le corpus enregistré à l'aide du scénario Sound Teacher : en effet une bonne précision temporelle des annotations pourrait ainsi être atteinte. Cependant étant donné que l'induction réalisée dans le scénario Sound Teacher n'est pas aussi extrême que dans l'expérience de Rosenberg & Ekman (1994), ni aussi dirigée vers des émotions spécifiques, la cohérence catégorielle entre les émotions ressenties et leurs expressions peut être moins systématique dans les tâches proposées dans le scénario Sound Teacher.

Walbott & Scherer (1989) soulignent qu'une description des émotions par les sujets eux-mêmes, en dépit des biais possibles ainsi que des difficultés pouvant être rencontrées lors de sa mise en place, est la seule qui permette d'accéder à certains aspects de l'expérience émotionnelle, définie comme la représentation consciente des changements d'états physiologiques et cognitifs induits par une situation à même de provoquer la survenue d'une émotion. Les auteurs relèvent qu'une technique fondée sur des questionnaires anonymes est plus appropriée pour obtenir des descriptions sincères et non-biaisées des expériences émotionnelles vécues par les sujets que des entrevues individualisées. De telles entrevues, bien que permettant théoriquement d'affiner les descriptions recueillies, sont aussi susceptibles de motiver des refus plus nombreux ainsi que l'apparition de biais, tels que celui de désirabilité sociale, et ce d'autant plus que l'enquêteur est proche du sujet et physiquement présent lors de l'entretien.

[13] Traduction personnelle de la citation originale : « *[…] experimental tests based on self-reports of past emotions, and testimony concerning the emotional impact of past events, should be interpreted with caution, particularly when an individual's report follows major changes in his or her goals and beliefs.* »

5.5.4.4 Protocole retenu pour l'auto-annotation

Rapidement après l'enregistrement (avec un délai de l'ordre de quelques jours), il est donc demandé à chaque sujet piégé à l'aide de Sound Teacher de visionner seul et dans un environnement calme, l'enregistrement vidéo de sa participation au scénario, avec l'instruction d'indiquer sur une grille les états émotionnels ressentis (et non seulement exprimés) aux différentes étapes du scénario. Cette grille, dont des exemplaires complétés sont reproduits en Annexe II :, indique des instants précis correspondant aux tâches successives proposées dans le scénario, chacun de ces instants étant associé à un champ vide permettant au sujet de décrire le ou les affects ressentis, sans que cette description ne soit figée en aucune sorte : le sujet est ainsi libre d'avoir recours à des phrases, des symboles, voire des dessins ou tout autre moyen de description qu'il juge utile. De plus ce protocole d'annotation autorise les sujets à visionner l'enregistrement vidéo autant de fois que nécessaire.

La méthode que nous proposons présente un certain nombre d'inconvénients potentiels, au-delà des biais susceptibles d'affecter la mémoire des expériences émotionnelle mentionnés ci-dessus :

- Comme relevé par Walbott & Scherer (1998), le recours à des annotations réalisées par les sujets eux-mêmes quant à leur ressenti émotionnel peut déboucher sur des biais, notamment celui de désirabilité sociale. Ainsi les sujets placés dans ce type de situation peuvent, consciemment ou non, ne pas décrire leurs propres émotions de façon totalement sincère, mais les adapter en fonction de ce qu'ils jugent valorisant selon les normes sociales établies par leur culture. Bien qu'il ne nous soit pas possible d'attester de la sincérité des sujets lors de cette tâche d'étiquetage, nous nous sommes efforcé de limiter ce problème potentiel. Il a donc été décidé de demander aux sujets de s'isoler pour étiqueter leurs productions afin d'éviter d'être influencés par des personnes extérieures, ainsi que de porter un accent particulier sur l'importance de la sincérité des annotations lors de l'explication de la procédure d'annotation. De plus il nous a semblé préférable, afin de limiter la possibilité d'apparition de ce type de biais, d'éviter de recruter les sujets dans l'entourage des expérimentateurs.
- D'après Scherer (2004a), la représentation consciente de l'émotion, qui correspondrait à la notion de *qualia* souvent mise en avant en philosophie et à ce qui est le plus souvent désigné par le terme de « sentiment », ne recoupe que partiellement les mécanismes inconscients à la base de l'émotion elle-même, au sein du système nerveux central. En effet, certains des processus impliqués dans l'émotion demeureraient inaccessibles au champ de la conscience, tandis que la partie consciente dépasserait le niveau d'une simple représentation consciente de

mécanismes inconsciemment mis en jeu en faisant intervenir la représentation que le sujet se fait de lui-même, ainsi que des contraintes socio-normatives. Il considère également que la communication d'une expérience émotionnelle par un sujet n'est qu'en partie confondue avec sa représentation consciente, non seulement en raison de la possibilité évoquée ci-dessus que le sujet souhaite masquer certains aspects de cette expérience émotionnelle, mais également en raison du recours au langage pour verbaliser cette expérience. Selon lui, le langage peut ne pas être à même d'exprimer en totalité la composante consciente de l'expérience émotionnelle, en partie en raison de l'absence d'éléments lexicaux appropriés mais, surtout, de la richesse de l'expérience émotionnelle consciente qui pourrait difficilement être capturée et restituée dans sa totalité grâce au langage. Selon Scherer (2004a), cette médiation langagière pourrait conduire à une catégorisation forcée qui appauvrirait l'expérience émotionnelle et aurait tendance à la réduire à des schémas socioculturels préétablis. De plus, les connotations associées aux éléments langagiers utilisés pour décrire l'expérience émotionnelle seraient à même d'ajouter une part d'information absente de la composante consciente de l'émotion. Cette conception des interactions entre composantes conscientes et inconscientes et verbalisation des expériences émotionnelles est illustrée Figure 10.

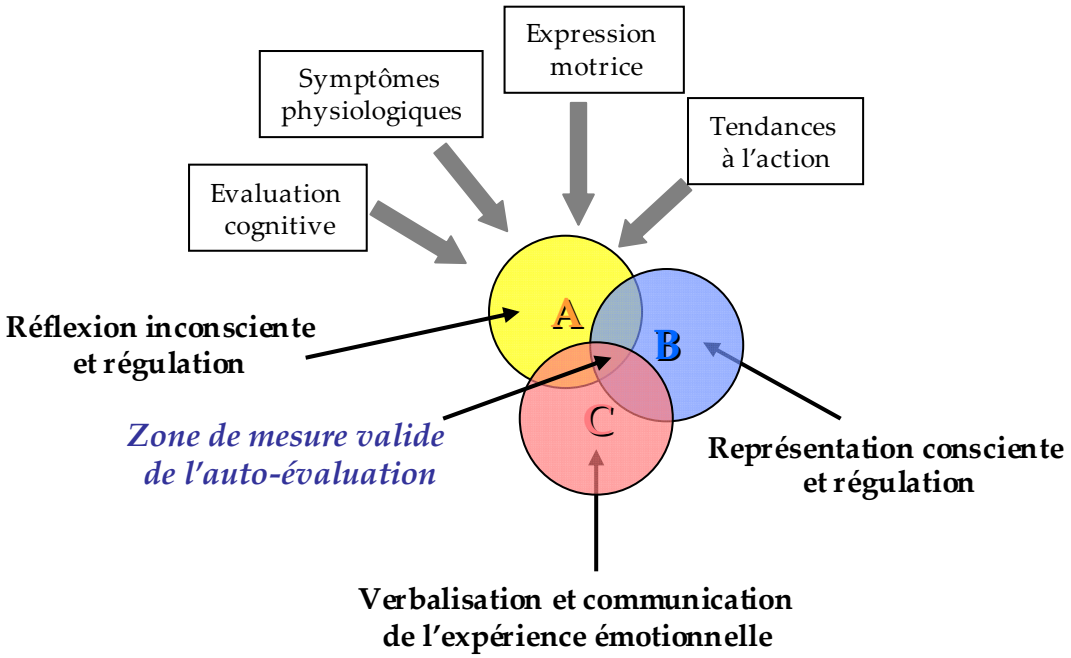


Figure 10 : Représentation schématique de l'hypothèse de distribution en 3 types des représentations centrales des processus composants[14] dans le modèle CPM de Scherer. D'après (Scherer, 2004a).

[14] Le terme de « réflexion inconsciente », traduction de l'anglais *unconscious reflection*, est ici à comprendre dans le sens de renvoi d'une image.

- Au-delà des réserves émises par Scherer quant aux limites de la médiation des expériences émotionnelles par le langage, notons que la liberté laissée aux sujets dans la forme de leurs annotations peut présenter l'inconvénient d'une généricité réduite entre annotateurs en raison de la relativité conceptuelle du langage pour la description d'événements émotionnels. En effet, en l'absence d'une définition claire et dénuée d'ambiguïtés de la monnaie d'échange utilisée pour l'annotation, rien ne permet *a priori* de considérer qu'une même étiquette utilisée par deux locuteurs différents fait référence au même phénomène. De plus, la forme des annotations produites par les sujets n'étant aucunement contrainte, les étiquettes ainsi recueillies peuvent être complexes et donc difficiles à mettre en relation avec des étiquettes de référence. Le traitement des étiquettes recueillies grâce à cette méthode est néanmoins facilité dans notre cas par le scénario figé : en effet le seul facteur de variation des réactions des sujets à l'induction réalisée dans Sound Teacher est leur profil psychologique. Ces contraintes, issues du scénario, permettent donc non seulement de faciliter la tâche d'étiquetage des émotions ressenties par les sujets, mais également de guider les expérimentateurs pour la définition *a posteriori* de classes d'étiquettes.
- En dépit de ces limitations, la méthode d'auto-annotation des affects que nous avons retenue peut donc être considérée comme complémentaire de la méthode plus classique d'annotation experte. Néanmoins, cette méthode d'annotation peut être difficile à généraliser à d'autres méthodes de recueil de données : en effet elle n'est applicable que lorsqu'il est possible de solliciter les sujets pour leur faire effectuer l'étiquetage peu de temps après l'enregistrement des données, ce qui n'est pas toujours possible dans le cas de données issues de situations de la vie courante, et la comparaison des annotations réalisées par différents sujets peut s'avérer plus difficile si les situations dans lesquelles ceux-ci sont placés deviennent plus variables. Lorsque l'étiquetage doit être défini de façon très précise dans l'optique d'analyses futures, on peut ainsi imaginer de cumuler ces deux méthodes pour ne conserver que les stimuli pour lesquels les jugements experts sont en accord avec l'auto-annotation.

Quoi qu'il en soit, une dernière étape de validation perceptive par des locuteurs naïfs des valeurs d'affect véhiculées par les stimuli sélectionnés demeure nécessaire, de même que dans le cas d'une annotation experte. En effet, la méthode que nous proposons est destinée à annoter les émotions exprimées par le locuteur et perceptibles par des tiers, mais n'est pas un moyen d'accéder à des émotions ressenties mais non exprimées. Nous considérons donc cette méthode d'auto-annotation comme une première étape d'un processus permettant de sélectionner les énoncés pour lesquels les valeurs d'affects

perçues par des auditeurs naïfs correspondent bien à ce que le locuteur considère avoir exprimé.

6. Le corpus Sound Teacher/E-Wiz

Nous présentons dans cette section la mise en œuvre du scénario Sound Teacher pour l'enregistrement d'un corpus multimodal de 17 locuteurs (10 femmes, 7 hommes) dont 7 acteurs, sur lequel nous avons fondé les analyses présentées dans cette thèse.

Nous détaillons les critères adoptés pour le recrutement des locuteurs enregistrés, les différents signaux enregistrés (audio, vidéo, électroglottographique et bio-physiologiques), ainsi que leur synchronisation et leur prétraitement. Nous donnons également un aperçu des autres études menées sur ce corpus au sein de l'équipe de recherche dans laquelle nous avons effectué nos travaux.

6.1. Sujets recrutés

Nous avons sollicité 17 sujets au total (10 femmes et 7 hommes) au cours du printemps 2003, avec comme prétexte invoqué leur participation à l'évaluation du logiciel Sound Teacher (Audibert, 2003 ; Audibert, Aubergé & Rillard, 2004). A l'exception de trois locuteurs masculins ne possédant pas de compétence, activité ou intérêt particulier pour l'apprentissage des langues étrangères, les discussions informelles préalables nous ont permis de vérifier que les sujets présentaient un intérêt pour ce domaine.

Les deux premiers sujets ont été recrutés afin d'effectuer un pré-test et de pouvoir corriger le cas échéants certains aspects du déroulement du scénario que nous n'aurions pas prévus. Bien que possédant une pratique de la langue française depuis l'enfance, ces deux premiers sujets féminins n'étaient pas de langue maternelle française. Leurs productions ne peuvent donc être analysées au même titre que celles des autres sujets.

Parmi les 17 sujets, sept étaient des sujets féminins (dont les deux premières dont le cas est évoqué ci-dessus) étudiantes en Sciences du Langage. Trois autres étaient les locuteurs masculins dont le cas est exposé ci-dessus.

Les sept sujets restants étaient des acteurs (3 hommes et 4 femmes), pour la plupart semi-professionnels, recrutés par Aude Noiray avec le prétexte de l'enregistrement d'un corpus d'émotions actées (Noiray, 2003) selon le protocole défini en complément du scénario Sound Teacher. Le critère principal de recrutement pour ces acteurs était une pratique confirmée d'un jeu de type ligue d'improvisation. Ils avaient de plus été interrogés oralement sur leur intérêt pour l'apprentissage des langues étrangères, afin de s'assurer *a priori* de leur implication potentielle dans la tâche proposée dans le scénario Sound Teacher. Les informations sur les acteurs recueillies par Noiray (2003) sont présentées en Annexe II :. Bien que les sept acteurs recrutés se soient, à des degrés variables, montrés surpris par l'intrusion non prévue d'un expérimentateur leur demandant de participer à l'évaluation d'un logiciel d'aide à l'apprentissage des langues du monde, tous se sont laissés convaincre et ont accepté de participer.

Enfin, deux autres sujets féminins ont été recrutés par Myrtille Terrier et ont participé à l'enregistrement selon un protocole légèrement différent : ces deux sujets avaient en effet pour instruction de fournir leurs commentaires libres par écrit au moyen d'un clavier mis à leur disposition (Terrier, 2003). Nous avons pour ce faire développé une fonctionnalité additionnelle d'E-Wiz permettant au magicien de déclencher l'apparition d'un champ de texte éditable et d'enregistrer son contenu.

6.2. Quelles mesures pour quels paramètres ?

La mise en place et la réalisation effective des enregistrements décrits dans cette section ont été réalisées en étroite collaboration avec Albert Rilliard, chargé de recherches au CNRS, Christophe Savariaux, ingénieur de recherche CNRS, et Alain Arnal, technicien audiovisuel à l'Université Stendhal-Grenoble 3, tous trois affiliés à l'époque de l'enregistrement du corpus à l'Institut de la Communication Parlée, devenu depuis le département Parole & Cognition de GIPSA-lab.

6.2.1 Le signal acoustique

La modalité à laquelle nous nous intéressons en priorité est celle de la parole acoustique. Le signal acoustique doit donc être de la meilleure qualité possible, ce qui nécessite l'utilisation d'un dispositif d'enregistrement professionnel, réalisé en chambre sourde avec un microphone de bonne qualité.

L'ensemble de la production vocale des sujets (commandes vocales et commentaires) est donc enregistré à l'aide d'un enregistreur numérique DAT portable Tascam DA-P1, à une fréquence d'échantillonnage de 44100 Hz et une quantification de 16 bits par échantillon, à l'aide d'un microphone AKG C1000S.

Nous enregistrons également les autres signaux de référence énumérés ci-dessous. En raison de l'impossibilité d'enregistrer directement l'ensemble de ces signaux avec une synchronisation fine, un signal spécifique permettant de réaliser cette synchronisation *a posteriori* a également été recueilli.

6.2.2 Le signal visuel

Au cours de l'enregistrement, les sujets sont assis face à l'écran sur lequel les différentes étapes du scénario Sound Teacher sont présentées, et leur regard est dirigé en direction de cet écran. La présence de la caméra n'est pas explicitement mentionnée, celle-ci étant justifiée auprès des sujets qui posent la question à l'expérimentateur comme une mesure de sécurité obligatoire pour permettre d'intervenir en cas de malaise. Les différents capteurs sont posés de telle façon qu'ils ne puissent pas masquer les gestes faciaux. La vidéo a été enregistrée sur fond bleu. Pour les cinq premiers locuteurs, l'enregistrement vidéo a été réalisé à l'aide d'un dispositif professionnel d'enregistrement au format Betacam SP de marque Sony. Par souci de commodité, l'enregistrement vidéo pour les sujets suivants a été réalisé au format mini DV à l'aide d'un caméscope Canon MV20i.

Les deux types de bande vidéo (Betacam SP et mini DV) ont ensuite été numérisées à l'aide du logiciel Reality, développé par la société DPS. Ce logiciel nous a permis de

compresser les vidéos numérisées au format AVI, en utilisant un algorithme de compression/décompression (codec) compatible avec les systèmes d'exploitation les plus utilisés au laboratoire (Windows, MacOS X).

6.2.3 Les capteurs bio-physiologiques

Les capteurs bio-physiologiques ne mesurent pas des expressions émotionnelles, mais suivent les marqueurs de l'état somatique qui varie avec l'état émotionnel du sujet. Si le pari fait par Lange (1885) de patrons distincts de réactions physiologiques selon les émotions s'est avéré non vérifié (voir par exemple Sander (2008) pour une courte discussion sur certains de ces aspects), des changements importants d'état émotionnel sont détectables via les pics de réflexe galvanique (Vernet-Maury, 2001). Ces pics de réflexe galvanique permettent seulement de détecter un changement d'état émotionnel, et non de déterminer l'état dans lequel se trouve le sujet avant et après l'apparition de ce pic.

Le système que nous avons adopté est celui utilisé au MIT par l'équipe de Rosalind Picard et distribué par la société ProComp. Les capteurs installés sur chaque sujet mesurent les activités suivantes :

- Le réflexe galvanique (augmentation de la transpiration à la surface de la peau) : il est mesuré à l'aide de deux électrodes appliquant une imperceptible tension à travers la peau, qui se fixent à l'intérieur de la deuxième phalange de deux doigts adjacents du sujet.
- L'amplitude respiratoire : ce capteur se place au-dessus du sternum pour la surveillance thoracique. Ce capteur est constitué d'une grande ceinture Velcro entourant le torse et d'un plétismographe qui s'étend lorsque le sujet expire. Cet allongement est mesuré et enregistré en fonction du temps, ce qui donne une courbe permettant de connaître l'intensité du mouvement respiratoire du sujet ainsi que sa cadence respiratoire.
- L'électromyogramme (mesure de l'activité musculaire) : cette mesure se fait à l'aide de trois électrodes groupées qui se posent sur un muscle au repos. Nous avons choisi de les appliquer sur le biceps du bras gauche du sujet, qu'il avait pour consigne de déplacer le moins possible au cours de l'expérience.
- La température de la peau : cette mesure est faite à l'aide d'un thermomètre spécial de petite taille qui s'applique sur la peau. Nous avons fixé ce thermomètre au contact du poignet du sujet.

- La fréquence cardiaque : elle se mesure grâce à un capteur fixé au pouce du sujet. La courbe obtenue en sortie représente l'évolution du nombre de battements en fonction du temps.

Les signaux biophysiologiques ont été enregistrés sous environnement Windows 98 à l'aide du logiciel *BioGraph*. Cet enregistrement a été réalisé sur un PC dédié à cette mesure, à l'aide d'une connexion par câble optique. Un canal a été réservé pour chaque capteur, l'un des canaux étant destiné au signal de synchronisation.

6.2.4 Mesure électroglottographique

Pour deux sujets (un acteur et un non-acteur), nous avons enregistré des mesures indirectes des mouvements glottiques à l'aide d'un électroglottographe intégré dans le système EVA2 (Evaluation Vocale Assistée), développé par la société SQ Lab. Cet électroglottographe permet d'enregistrer simultanément, à une fréquence d'échantillonnage de 25 kHz, le signal EGG ainsi que le canal audio.

Nous présentons plus en détails cette technique de mesure et l'exploitation des données que nous avons recueillies dans le chapitre traitant de la qualité de voix (section 7). Cette mesure, quoique n'étant pas invasive comme les autres méthodes de mesure de paramètres glottiques, constitue tout de même une charge supplémentaire pour le sujet, ce qui peut nuire au déroulement de la tâche. Son usage a donc été limité à deux sujets afin d'éviter que cet appareillage ne risque de perturber la qualité des données recueillies pour l'ensemble des corpus. Ces deux sujets ont été choisis parmi les sujets masculins. En effet les signaux EGG sont plus aisés à analyser pour les voix d'hommes que de femmes (Henrich, 1999). De plus l'appareil de mesure est également plus simple à installer dans le cas d'un homme, en raison de la présence de la pomme d'Adam qui permet de repérer sans difficultés la position du larynx.

6.2.5 Synchronisation des signaux

Il n'aurait pas été possible d'enregistrer directement de façon synchrone l'ensemble des signaux que nous avons collecté. En effet, outre la difficulté à déclencher simultanément tous les appareils de mesure, chacun de ces appareils est régi par un quartz spécifique et la fréquence d'échantillonnage effective des enregistrements réalisés peut dévier légèrement de la fréquence théorique, de façon constante ou non (phénomène de dérive des quartz). En conséquence, sur un enregistrement d'une durée telle que ceux que nous avons effectué, un décalage suffisamment conséquent pour biaiser les mesures peut être observé, qui ne peut être corrigé entièrement par une simple translation du signal en fonction du temps de déclenchement de la mesure.

Afin de pallier cet inconvénient et permettre une resynchronisation *a posteriori*, un bip de synchronisation d'une durée de 20 ms pour une fréquence de 2 kHz, déclenché par l'expérimentateur suite à chaque occurrence de « *Page suivante* », a été enregistré simultanément sur l'ensemble des canaux présentés ci-dessus. Le bip est déclenché à l'instant du changement de trame sur le canal vidéo consécutif à une pression sur le boîtier. Les portions de signal enregistrés sur les différents canaux et comprises entre deux bips consécutifs, dont la durée n'excède pas une minute, peuvent ainsi être ensuite synchronisées.

Un schéma complet du montage utilisé est présenté en Annexe I :. Les éléments composant ce montage ont été centralisés grâce à une table de mixage réalisée spécialement par Alain Arnal dans le cadre du projet Expressive Speech Processing du JST/CREST (Arnal, 2003). Afin d'éviter la présence sur l'enregistrement acoustique de bruits parasites causés par le système de ventilation du poste client, ce dernier a été placé à l'extérieur de la chambre sourde, sa sortie vidéo étant dupliquée vers un écran situé à l'intérieur.

6.3. Mise en œuvre de l'auto-annotation

Quelques jours après le passage de l'expérience, nous avons donné à chaque sujet l'enregistrement vidéo au format VHS, ainsi qu'une fiche récapitulative des différentes étapes du scénario, selon le protocole décrit en section 5.5.4.4. Il lui était demandé de visionner attentivement la vidéo de ses performances et d'annoter finement ce qu'il pensait avoir ressenti au cours de l'expérience. La possibilité a été laissée aux sujets qui disposaient du matériel nécessaire de réaliser cette tâche d'annotation à domicile dans un environnement calme, tandis que les autres l'ont effectuée au laboratoire.

Les fiches d'auto-annotation de cinq des six sujets dont nous avons analysées les productions en tenant compte de l'étiquetage des affects exprimés sont présentées en Annexe II :. Ces sujets seront désignés dans la suite par les sigles M1, M2 et M3 pour les sujets masculins, et par les sigles F1, F2 et F3 pour les sujets féminins.

La locutrice F2 ne nous ayant pas rendu sa fiche d'auto-annotation, les données correspondantes sont manquantes. Etant donné que cette locutrice semble toutefois empiriquement être la meilleure actrice parmi les locuteurs enregistrés, nous avons décidé de conserver ses productions en sélectionnant par écoute experte (par consensus entre deux expérimentateurs) les stimuli qui nous semblaient exprimer de façon non ambiguë des affects relevant de l'une trois grandes classes considérées pour les expériences présentées en section 6.

6.4. Prétraitement et étiquetage des données collectées

Un enregistrement d'une durée totale d'environ 50 minutes a ainsi été collecté pour chaque sujet. Chacun de ces enregistrements comprenait au minimum 52 occurrences de la commande « *Page suivante* », de 11 à 49 occurrences de chaque nom de couleur selon les réponses apportées par les sujets, ainsi que des commentaires libres plus ou moins fournis suite à chacune des tâches de Sound Teacher.

Nous avons converti les enregistrements audio sous forme de fichiers son échantillonnés à l'aide du logiciel CoolEdit 2000, en conservant les caractéristiques originales de fréquence d'échantillonnage et de quantification. Nous avons ensuite étiqueté à l'aide du logiciel de traitement phonétique Praat (Boersma & Weenink, 1992–2008) les productions des locuteurs, afin d'extraire les parties exploitables des corpus : commandes vocales et commentaires. Pour dix des locuteurs, dont les sept acteurs, nous avons également réalisé un étiquetage des frontières des frontières phonémiques. Pour chaque stimulus un fichier son échantillonné ainsi que le fichier d'étiquettes phonétiques associé ont été extraits au moyen de scripts Praat.

L'étiquetage phonétique a été centré sur les voyelles : nous avons placé les bornes des frontières des voyelles de façon à ne pas inclure de façon marquée les phénomènes de covariation avec les consonnes adjacentes : seule la partie stable des voyelles a ainsi été incluse, en nous appuyant sur l'évolution des valeurs formantiques observée sur les spectrogrammes. Toutefois, l'étiquetage phonétique s'est avéré délicat pour certains stimuli en raison de la variabilité importante observée dans les productions des sujets (dévoisement, présence d'approximantes, de schwas et de souffle en fin d'énoncé).

En particulier, des schwas d'intensité et de fréquence fondamentale (lorsque le voisement pouvait être mesuré de façon fiable, ce qui n'était pas toujours le cas) faibles ont très fréquemment été ajoutés en fin d'énoncé. Selon la taxonomie de l'épithèse vocalique (aussi appelée paragoge) en français contemporain établie par Carton (1999), de tels schwas ajoutés en fin d'énoncé constituent un [ə] de détente et ne changent pas la syllabicité, contrairement au [ə] du français méridional et à celui fortement accentué, fréquemment ajouté en fin de mot comme un marqueur de l'expressivité. Ainsi, à l'exception de rares exemples observés dans le corpus acté, principalement sur les expressions de colère réalisées par les actrices F1 et F4, pour lesquelles un [ə] final fortement accentué est réalisé, il semble raisonnable de considérer que la grande majorité des énoncés supposés monosyllabiques que nous avons recueilli le sont effectivement.

Tout étiquetage phonétique comporte une part de choix arbitraires, qui peuvent s'avérer fortement divergents d'un annotateur à l'autre (voir par exemple Cole *et al.* (1994)). Dans

le cas de notre corpus, ceci est particulièrement sensible dans le cas évoqué ci-dessus des schwas de détente. En effet ceux-ci correspondent au relâchement de la pression de l'air expiré par les poumons avec la bouche restée ouverte. L'énergie de ces schwas de détente diminue donc très graduellement et il est difficile de fixer objectivement une frontière finale. Afin de limiter l'impact de ces choix arbitraires sur les mesures réalisées par la suite, nous avons fait en sorte d'appliquer autant que possible les mêmes critères à chacun des stimuli étiquetés. Ces schwas de détente ont donc été étiquetés en tant que tels jusqu'à la limite à laquelle le schwa était audible, avant que nous ne le percevions comme un simple souffle.

Pour les deux locuteurs pour lesquels les mesures EGG étaient disponibles, nous avons synchronisé cet enregistrement *a posteriori* avec l'enregistrement audio, pour chaque portion de signal acoustique comprise entre deux bips de synchronisation consécutifs. La piste audio de l'enregistrement EGG, synchrone avec le signal EGG proprement dit, n'a pu être utilisée directement pour les analyses acoustiques en raison d'un bruit important. La synchronisation de ces signaux a été effectuée grâce à des scripts développés avec Matlab, afin de détecter les bips dans les portions de signal à traiter et d'extraire un signal EGG synchronisé avec la portion de signal acoustique correspondante.

6.5. Autres travaux réalisés sur le corpus recueilli

Dans les travaux que nous présentons dans la suite de ce manuscrit, nous nous sommes focalisé sur les pseudo-commandes vocales prononcées par les sujets (noms de couleur monosyllabiques et commande « *page suivante* »). Nous avons principalement étudié la modalité audio, et l'enregistrement vidéo dans notre étude de la discrimination des expressions actées vs. spontanées (cf. section 6).

Cependant, au-delà de ces pseudo-commandes vocales, la partie spontanée du corpus que nous avons recueilli est très riche en informations, et a permis d'initier des études variées au sein de l'équipe de recherche dans laquelle nous avons effectué nos travaux.

Comme nous l'avons évoqué ci-dessus, Terrier (2003) a enregistré en modalité écrite les commentaires libres de deux sujets en complément des 17 sujets que nous avons initialement recrutés. Ceci lui a permis de comparer les structures syntaxiques et le choix du lexique utilisés dans l'expressivité écrite et orale.

Bisognin (2005) a prolongé les travaux de Terrier (2003) en étudiant l'expressivité véhiculée par les structures syntaxiques adoptées par différents locuteurs dans leurs commentaires libres.

Bougé *et al.* (2006) ont proposé une esquisse de solution pour caractériser les instants de changement d'état émotionnel à partir des données biophysiologiques, et en particulier

le réflexe galvanique. Ils ont mis en évidence pour l'induction « cognitive » réalisée dans le corpus Sound Teacher un temps de latence sensiblement plus important entre l'instant d'induction et les manifestations physiologiques de cette induction que celui observé dans les expériences d'induction émotionnelle de plus bas niveau (tel que l'induction de peur chez des sujets phobiques). Dans la mesure où cette étude s'est attachée à caractériser via l'analyse de l'évolution des marqueurs somatiques les instants auxquels les effets de l'induction d'états émotionnels se font ressentir, elle pourrait présenter un intérêt direct pour nos travaux. Cependant, les données sélectionnées pour servir de base aux études présentées dans la suite de ce manuscrit étant suffisamment éloignées des instants d'induction émotionnelle, l'exploitation des données biophysiologiques n'aurait pas apporté d'éléments complémentaires pour la caractérisation des états exprimés dans les sous-corpus retenus.

Loyau (2007), à partir d'une annotation minutieuse de la gestualité de deux locutrices hors tours de parole, a proposé un ensemble d'icônes gestuelles permettant de caractériser la notion de « *feeling of thinking* » comme une extension de la notion de « *feeling of knowing* » proposée par Swerts & Krahmer (2005).

Vanpé & Aubergé (2008), s'appuyant sur les travaux de Loyau (2007), adoptent une démarche d'éthologues et s'attachent quant à elles à caractériser l'utilisation que font ces deux locutrices de ces icônes gestuelles dans leur expressivité hors tours de parole. Elles s'intéressent plus particulièrement à l'influence relative des informations portées par différentes zones faciales ainsi qu'à la dynamicité et à la rythmicité des icônes du « *feeling of thinking* ».

7. La qualité de voix

Dans cette section, nous abordons la notion de qualité de voix et commençons par discuter sa définition et sa fonction dans la communication parlée. Nous nous penchons ensuite sur la description phonétique des types de phonation les plus couramment étudiés dans la littérature, en nous appuyant sur les travaux fondateurs de Laver.

Après un bref aperçu des méthodes proposées pour l'évaluation perceptive de ces types de phonation, nous présentons différentes mesures acoustiques supposées rendre compte de phénomènes propres à la qualité de voix, et notamment les mesures de pente spectrale et de différence d'amplitude entre harmoniques. Nous mettons l'accent sur les limites que présente l'étude de la qualité de voix à partir de tels paramètres, et en particulier la difficulté d'étendre de telles mesures à des comparaisons inter-voyelles.

Nous présentons ensuite les modèles de l'onde de débit glottique supposés rendre compte du comportement de la glotte pendant la phonation et permettre ainsi de caractériser les différents types de phonation, et détaillons la technique de l'électroglottographie (EGG) qui permet de mesurer directement une partie du fonctionnement de la glotte. Après avoir brièvement abordé la problématique de la synthèse de la qualité de voix, notamment à partir de tels modèles du débit glottique, nous présentons le principe du filtrage inverse qui se propose d'évaluer l'onde de débit glottique à partir du signal acoustique en tirant parti des postulats de la théorie source-filtre de Fant.

Enfin, nous présentons les résultats d'une expérience visant à évaluer, sur les productions spontanées de deux locuteurs masculins, la validité pour des mesures fines de type suivi de contour d'un algorithme d'extraction automatique du paramètre NAQ opérant par filtrage inverse et auparavant validé sur un corpus expressif spontané de très grande taille, en confrontant les estimations réalisées par cet algorithme à des mesures électroglottographiques.

7.1. Introduction

La structure acoustique des sons de parole, qu'aucun des modèles proposés dans la littérature ne s'est avéré apte à représenter dans sa complétude, véhicule des indices relatifs à la voix du sujet parlant et à la structuration phonétique du message transmis.

Ainsi, en première approximation, la phonétique et la phonologie proposent des modèles des systèmes vocaliques, validés en perception et production, qui mettent en œuvre les deux ou trois premiers formants des voyelles, démontrant que ces paramètres sont nécessaires et suffisants. Pourtant les autres paramètres qui organisent la structure sonore (les harmoniques supérieures ou plus généralement le spectre de parole) sont traités par le système perceptif, et ne constituent pas uniquement une redondance de l'information transmise par les premiers formants.

La structure acoustique, qui transmet à la fois les informations liées à la chaine phonémique et aux indices idiosyncratique et expressifs, est souvent désignée par le terme de « timbre », timbre de la voix ou de la voyelle. Si la notion de timbre, issue des études musicales, désigne en premier lieu la perception qualitative qu'en ont les auditeurs, ce même terme est également et très largement utilisé pour décrire la description acoustique de ces phénomènes. Castellengo & Dubois (2005, p.1) reviennent sur la notion de timbre et mettent en avant la difficulté à la définir. Elles soulignent que la définition consensuelle parmi la communauté des acousticiens est une définition par élimination. Celle-ci, donnée par l'A.S.A. (*Acoustical Society of America*), indique que le timbre est « *ce qui différencie deux sons de même hauteur, de même durée et de même intensité* ».

Le terme « qualité de voix », fréquemment utilisé comme synonyme de timbre de la voix, décrit aussi les aspects acoustiques de l'ensemble des phénomènes qui, outre le phonème, donnent des informations sur :

- L'environnement du locuteur : vent, bruit, conduction dans un gaz autre que l'air (par exemple l'hélium) ou à travers un appareil tel qu'un masque de plongée (voir par exemple Ghio *et al.* (2006)).
- Le locuteur lui-même : changement perçu de registre, identité, âge, sexe, état de santé, états affectifs du locuteur…

La fonction de la qualité de voix à laquelle nous nous intéressons spécifiquement ici est celle de la transmission des états affectifs du locuteur. Cependant, les outils de description de la qualité de voix, et même plus généralement les modèles « phonétiques » au sens large, sous-entendent le pointage sur une information particulière, par exemple le lien entre position des formants et phonèmes. Ainsi, la description phonétique pionnière de la qualité de voix proposée par Laver (1980) que

nous développons ci-après, et sur laquelle se sont appuyés la plupart des travaux développés par la suite, donne une amorce des fonctions pragmatiques de la qualité de voix en liant notamment la qualité soufflée à la notion d'intimité. Cette notion a par la suite été étendue et surtout remise en contexte par Campbell (2003a), qui a établi à partir d'une grande quantité de données, un lien entre la tendance à utiliser cette qualité de voix soufflée et le « degré d'attention » (*degree of care*) porté à l'interlocuteur.

Construire une phonologie affective de la qualité de voix impliquerait deux étapes :

- Passer du signal à une caractéristique discrète (phonétique), qui pourrait être ici le registre de phonation ou plus généralement la qualité de voix particulière résultant de la configuration des articulateurs mis en œuvre. Etant donné la correspondance quasi-directe entre contrôle articulatoire et qualité de voix perçue mise en avant par Laver (1980), il est possible de supposer que les éventuels artefacts acoustiques peuvent être négligés et qu'une telle phonologie serait compatible avec l'ensemble des théories de la perception la parole, dont la théorie motrice de la perception (voir par exemple Liberman & Mattingly (1985)).
- Montrer que de tels traits discrets renvoient à des fonctions. « Accepter » une telle modélisation par les types de qualité de voix comme une interface entre phonétique et phonologie nécessiterait de montrer la correspondance, sous une forme mathématique quelconque, entre l'ensemble d'indices décrivant les qualités de voix particulières et certaines valeurs affectives considérées dans leur sens le plus large. Dans cette optique, les valeurs affectives considérées pourraient être des notions comme l'intimité ou la sympathie issues de théories sociales et ne faisant sens que considérées en contexte, aussi bien que des émotions dites « de base », telles que la colère. La description des affects en tant que catégories ou réparties selon un ou plusieurs continuums ne serait alors que secondaire.

Etant données d'une part l'état actuel des connaissances relatives à la qualité de voix et notamment de leur caractérisation à partir du signal acoustique (d'Alessandro, 2006), sur lesquelles nous revenons dans la suite de cette section et, d'autre part, les difficultés méthodologiques liées à la caractérisation des affects, l'ampleur d'une telle tâche dépasse très largement l'objectif que nous nous fixons dans nos travaux. Nous nous contenterons donc ici de passer en revue et d'évaluer certains des outils proposés pour l'étude de la qualité de voix à partir du signal acoustique, afin de déterminer dans quelle mesure et sous quelles conditions il est envisageable de passer du signal au trait, préalable nécessaire à la construction d'une telle phonologie affective de la qualité de voix.

7.2. Qu'est ce que la qualité de voix ?

Comme nous l'avons évoqué ci-dessus, il est difficile de décrire objectivement et dans sa complétude la qualité de voix, qui se rapporte à l'ensemble des phénomènes acoustiques relatifs au timbre. Cette difficulté de description est particulièrement sensible si l'on en retient la définition « en creux » qui considère la qualité de voix comme l'ensemble des phénomènes relevant d'une définition étendue de la prosodie mais ne consistant en aucun des paramètres considérés classiquement comme constituant la prosodie. Kreiman *et al.* (2004) soulignent ainsi la difficulté à étudier la qualité de voix en tant qu'objet défini par ce qu'il n'est pas.

Tous les phénomènes pouvant être considérés comme relevant de la qualité de voix ne sont pas issus de mouvements glottiques. Certains sont en effet dus aux modulations d'autres parties du conduit vocal. Ainsi, l'analyse auditive des « profils vocaux » (*Voice Profile Analysis*) proposée par Laver (1980, 2000) inclut comme composantes à part entière de la qualité de voix la configuration des lèvres, des diverses parties de la langue, du pharynx, ainsi que la hauteur du larynx. Cependant, la dénomination de qualité de voix est essentiellement utilisée en référence aux modes de phonation, qui désignent le comportement de la source glottique lors de l'émission de sons de parole. En effet, les plis vocaux[15] peuvent vibrer selon différents modes correspondant à des configurations glottiques particulières, qui produisent des qualités de voix différentes (Laver, 1980). Tout en gardant à l'esprit le fait que la qualité de voix ne se réduit pas à ces phénomènes, sauf mention contraire nous utiliserons donc par la suite ce terme pour désigner les régimes vibratoires de la source glottique.

Plus largement que dans la parole, ces régimes vibratoires ont été décrits dans le domaine de l'étude la voix chantée, dans lequel ils sont décrits en tant que registres phonatoires liés de façon plus ou moins étroite à la plage de fréquence fondamentale associée (voir notamment Henrich (2006) pour un panorama très complet).

Dans les langues largement étudiées que sont l'anglais, le français, l'allemand, le japonais et le mandarin, les différents régimes vibratoires de la glotte n'ont pas de statut phonologique particulier. En conséquence, et à de rares exceptions près, leur étude a été confinée jusqu'à une époque récente au champ de la phonétique clinique, les vibrations

[15] Bien que le terme de « cordes vocales », issu d'une analogie erronée avec les instruments de musique à cordes, soit encore très couramment utilisé (de même que son équivalent *vocal cords* dans la littérature anglophone), le terme de « plis vocaux » (en anglais *vocal folds*), consacré dans la littérature médicale (Doyon *et al.*, 1998) et plus conforme à la réalité anatomique, lui est préférable.

non modales des plis vocaux étant généralement considérées comme marquant des dysphonies.

Cependant Gordon & Ladefoged (2001), à partir de la taxonomie des sons des langues du monde la plus exhaustive à ce jour (Ladefoged & Maddieson, 1996), montrent que ces régimes vibratoires peuvent également constituer des traits contrastifs, donnant un statut phonologique particulier à ces modes de vibration. Gordon & Ladefoged (2001, p. 383) résument ceci de façon simplifiée, un trait phonologique ne pouvant être directement assimilé à la catégorie que constitue un phonème, par :

« (...) la pathologie vocale d'une personne peut être le phonème d'une autre personne »[16].

La qualité de voix, considérée comme n'intervenant pas de façon claire dans les fonctions linguistiques de la prosodie, a longtemps été le parent pauvre des études prosodiques. Cependant quelques recherches dans ce domaine peuvent être mentionnées. Ainsi Beckman & Campbell (1997) ont étudié les variations spectrales liées à différents types d'accentuation en anglais américain. Slifka (2007) note quant à elle que la phonation irrégulière constitue un indice privilégié de l'apparition de silence dans divers systèmes phonologiques.

A la suite de Laver (1980), qui parmi les premiers a proposé de lier changements de qualité de voix et expressivité, l'intérêt croissant pour la composante expressive de la parole a mené à un regain d'intérêt pour l'étude des régimes vibratoires de l'appareil phonatoire en dehors de l'étude des phonations pathologiques.

7.3. Description phonétique

7.3.1 La taxonomie de Laver (1980)

Ces modes de vibration, associés à une description impressionniste des qualités de voix résultantes, ont été décrits par Laver (1980) en fonction de trois paramètres de tension musculaire, schématisés dans la Figure 11.

- La tension adductive, définie comme la force avec laquelle les cartilages aryténoïdes sont compressés l'un contre l'autre, contrôlée par le muscle inter-aryténoïdien.
- La compression médiale, qui est la force de fermeture de la glotte ligamentaire, essentiellement contrôlée par le muscle crico-aryténoïdien latéral.

[16] Traduction personnelle de la citation originale : « *(…) one person's voice disorder might be another person's phoneme* ».

- La tension longitudinale, c'est-à-dire la tension des plis vocaux, qui résulte principalement de la contraction des muscles crico-tyroïdiens et du muscle vocal (*vocalis)*. Ce troisième paramètre musculaire permet en premier lieu de moduler la fréquence fondamentale.

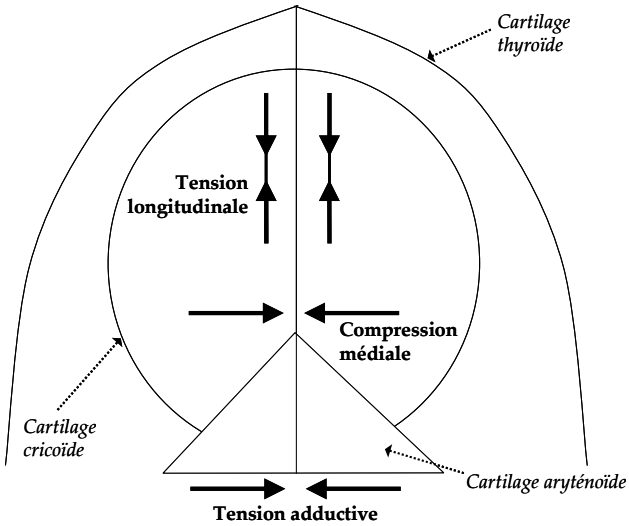


Figure 11 : Les trois paramètres musculaires décrivant les modes de vibration des plis vocaux. D'après Ní Chasaide & Gobl (1997).

Laver (1980) distingue notamment les qualités de voix suivantes :

- La voix modale (*modal voice)* correspond à une phonation neutre, et est utilisée comme référence pour l'étude des autres qualités de voix. Dans ce mode de phonation, la tension adductive, la compression médiale et la tension longitudinale sont modérées, ce qui correspond schématiquement à l'illustration de la Figure 11. Les plis vocaux vibrent périodiquement et efficacement, la fermeture glottique est complète, et aucun bruit de friction glottique n'est perceptible.
- La voix soufflée (*breathy voice*, Figure 12) se réfère à une tension adductive minimale, avec une compression médiale et une tension longitudinale faibles. Dans ce mode, les plis vocaux vibrent de façon extrêmement inefficace et ne se trouvent jamais entièrement en contact. Ceci a pour conséquence une fuite glottique importante, d'où un bruit de friction audible.

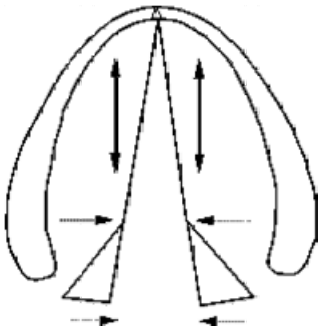


Figure 12 : Configuration de la glotte en voix soufflée. D'après (Eckert & Laver, 1994).

- La voix murmurée (*whispery voice*, Figure 13), est caractérisée par une tension adductive très faible, avec une compression médiale et une tension longitudinale relativement importantes. Ceci se traduit au niveau physiologique par une ouverture triangulaire des cartilages aryténoïdes. La vibration des plis vocaux, confinée à la partie compressée de la glotte ligamentaire, est très inefficace, et le bruit de friction perceptible est considérable.

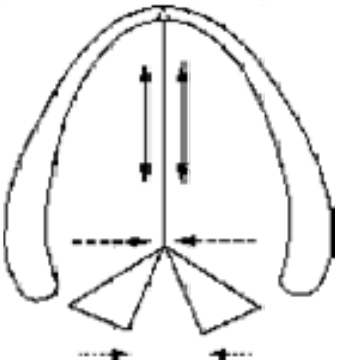


Figure 13 : Configuration de la glotte en voix murmurée. D'après (Eckert & Laver, 1994).

- Dans la taxonomie de Laver (1980), la voix craquée (*vocal fry*[17], Figure 14) correspond à une tension adductive et une compression médiale importantes, avec une tension longitudinale faible. Ce mode de phonation est associé à une fréquence fondamentale très basse. De plus, la fréquence fondamentale et l'amplitude des pulsations glottiques consécutives sont extrêmement irrégulières. La vibration des plis vocaux ne se produit que dans leur partie ligamentaire.

[17] En raison notamment de l'assimilation très fréquente à la voix craquée, aucune traduction française du terme *vocal fry* n'a été proposée dans la littérature à notre connaissance et il nous semble peu approprié de le traduire par un terme tel que « friture vocale » qui, s'il peut évoquer l'irrégularité de la phonation, rend moins bien compte de la fréquence fondamentale très basse associée. Nous conserverons donc le terme anglais.

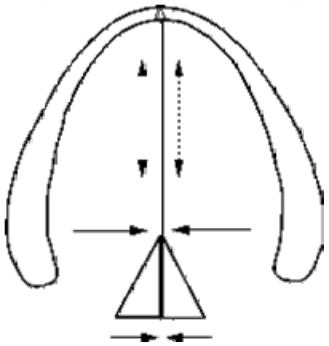


Figure 14 : Configuration de la glotte en voix craquée. D'après (Eckert & Laver, 1994).

- La voix pressée (*pressed voice*), fréquemment appelée voix tendue (*tense voice*) implique une tension de l'ensemble du conduit vocal plus importante que pour la voix modale. Au niveau du larynx, cette augmentation de la tension musculaire est particulièrement sensible pour la tension adductive et la compression médiale. Dans ce mode de phonation, la pulsation glottique est plus asymétrique que dans le cas d'une voix modale, et on observe une phase fermée plus longue.
- La voix relâchée (*lax voice*), qui s'oppose à la voix pressée, implique une tension plus faible de l'ensemble du conduit vocal, et particulièrement un relâchement de la tension adductive et de la compression médiale. Cette tension musculaire reste toutefois plus importante que dans le cas de la voix soufflée.
- La voix de fausset (*falsetto voice*, figure 15) caractérisée principalement par une tension longitudinale très importante qui a pour effet de rendre les plis vocaux plus fins. Dans cette configuration la tension adductive et la compression médiale sont également importantes Ce type de phonation est associé à une fréquence fondamentale très élevée.

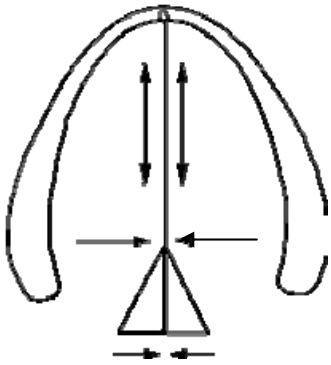


Figure 16 : Configuration de la glotte en voix de fausset. D'après (Eckert & Laver, 1994).

7.3.2 Quelques raffinements de la taxonomie de Laver

Si la taxonomie établie par Laver (1980) est la plus couramment utilisée pour désigner les différentes qualités de voix produites dans une phonation normale, plusieurs auteurs en ont proposé des raffinements. En particulier, Ní Chasaide & Gobl (1997) ont proposé à partir de leurs observations de compléter la taxonomie de Laver (1980) par la configuration relâchée-craquée (*lax creaky*).

A la suite de Laver (1980), les termes de *vocal fry* et de voix craquée ont fréquemment été utilisés l'un pour l'autre, et parfois désignés par des termes tels que glottalisation ou encore phonation diplophonique, sans qu'une distinction claire ne soit établie entre ces modes de phonation. Geratt & Kreiman (2001), notant qu'une grande confusion règne dans la littérature quant à la désignation des différents types de phonation irrégulière, montrent que le mécanisme de *vocal fry*, la diplophonie et les modulations de l'amplitude possèdent des propriétés perceptives distinctes et proposent de considérer ces types de phonation de façon distincte. Le mécanisme de *vocal fry* a été décrit de façon détaillée par Blomgren *et al.* (1998) à partir de mesures acoustiques, aérodynamiques et électroglottographiques, qui ont confirmé que ce mécanisme était associé à une plage de valeurs de fréquence fondamentale nettement inférieure à celle de la phonation modale, ainsi qu'à des microvariations de la fréquence fondamentale (*jitter*) ou de l'intensité (*shimmer*) supérieures à celles mesurées en phonation modale. Yegnanarayana *et al.* (1998) ont de plus montré que, outre le jitter et le shimmer, les apériodicités perceptibles du signal acoustique peuvent provenir d'une fuite glottique. De plus, d'Alessandro (2006) indique que la voix craquée peut résulter d'une vibration non des plis vocaux mais des bandes ventriculaires (aussi appelées « fausses cordes vocales »).

Il faut noter toutefois que les catégories de qualités de voix étiquetées par Laver ne sont pas mutuellement exclusives. Ainsi, selon Laver (1980), il est possible d'observer des qualités de voix relevant de plusieurs de ces catégories, comme un *vocal fry* combiné avec une voix murmurée. Ainsi, bien que les descriptions articulatoires données par Laver pour ces deux types de qualité de voix soient proches, Shochi (2008) observe perceptivement la réalisation conjointe d'une voix soufflée et tendue pour l'expression de la surprise japonaise. L'existence de ces qualités de voix composées est toutefois remise en cause par Roach (2000), pour qui la définition d'étiquettes complémentaires telles que la notion d'« attaque glottale » est nécessaire.

7.3.3 Descriptions dimensionnelle de la qualité de voix

Les descriptions données par Laver (1980) et résumées ci-dessus ne constituent cependant que des points particuliers dans l'espace des qualités de voix existantes. Ainsi, les types de voix soufflée et murmurée forment un continuum auditif, sans frontière claire entre eux. Ladefoged (1983), tout en admettant qu'une telle projection constitue une simplification à l'extrême des mécanismes mis en œuvre, propose de projeter les qualités de voix selon un continuum allant de l'ouverture maximale de la glotte permettant la phonation avec un voisement de faible intensité (voix murmurée) à l'ouverture minimale (voix craquée). Bien qu'une telle projection puisse vraisemblablement rendre compte d'une part importante des variations perçues, elle revient à ne prendre en considération que la dimension tendue-relâchée, c'est-à-dire la variation conjointe de la tension adductive et de la compression médiale.

Comme nous l'avons mentionné, le foisonnement dans la littérature de termes souvent divergents et de descriptions ne se recouvrant que partiellement rend difficile une description claire de la qualité de voix. La description la plus aboutie à notre connaissance dans l'optique d'études prosodiques étendues à la qualité de voix est celle donnée par d'Alessandro (2006). Il propose une description de la qualité de voix en considérant quatre dimensions de description de la source glottique, et donne quelques indications de l'usage qui est fait de chacune en parole :

- Les registres vocaux, décrits notamment par Roubeau *et al.* (1997) et consistant en quatre mécanismes, dont les trois premiers sont utilisés dans la parole. Dans le mécanisme 0, qui correspond au *vocal fry*, les plis vocaux épais et lourds vibrent irrégulièrement à une fréquence très basse en raison de la tension longitudinale faible. Dans le mécanisme 1, qui correspond à la voix modale, les plis vocaux épais et lourds vibrent régulièrement sur toute leur longueur. Dans le mécanisme 2 qui correspond à la voix de fausset les plis vocaux rendus plus légers et plus fins par la tension musculaire plus importante ne vibrent que sur les deux tiers de leur partie antérieure.
- Les apériodicités présentes dans le signal, dues à un bruit additionnel continu ou à un bruit structurel comme une variation de la période (*jitter*) ou de l'amplitude (*shimmer*) des vibrations. Les bruits additionnels, responsables de l'essentiel des apériodicités en parole non-pathologique, consistent en un flux turbulent au niveau de la constriction glottique observé en voix murmurée et voix soufflée.
- La dimension voix relâchée-voix tendue, directement liée à la position des cartilages aryténoïdes.

- L'effort vocal qui se traduit par la sonie, et dont les mécanismes articulatoires sont d'après d'Alessandro (2006) encore mal connus. L'effort vocal pourrait être lié à l'action des cartilages crico-thyroïdes ainsi qu'à la contraction du muscle vocal (*vocalis*), ainsi qu'à l'augmentation de la pression subglottique.

  Ainsi, les qualités de voix décrites par Laver (1980) et présentées ci-dessus se répartiraient selon les dimensions du registre de phonation, des apériodicités présentes dans le signal acoustique, et de la dimension relâché-tendu, la dimension d'effort vocal n'étant pas prise en compte par Laver.

7.4. Evaluation perceptive de la qualité de voix

Outre l'analyse par la synthèse de la qualité de voix, qui ne peut toutefois systématiquement remplacer l'analyse de données, une alternative pourrait être de recourir à des mesures perceptives de la qualité de voix. Ainsi, le protocole de jugement perceptif de la qualité de voix VPAS introduit par Laver (2000) fournit à des évaluateurs entraînés spécifiquement une grille d'analyse leur permettant de juger une qualité de voix selon un ensemble très complet de configurations laryngées et supralaryngées. Bien que ce protocole soit initialement destiné à l'évaluation de la voix pathologique en orthophonie, il a également dans quelques cas été appliqué avec succès à l'évaluation de la qualité de voix en phonation normale. Coadou (2006) l'a ainsi appliqué à la caractérisation de l'accent de l'accent de la région de Liverpool comparativement à d'autres accents régionaux britanniques. Viola & Madureira (2007) l'ont quant à elles utilisé pour décrire les variations de qualité de voix liés à l'expressivité dans le conte en portugais brésilien.

Cependant, ce protocole n'a pas à notre connaissance été utilisé dans le cadre de la description de variations locales de la qualité de voix, et il semble difficilement réalisable de l'appliquer à la mesure de variations fines permettant d'obtenir des « contours de qualité de voix » d'une résolution comparable à celle de contours de fréquence fondamentale tels que ceux que nous examinons en section 8.

Une autre limitation de l'évaluation perceptive de la qualité de voix est que, faute de consensus dans le langage courant quant à la définition des traits perçus, elle ne peut que difficilement être l'objet d'une évaluation perceptive par des auditeurs naïfs. Kreiman *et al.* (1992) ont de plus montré à partir d'une tâche de jugements de similarité par paires une importante variabilité interindividuels dans l'évaluation de la qualité de voix normale et pathologique, tant par des juges naïfs que par des experts. Ce dernier résultat est toutefois en désaccord avec les conclusions de Bänziger & Scherer (2003), qui ont fait évaluer par des auditeurs naïfs pour des expressions émotionnelles actées un ensemble

d'attributs supposés liés à la qualité de voix mais formulés de façon à être facilement interprétables. Les indices de fiabilité inter-juge qu'ils obtiennent sont élevés, cependant ils indiquent que cela pourrait être lié au caractère prototypique des expressions émotionnelles évaluées. De plus, les auteurs concluent qu'il n'est pas possible l'hypothèse selon laquelle les jugements d'attributs liés aux caractéristiques de la voix seraient influencés par les émotions exprimés par les acteurs et perçues par les auditeurs.

7.5. Mesures acoustiques de la qualité de voix

7.5.1 Mesures spectrales

Comme nous l'avons évoqué en introduction de cette section, il ne fait aucun doute que des corrélats acoustiques de la qualité de voix peuvent être trouvés dans l'organisation du spectre des signaux de parole. De plus, des éléments convergents indiquent que les variations de certaines mesures spectrales peuvent être liées aux émotions exprimées. En particulier, l'étude menée par Banse & Scherer (1996) sur un corpus multilocuteurs d'expressions émotionnelles actées indique que certains des paramètres spectraux calculés sur la partie voisée permettent d'expliquer une part importante de la variance, indépendamment des paramètres prosodiques plus « classiques ». Ces paramètres spectraux sont l'indice de Hammarberg *et al.* (1980) défini comme la différence entre les valeurs maximales d'énergie dans les bandes de fréquence 0-2000 Hz et 2000-5000 Hz, la proportion relative d'énergie au dessus et en dessous de 1000 Hz notée PE1000, et le gradient de l'approximation par la méthode des moindres carrés de la pente spectrale au dessus de 1000 Hz notée DO1000.

L'étude de Banse & Scherer (1996) a été menée de façon très contrôlée sur des énoncés identiques d'une émotion à l'autre et reste parfaitement valide. Cependant des mesures telles que l'indice de Hammarberg *et al.* (1980) ou les paramètres PE1000 et DO1000 présentent l'inconvénient de ne pas permettre de comparaison inter-voyelles, contrairement aux paramètres prosodiques classiques. En effet, ces paramètres classiques, s'ils sont également sujets à des variations intrinsèques en fonction de la qualité de la voyelle (voir par exemple Rossi *et al.* (1981, p. 22) pour la fréquence fondamentale ou Rossi (1971) pour l'intensité), varient dans une beaucoup plus large mesure en fonction de variations linguistiques ou affectives, ce qui permet par exemple de comparer des variations de fréquence fondamentale sur des énoncés construits sur des substrats lexicaux différents.

A l'inverse, des mesures spectrales fondées sur des proportions relatives d'énergie dans diverses bandes fréquentielles sont extrêmement affectées par la qualité de la voyelle. En

effet, par définition leurs valeurs varient de façon très sensible en fonction de la position des formants. Lorsque les corrélats acoustiques des variations émotionnelles sont mesurés sur des énoncés identiques, ces mesures peuvent être considérées comme fiables. Il a été montré que la variation de la position des formants, et notamment celle de F1 (voir la revue de littérature réalisée par Juslin & Laukka (2003)) est liée à la variation émotionnelle. Ceci rejoint la proposition de Pfitzinger (2006) d'une prosodie étendue au degré d'articulation, testée spécifiquement par Beller *et al.* (2008) dans le cas d'expressions émotionnelles actées.

De plus, bien qu'à qualité de voyelle constante de telles mesures spectrales permettent certainement de capturer des variations relatives à la qualité de voix, elles peuvent difficilement être liées directement à des phénomènes articulatoires ou perceptifs (Doval *et al.*, 2006).

D'Alessandro (2006) indique que l'énergie de la source est principalement concentrée dans les basses fréquences. Il note que l'un des principaux corrélats spectraux de l'excitation glottique est le maximum d'énergie de la source appelé « formant glottique », dont la fréquence est inférieure à celle du premier formant. Ce pic est souvent repérable dans le spectrogramme du signal de parole, dans lequel il est désigné sous le terme de « barre de voisement ».

Il est alors possible d'étudier les propriétés de la source via les propriétés de ce formant glottique. Doval & d'Alessandro (1997) montrent ainsi que la fréquence du formant glottique est inversement proportionnelle au quotient ouvert, cette fréquence étant donc d'autant plus basse que la voix est relâchée. De plus, et bien que la relation soit complexe, il est possible de considérer qu'une onde de débit glottique asymétrique a pour conséquence un formant glottique plus large et de fréquence plus élevée.

D'Alessandro (2006) indique ainsi que le formant glottique est localisé à une fréquence légèrement inférieure à la fondamentale pour des valeurs typiques du coefficient d'asymétrie (c'est-à-dire une valeur de 2/3, qui est la valeur fixée par le modèle de source KLGLOTT88 (Klatt & Klatt, 1990) pour ce coefficient) et un quotient ouvert compris entre 0.5 et 1. Il note qu'en revanche ce formant glottique peut s'approcher du quatrième harmonique pour une valeur de quotient ouvert de 0.4 et un coefficient d'asymétrie de 0.9.

Enfin il souligne que la pente en hautes fréquences de ce formant glottique, liée à la notion de pente spectrale que nous développons dans la section suivante, reflète la vitesse de fermeture de la glotte.

De très nombreuses mesures spectrales supposées rendre compte de phénomènes de qualité de voix ont été proposées dans la littérature. Cependant, comme le montrent Kreiman *et al.* (2007) à partir de l'étude de voix normales et pathologiques de 70 locuteurs, la plupart de ces mesures sont fortement redondantes. De plus les auteurs concluent que, à l'exception des mesures globales de l'enveloppe spectrale et de la différence d'amplitude entre les deux premiers harmoniques, ces mesures ne permettent pas de rendre compte de façon satisfaisante des variations de la forme de l'onde de débit glottique. En conséquence, nous n'aborderons que le cas des mesures globales de l'enveloppe spectrale et de la différence d'amplitude entre les deux premiers harmoniques.

7.5.2 Mesures de la pente spectrale en hautes et basses fréquences

De nombreuses mesures différentes de la pente spectrale (dénotée selon les auteurs par les termes *spectral slope* ou *spectral tilt* dans la littérature anglophone) ont été proposées. L'une des raisons de la divergence des définitions de la pente spectrale est le manque de consensus quant aux corrélats acoustiques de la forme de l'onde de débit glottique.

Le terme de *spectral slope* a principalement été utilisé pour désigner diverses approximations de la forme générale de l'enveloppe spectrale, prenant également en compte les basses fréquences. Alku *et al.* (1997) ont ainsi introduit le paramètre PSP, calculé à partir d'une approximation polynomiale du second ordre de l'enveloppe spectrale. Childers & Lee (1991) ont pour leur part proposé de calculer le facteur de richesse harmonique HRF comme le rapport entre la somme des amplitudes des harmoniques de fréquence plus élevée que F0 et l'amplitude du premier harmonique H1.

Doval *et al.* (2006), qui ont apporté la contribution théorique la plus complète à la description du spectre des modèles de débit glottique, utilisent le terme de *spectral tilt* pour désigner l'atténuation de l'énergie dans les hautes fréquences due à une fermeture moins abrupte de la glotte, déjà notée par Fant *et al.* (1985).

Cependant le terme de *spectral tilt* a également été très largement utilisé pour désigner les caractéristiques spectrales en basses fréquences. En particulier, il est utilisé par certains auteurs (voir par exemple Beckman & Campbell (1997), Swerts & Veldhuis (2001)) pour désigner la différence entre les amplitudes des deux premiers harmoniques, H1-H2[18].

[18] Bien que le décalage des indices utilisés puisse prêter à confusion, l'amplitude H1 du premier harmonique désigne l'amplitude du spectre à la fréquence F0, et l'amplitude H2 du deuxième harmonique l'amplitude du spectre à la fréquence 2*F0.

Afin d'éviter toute confusion, nous utiliserons dans la suite le terme de pente spectrale pour désigner l'atténuation de l'énergie, et conserverons le terme de différence entre les amplitudes des deux premiers harmoniques ou sa notation abrégée H1-H2 pour désigner cette quantité.

Les corrélats perceptifs de ce paramètre sont bien établis. Beckman & Campbell (1997) ont par exemple montré que ses variations pouvaient être reliées aux différents types d'accentuation en anglais américain.

La différence entre les amplitudes des deux premiers harmoniques a été liée de façon claire au quotient ouvert. Hanson (1995), ainsi que (Holmberg *et al.*, 1995) et Fant (1997), établissent ainsi une relation univoque entre le quotient ouvert et ce paramètre, qui permettrait ainsi de caractériser le continuum voix relâchée-voix tendue. Cependant Swerts & Veldhuis (2001) relèvent que la quantité H1-H2 dépend non seulement du quotient ouvert, mais également du coefficient d'asymétrie. Doval *et al.* (2006) renforcent ce résultat en établissant à partir des modèles de débit glottique la relation entre la différence d'amplitude entre les deux premiers harmoniques, le quotient ouvert et le coefficient d'asymétrie. Ils notent que la principale raison pour laquelle Hanson (1995) a trouvé une relation univoque entre H1-H2 et le quotient ouvert est qu'elle s'est appuyée sur le modèle de débit glottique KLGLOTT88 (Klatt & Klatt, 1990) dans lequel le coefficient d'asymétrie est fixé à une valeur constante de 2/3.

De même que la plupart des autres paramètres spectraux proposés dans la littérature et en particulier ceux relatifs aux basses fréquences, la différence H1-H2 est sensible à la qualité de la voyelle. Afin de pallier cette limitation, Hanson (1995) propose une mesure corrigée H1*-H2* dans laquelle les amplitudes des deux premiers harmoniques sont corrigées de la valeur du premier formant. H1* et H2* sont obtenus par les équations :

$$H1^* = H1 - 20\log_{10}\frac{F1^2}{F1^2 - F0^2}, \text{ et } H2^* = H2 - 20\log_{10}\frac{F1^2}{F1^2 - 2F0^2}$$

Hanson (1995) reste cependant prudente dans ses conclusions, et précise que la formule de correction qu'elle propose n'est valable que pour les voyelles basses, seules recueillies et testées dans son corpus. Il semble hasardeux d'étendre cette correction à des voyelles hautes, tout particulièrement dans le cas d'une fréquence fondamentale élevée dont résulterait une fréquence du second harmonique proche de la position du premier formant. En effet les formules proposées reposent sur une hypothèse simplificatrice dans laquelle les formants sont assimilés à une résonance « idéale » de largeur de bande nulle, qui suppose pour être considérée comme valide que la fréquence fondamentale soit très inférieure à la position du premier formant.

Le terme correctif soustrait de l'amplitude du second harmonique a un comportement asymptotique lorsque cette hypothèse n'est pas vérifiée et que F1 est proche de la fréquence fondamentale ou du double de la fréquence fondamentale. Dans le cas des voyelles du français, Tubach (1989) indique pour les voyelles [i], [u] et [y] des valeurs moyennes de F1 comprises entre 300 Hz et 315 Hz selon la voyelle et le genre du locuteur. Ainsi une fondamentale de l'ordre de 150 à 160 Hz, courante pour une voix d'homme en parole conversationnelle, ou une fondamentale proche de ces valeurs de F1, courante pour une voix de femme (Henrich, 2001) aura pour conséquence des valeurs totalement irréalistes de H1*-H2* sur une voyelle haute.

Iseli & Alwan (2004) proposent une généralisation de la formule de correction proposée par Hanson (1995) en prenant en compte l'intégralité des formants et leurs largeurs de bande. Selon leurs propositions, l'amplitude corrigée de l'influence des formants de chacun des harmoniques considéré pourrait ainsi être calculée quelle que soit la voyelle et le niveau de la fréquence fondamentale comme :

$$H_k^* = H_k - \sum_{i=1}^{n} 10\log_{10} \frac{(r_i{}^2 + 1 - 2 \cdot r_i \cdot \cos(\omega_i))^2}{(r_i{}^2 + 1 - 2 \cdot r_i \cdot \cos(\omega_i + \omega)) \cdot (r_i{}^2 + 1 - 2 \cdot r_i \cdot \cos(\omega_i - \omega))},$$

où $r_i = e^{-\pi \frac{B_i}{Fs}}$, $\omega_i = \frac{2\pi \cdot F_i}{Fs}$ et $\omega = \frac{2\pi \cdot k \cdot F_0}{Fs}$, n étant le nombre de formants considérés, Fs la fréquence d'échantillonnage, F0 la fréquence fondamentale, F_i la fréquence du $i^{ème}$ formant et B_i sa largeur de bande.

La mesure des largeurs de bande des formants, et en particulier de celle du premier formant dont l'influence est ici prépondérante, revêt donc un caractère crucial pour l'estimation fiable du terme correctif appliqué aux amplitudes mesurées des harmoniques. Sans remettre en cause la validité de la démonstration d'Iseli & Alwan (2004) sur les modèles de débit glottique, une telle approche nécessite donc de pouvoir mesurer avec précision la largeur de bande du premier formant, ce qui peut se révéler difficile lorsque la fréquence fondamentale est proche de la fréquence de ce premier formant, notamment dans le cas de voyelles hautes.

En conséquence, il semble difficile de réaliser des comparaisons inter-voyelles de valeurs de différence d'amplitude entre les deux premiers harmoniques. Il paraît donc impératif que les comparaisons des valeurs prises par ces paramètres spectraux soient effectuées sur les mêmes voyelles. Bien que dans ces conditions l'application de termes correctifs tels que celui proposé par Hanson (1995) puisse de prime abord sembler moins crucial, de telles corrections revêtent toutefois un intérêt indéniable. En effet, comme nous l'avons souligné ci-dessus, l'un des effets de l'expression d'affects dans la parole est une variabilité accrue de la réalisation des cibles formantiques pour une même voyelle. Ces

corrections, à condition d'être appliquées à des voyelles suffisamment basses pour que les mesures intermédiaires soient réalisables, pourraient donc permettre d'obtenir une estimation du quotient ouvert et de l'asymétrie de l'onde de débit glottique indépendamment des variations dans l'articulation supraglottique.

7.5.3 Mesure des apériodicités

Les mesures acoustiques les plus classiques des apériodicités évaluent les apériodicités structurelles que constituent les variations de période (*jitter*) ou d'amplitude (*shimmer*). Diverses mesures du jitter et du shimmer existent, notamment en fonction du nombre de périodes consécutives prises en compte dans le calcul. Fourcin & Abberton (2008) indiquent toutefois que la définition la plus commune du taux de jitter est la différence moyenne entre les durées de deux périodes consécutives divisée par la durée moyenne d'une période.

Une seule étude à notre connaissance s'est attachée à évaluer les variations intrinsèques du jitter et du shimmer ainsi que d'autres mesures de perturbations en fonction de la voyelle sur laquelle ces mesures sont effectuées (Akif Kiliç *et al.*, 2004). Les auteurs, qui ont évalué ces mesures en phonation non-pathologique sur les différentes voyelles du turc ont mis en évidence un effet de la voyelle sur l'ensemble des mesures de perturbation, ainsi qu'une faible corrélation inverse avec la position du premier formant.

D'Alessandro (2006) indique que le degré de voisement fournit une bonne estimation des bruits ajoutés au niveau de la source glottique, et donc des apériodicités : en effet plus le voisement est important, plus le signal résultant est périodique. Il indique également que l'inclinaison spectrale peut être un bon indicateur de ce type d'apériodicités, de même que la mesure du bruit inter-harmoniques (Childers & Lee, 1991).

La mesure la plus classique de la proportion d'apériodicités est le rapport de l'énergie présente dans les parties harmoniques sur celle présente dans les parties inharmoniques du signal, aussi appelée HNR (*Harmonics to Noise Ratio*). Ce rapport (voir par exemple Bailly (2001) pour un état de l'art et une évaluation des mesures de HNR proposées dans la littérature) permet notamment de caractériser la phonation soufflée (de Krom, 1995) ainsi que l'enrouement[19] de la voix (Yumoto *et al.*, 1982).

La variation intrinsèque du ratio HNR en fonction de la voyelle n'a à notre connaissance pas été spécifiquement étudiée. Cependant Bailly (2001) indique des ordres de grandeur proches pour les voyelles du français [a], [i], [y] et [u] à partir de 24 mesures par voyelle,

[19] Traduction du terme anglais *hoarseness*.

ce qui tendrait à indiquer que, contrairement à la majorité des autres mesures acoustiques de qualité de voix, ce ratio peut être appliqué à des comparaisons inter-voyelles.

7.6. Paramétrisation de la qualité de voix

7.6.1 Description de l'onde de débit glottique

L'onde de débit glottique permet de caractériser les types de phonation impliqués dans la qualité de voix, aucun instrument ne permettant une mesure directe des paramètres de tension musculaire décrits par Laver sans perturbation importante de la phonation. Des variations de la forme de l'onde de débit glottique ont pu être liées à l'expression de divers affects, notamment par Cummings & Clements (1995) à partir de l'analyse d'expressions actées, et par Ní Chasaide & Gobl (2003) dans une méthodologie d'analyse par la synthèse.

Divers modèles ont été proposés pour caractériser cette onde de débit ainsi que sa dérivée, en particulier le modèle LF de Liljencrants et Fant (Fant *et al*., 1985) et le modèle KLGLOTT88 (Klatt & Klatt, 1990). D'Alessandro (2006) relève que, parmi ces modèles, le plus largement utilisé a été le modèle LF.

Un ensemble équivalent de cinq paramètres peut être utilisé pour décrire l'onde de débit glottique, quel que soit le modèle retenu (Doval & d'Alessandro, 1999 ; Ní Chasaide & Gobl, 2000). Doval *et al.* (2006) proposent de retenir le jeu de paramètres schématisés dans la Figure 17, et définis comme suit :

- La vitesse de fermeture E, exprimée en $m^3.s^{-2}$, qui correspond à la vitesse maximale du débit glottique à l'instant de fermeture et est mesurée par l'amplitude maximale de l'onde de débit dérivée. Ce pic, situé à l'instant Te = OQ*T0, est considéré par Doval *et al.* (2006) comme plus pertinent que l'amplitude de voisement Av exprimée en $m^3.s^{-1}$ plus classiquement utilisée.
- La période fondamentale T0 exprimée en secondes et définie comme l'inverse de F0.
- Le quotient ouvert OQ (o*pen quotient*), quantité sans dimension définie comme le rapport entre la durée de la phase ouverte et la période fondamentale. OQ peut théoriquement prendre des valeurs comprises entre 0 et 1. Bien que n'étant pas lié directement à la tension musculaire, le quotient ouvert fournit une bonne approximation du continuum relâché-tendu. En pratique des valeurs comprises entre 0.30 dans le cas d'une phonation très tendue et 0.98 pour une phonation très relâchée sont observées pour ce paramètre. Pour une voix modale, la valeur prise par OQ est

de l'ordre de 0.50 (Henrich, 2001). Le quotient ouvert est lié à l'instant de fermeture glottique Te par la relation Te = OQ*T0, et est d'un point de vue perceptif principalement lié à la dimension tendu-relâché. Dans le domaine de la parole expressive, la mesure du quotient ouvert par électroglottographie (voir dans la suite de cette section pour un exposé de cette technique) a notamment permis à Gendrot (2004) de séparer clairement des expression actées de joie et de colère, réputées proches d'un point de vue acoustique.

- Le coefficient d'asymétrie α_m, sans dimension, défini comme le rapport entre le temps d'ouverture et la durée de la phase ouverte. Ce coefficient permet en relation avec le quotient ouvert de déterminer l'instant Tp du pic de l'onde de débit glottique par la relation Tp = α_m *OQ*T0 (cf. Figure 17). D'après d'Alessandro (2006), une fermeture abrupte est liée à un coefficient d'asymétrie élevé.
- Le quotient de la phase de retour Qa, sans dimension, défini comme le rapport entre la durée effective Ta de la phase de retour et la durée de la phase fermée. Ce quotient, nul dans le cas d'une fermeture abrupte de la glotte, est défini comme :

$$Q_a = \frac{T_a}{(1 - OQ) \cdot T0}$$

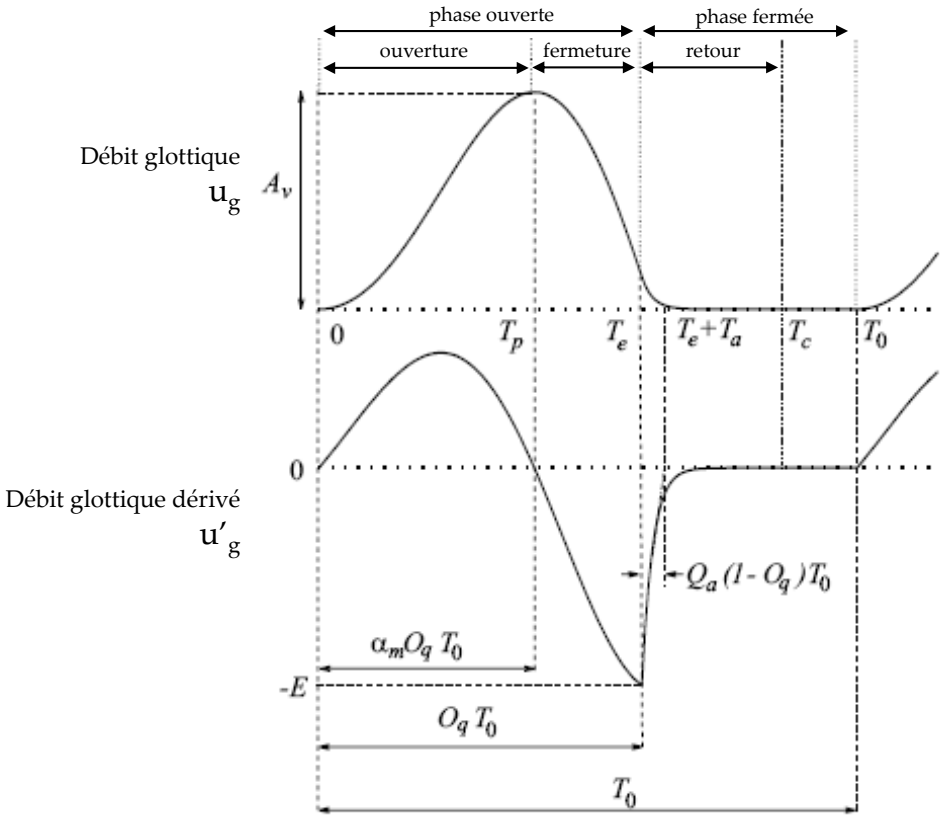


Figure 17 : Paramétrisation de l'onde de débit glottique u_g et de sa dérivée u'_g. D'après Doval *et al.* (2006).

Deux autres quantités, qui peuvent être calculées à partir de ces cinq paramètres, sont fréquemment rencontrées dans la littérature :

- La durée effective Ta (exprimée en secondes) de la phase de retour, également désignée dans la littérature par le terme de « temps de déclinaison » (Fant, 1997). Cette durée est équivalente au paramètre AQ (*Amplitude Quotient*) estimé comme le rapport entre l'amplitude de voisement Av et le pic E de l'onde de débit dérivée (Alku & Vilkman, 1996), sur lequel nous revenons dans la suite de cette section.
- Le quotient de vitesse SQ (S*peed Quotient*), défini comme le rapport entre la durée de la phase ouverte et la durée de la phase de fermeture, est une mesure de l'asymétrie de l'impulsion glottique. Ses valeurs, qui en théorie peuvent être comprises entre 1 et 9, varient en pratique entre 2 et 3 (Henrich *et al.*, 2001).

7.6.2 Mesure par électroglottographie (EGG)

7.6.2.1 Principe

Un ensemble de techniques, dont un tour d'horizon est donné par Airas (2008), ont été proposées pour mesurer les vibrations des plis vocaux au cours de la phonation. D'autres méthodes telles que l'usage d'un masque spécial appelé « masque de Rosenberg » (Sundberg, 2002) permettent de mesurer le débit d'air expiré afin de faciliter l'estimation du débit au niveau glottique.

Cependant, outre le caractère plus ou moins invasif et susceptible de perturber la phonation de la plupart ces techniques, une mesure directe de l'onde de débit glottique est impossible en pratique. Airas (2008, p. 52-53) note ainsi dans le contexte de l'évaluation du filtrage inverse (technique d'estimation de l'onde de débit glottique dont nous exposons brièvement le principe dans la suite de cette section) que :

« *Bien que [les diverses techniques utilisées pour valider les méthodes de filtrage inverse] permettent d'attester le comportement des plis vocaux, elles ne mesurent pas directement le débit glottique, mais uniquement certaines quantités qui lui sont liées. Aucune correspondance biunivoque n'existe entre le débit glottique et ces valeurs, et les informations qu'elles donnent sont donc toujours limitées.* »[20]

[20] Traduction personnelle de la citation originale « *While [the techniques used to assess the correctness of inverse filtering methods] assess the vocal fold behaviour, they do not measure the glottal airflow, but just some related quantities. No one-to-one mapping exists between the glottal flow and their values, and therefore information given by them is always limited.* »

L'électroglottographie est une technique non invasive de mesure de l'accolement des plis vocaux (voir notamment Fourcin & Abberton (1971) qui ont développé et popularisé cette technique de mesure). L'électroglottographe, qui ne perturbe pas la phonation, consiste en un collier contenant deux électrodes placées sur le cou, de chaque côté du cartilage thyroïde, entre lesquelles circule un courant de 4 MHz. Le contact des plis vocaux, ou adduction, diminue la résistance et amène la conductance à un maximum, et inversement.

Le tracé de la variation d'impédance obtenue est appelé glottogramme ou simplement signal EGG. Le signal EGG peut se décomposer en une composante haute fréquence appelée Lx qui est la mesure des variations d'impédance résultant des mouvements de plis vocaux présentée ci-dessus, et une composante basse fréquence appelée Rx qui résulte des mouvement plus lents du larynx lors du processus de phonation. Seule la composante Lx est pertinente pour le calcul de paramètres liés à la qualité de voix, et il convient donc d'éliminer la composante Rx préalablement à tout traitement. Ceci peut-être effectué en mesurant les mouvements du larynx grâce à des électrodes placées sur le cou du sujet, ou plus communément par un filtrage passe-haut (Rothenberg, 1992).

Comparativement aux autres techniques existantes, le caractère non-invasif de l'électroglottographie constitue un avantage certain pour la mesure des variations de qualité de voix liées à l'expressivité. De plus, cette technique permet de mesurer de façon fiable la fréquence fondamentale, y compris dans le cas de régimes phonatoires pour lesquels la mesure à partir du seul signal acoustique peut poser problème. L'électroglottographie apporte donc des renseignements très précis sur les phases fermée, de fermeture et d'ouverture, mais elle ne permet pas d'obtenir des informations fiables sur la phase ouverte du cycle de vibration des plis vocaux, car les plis vocaux n'étant alors pas en contact, il n'enregistre plus de variation du signal électrique. Henrich (2001) donne une description détaillée de cette technique, de ses possibilités et de ses limites. Le signal EGG permet ainsi de mesurer le quotient d'ouverture, via la mesure de la durée de la phase ouverte T1, mais des mesures telles que celle de la durée de la phase de fermeture sont impossibles (Henrich, 2001, p. 96).

La Figure 18 présente la correspondance d'un enregistrement EGG et de sa dérivée avec une onde de débit, et l'extraction du paramètre T1 utilisé dans le calcul du quotient ouvert selon l'algorithme fondé sur une intercorrélation entre le signal EGG et le signal acoustique développé par Henrich (2001).

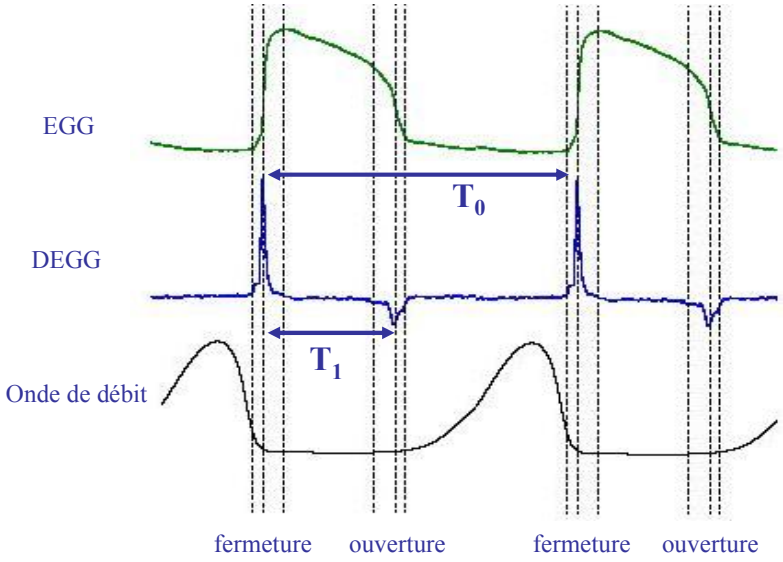


Figure 18 : Exemple de signal EGG, de sa dérivée DEGG et de l'onde de débit correspondante. Les phases d'ouverture et de fermeture ainsi que les pics correspondants sont indiqués par les lignes pointillées. D'après Henrich (2001).

L'électroglottographie a peu été utilisée dans le domaine de l'étude de la parole émotionnelle. Cependant l'étude de Gendrot (2004) peut être mentionnée. Il a pu grâce à des mesures électroglottographiques du quotient ouvert séparer clairement des expressions actées de joie et de colère réputées mal discriminées par des mesures de paramètres prosodiques plus classiques.

7.7. Synthèse de la qualité de voix

Nous avons principalement présenté ci-dessus les aspects relatifs à l'analyse de la qualité de voix dans le cadre d'une méthodologie hypothético-déductive. Cependant, il est également envisageable d'adopter une méthodologie déductive pour étudier la qualité de voix en tant que telle ou les fonctions qu'elle réalise.

Bőhm & Shattuck-Hufnagel (2007), partant du constat que la voix craquée ou glottalisation se traduit par des irrégularités dans le signal acoustique, proposent une méthode simple de rééchelonnage de périodes individuelles afin de produire une voix craquée artificielle. Ils appliquent cette méthode à la reconnaissance par des auditeurs naïfs de la voix de locuteurs natifs de l'anglais américain à partir de sa tendance à produire une telle glottalisation en fin d'énoncé. En collaboration avec Tamas Bőhm (Bőhm, Audibert, Shattuck-Hufnagel, Nemeth & Aubergé, 2008), nous avons proposé une semi-automatisation de cette méthode afin de permettre d'appliquer un patron de glottalisation extrait d'une portion d'un énoncé naturellement glottalisé à une portion

sélectionnée d'un énoncé produit avec une phonation modale[21]. Les jugements perceptifs de naturalité et d'enrouement, de même que la plupart des mesures acoustiques effectuées, n'ont par mis en évidence de différence entre les stimuli glottalisés artificiellement ou naturellement.

Kreiman *et al.* (2001) ont ainsi proposé une méthode d'analyse/synthèse en temps réel de caractéristiques de la qualité de voix, afin d'aboutir à une évaluation plus fiable des caractéristiques de la voix pathologique et contourner ainsi les limites liées à l'analyse spectrale ou à l'évaluation perceptive de la qualité de voix.

Gobl & Ní Chasaide (2003a) ont quant à eux synthétisé des stimuli en modifiant la forme de l'onde de débit paramétrisée par le modèle LF par rapport à celle estimée pour une phonation modale, afin de correspondre aux différents modes de phonation décrits par Ní Chasaide & Gobl (2000) et constituant un prolongement de la taxonomie proposée par Laver (1980). Ils ont ensuite fait évaluer par un ensemble d'auditeurs les affects perçus comme véhiculés par ces différents stimuli. Nous revenons sur les résultats de l'évaluation perceptive de ces stimuli en section 9.2.

7.8. Estimation par filtrage inverse

7.8.1 La théorie source-filtre (Fant, 1960)

La théorie source-filtre de production de la parole (Fant, 1960) considère le signal acoustique d'un son de parole comme la résultante du son produit au niveau de la glotte par le passage de l'air en provenance des poumons, modifié lors du passage au travers du conduit vocal, puis à nouveau lors du passage par les lèvres. La pente spectrale de l'excitation glottique est en moyenne de -12 décibels par octave. Cette pente spectrale est relevée de 6 décibels par octave par la radiation aux lèvres.

Dans cette théorie, le conduit vocal est considéré comme un filtre linéaire, dont les propriétés seraient indépendantes de celles de la source glottique. Plus généralement, les trois étapes dont résulterait le signal de parole pourraient être étudiées indépendamment. La Figure 19 présente schématiquement les spectres correspondant aux trois étapes considérées par la théorie source-filtre, ainsi que le spectre du signal de parole résultant.

[21] Le programme que nous avons développé sous environnement Matlab®, nommé Glottalizer, est distribué librement sous licence Creative Commons et téléchargeable avec sa documentation à l'URL http://www.bohm.hu/glottalizer

Considérant ce mécanisme de production comme assimilable à un circuit acoustique élémentaire, la théorie source-filtre de production de la parole peut être traduite sous forme d'une équation dans l'espace des nombres complexes. Ainsi le signal de parole S(z) peut être codé sous la forme :

$$S(z) = G(z) \cdot V(z) \cdot R(z),$$

où G(z) est le signal correspondant au flux glottique, V(z) est le filtre modélisant le conduit vocal, et R(z) est la radiation aux lèvres.

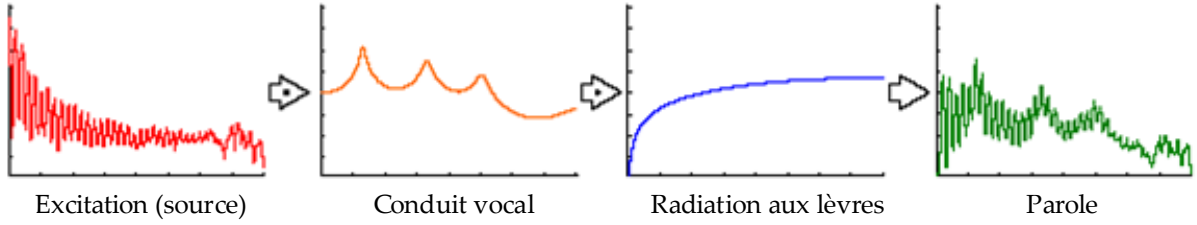


Figure 19 : Représentation schématique des spectres correspondant aux différentes composantes considérées par la théorie source-filtre (Fant, 1960), et spectre du signal de parole résultant de la modification de la source par le conduit vocal puis la radiation aux lèvres[22]. D'après Airas (2008).

La théorie source-filtre, qui permet de rendre compte d'un grand nombre de phénomènes mesurables dans la parole, s'appuie sur l'hypothèse forte d'une absence d'interactions entre les mouvements de la glotte et le conduit vocal. Or, il existe un couplage physique entre la glotte et le conduit vocal (voir par exemple Fant & Lin (1987)). Comme l'indique Henrich (2008), la première conséquence de ce couplage est une asymétrie de l'onde de débit glottique, qui est toutefois généralement prise en compte dans les modèles de débit glottique. Elle mentionne également un second aspect de l'interaction source-filtre qui est une perturbation du débit glottique provoquant un phénomène d'oscillation de l'onde de débit dérivée ou *ripple* dans le cas d'une proximité de la fréquence fondamentale et du premier formant, tout en soulignant qu'étant donné les plages de F0 utilisées dans la parole, cet aspect n'est pas problématique en voix parlée pour la plupart des voyelles. Enfin le troisième aspect de l'interaction source-filtre est l'influence de l'impédance de la glotte et du couplage avec le conduit subglottal sur la fonction de transfert du conduit vocal (Barney *et al.*, 2007). Bien que les limites de la théorie source-filtre liées à l'interaction entre source et filtre puissent être sensibles dans le cas de la voix chantée (Henrich, 2008), cette théorie est considérée comme une approximation satisfaisante des mécanismes de production de la parole, en raison

[22] Les abscisses sont ici représentées en Hertz sur une échelle linéaire, d'où la forme du spectre correspondant à la radiation aux lèvres. Ce spectre est le plus souvent représenté sur une échelle logarithmique sous forme d'une droite de pente 6 dB/octave.

notamment de l'impédance élevée de la glotte qui permet d'assimiler la source glottique à une source parfaite indépendante de la pression supraglottique (voir par exemple Vincent (2007)).

7.8.2 La problématique de l'inversion

Considérant, selon le principe de la théorie source-filtre, la production de parole comme résultant des effets successifs et linéairement indépendants de la source glottique, du conduit vocal et de la radiation aux lèvres, il est possible en définissant des contraintes liées à l'anatomie et à la physiologie de l'appareil phonatoire de calculer à partir du signal acoustique les paramètres mis en œuvre dans le mécanisme de production de parole.

Nous ne décrirons pas de façon détaillée les techniques permettant de retrouver la configuration articulatoire du conduit vocal dans le cadre de la production de certaines voyelles ou consonnes. Soulignons cependant que ces techniques ont fait la preuve de leur efficacité. Ainsi Mawass *et al.* (2000) ont validé par cinéradiographie ainsi que par confrontation aux données audio l'inversion de consonnes fricatives du français, ce qui leur a permis d'appliquer leurs résultats en synthèse.

7.8.3 Le filtrage inverse

La technique du filtrage inverse consiste en une estimation de l'onde de débit glottique à partir du signal acoustique, en s'appuyant sur les hypothèses de la théorie source-filtre de production de la parole. Ainsi, le filtrage inverse consiste en un passage du signal acoustique à travers un filtre dont la fonction de transfert est l'inverse de la fonction de transfert supraglottale. Le filtrage inverse est complété par l'annulation de la radiation des lèvres, réalisé par une simple intégration du signal en considérant cette radiation comme assimilable à l'effet d'un filtre dérivateur. Une estimation de l'onde de débit glottique ou de sa dérivée est ainsi obtenue, ce qui permet de calculer les paramètres de la source glottique.

En reprenant l'équation présentée ci-dessus, étant donné que ses trois composantes sont considérées comme linéairement indépendantes, ce problème peut se traduire (voir Airas (2008)) comme la résolution de l'équation :

$$Q(z) = G(z) \cdot R(z) = \frac{S(z)}{V(z)},$$

où Q(z) est la dérivée de l'onde de débit glottique considérée comme équivalente au produit de la source glottique et de la radiation aux lèvres. V(z) étant également inconnu, la résolution de cette équation est un problème de déconvolution aveugle.

Diverses méthodes, dont le principe général est fondé sur la mesure de la position des formants et de leur largeur de bande ont été proposées pour résoudre cette équation en séparant l'influence du conduit vocal de celle du débit glottique dérivé. Airas (2008) passe en revue la plupart de ces techniques. La première a été proposée par Miller (1959) antérieurement à la publication de la théorie source-filtre (Fant, 1960) et consiste en une mesure manuelle des positions et largeurs de bande des deux premiers formants. La plus couramment utilisée est celle de Wong *et al.* (1979) qui s'appuie sur la phase fermée du cycle glottique au cours de laquelle l'estimation du filtre V(z) modélisant les résonances du conduit vocal est plus aisée. Cependant, comme le remarque Airas (2008), cette méthode présente l'inconvénient de ne pas être adaptée aux cas extrêmes de voix soufflées. En effet, dans ces types de phonation les plis vocaux ne se referment pas entièrement, et il n'est donc pas possible de s'appuyer sur une phase fermée clairement marquée.

Une méthode alternative consiste en une mise en relation de la sortie du filtre inverse avec un modèle de source glottique, les paramètres de la source étant dérivés du modèle d'onde glottique résultant. C'est par exemple la méthode adoptée par Arroabarren & Carlosena (2003). Pour que cette méthode soit efficace, le modèle de source utilisé doit à la fois permettre de représenter de façon fidèle la véritable onde de débit glottique, et être suffisamment souple pour capturer les différents modes de phonation.

La méthode du filtrage inverse montre ses limites dans le cas des modes de phonation autres que la voix modale, et lorsque la fonction de transfert du conduit vocal subit des changements rapides. En effet, l'efficacité de cette méthode repose sur une estimation précise de cette fonction de transfert fondée sur l'analyse de la position des formants et de leur largeur de bande, qui selon Ní Chasaide & Gobl (2000), ne peut en pratique être réalisée de manière entièrement automatique que pour une phonation modale.

7.9. Le Quotient d'Amplitude Normalisé (NAQ)

7.9.1 Définition

Le Quotient d'Amplitude Normalisé (NAQ) proposé par Alku *et al.* (2002), peut être considéré comme une normalisation du « temps de déclinaison » défini par Fant (1997) et nommé quotient d'amplitude par Alku & Vilkman (1996). Ce paramètre, proposé comme un bon descripteur du continuum entre voix soufflée et voix tendue, est défini comme le rapport de l'amplitude crête à crête de l'onde de débit glottique (UP) et du pic négatif maximal de sa dérivée (EE), normalisé par la période fondamentale (T0). Son calcul s'exprime donc comme :

$$NAQ = \frac{UP}{EE \times T0}$$

La Figure 20 présente les paramètres extraits de l'onde de débit glottique et nécessaires au calcul de NAQ.

D'après les auteurs, le calcul de ce paramètre à partir de mesures d'amplitude de l'onde de débit glottique estimée permet d'obtenir des mesures plus fiables que celles que permettraient d'obtenir des mesures temporelles effectuées directement à partir de l'onde estimée.

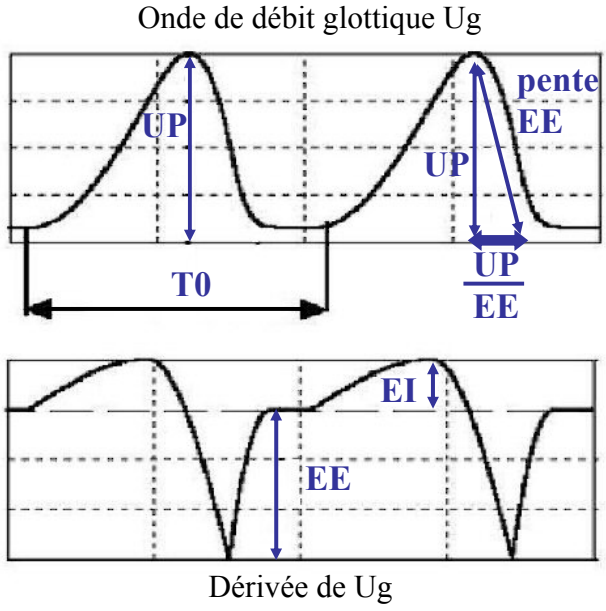


Figure 20 : Paramètres tirés de l'onde de débit et de sa dérivée nécessaires au calcul de NAQ. D'après (Eckert & Laver, 1994).

7.9.2 NAQ : la 4ème dimension prosodique (Campbell, 2003) ?

Campbell & Mokhtari (2003) ont calculé le paramètre NAQ sur un très large corpus comprenant d'importantes variations d'état émotionnel (une locutrice japonaise enregistrée dans des situations de vie courante pendant plusieurs mois), et ont constaté que ce paramètre n'était pas corrélé aux paramètres classiques de la prosodie que sont la fréquence fondamentale, la durée et l'intensité. A partir de ces observations, ils proposent de considérer le paramètre NAQ comme la quatrième dimension prosodique. Cette dimension correspondrait au « degré d'attention » porté à l'interlocuteur et véhiculé par la voix.

Le calcul du quotient d'amplitude AQ et de sa forme normalisée NAQ a été implémenté au sein d'un algorithme développé aux laboratoires ATR (Japon) par Parham Mokhtari, qui se propose de calculer les valeurs du quotient d'amplitude (AQ) et du quotient d'amplitude normalisé à partir d'un signal acoustique non étiqueté (Mokhtari & Campbell, 2002). Le calcul de valeurs de quotient d'amplitude n'ayant de sens que pour les portions de signal acoustique correspondant à la réalisation de vocoïdes, ce logiciel réalise un étiquetage automatique des parties du signal à traiter, à partir de la détection de centres de confiance (*reliability centres*). Cette détection est fondée sur le fait que les vocoïdes présentent d'un point de vue acoustique à la fois une stabilité spectrale et un niveau élevé d'énergie acoustique (Lea *et al.*, 1975). Les intervalles de confiance potentiels sont ainsi définis comme les portions de signal présentant de faibles variations spectrales, ainsi qu'une énergie supérieure à un certain seuil. Une estimation des formants, qui sert de base au filtrage inverse, est ensuite réalisée. Le cepstre régénéré à partir de cette estimation formantique est alors comparé à celui déduit de la transformée de Fourier du signal acoustique original, par mesure d'une distance cepstrale. Seuls les centres de confiance potentiels pour lesquels cette distance ceptrale est suffisamment faible sont conservés pour le calcul du quotient d'amplitude.

Le calcul proprement dit du quotient d'amplitude AQ et de sa forme normalisée par la période fondamentale NAQ se fait par une estimation par filtrage inverse du débit glottique et de sa dérivée sur une portion de signal constituée de 5 trames de 30 ms, réparties autour du centre de confiance. Le quotient d'amplitude est calculé par une mesure sur cette portion de signal des valeurs minimale et maximale des courbes estimées du débit et de sa dérivée, dont sont déduites les amplitudes des deux fonctions (Alku & Vilkman, 1996).

En collaboration avec Solange Rossato (Rossato, Audibert & Aubergé, 2004) et dans le cadre du projet Expressive Speech Processing du JST/CREST[23] dirigé par Nick Campbell, nous avons voulu évaluer dans quelle mesure cet algorithme pourrait, au-delà d'une utilisation sur un très large corpus afin d'identifier des tendances, permettre une analyse fine des valeurs de NAQ sur des énoncés individuels et plus spécifiquement un suivi de contours de NAQ.

7.9.3 Calcul de NAQ

Le calcul du quotient d'amplitude normalisé a été effectué à l'aide l'algorithme de Mokhtari (2002), que nous avons adapté afin d'extraire pour chaque centre de confiance trouvé :

- La position du centre de confiance
- La fréquence fondamentale estimée
- Le quotient d'amplitude (AQ) de la portion de signal longue de 80 ms centrée sur le centre de confiance détecté.
- Le quotient d'amplitude normalisé (NAQ) correspondant.

Nous avons ensuite apparié ces informations avec l'étiquette du phonème correspondant à chaque centre de confiance, défini par l'étiquetage manuel réalisé. Lorsque plusieurs centres de confiance étaient détectés au sein d'un seul et même phonème, et donc plusieurs valeurs extraites pour ce phonème, la moyenne de ces valeurs a été retenue pour chaque paramètre comme valeur unique pour les analyses ultérieures.

7.9.4 Traitement des signaux EGG

Nous avons ensuite traité au moyen de routines Matlab les signaux EGG. Nous avons extrait les paramètres suivants de ces signaux, pour chaque phonème retenu, suivant le mode de calcul décrit par Henrich (2001) :

- La fréquence fondamentale F0 mesurée en Hertz.
- La période fondamentale T0 mesurée en millisecondes.
- T1, le temps entre l'ouverture et la fermeture des plis vocaux en millisecondes, calculé par la méthode d'intercorrélation de Henrich (2001).
- Le quotient ouvert OQ, calculé comme le rapport T1/T0.

[23] Ministère japonais de la recherche

- AVF, la vitesse du débit à l'instant de fermeture des plis vocaux, définie comme l'amplitude positive maximum de la dérivée du signal EGG.
- AVO, la vitesse du débit à l'instant d'ouverture des plis vocaux, définie comme l'amplitude négative maximum de la dérivée du signal EGG.

Les signaux ont été traités pour chaque portion de signal correspondant à une voyelle, ainsi qu'au phonème [n], en fonction des bornes données par les fichiers d'étiquettes. En effet, la mesure de paramètres glottiques n'est pertinente que pour les portions voisées. D'autre part, les mesures EGG étant destinées à évaluer l'algorithme de calcul du quotient d'amplitude de Mokhtari (2002), ces mesures devaient être effectuées pour les portions de signal correspondant aux centres de confiance détectés par cet algorithme, parmi lesquels le [n] se trouvait largement représenté. Les paramètres glottiques ont donc été calculés pour ce phonème en complément des voyelles. Lors de l'extraction de ces paramètres, un filtrage passe-haut a été appliqué aux stimuli préalablement au découpage en phonèmes du signal, afin d'éliminer du signal EGG la composante Rx issue des mouvements du larynx (Rothenberg, 1992).

7.9.5 Evaluation de l'algorithme de calcul du paramètre NAQ

Bien que l'algorithme effectuant le calcul du paramètre NAQ (Mokhtari, 2002) ait déjà été évalué sur des ensembles très larges de données, en particulier par Campbell & Mokhtari (2003), des pré-tests réalisés sur quelques signaux de parole nous ont montré que les variations fortes de NAQ liées au contrôle de la qualité de voix risquaient d'être du même ordre de grandeur que celles dues aux variations d'ordre phonétique. En particulier se pose le problème des voyelles nasales qui intrinsèquement en phonation modale font varier NAQ dans des ordres de grandeur correspondant à des changements de qualité de voix pour les autres voyelles. Si de telles variations peuvent être absorbées par une large représentation de tous les contextes phonétiques dans les grandes bases de données du type de celle utilisées par Campbell & Mokhtari (2003), qui extraient de ces corpus des données statiques globales de NAQ, notre approche basée sur des données plus détaillées nécessite un suivi fin des contours de NAQ sur le signal de parole. Dans notre approche le paramètre NAQ doit être finement évalué, afin de connaître les effets intrinsèques et co-intrinsèques (effets non prosodiques) de la variation de NAQ, et d'identifier les éventuels problèmes liés à l'inversion du signal acoustique, et en particulier dans le cas des nasales qui seraient susceptibles de produire les mêmes effets acoustiques qu'une phonation soufflée. Vaissière (1995, p. 13) note ainsi dans le cadre de la production d'un [h] :

« *La voix* breathy *peut également donner l'impression d'être nasalisée. L'ouverture de la glotte pendant h aspiré provoque un affaiblissement considérable de l'énergie dans la zone du premier formant, une caractéristique commune avec l'ouverture du port nasal.* »

Afin de vérifier et calibrer cet algorithme, nous l'avons utilisé pour calculer les valeurs de NAQ sur un sous-ensemble du corpus spontané « Sound Teacher », qui présente des stimuli monosyllabiques équilibrés dans l'espace phonologique des voyelles. Les commandes (« page suivante » et les noms de couleurs monosyllabiques) produites dans la partie spontanée du corpus par les deux locuteurs masculins (M1 et M4) pour lesquels le signal électroglottographique a été enregistré ont été sélectionnées pour cette étude, soit un corpus de 373 stimuli d'une durée utile totale de 204 secondes. Les enregistrements du signal EGG présentent l'avantage de fournir des données fiables sur une partie du cycle glottique (sauf la phase ouverte de la glotte), permettant notamment de calculer aisément des valeurs de référence pour la fréquence fondamentale et le quotient ouvert.

Lors de l'étiquetage phonétique, un nombre considérable de stimuli parmi les productions traitées ont révélé la présence d'un schwa à la fin des énoncés supposés monosyllabiques (par exemple [ʒɔnə] au lieu de [ʒɔn]). Les schwas ont été inclus dans les analyses, au même titre que les autres voyelles.

7.9.5.1 Détection des centres de confiance

Lorsqu'il est calculé à partir de parole continue non étiquetée, NAQ ne peut être obtenu qu'au niveau des centres de confiance, c'est-à-dire des vocoïdes définis par Mokhtari (2002). La localisation de ces centres de confiance a donc également été extraite, et mise en correspondance avec l'étiquetage phonétique du corpus, afin de s'assurer de leur statut de vocoïdes. 68% des centres de confiance sont trouvés dans des voyelles, contre 15% dans des consonnes voisées qui satisfont aux critères d'énergie des vocoïdes et 17 % dans d'autres consonnes. La consonne nasale [n], fréquemment détectée comme centre de confiance, a également été prise en compte dans la suite des analyses. Il ressort de la Table 2, qui présente la répartition des centres de confiance en fonction des étiquettes phonétiques, que les distributions des phonèmes étudiés sont comparables, à l'exception du schwa non systématiquement réalisé.

i	ɛ	a	o	u	ə	ɑ̃	n	Autres
9.4%	11.6%	14.7%	7.3%	8.8%	3.0%	13.2%	8.3%	23.7%

Table 2 : Répartition par phonème des centres de confiance détectés par l'algorithme d'extraction du paramètre NAQ.

7.9.5.2 Influence du phonème sur NAQ

Nous avons effectué une analyse de variance à un facteur afin d'évaluer, sur ces données, l'influence du phonème sur les valeurs de Quotient d'Amplitude Normalisé. La Figure 21 présente les valeurs moyennes et l'intervalle de confiance de NAQ par phonème, pour l'ensemble des 373 stimuli analysés. Les valeurs de NAQ sont comprises entre 0.07 et 0.32 ce qui, en comparant aux valeurs obtenues par Alku *et al.* (2002) pour cinq locuteurs masculins, signifierait que les stimuli analysés se répartissent sur l'ensemble du continuum voix soufflée-voix pressée. Les valeurs moyennes de NAQ semblent plus élevées pour les voyelles orales hautes, bien que cette tendance ne soit pas significative.

Le phonème [ə] présente en outre un NAQ moyen plus élevé, mais présente une répartition clairement bimodale des valeurs de NAQ. Ainsi le locuteur N. ajoute [ə] sur les fins de mots avec des valeurs de NAQ élevées (0.28 en moyenne), correspondant à une voix soufflée d'après les valeurs données par Alku *et al.* (2002). Le locuteur R., quant à lui, ajoute [ə] avec une voix modale (valeurs de NAQ autour de 0.12 de même que pour [ɛ]). Le choix de l'ajout ou non d'un schwa final semble relever de stratégies relatives aux valeurs expressives des actes de langage. A noter que les deux locuteurs présentent des fréquences d'ajout de [ə] voisines: 36.8% des stimuli du locuteur N., contre 42.9% chez le locuteur R. Tandis que la voyelle nasale [ɑ̃] présente des valeurs de NAQ similaires à celles des voyelles hautes, la consonne nasale [n] a des valeurs de NAQ correspondant à une voix soufflée (0.19). Toutes les différences entre les valeurs moyennes de NAQ par phonème sont significatives à l'exception de la différence entre [n] et [ə] (test *post-hoc* de Tukey, $p<0.01$). Il paraît irréaliste que le phonème [n] de [ʒon] soit systématiquement réalisé avec une phonation soufflée alors que le [o] ne l'est pas. On pourrait arguer que cela est dû à sa position finale, mais ceci reste observable lorsque [ə] est ajouté en fin d'énoncé. Une explication possible serait que la nasalité produit essentiellement des basses fréquences, ce qui augmentent la pente spectrale (que l'on peut considérer comme la différence entre l'énergie présente dans les hautes fréquences du signal et celle dans les basses fréquences) en atténuant les hautes fréquences. En effet les mouvements supra-laryngés dans le cas de la nasalité, et laryngés pour le caractère soufflée produisent le même effet acoustique, à savoir une augmentation de la pente spectrale. On a donc dans le cas du [n] une mauvaise interprétation de cette augmentation de la pente spectrale : ici, un effet supra-laryngé est en effet attribué à une voix soufflée.

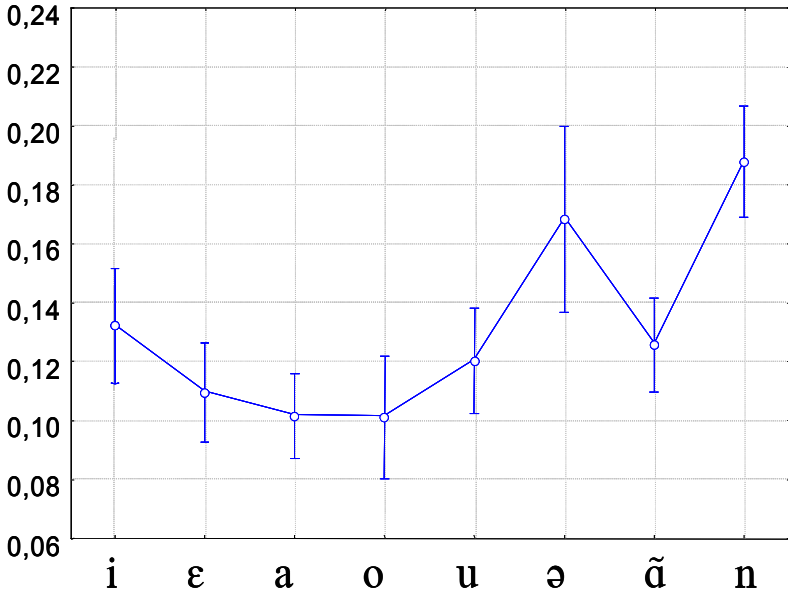


Figure 21 : Valeur moyenne et intervalle de confiance (99%) du quotient d'amplitude normalisé NAQ pour chaque voyelle et pour la consonne nasale [n], extraite des productions des locuteurs M1 et M4.

7.9.5.3 Estimations de F0

La plupart des paramètres basés sur l'amplitude étant normalisés par la fréquence fondamentale, cela implique que les erreurs dans son estimation sont répercutées sur l'estimation de tous les autres paramètres. Nous avons donc comparé $F0_A$, estimée par l'algorithme de calcul des paramètres basés sur l'amplitude, à $F0_{EGG}$ (Figure 22), extraite par autocorrélation du signal EGG selon la méthode proposée par Henrich (2001). Ces deux valeurs ont été calculées sur les mêmes portions de signal, centrées sur les centres de confiance détectés. La corrélation entre ces deux mesures est de $r = 0.8$.

Il ressort en outre de cette comparaison que les valeurs de fréquence fondamentale utilisées pour normaliser les paramètres basés sur l'amplitude tendent à être sous-estimées par rapport à la valeur de référence que constitue celle mesurée à l'aide de l'EGG, ce qui implique que les valeurs de ces paramètres devraient également être sous-estimées.

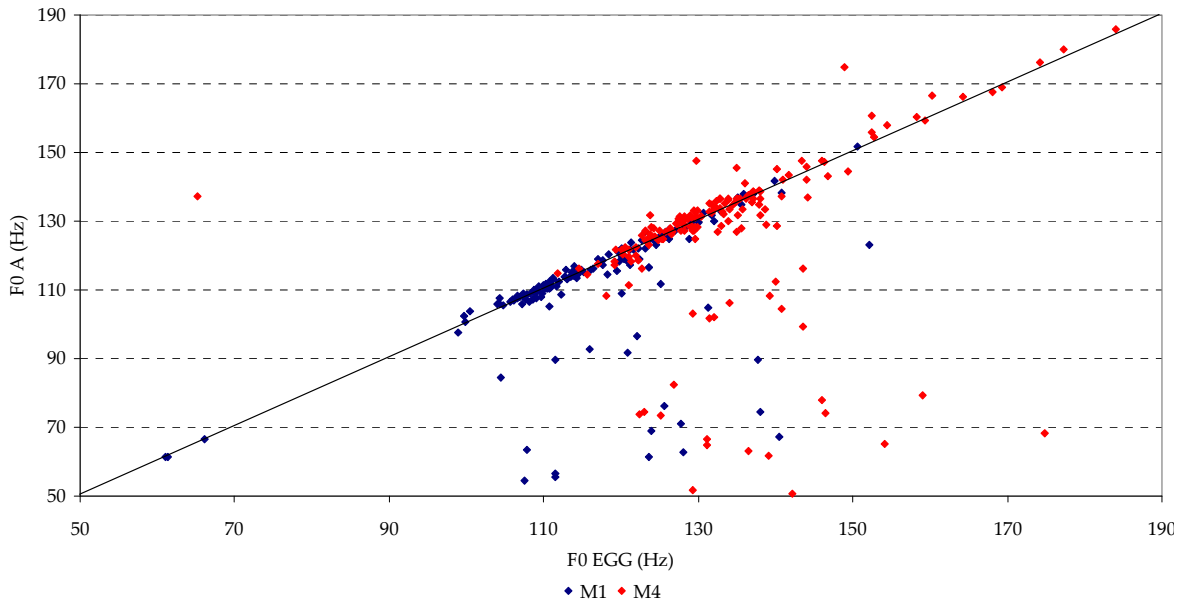


Figure 22 : Comparaison des valeurs de $F0_A$ estimées par l'algorithme de calcul des paramètres basés sur l'amplitude et de $F0_{EGG}$ extraites par autocorrélation du signal EGG, pour les locuteurs M1 (bleu) et M4 (rouge), corrélées à r=0.8. 59% des valeurs de $F0_{EGG}$ pour le locuteur M1 sont comprises entre 110 Hz et 140 Hz, avec une valeur de $F0_A$ divergente de moins de 5 Hz. Les points correspondants sont en grande partie masqués par ceux correspondant au locuteur M4 dans cette figure.

Dans notre corpus, pour les deux locuteurs masculins sélectionnés, la fréquence fondamentale présente des valeurs significativement supérieures pour [ə] réalisé par le locuteur R., accompagnées de valeurs élevées de NAQ.

7.9.5.4 Evaluation du filtrage inverse : OQ_A vs. OQ_{EGG}

Gobl & Ni Chasaíde (2003b) proposent d'étendre l'idée originale d'Alku & Vilkman (2000) d'une estimation des paramètres temporels de forme de l'onde de débit glottique à partir des amplitudes. En particulier, ils ont introduisent le paramètre OQ_A, estimation du quotient ouvert défini comme (cf. Figure 20) :

$$OQ_A = \frac{UP}{T0} \cdot \left(\frac{\pi}{2EI} + \frac{1}{EE}\right)$$

Autrement dit Gobl & Ni Chasaíde (2003b) proposent une estimation à partir des amplitudes de la durée de la phase ouverte de la glotte définie comme :

$$T1_A = UP \cdot \left(\frac{\pi}{2EI} + \frac{1}{EE}\right)$$

Le quotient ouvert représente la part de la durée la phase ouverte dans un cycle glottique, c'est-à-dire la somme de la durée de la phase d'ouverture et de celle de la

phase de fermeture, normalisée par la période fondamentale T0 (la durée d'un cycle glottique). Son estimation basée sur l'amplitude OQ_A devrait en conséquence être partiellement corrélée à NAQ, lié à la phase de fermeture (Alku & Vilkman, 2000). Dans notre corpus, cette corrélation est élevée (r=0.96) ce qui tendrait à montrer que la phase de fermeture explique l'essentiel de la variance du quotient ouvert, l'asymétrie entre les phases d'ouverture et de fermeture de la glotte étant de moindre importance.

La corrélation entre OQ_A et $F0_{EGG}$ que nous observons est faible (r=0.53). La fréquence fondamentale ne peut donc expliquer la variation de durée de la phase ouverte, qui semble clairement indépendante des autres paramètres prosodiques. Les valeurs de quotient ouvert extraites du signal EGG, OQ_{EGG}, ne présentent pas de corrélation avec F0. Ces résultats doivent être comparés à ceux obtenus par Henrich (2001) en voix chantée, qui a trouvé une corrélation entre F0 et OQ chez les chanteurs utilisant le mécanisme laryngé II, mais pas pour le mécanisme I qui est le plus fréquemment utilisé par les sujets masculins en voix parlée.

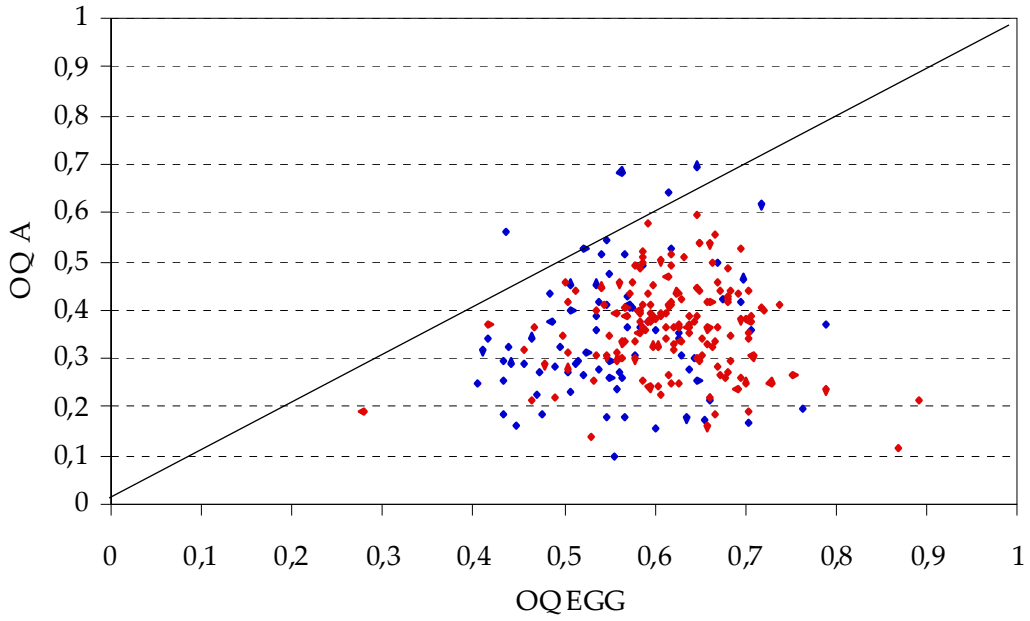


Figure 23 : Répartitions relatives des valeurs de OQ_A estimées à partir des valeurs données par l'algorithme d'extraction de paramètres glottiques à partir de l'amplitude, et de OQ_{EGG}. dérivées du signal EGG pour les locuteurs M1 (bleu) et M4 (rouge)

La Figure 23 montre la répartition des valeurs de OQ_A par rapport à celles de OQ_{EGG}. Il apparaît que ces valeurs, pourtant supposées mesurer la même quantité, sont totalement décorrélées (r=-0.24). Les valeurs de OQ_A sont moins élevées, ce qui s'explique en partie par la sous-estimation de $F0_A$. Toutefois on observe une répartition similaire entre $T1_A$ et $T1_{EGG}$, bien que F0 n'intervienne pas dans leur calcul. En effet, les valeurs de $T1_A$ sont toujours plus faibles et on n'observe pas plus de corrélation entre $T1_A$ et $T1_{EGG}$ qu'entre OQ_A et OQ_{EGG}, quand bien même on considère chaque phonème séparément.

Les quotients NAQ et OQ_A sont fortement corrélés dans notre corpus, ce qui n'est pas surprenant étant donné qu'ils décrivent des phénomènes liés l'un à l'autre. Il convient cependant de souligner que ces quotients les phénomènes décrits par ces quotients, respectivement la part de la phase de fermeture et de la durée relative de la phase ouverte de la glotte dans la durée totale du cycle glottique, sont distincts.

Les estimations de NAQ sont vraisemblablement plus fiables que celles du quotient ouvert estimé OQ_A. En effet le calcul de OQ_A requière l'estimation d'un paramètre de plus que celui de NAQ, à savoir EI (cf. Figure 20), ce qui introduit une source d'erreur supplémentaire.

A la lumière des résultats de Gobl & Ní Chasaide (2003b), qui obtiennent une corrélation de r=0.76 entre OQ_A et la valeur du quotient ouvert extraite du domaine temporel de l'onde de débit glottique estimée, il est toutefois surprenant que OQ_A et OQ_{EGG} soient si faiblement corrélés. Une explication pourrait venir d'une inadéquation du filtre inverse utilisé pour l'estimation de l'onde de débit glottique. En effet nous avons calculé automatiquement les paramètres basés sur l'amplitude, sans adaptation particulière au locuteur, tandis que les résultats de Gobl & Ní Chasaide (2003b) ont été obtenus par une mesure manuelle experte des formants. Etant donné qu'aucune méthode ne donne de mesure directe du débit glottique, la meilleure solution pour assurer un filtrage inverse adéquat semble être la supervision par un expert. Contrairement à Gobl & Ní Chasaide (2003b) qui ont mesuré l'amplitude EI directement à partir du signal plutôt qu'à partir de l'onde de débit estimée pour s'affranchir des contraintes internes du modèle LF (Fant *et al.*, 1985), nous avons de plus mesuré cette composante à partir de l'onde de débit estimée. Une autre source d'erreur potentielle pourrait provenir de la mesure de OQ_{EGG}. En effet nous avons utilisé un électroglottographe ne comprenant que deux électrodes et avons éliminé la composante basse fréquence correspondant aux mouvements du larynx par un filtrage passe-haut, ce qui ne permet pas une mesure aussi fine que celle obtenue avec un appareil de mesure plus élaboré dans lequel des électrodes additionnelles permettent de prendre directement en considération ces mouvements.

7.9.5.5 Conséquences

Nous ne pouvons donc pas lier directement les mesures articulatoires issues du signal EGG aux estimations basées sur l'amplitude du débit obtenu par filtrage inverse du signal acoustique. Il a cependant été montré que les caractéristiques de l'onde de débit glottique influencent le jugement perceptif émotionnel (voir par exemple Gobl & Ní Chasaide (2003a)), et que l'on peut lier NAQ au degré « d'attention » porté à la voix, comme l'a montré Campbell (2003). Ainsi NAQ apparaît clairement comme un paramètre extrait du signal acoustique qui est porteur d'informations sur la qualité de

voix. Le postulat d'Alku & Vilkman (2000) selon lequel NAQ capturerait un continuum voix tendue-voix soufflée semble cependant devoir être nuancé. Doval *et al.* (2006) considèrent en effet le paramètre NAQ comme un bon descripteur du continuum voix tendu-voix relâchée, mais ne mentionnent pas le cas de la voix soufflée.

Etant donné les configurations articulatoires correspondant aux modes de phonation relâché et soufflé, il est vraisemblable que ces modes de phonation soient fréquemment co-occurrents. Cependant les descriptions les plus complètes à ce jour de la qualité de voix (voir notamment d'Alessandro (2006)) considèrent la dimension voix tendue-voix relâchée et l'ajout d'un flux turbulent au niveau de la constriction glottique (voix soufflée) comme des dimensions au moins partiellement indépendantes. Il est ainsi possible d'observer une qualité de voix à la fois tendue et soufflée, comme par exemple dans le cas de l'attitude prosodique japonaise de surprise (Shochi, 2008).

Les résultats que nous obtenons sur nos données ne remettent pas en cause la validité d'une mesure globale de NAQ effectuée sur de larges corpus, et en particulier les conclusions de Campbell & Mokhtari (2003) sur le lien entre valeurs de NAQ et degré d'attention porté. En revanche, ils nous indiquent que nous ne pouvons directement, en calculant NAQ selon la méthode proposée par Alku (2002) et mise en œuvre par Mokhtari & Campbell (2002), effectuer de suivi de contours sur ce paramètre.

7.10. Vers un suivi des contours de qualité de voix ?

Il ressort de la revue de littérature et de l'évaluation d'un outil automatisé d'extraction des valeurs du paramètre NAQ présentées ci-dessus qu'aucun des outils dont nous disposons ne semble apte à mesurer de façon fiable l'évolution de la configuration de la source glottique au cours d'un énoncé, et en particulier de rendre compte du continuum voix tendue-voix relâchée, probablement le plus pertinent d'un point de vue perceptif. La problématique de la mesure sur les voyelles hautes de paramètres liés au comportement de la source glottique et plus particulièrement au degré tension musculaire semble se poser de façon récurrente, en raison de la proximité de la fondamentale et du premier formant.

Bien que nous ne soyons pas en mesure de calculer les variations fines du paramètre NAQ, ce dernier apparaît néanmoins comme le meilleur candidat pour capturer le continuum voix tendue-voix relâchée, ce qui est également l'une des conclusions de Doval *et al.* (2006). En dépit des limites inhérentes à la technique du filtrage inverse, une extraction supervisée des paramètres de forme de l'onde de débit, semblable à celle appliquée par Gobl & Ní Chasaide (2003b), pourrait permettre un suivi fin de ce

paramètre. Cependant une telle méthode est extrêmement coûteuse en temps, et il ne semble donc pas réaliste de l'appliquer à large échelle.

8. Analyses acoustiques et perceptives du corpus recueilli

Dans cette section, nous proposons une analyse perceptive et acoustique d'énoncés monosyllabiques sélectionnés à partir des productions de l'un des locuteurs masculins. Nous présentons tout d'abord les critères nous ayant mené à la sélection des productions de ce locuteur. Ensuite, après avoir introduit le débat entre gradience et contours comme support de la prosodie affective, nous étudions les contours de fréquence fondamentale de certaines des productions actées et spontanées de ce locuteur et montrons en particulier qu'une forme de contour sans valeur affective peut être isolée.

Nous présentons ensuite les principaux outils d'analyse statistique auxquels nous avons recours pour l'analyse des résultats des évaluations perceptives que nous présentons dans cette section et les suivantes. Nous détaillons plus particulièrement les tests d'hypothèses permettant d'évaluer la significativité de taux de reconnaissance à partir d'évaluations catégorielles, et l'application de la classification hiérarchique à la visualisation des principales confusions entre catégories.

Enfin, nous présentons les résultats d'une évaluation perceptive d'énoncés actés de diverses longueurs produits par le locuteur sélectionné, en conditions audio et audiovisuelle. Cette évaluation montre notamment que les énoncés monosyllabiques sont aussi bien reconnus que ceux plus longs, quelle que soit la modalité de présentation, et fait apparaître un effet inter-juge pour la perception du dégoût en modalité audio.

8.1. Sélection d'un sous-corpus audiovisuel pertinent

Nous avons choisi dans un premier temps de nous concentrer sur les productions d'un locuteur masculin, le locuteur M2. Le choix de ce locuteur a été guidé par divers critères, le premier d'entre eux étant que ce locuteur montrait une expressivité apparente supérieure à celle des autres locuteurs, notamment dans la partie spontanée de son corpus.

De plus, contrairement à celles de nombreux locuteurs parmi ceux que nous avons enregistrés, les auto-annotations fournies par le locuteur M2 comprennent peu de mélanges d'affects. En outre, ces auto-annotations incluent des passages étiquetés « rien », candidats potentiels pour fournir des mesures de référence pour la comparaison avec des énoncés porteurs d'expressions d'affects. Nous avons interprété les autres annotations indiquées par ce locuteur comme : confiance, concentration positive, joie/surprise, joie, concentration négative, déception/surprise, anxiété, anxiété/peur et lassitude.

Enfin, et bien que tels mouvements puissent être observés par moments pour ce locuteur d'après l'enregistrement vidéo, il effectue peu de mouvements du buste à même de faire varier la distance entre sa bouche et le microphone et donc de biaiser de façon importante les mesures d'intensité.

8.2. Des contours de F0 porteurs d'information

8.2.1 Gradience vs. contours

La question de la description de la prosodie émotionnelle est en particulier des variations de fréquence fondamentale par des valeurs globales à l'énoncé (gradience) ou par la dynamique de ses variations (contours) est à l'intersection de la linguistique et de la psychologie expérimentale. En effet, les modèles de description de la prosodie linguistique s'appuient sur des représentations en contours de l'intonation. Des auteurs tels que Fonagy (1983) ont ainsi étendu cette notion à la fonction expressive de la prosodie en proposant la notion de « clichés mélodiques », qui seraient liés de façon univoque à des émotions. Cette approche a cependant été critiquée lorsque la psychologie s'est emparée du champ des expressions vocales des émotions. Ainsi, Pakosz (1983) évalue l'influence des contours sur les jugements émotionnels et conclut que ceux-ci n'indiquent que l'excitation. Frick (1985) revient sur l'approche phonostylistique de la prosodie émotionnelle et notamment sur les propositions de Fonagy & Magdics (1963), et souligne que le contenu verbal des énoncés sélectionnés est susceptible d'influer sur les jugements émotionnels. Ladd *et al.* (1985) testent

systématiquement des indices de gradience et la forme générale du contour et concluent à une influence majoritaire de l'étendue de la fréquence fondamentale sur les jugements d'intensité de l'émotion, au-delà de la forme générale du contour. Bänziger & Scherer (2005) proposent une méthode de stylisation des contours de fréquence fondamentale et ont recours à la méthode de l'analyse/synthèse pour évaluer l'impact de ces contours sur les attributions émotionnelles. Ils concluent que la forme des contours a peu d'influence, mais que le niveau global et l'étendue de la fréquence fondamentale constituent des indicateurs déterminants de l'activation de l'expression émotionnelle.

8.2.2 Sélection des stimuli

Parmi les 136 stimuli produits par le locuteur M2 (86 stimuli actés et 50 stimuli authentiques pour lesquels nous avons réinterprété les étiquettes issues de l'auto-annotation du corpus spontané), nous avons sélectionné par écoute experte un exemplaire nous semblant représentatif de chacune des étiquettes émotionnelles. Calculer un contour-moyen (au sens d'Aubergé, 1992) aurait ici peu de validité, tout d'abord car nous disposons d'un nombre réduit d'exemplaire par chaque émotion, mais surtout car nous n'avons pas de critère perceptif ou morphologique qui permette d'attribuer une intensité à l'émotion de chaque stimulus. Pour cette raison, nous avons opté pour une approche typiquement phonologique, dans laquelle nous avons sélectionné des exemplaires que nous avons jugés comme prototypiques.

- Les expressions de satisfaction, surprise, attente, inquiétude, anxiété, déception, joie, tristesse, colère chaude, dégoût et peur, ainsi qu'une expression neutre ont été sélectionnées pour les expressions actées. Notons que l'expression de surprise, qui peut avoir un caractère ambivalent, nous a paru correspondre plus spécifiquement à une surprise positive. L'expression actée de colère est produite par ce locuteur comme de la colère chaude. L'étiquette « attente », indiquée par le locuteur à la suite de l'expérience d'induction, correspondrait aux passages étiquetés comme « lassitude » par le locuteur dans son auto-annotation.
- Les expressions de confiance, concentration positive, joie/surprise, joie, concentration négative, déception/surprise, anxiété, anxiété/peur, lassitude et rien ont été sélectionnées pour les expressions spontanées.

8.2.3 Caractéristiques des contours de fréquence fondamentale

Chaque stimulus est segmenté en phonèmes, syllabes et mot à l'aide du logiciel Praat (Boersma & Weenink, 2008). Les contours de F0 ont été extraits sous environnement Matlab® pour chaque segment vocalique délimité par les frontières phonémiques

définies lors de l'étiquetage. L'algorithme de calcul de F0 utilisé (Feng, 1983) est fondé sur une détection du nombre de passage par un seuil (fixé par défaut à 10 % de l'amplitude du signal) du signal de parole filtré. Les contours de F0 lissés, moyennés sur des fenêtres de 32 ms avec un décalage de 10 ms, ont été extraits à partir des valeurs calculées par cet algorithme. Ensuite, les contours ont été normalisés sur 10 points pour chaque voyelle afin de permettre des comparaisons de formes de contours indépendamment de la durée de la voyelle.

Les durées des voyelles, extraites de la segmentation en phonèmes, ont été converties en durées relatives en calculant leur différence par rapport à la durée moyenne de la même voyelle dans notre corpus, pour le locuteur sélectionné, afin de pouvoir effectuer des comparaisons inter-voyelles selon une normalisation similaire à celle réalisée par Pasdeloup (1992) sur les syllabes. Les valeurs de F0 mesurées à 10 et 90% de la durée de chaque voyelle ont également été extraites et sont utilisées pour calculer la différence entre niveau de F0 en attaque et en finale de la voyelle, que nous désignerons ici par le terme de déclinaison de F0. La dynamique de F0 est quant à elle définie comme la différence entre les valeurs minimum et maximum de F0 mesurée sur chaque énoncé. Toutes les valeurs de F0 sont exprimées en demi-tons par rapport à la fréquence fondamentale moyenne du locuteur dans l'ensemble du corpus (96.8 Hz). La moyenne et l'écart type des valeurs d'attaque, de finale et de durée ont été calculées pour chaque étiquette émotionnelle. Ces valeurs sont récapitulées dans la Table 3.

Affect	Condition	Valence	Attaque F0	Déclinaison F0	Dynamique F0	Durée normalisée (%)
anxiété	Acté	N	10	-1	1	-15.9
déception	Acté	N	1	1	1.5	85.6
dégoût	Acté	N	3	0	1	142
peur	Acté	N	-4	6	6	14.5
colère	Acté	P	15	3	3	29.2
joie	Acté	P	11	0	1.5	16.2
attente	Acté	P	10	-2	3	18.6
surprise	Acté	P	-2	8	8	30.2
résignation	Acté	N	8	1	1	-2.9
tristesse	Acté	N	10	-3	3	0.4
satisfaction	Acté	P	21	-3	7	77.7
inquiétude	Acté	N	0	11	11	17.9
neutre	Acté	-	0	0	0.5	1.2
anxiété/peur	Spontané	N	2	7	7	-6.6
confiance	Spontané	P	3	-5	6	23.4
joie/surprise	Spontané	P	-1	5	5	-12.6
lassitude	Spontané	N	-3	2	2	-14.3
concentration nég.	Spontané	N	2	3	3	-20.6
rien	Spontané	-	0	-2	2	-14.1
concentration pos.	Spontané	P	1	-4	6	-1.3
joie	Spontané	P	1	5.5	5.5	-5.5
déception/surprise	Spontané	P	-1.5	7.5	7.5	-26.6
anxiété	Spontané	N	1	7.5	7.5	-7.5

Table 3 : Valeurs caractéristiques des contours des stimuli sélectionnés parmi les productions du locuteur M2 : attaque, déclinaison, dynamique, durée normalisée. Toutes les valeurs de F0 sont données en demi-tons. N indique une valence négative, P une valence positive.

8.2.4 Contours de fréquence fondamentale

La première tendance qui peut être remarquée est que les contours étiquetés « neutre » (pour les émotions actées) et « rien » (pour les émotions authentiques) semblent correspondre à une intonation minimale (c'est-à-dire à une réalisation réduite des fonctions de segmentation/hiérarchisation et de focalisation). En effet, l'attaque de ces deux expressions est au même niveau que la fréquence fondamentale moyenne du locuteur, définie par Aubergé (2002) comme le niveau de référence de F0, l'attaque étant considérée comme le point d'ancrage des contours. La forme du contour semble donc réduite à la ligne de base de déclinaison due au mécanisme d'éjection de l'air contenu dans les poumons indépendamment du contrôle moteur du larynx responsable des variations locales de F0, selon l'hypothèse de la poussée trachéale formulée par Maeda (1976, p. 263-270). Une telle forme de contour indiquerait ainsi un effort articulatoire minimal sur de telles monosyllabes, qui correspondrait à la modalité déclarative.

Une analyse des valeurs moyennes sur l'ensemble des stimuli acté et spontanés permet de relever les faits suivants. La dynamique moyenne des contours actés s'avère moins élevée (3.7 demi-tons) que celle des contours authentiques (5.2 demi-tons). Le niveau moyen de l'attaque des contours actés est plus élevé (6.4 demi-tons) que celui des contours authentiques dont la moyenne est proche de 0 (avec toutefois d'importantes variations). En outre, la durée moyenne des voyelles est notablement plus élevée pour les expressions actées qu'authentiques (32 % vs. -8.6 % par rapport à la durée moyenne des voyelles sur l'ensemble du corpus).

Les contours actés de F0 normalisés sur dix points sont présentés en Figure 24, Figure 25 et en Figure 26. Parmi les contours présentés en Figure 24 et en Figure 25, seul ceux de la fréquence fondamentale et de la satisfaction se détachent clairement, tandis que les contours de la joie, de l'anxiété, de la tristesse et de l'attente sont proches du neutre. Le dégoût, la résignation et la déception ne présentent pas d'excursions particulières, mais ne suivent pas la ligne de déclinaison de base de l'expression neutre.

Peur, surprise et inquiétude actées présentent une forme similaire avec une proéminence finale (Figure 26).

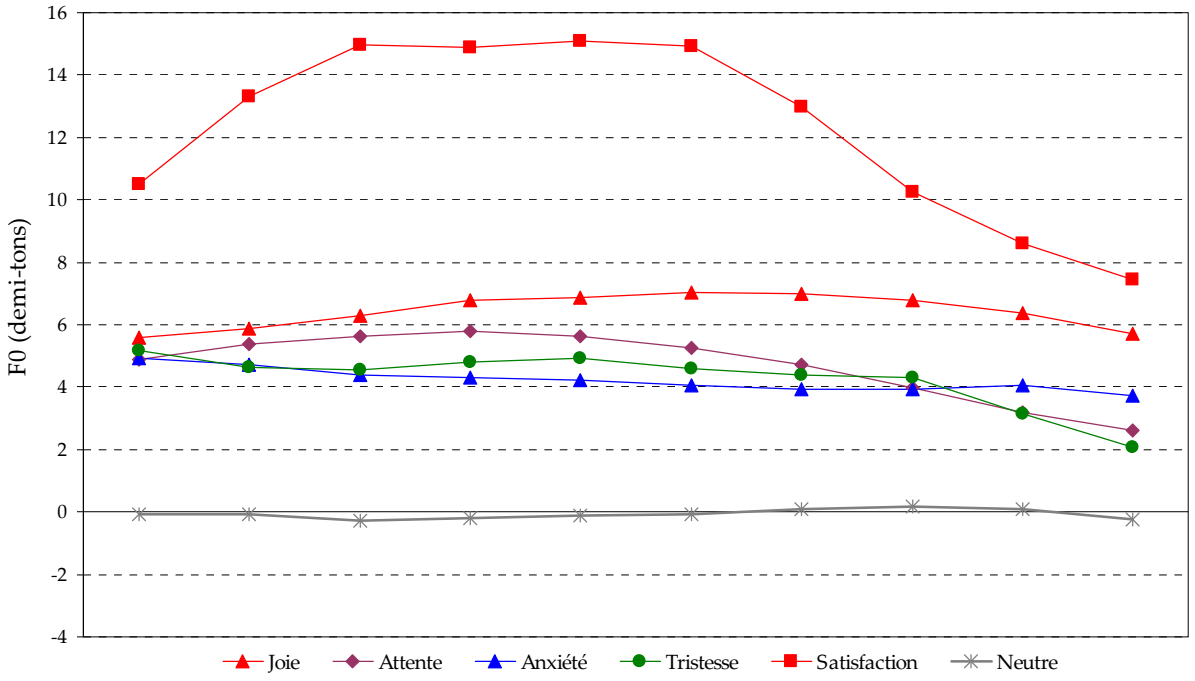


Figure 24 : Contours de fréquence fondamentale des expressions actées de joie, attente, anxiété, tristesse, satisfaction, et de l'expression neutre sélectionnées parmi les productions du locuteur M2.

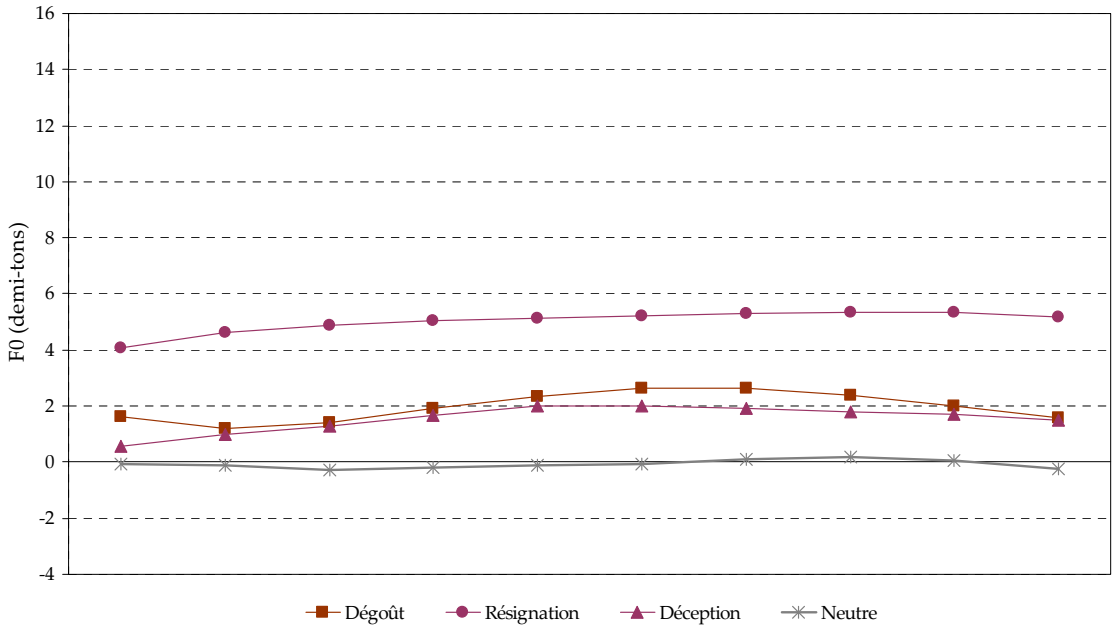


Figure 25 : Contours de fréquence fondamentale des expressions actées de dégoût, résignation, déception, et de l'expression neutre sélectionnées parmi les productions du locuteur M2.

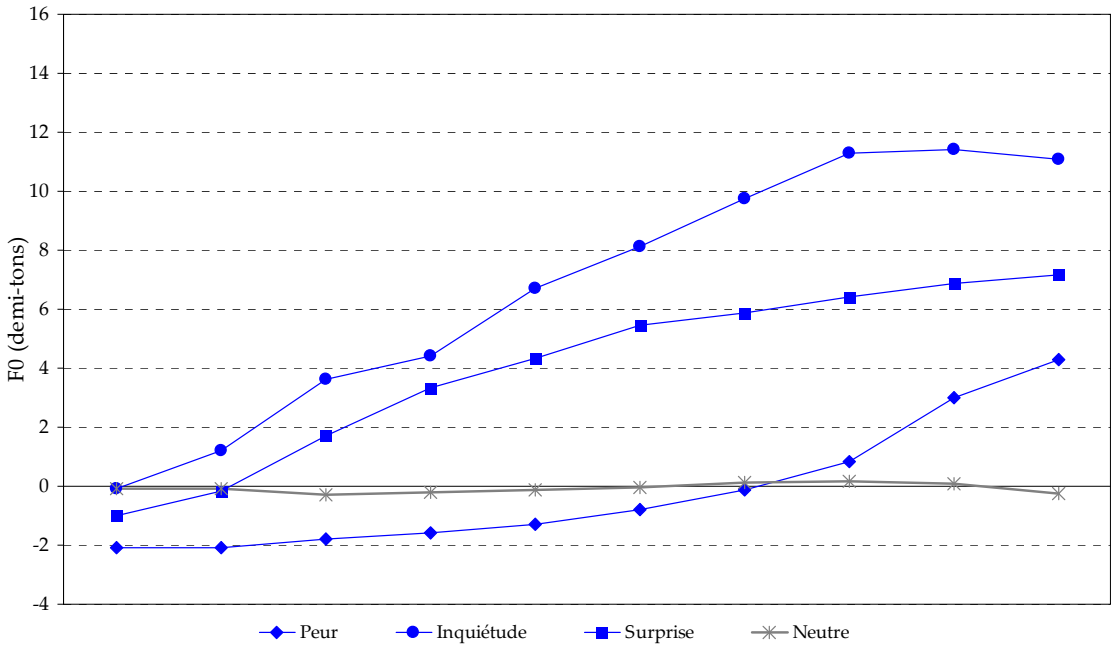


Figure 26 : Contours de fréquence fondamentale des expressions actées de peur, inquiétude, surprise, et de l'expression neutre sélectionnées parmi les productions du locuteur M2.

Les contours des expressions spontanées sélectionnées sont présentés en Figure 27 et Figure 28. La plupart des stimuli testés ici (Figure 27) sont caractérisés par un contour de fréquence fondamentale comprenant une proéminence finale.

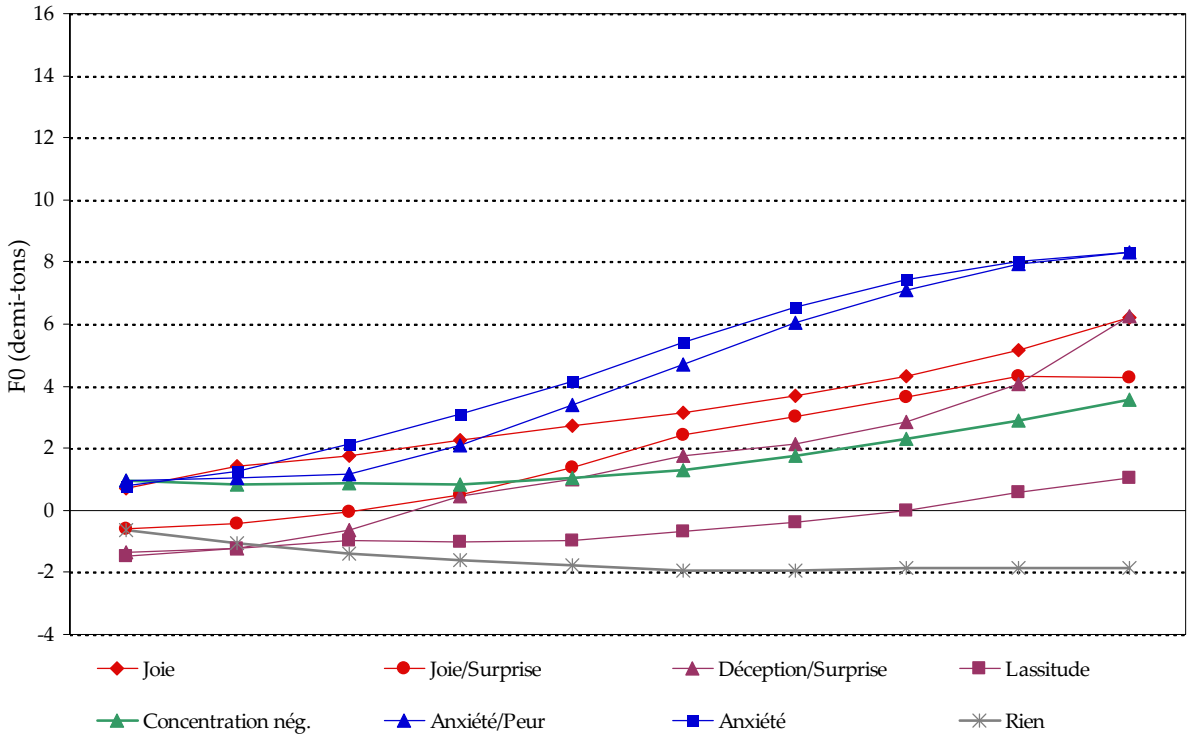


Figure 27 : Contours de fréquence fondamentale des expressions spontanées de déception/surprise, joie, joie/surprise, anxiété/peur, anxiété, concentration négative et lassitude, et de l'expression « rien » sélectionnées parmi les productions du locuteur M2.

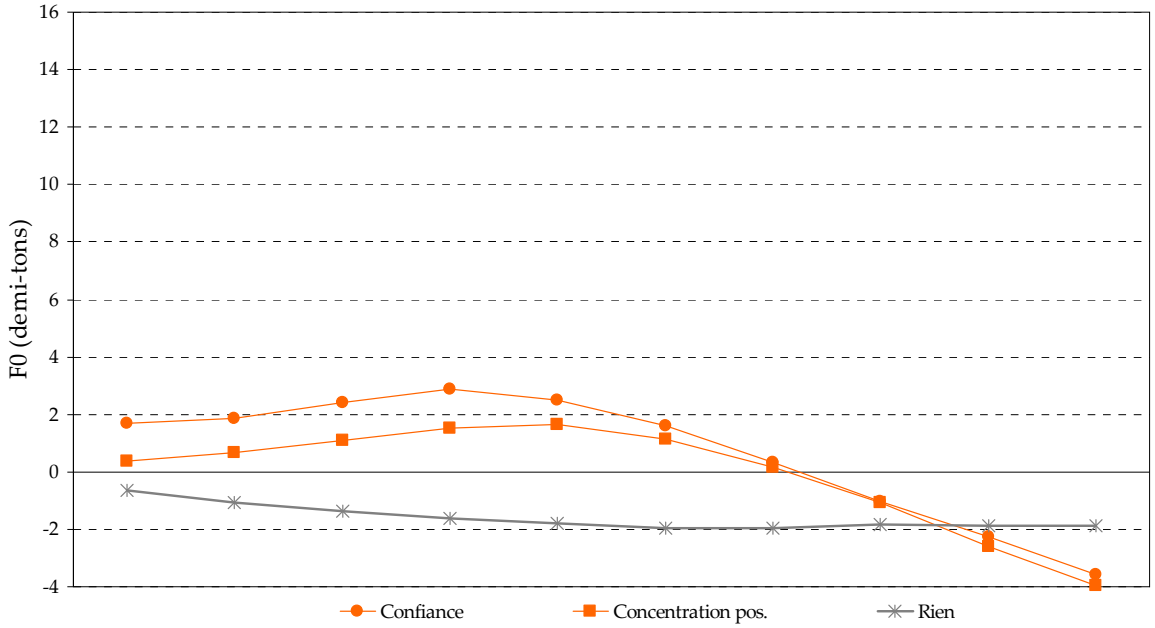


Figure 28 : Contours de fréquence fondamentale des expressions spontanées de confiance et concentration positive, et de l'expression « rien » sélectionnées parmi les productions du locuteur M2..

8.2.5 Discussion

Bänziger & Scherer (2005) ont obtenu des résultats perceptifs montrant que seul le niveau général de F0 est pertinent pour la perception des expressions émotionnelles, et confirment que la hauteur moyenne peut être liée aux valeurs d'activation et d'arousal. Quoique ces caractéristiques soient vraisemblablement les plus pertinentes, nous suggérons que des indices extra-linguistiques plus fins, véhiculés par la forme des contours, ont pu être masqués dans l'expérience qu'ils ont mené par les contours prosodiques porteurs d'informations linguistiques, par exemple de démarcation. Les énoncés retenus dans cette expérience étant identiques d'une émotion à l'autre, les contours prosodiques correspondant à la réalisation des fonctions linguistiques sur ces énoncés sont supposés identiques. Cependant, en l'absence d'une modélisation de la prosodie linguistique suffisamment fine pour permettre de mesurer par soustraction les variations prosodiques dues aux expressions émotionnelles, il n'est pas possible de s'assurer que la variabilité de réalisation de la prosodie linguistique permet de mesurer d'éventuelles variations plus fines de la prosodie émotionnelle.

Il serait donc intéressant de répliquer une étude similaire sur des unités linguistiques minimales, en contraignant également les variations pragmatiques. Les contours de F0 stylisés et paramétrisés par un ensemble de valeurs d'ancrage qui sont présentés ici permettent d'observer différentes classes de contours lorsqu'on les compare aux contours de référence (pour la parole actée et authentique). Bien que le lien ne soit pas

établi ici entre l'étiquetage des expressions et la morphologie des contours, ces contours sont suffisamment simples pour tester différents types de stylisation dans le cadre d'expériences perceptives fondées sur un processus de resynthèse de la prosodie. Dans cette optique il serait nécessaire d'évaluer en perception quels indices de ces contours (par exemple l'alignement du pic de F0) sont sensibles aux différentes étiquettes. Préalablement à cette validation perceptive, des premières conclusions peuvent être tirées de cette étude. Tout d'abord il existe bien un contour « sans valeur émotionnelle », aussi bien pour les expressions actées qu'authentiques. De plus la forme de ce contour correspond de façon frappante à celle attendue sur de telles unités linguistiques minimales.

8.3. Méthodologie d'analyse statistique

8.3.1 Reconnaissance et confusions

Le choix méthodologique de nous appuyer principalement sur des tests perceptifs d'expressions affectives par des sujets « naïfs », tests conduits sur un ensemble restreint de stimuli, nous amène à détailler ici les méthodes statistiques qui nous permettrons d'analyser ces données perceptives. En particulier, nous serons principalement amenés à analyser des tâches de catégorisation perceptive d'expressions émotionnelles réalisées sur un ensemble fermé d'étiquettes. Il nous faut donc choisir une méthode d'analyse statistique appropriée à l'analyse des matrices de confusion résultant de ces tâches de catégorisation perceptive.

Les méthodes d'analyse statistique classiques et notamment les tests paramétriques, qui évaluent la significativité d'une différence en deux moyennes, ne permettent pas l'analyse d'une distribution multidimensionnelle. Une analyse de la variance induite par les différents facteurs d'un test de perception est cependant possible à condition de projeter les données pour ne travailler que sur les diagonales des matrices, c'est-à-dire sur le taux de catégorisation correcte.

Bien que le taux de catégorisation correcte ait été très largement utilisé pour rendre compte des résultats de tâches de catégorisation, notamment dans le cadre de l'étude des expressions faciales des émotions, la validité de cette mesure est controversée. Johnstone & Scherer (2000) notent ainsi que, lorsque le nombre d'alternatives de réponses est restreint, une telle tâche de catégorisation mesure plutôt la discrimination entre émotions que leur reconnaissance à proprement parler. Wagner (1993) souligne quant à lui que dans le cas de juges catégorisant systématiquement l'ensemble des stimuli présentés comme des expressions de colère, le taux de catégorisation correcte de la colère sera de

100%, sans qu'il soit possible d'affirmer que ces sujets discriminent les expressions de colère des autres stimuli présentés. Afin de corriger de tels biais systématiques dans les réponses fournies par les juges, Wagner (1993) propose de mesurer la reconnaissance des catégories dans le cadre d'un choix fermé au moyen d'un taux de catégorisation correcte sans biais *Hu* défini comme Hu = H x (1-F), où H est le taux brut de catégorisation correcte, et F le taux de fausses alarmes défini comme la moyenne des reports vers l'étiquette considérée. Si cette méthode permet d'étudier les biais systématiques de réponses à la présentation de stimuli relevant de catégories très distinctes (voir par exemple Elfenbein *et al.* (2002) pour une application à l'étude comparative des biais de réponse à la présentation d'expressions faciales d'émotions dans les cultures américaine, indienne et japonaise), Kätsyri (2006) souligne qu'une telle correction ne permet pas de faire la distinction entre biais de réponses systématiques et confusions dues à la nature des stimuli présentés. En effet, en dehors de la catégorisation de stimuli supposés prototypiques, ce qui est généralement le cas dans les études d'expressions faciales d'émotions dites « de base » produites par des acteurs, la présence de stimuli moins prototypiques, comme peuvent l'être des expressions vocales recueillies en condition écologique, est susceptible de faire baisser artificiellement les valeurs de Hu. Pour reprendre l'exemple donné par Kätsyri (2006), une telle correction appliquée à une tâche de catégorisation de couleurs aurait pour effet de rendre la couleur rouge moins rouge du fait de la présence de confusions avec la couleur orange.

D'autres mesures destinées à corriger le taux de reconnaissance ont également été proposées. Ainsi, Ekman (1994) suggère de corriger le niveau théorique du hasard pour prendre en compte le déséquilibre entre émotions positives et négatives. En effet, les expressions d'émotions positives sont, dans la grande majorité des études, sous-représentées parmi les choix de réponse proposées, et souvent restreintes à la seule expression de joie. Bien que cette correction ait été proposée dans le cadre de l'étude des expressions faciales, le déséquilibre entre émotions positives et négatives proposées comme alternatives de réponse est également patent dans la plupart des études portant sur les expressions vocales, et la nôtre ne déroge pas à cette règle. Cependant Scherer *et al.* (2003) reviennent sur la suggestion d'Ekman (1994) et soulignent que, si elle fait sens dans le cas des expressions faciales pour lesquelles l'action du zygomatique majeur est généralement un indicateur clair et non-ambigu d'expressions positives, une telle correction peut être moins adéquate dans le cas des expressions vocales pour lesquelles les propriétés acoustiques de différentes émotions ont tendance à se chevaucher.

En l'absence de règles universelles permettant de corriger les taux de reconnaissance, il semble donc préférable de ne travailler sur les taux de catégorisation correcte corrigés que lorsque les catégories présentées sont très clairement distinctes. La prise en compte

explicite des confusions, suggérée par Johnstone & Scherer (2000), apparaît dès lors comme une règle raisonnable.

8.3.2 Tests d'hypothèses pour les comparaisons de moyennes

8.3.2.1 Prise en compte de données dichotomiques dans une ANOVA à mesures répétées

Les méthodes d'analyse de la variance (ANOVA) permettent de tester la fiabilité d'une hypothèse à partir d'une mesure continue, affectée par un ensemble déterminé de facteurs. Le principe d'une telle analyse est d'établir la probabilité de rejeter à tort l'hypothèse nulle (H0), hypothèse selon laquelle les variations des valeurs moyennes de la mesure considérée (la variable dépendante) sont négligeables (autrement dit peuvent être dues au hasard d'échantillonnage) pour chacune des modalités des facteurs considérés. Ceci est réalisé en calculant les valeurs prises par la fonction F de Fischer en fonction du nombre de degrés de liberté découlant de la structure des données analysées. Autrement dit, une ANOVA permet de tester la probabilité que les variations de la valeur moyenne de la variable observée entre différents groupes soient le fruit du hasard (ce qu'exprime H0). Lorsque cette probabilité de rejeter H0 sans erreur est inférieure à un certain seuil (classiquement fixé à 5%, 1% ou 0,1% selon le degré de significativité considéré) l'effet du facteur qui définit ces groupes est dit significatif. Pour une explication détaillée des méthodes statistiques appliquées aux sciences de la parole, on pourra se reporter par exemple à Rietveld & van Hout (2005).

Un cas particulier de l'analyse de variance est celui du modèle à mesures répétées. En effet, l'une des conditions d'application de l'analyse de variance est l'indépendance temporelle des mesures de la variable dépendante. Ceci n'est pas vérifié dans le cas (fréquent en sciences de la parole, comme le soulignent Max & Onghena (1999)) de sujets auxquels il est demandé d'émettre une série de jugements successifs sur un même ensemble de stimuli. Il est alors possible d'appliquer une analyse de variance à mesures répétées, qui permet de prendre en compte spécifiquement cette dépendance temporelle entre les mesures. Avec ce type de modèle à mesures répétées, il est important de vérifier si l'hypothèse de sphéricité (pour plus de détails, voir par exemple Rietveld & van Hout, 2005), qui signifie que les données ne sont pas corrélées, est vérifiée. Si cette hypothèse de sphéricité est violée, il est alors nécessaire d'appliquer une correction (les plus courantes sont celles de Huynh-Feldt et de Greenhouse-Geisser) afin de diminuer le nombre de degrés de libertés et ainsi compenser la déviation à la sphéricité estimée.

La condition de sphéricité est classiquement vérifiée à l'aide du test de Mauchly. Cependant Huynh & Mandeville (1979) insistent sur le manque de robustesse de ce test,

notamment dans le cas d'une distribution non normale. De plus, plusieurs auteurs, parmi lesquels Max & Onghena (1999), mettent en avant le problème conceptuel lié à l'usage de tests d'hypothèse destinés à évaluer les conditions d'application d'autres tests d'hypothèse. En effet, l'application stricte de telles règles impliquerait de devoir tester également par un test d'hypothèse ce second test, et ainsi de suite. Le terme correctif de Huynh-Feldt (Huynh & Feldt, 1976) estime la déviation à la sphéricité par un coefficient ε d'autant plus proche de 1 que la distribution des données satisfait la condition de sphéricité, par lequel sont multipliés le nombre de degrés de liberté de F et le terme d'erreur. En conséquence, Max & Onghena (1999) recommandent dans une ANOVA à mesures répétées d'appliquer systématiquement cette correction plutôt que de tester la condition de sphéricité par un test d'hypothèse.

Les modèles d'ANOVA supposent d'une part que les données analysées suivent une distribution normale, et, d'autre part, que les variances dans les différents groupes soient homogènes. La première condition peut être évaluée par un test de normalité comme celui de Kolmogorov-Smirnoff, tandis que la seconde peut être établie à l'aide du test d'égalité des variances de Levene. Glass *et al.* (1972) ont réalisé un état de l'art d'études empiriques fondées sur des distributions simulées et établissent une taxonomie des conséquences sur les différents types d'erreur du recours à une analyse de variance à effets fixés. En caractérisant la déviation à la distribution normale par les valeurs des coefficients de dissymétrie et d'aplatissement (respectivement de 0 et de 3 dans le cas d'une distribution normale), ils concluent à un effet très marginal de la dissymétrie. En revanche, les auteurs relèvent que les effets du coefficient d'aplatissement et de la violation de l'hypothèse d'homogénéité des variances sur l'erreur de type I, c'est-à-dire la probabilité de rejeter à tort H0, minimes dans la plupart des cas, peuvent être substantiels dans le cas d'un faible nombre de mesures et plus encore de groupes non équilibrés.

Un cas extrême de déviation d'une distribution normale est celui d'une variable dichotomique, codée par exemple sous forme de 0 ou de 1 pour chaque observation. Ce cas de figure a été étudié par Lunney (1970), qui note qu'il peut être particulièrement utile de recourir à un tel codage des données dans le cas de certaines études développementales s'adressant à des sujets trop jeunes pour pouvoir fournir des réponses autres que binaires, ou encore lorsque les sujets d'étude fournissent des réponses qui peuvent être considérées comme justes ou fausses. Ce dernier cas de figure revêt pour nous un intérêt tout particulier. En effet, si dans le cadre d'une tâche de catégorisation un tel recodage des données induit une perte de l'information relative aux confusions entre catégories, il permet également d'évaluer de manière standardisée la significativité des différences entre taux d'identification des différentes catégories.

Lunney (1970) a évalué, pour des niveaux de probabilité identiques, l'écart entre les résultats obtenus en termes de taille et de puissance de l'effet pour une distribution dichotomique comparativement à une distribution normale, en simulant des séries de 1000 distributions pour chaque type de distribution étudié à l'aide de techniques de Monte Carlo. Il en conclut que plus la probabilité de succès dans chacune des catégories considérées est proche de 50%, plus la distribution dichotomique peut être assimilée à une distribution normale et moins la taille de l'échantillon et le nombre de degrés de liberté du terme d'erreur ne nécessitent d'être élevés. Afin de fournir un cadre plus précis pour l'analyse, Lunney préconise un minimum de 20 degrés de libertés pour le terme d'erreur lorsque la proportion de réponses positives (respectivement négatives) dans la catégorie pour laquelle cette proportion est la plus faible est d'au moins 20%, et un minimum de 40 degrés de liberté pour le terme d'erreur lorsque cette proportion est inférieure à 20%. D'Agostino (1971) prolonge les travaux de Lunney (1970) et évalue l'impact sur les résultats des analyses de variance de l'utilisation de transformations logit et arcsinus des proportions de réponses en lieu et place des valeurs originales de la variable dichotomique. Il conclut à un intérêt limité de ces transformations dans la majorité des cas car, si elles permettent de stabiliser la variance lorsque les proportions sont inférieures à 30% ou supérieures à 70%, elles ne permettent pas en contrepartie de connaître la véritable valeur de l'erreur de type I. Il préconise donc de travailler directement sur les valeurs de la variable dichotomique en suivant les recommandations formulées par Lunney (1970).

Les travaux de Glass *et al.* (1972) et de Lunney (1970) ne mentionnent pas l'utilisation d'un modèle à mesures répétées. Cependant, étant donné que ce modèle impose un plan d'expérience factoriel, dans lequel les groupes sont équilibrés, la condition dans laquelle les implications d'une violation des conditions d'application sont les plus importantes est, de fait, éliminée. Afin de pouvoir appliquer une ANOVA à mesures répétées aux réponses données individuellement par les sujets et recodées sous la forme de bonnes ou mauvaises réponses, il convient de s'assurer que cette analyse n'est pas effectuée avec un nombre de sujets trop faible. Une telle précaution permet de s'assurer de la cardinalité suffisante des degrés de liberté du terme d'erreur, tout particulièrement lorsque la proportion de réponses correctes dans au moins l'une des catégories est inférieure à 20% ou supérieure à 80%.

Si une analyse de variance après recodage des données sous forme de variable dichotomique permet d'évaluer la significativité de différences observées entre groupes sur la base des performances de reconnaissance, cette méthode impose d'éliminer les données présentes en dehors des diagonales des matrices de confusion étudiées et correspondant aux confusions. Bien que ne permettant pas de tester la significativité de

ces confusions, le recours à des méthodes exploratoires permettant de rendre compte de la complexité des données peut donc s'avérer utile.

8.3.2.2 Comparaisons multiples

Au-delà des effets principaux des facteurs fixés qui peuvent être mis en évidence par une analyse de variance, il est souvent intéressant de pouvoir comparer deux à deux au moyen de tests *post hoc* les valeurs moyennes prises par la variable dépendante pour chaque valeur des variables catégorielles que constituent les facteurs fixés. Ces tests permettent d'évaluer la significativité de ces différences en prenant en compte le fait que la probabilité de rejeter à tort au moins l'une des hypothèses nulles augmente avec le nombre de comparaisons réalisées : en comparant ces groupes au moyen de tests t sans une telle correction, il suffirait dans le cas d'un seuil fixé à 5% pour le rejet de l'hypothèse nulle de 20 comparaisons deux à deux pour que la probabilité de rejeter de façon erronée au moins l'une des 20 hypothèses nulles soit de 1. La correction la plus simple et la plus classiquement utilisée est celle de Bonferroni, dont le principe est de diviser le seuil α de probabilité d'erreur de type I (fixé avant correction à 5%, 1% ou 0,1% selon le degré de significativité considéré) par le nombre de comparaisons effectuées.

Cependant la correction de Bonferroni est plus stricte que ne le nécessite le contrôle du taux de faux rejet d'hypothèses nulles, pouvant occasionner une perte de puissance statistique substantielle (voir par exemple Benjamini & Hochberg (1995)). Une solution classique pour évaluer la significativité de toutes les différences par paire avec un contrôle adapté du taux de faux rejet d'hypothèses nulles, lorsque l'analyse de variance a révélé un effet du facteur correspondant, est le recours à un test *post hoc* comme celui de Tukey ou celui de Scheffé. Toutefois le logiciel SPSS que nous utilisons pour les analyses statistiques ne permet pas d'appliquer de tels tests aux facteurs intra-sujets, dans le cas d'une analyse de variance à mesures répétées.

Benjamini & Hochberg (1995) ont montré que leur méthode séquentielle de type Bonferroni, adaptée de la méthode de Holm (1979), permettait un contrôle du taux de faux rejet d'hypothèses nulles équivalent à celui obtenu avec la méthode de Bonferroni classique avec un gain important de puissance statistique. Le principe de cette méthode, que nous appliquerons aux comparaisons par paires effectuées, est d'ordonner les valeurs de p non corrigées obtenues pour l'ensemble des N comparaisons par paires effectuées (typiquement au moyen d'un test t) et de considérer que toutes les hypothèses nulles H0(i) pour lesquelles la probabilité p_i de rejeter à tort l'hypothèse nulle est telle que :

$$p_i \geq \frac{i}{N}\alpha$$

8.3.2.3 Comparaison au niveau du hasard

Comme nous l'avons mentionné ci-dessus, il est difficile de définir de façon univoque une correction des taux de reconnaissance bruts issus d'une tâche de catégorisation d'expressions émotionnelles prenant en compte à la fois les confusions et le niveau de reconnaissance au hasard, et nous avons en conséquence opté pour conserver les taux de reconnaissance bruts. Dans ces conditions, il est cependant utile de pouvoir évaluer l'écart de ces taux de reconnaissance bruts par rapport au niveau du hasard.

Une telle comparaison lorsque le données consistent en des proportions de réponse se fait classiquement au moyen de tests binomiaux pour échantillons uniques (voir par exemple Franck & Stennett (2001) pour une application à la catégorisation d'expressions faciales d'émotions). Etant donné que nous testons au moyen d'analyses de variance à mesures répétées les effets principaux des facteurs considérées dans nos évaluations perceptives, dans un souci de cohérence nous opterons pour une évaluation de l'écart entre le taux de reconnaissance observé et le niveau théorique du hasard par des tests t pour échantillon unique. En effet, de même que les analyses de variance, ce test repose sur la distribution de la fonction F de Fisher.

L'intérêt d'une correction de type Bonferroni lorsqu'une série de comparaisons de taux de reconnaissance avec le niveau du hasard est réalisée peut être discuté. En effet le point de vue selon lequel un tel cas de figure est équivalent à celui de comparaisons multiples est défendable. Dans les résultats que nous présentons de comparaisons de taux de reconnaissance avec le niveau du hasard, nous n'avons pas effectué de correction de ce type. Cependant, pour autant que nous les ayons testées, ces corrections ne modifient que de façon très marginale les niveaux de significativité observés.

8.3.3 Classification hiérarchique

Un outil intéressant pour l'analyse des confusions qui résultent d'une tâche de catégorisation peut être trouvé du côté des méthodes de classification automatique, qui permettent de définir une partition d'un ensemble en fonction de mesures de similarité entre éléments. Appliquées au cas de matrices de confusions, ces méthodes permettent de repérer les ensembles au sein desquels les confusions mutuelles sont les plus importantes. L'une des méthodes de classification automatique les plus répandues est celle dite des K-means, pour laquelle plusieurs méthodes de calcul ont été proposées (voir par exemple Hartigan & Wong (1979)). Cependant cette méthode nécessite de prédéfinir le nombre de classes à obtenir, et n'est donc pas la plus adaptée à une approche exploratoire du type de celle présentée ici.

En revanche la méthode de classification hiérarchique dont Benzécri (1973) donne une description détaillée opère sans a priori quant au résultat à obtenir pour calculer un arbre binaire aussi appelé dendrogramme. Le principe général de la classification hiérarchique ascendante est d'assigner initialement un ensemble par élément et de regrouper deux à deux les ensembles les plus proches à chaque étape de la construction de l'arbre (grâce à la mesure d'une distance entre les deux ensembles), jusqu'à aboutir à la racine de l'arbre en les regroupant tous dans un même ensemble. Il existe également une méthode, moins répandue, de classification hiérarchique descendante qui suit le principe inverse en partant d'un ensemble contenant tous les éléments.

La mesure initiale de dissimilarité entre éléments est dépendante du type de distance choisi pour la mesurer. Nous opterons ici pour la plus classique, c'est à dire la distance euclidienne. Par ailleurs diverses méthodes d'agrégation peuvent être utilisées selon les propriétés attendues des clusters, ce qui revient à choisir une mesure de la distance entre ensembles, appliquée à chaque itération de l'algorithme de classification hiérarchique. Notre choix se portera sur la méthode de Ward, dont le principe est de minimiser la variance intra-classe.

8.4. Evaluation perceptive du sous-corpus sélectionné

Le but principal de ce test de perception est de valider les annotations émotionnelles indiquées par le locuteur M2 à la fin de la session d'enregistrement spontanée, puis actées par ce locuteur, que nous avons sélectionné comme locuteur de référence (cf. section 6). Ceci est nécessaire avant de pouvoir étendre notre étude aux productions spontanées de ce locuteur, car nous ne connaissons à priori ni la pertinence cognitive de ces annotations, celle des performances de l'acteur au regard de chacune de ces annotations. Cela revient à répondre aux questions suivantes : les annotations utilisées correspondent-elles aux expressions réalisées par l'acteur ? Deux réalisations expressives différentes renvoient-elles systématiquement à des annotations différentes ou certaines peuvent-elles être perçues comme équivalentes ? Inversement, chaque annotation renvoie-t-elle à une réalisation expressive différente ou certaines peuvent-elles être confondues ?

8.4.1 Sélection des stimuli

Pour chaque locuteur enregistré, 227 productions actées sont obtenues. Si chacun des stimuli était sélectionné, le test de validation en conditions audio et audiovisuelle comporterait 504 stimuli, d'où une charge cognitive que nous avons estimée trop importante pour les auditeurs. Afin de restreindre cet ensemble et éviter de biaiser une partie des résultats par un effet de lassitude des auditeurs, une présélection experte a été réalisée en choisissant les meilleures réalisations.

Cette présélection a été réalisée de façon concertée par deux juges, selon le protocole suivant. Les paires de stimuli (dans les deux conditions de présentation) ont été classées selon un ordre aléatoire. Chaque juge a alors évalué, pour chacune de ces paires, le stimulus en condition audio seul puis le stimulus audiovisuel, avant de s'accorder sur l'attribution d'une note entre 1 (mauvaise performance) et 4 (très bonne performance). Il convient de souligner que la problématique ici n'est pas directement comparable à celle de l'annotation experte de productions spontanées abordée en section 5. En effet, cette méthode plus classique d'évaluation est celle utilisée pour un corpus d'expressions émotionnelles actées dans lequel les émotions que l'acteur est supposé exprimer sont dictées par l'expérimentateur.

Seuls les stimuli ayant reçu une note de 3 ou 4 ont été présélectionnés. Parmi ceux-ci, un sous-ensemble de 70 stimuli a été retenu, selon les critères suivants.

- Une variation systématique de la longueur (stimuli de 1, 3, 5 et 7 syllabes) afin de tester les effets de longueur.
- La présence de toutes les émotions actées (soit14 émotions : amusement, colère, anxiété, déception, dégoût, attente, peur, joie, neutre, résignation, tristesse, satisfaction, surprise, inquiétude) pour chaque longueur de stimulus.
- Un énoncé identique (« page suivante ») représenté pour chaque expression, en complément de la phrase de 3 syllabes retenue. Cet énoncé est en effet très largement représenté parmi les expressions spontanées des locuteurs que nous avons sélectionnés.

8.4.2 Protocole expérimental

Une interface graphique permettant la présentation des stimuli et l'enregistrement des réponses des sujets a été développée à l'aide du logiciel *Revolution Studio*. Après l'affichage d'une page d'instructions détaillées, les stimuli sont présentés en condition audio ou audiovisuelle. Puis une page de réponse (cf. Figure 29), composée de 14 curseurs (un par étiquette proposée) placés initialement à 0, s'affiche. Cette représentation permet de noter l'intensité perçue pour chacune des émotions attribuée

par l'auditeur. Les sujets ne peuvent valider leur réponse (et donc écouter le stimulus suivant) qu'après avoir attribué une valeur non nulle à au moins un des 14 curseurs.

26 sujets (4 hommes, 22 femmes), âgés de 19 à 45 ans (25 en moyenne) ont participé à cette expérience. Leur tâche consistait à juger l'intensité émotionnelle perçue pour chacune des 14 émotions sur une échelle de 0 (l'émotion n'a pas été perçue) à 10 (l'émotion a été perçue très intensément). Pour cela, les sujets écoutaient tous les stimuli présentés, dans un ordre aléatoire contrôlé afin d'éviter deux présentations consécutives du même énoncé, et différent pour chacun des sujets. Ils devaient noter l'intensité émotionnelle de chacune des 14 émotions sur une échelle de 0 à 10 afin d'évaluer les éventuels mélanges d'émotions exprimés par un même stimulus et leur pondération. Les stimuli étaient d'abord présentés en condition audio seul (AS), puis dans une seconde condition expérimentale dans sa modalité audiovisuelle (AV). Les stimuli ne pouvaient être écoutés qu'une seule fois, et les sujets devaient donner leur réponse aussi spontanément que possible. Les sujets étaient placés dans une pièce silencieuse, et écoutaient les stimuli à l'aide d'un casque fermé de bonne qualité, réglé à un niveau d'écoute confortable avant le démarrage du test.

Dans un souci d'homogénéité de la présentation de la page de réponses, un curseur permettant de noter l'intensité perçue a été associé à l'étiquette *Neutre*, de même qu'aux autres étiquettes proposées. Cependant, la plupart des juges ayant donné des commentaires à la suite du test ont déclaré avoir été troublés par le fait de devoir associer une intensité à une absence d'expression émotionnelle perçue, et avoir systématiquement attribué la même valeur d'intensité (ou une valeur choisie aléatoirement pour certains d'entre eux).

Plus généralement, ces juges semblent avoir utilisé les curseurs non pas exclusivement pour donner une indication de l'intensité d'émotion perçue comme il le leur était demandé, mais également pour certains d'entre eux pour donner une indication de la proportion attribuée à chacune des étiquettes proposées indépendamment de la valeur de l'intensité perçue.

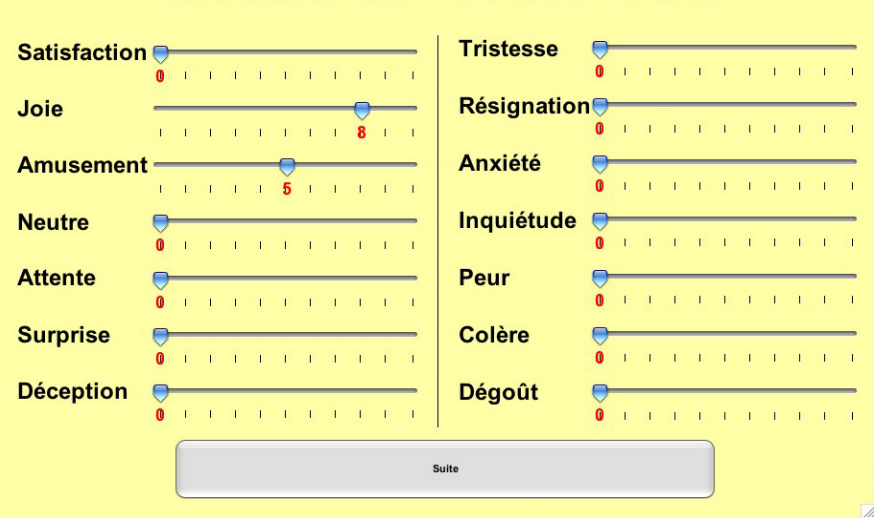


Figure 29 : Page de réponse du test utilisé pour l'évaluation perceptive en conditions audio et audiovisuelle de productions actées sélectionnées du locuteur M2.

8.4.3 Analyse qualitative

Afin d'analyser la reconnaissance de chacune des émotions présentées ainsi que les éventuelles confusions avec d'autres émotions, des matrices de confusion ont été extraites pour chacune des conditions de présentation. Les données de ces matrices de confusion sont présentées dans la Table 4 pour la condition AS et dans la Table 5 pour la condition AV.

Comme le souligne Bänziger (2004, p. 33), des données correspondant à l'évaluation perceptive de mélanges d'émotions peuvent difficilement faire l'objet d'un test d'hypothèse statistique. En effet, les méthodes classiques destinées à évaluer des différences de moyennes, comme l'analyse de variance (ANOVA), reposent sur un modèle statistique dans lequel il n'y a qu'une variable dépendante. Il n'est donc pas possible de les appliquer à de telles données sans éliminer une partie de l'information, par exemple en ne tenant compte que des réponses jugées correctes selon un critère donné.

Ainsi, la première étape d'analyse des réponses fournies par les juges a consisté en une analyse qualitative des matrices de confusion pour chacune des conditions de présentation, permettant de conserver toute l'information sur la distribution des réponses dans les différentes catégories proposées aux sujets. Comparativement à des matrices de confusions plus classiques résultant d'une tâche de catégorisation dans lesquelles une réponse unique catégorielle peut être donnée, qui indiquent la proportion de réponses attribuées à chacune des étiquettes proposées aux sujets, les matrices de confusion que nous obtenons indiquent l'intensité moyenne attribuée à chaque étiquette.

	amu.	anx.	att.	col.	déc.	dég.	inq.	joi.	neu.	peu.	rés.	sat.	sur	tri.
amusement	**3.73**	0.08	0.04	0.00	0.00	0.00	0.04	3.31	0.08	0.00	0.04	1.35	0.27	0.04
anxiété	0.00	**1.96**	0.58	0.62	0.35	0.15	2.12	0.04	0.62	1.23	0.54	0.08	0.46	0.23
attente	0.00	0.58	**0.69**	0.62	0.46	0.08	0.69	0.00	2.62	0.27	1.15	0.08	0.12	0.31
colère	0.00	0.27	0.50	**4.69**	0.35	0.27	0.12	0.00	0.31	0.12	0.46	0.04	0.08	0.15
déception	0.00	0.00	0.58	0.88	**1.96**	1.15	0.00	0.00	0.65	0.00	3.54	0.00	0.04	0.77
dégoût	0.42	0.04	0.38	0.27	1.15	**2.46**	0.04	0.15	0.42	0.00	1.85	0.38	0.15	0.38
inquiétude	0.00	1.04	0.88	0.31	0.69	0.19	**1.62**	0.00	0.88	0.54	0.96	0.00	0.88	0.23
joie	1.08	0.08	0.27	0.08	0.04	0.04	0.00	**1.00**	2.38	0.00	0.12	1.50	0.27	0.04
neutre	0.00	0.19	0.58	0.42	0.92	0.58	0.31	0.00	**2.92**	0.00	1.85	0.04	0.08	0.77
peur	0.00	2.04	0.69	0.19	0.62	0.42	1.65	0.00	0.58	**1.04**	0.96	0.04	0.23	0.77
résignation	0.00	0.15	0.42	2.08	0.69	0.23	0.19	0.00	2.00	0.00	**1.19**	0.12	0.27	0.46
satisfaction	1.27	0.12	0.23	0.12	0.27	0.15	0.08	2.65	0.23	0.04	0.15	**2.35**	0.92	0.00
surprise	0.15	0.27	0.73	0.31	0.50	0.08	0.54	0.04	0.35	0.04	0.73	0.23	**3.15**	0.27
tristesse	1.12	0.88	0.27	0.00	0.50	0.15	1.12	0.27	1.00	1.04	0.35	0.31	0.12	**1.42**

Table 4 : Matrice de confusion correspondant aux valeurs d'intensité attribuées en condition audio seule à chacune des étiquettes proposées comme choix de réponse possible pour catégoriser les productions du locuteur M2, moyennées par émotion présentée et sur les 26 sujets ayant participé au test perceptif.

	amu.	anx.	att.	col.	déc.	dég.	inq.	joi.	neu.	peu.	rés.	sat.	sur	tri.
amusement	**4.00**	0.00	0.04	0.00	0.04	0.00	0.00	2.73	0.08	0.00	0.00	1.54	0.50	0.00
anxiété	0.00	**2.31**	0.46	0.46	0.19	0.04	2.27	0.00	0.81	1.31	0.38	0.04	0.42	0.12
attente	0.00	0.69	**1.38**	0.15	0.23	0.04	1.42	0.00	1.73	0.15	0.50	0.04	1.35	0.08
colère	0.00	0.15	0.65	**5.00**	0.38	0.12	0.08	0.00	0.00	0.04	0.35	0.00	0.04	0.08
déception	0.00	0.00	1.12	1.04	**1.65**	1.08	0.00	0.00	0.12	0.00	4.12	0.00	0.00	0.23
dégoût	0.00	0.08	0.15	0.04	0.92	**4.96**	0.08	0.00	0.12	0.00	0.77	0.00	0.00	0.35
inquiétude	0.00	2.04	0.77	0.31	0.31	0.15	**2.54**	0.00	0.46	1.15	0.31	0.00	1.12	0.19
joie	1.54	0.04	0.27	0.04	0.00	0.00	0.00	**2.58**	0.62	0.00	0.12	3.00	0.31	0.00
neutre	0.04	0.19	0.58	0.27	0.54	0.08	0.15	0.00	**3.23**	0.00	1.65	0.04	0.00	0.85
peur	0.00	2.19	0.42	0.19	0.58	0.19	2.04	0.00	0.62	**1.88**	0.58	0.00	0.46	0.69
résignation	0.00	0.12	0.85	2.08	1.19	0.50	0.15	0.00	1.12	0.00	**2.46**	0.00	0.04	0.19
satisfaction	1.88	0.00	0.04	0.12	0.00	0.00	0.00	3.58	0.08	0.00	0.04	**3.08**	0.46	0.00
surprise	0.04	0.08	0.62	0.27	0.42	0.04	0.38	0.08	0.31	0.00	0.69	0.35	**4.04**	0.08
tristesse	0.00	1.19	0.23	0.08	0.50	0.23	1.23	0.00	0.54	0.96	0.50	0.08	0.04	**3.12**

Table 5 : Matrice de confusion correspondant aux valeurs d'intensité attribuées en condition audiovisuelle à chacune des étiquettes proposées comme choix de réponse possible pour catégoriser les productions du locuteur M2, moyennées par émotion présentée et sur les 26 sujets ayant participé au test perceptif.

La première observation qui peut être tirée des données recueillies est que les juges ont tiré parti de la possibilité d'indiquer des mélanges d'émotions. Ils ont ainsi utilisé en moyenne 1.66 étiquettes par réponse en condition AS (écart-type 0.90), et 1.69 en condition AV (écart-type 0.82). La variabilité inter-sujets est relativement importante. Tandis que 3 des 26 sujets n'ont indiqué aucun mélange d'émotion dans l'ensemble du test, 3 ont utilisé en moyenne plus de 3 étiquettes par réponse. La valeur élevée de l'alpha de Cronbach (α=0.95) indique néanmoins une stratégie globalement homogène des juges dans la tâche d'identification qui leur était proposée.

Une première observation, prévisible, est que les résultats en condition audiovisuelle sont toujours égaux ou meilleurs que ceux obtenus en condition audio seul. Néanmoins les résultats observés entre les deux conditions montrent une certaine cohérence. L'étude qualitative des différences entre ces deux conditions permet de tirer les observations suivantes.

- Certaines des étiquettes proposées n'ont pas été utilisées par les auditeurs. Ainsi, l'attente est quasi-systématiquement confondue avec le neutre, et la déception avec la résignation. De plus, pour ces deux catégories, les expressions faciales ne donnent pas d'informations supplémentaires, comme le montre la matrice de confusion pour les stimuli audiovisuels.
- En condition audio seul, les auditeurs ont largement utilisé l'étiquette *neutre* pour catégoriser les expressions émotionnelles qu'ils ne parvenaient pas à identifier. Nous avons pu observer cela en particulier pour les expressions émotionnelles avec une faible activation, telles que l'*attente*, la *joie* (l'expression de la joie actée par ce locuteur semble d'après une écoute experte très faiblement activée) ou encore la résignation.
- L'*anxiété*, la *peur* et l'*irritation* sont confondus dans les deux conditions proposées, les auditeurs pouvant difficilement faire la différence entre ces expressions. Les expressions d'*amusement*, de *joie* et de *satisfaction* sont également confondues mais ces confusions ne sont pas systématiques. L'*amusement* est bien reconnu en condition audiovisuelle, tandis que la *satisfaction* est reconnue tantôt comme de la *joie*, tantôt comme de la *satisfaction*. L'expression de la *joie*, quant à elle, est considérée comme *neutre* en condition audio seul, mais les réponses sont distribuées entre *amusement*, *joie* et *satisfaction* en condition audiovisuelle.
- Les expressions émotionnelles les mieux identifiées à partir des indices acoustiques seuls sont l'*amusement* (bien que parfois confondu avec la *joie*), l'*anxiété* (parfois confondue avec l'*irritation* et la *peur*), la *colère*, le *neutre*, la *satisfaction* (parfois confondue avec la *joie*) et la *surprise*.

- Le dégoût semble difficile à reconnaître en condition audio seul, tandis qu'il est parfaitement bien reconnu en condition audiovisuelle. Ces résultats sont en accord avec les conclusions de Scherer *et al.* (2003) ainsi qu'avec celles de Juslin & Laukka (2003). En effet, ils observent, à partir de méta-analyses, que l'expression du dégoût, globalement aussi bien reconnue que les expressions d'autres émotions dites « de base » dans sa modalité faciale, est reconnue en modalité audio avec un taux peu supérieur au niveau du hasard. Johnstone & Scherer (2000) suggèrent que cette reconnaissance plus faible du dégoût dans la parole pourrait s'expliquer en termes évolutionnistes. En considérant les émotions comme des réponses adaptatives à l'environnement, l'avantage adaptatif d'une expression de colère ou de peur pouvant être véhiculée sur une plus longue distance via la parole afin d'alerter les congénères est en effet clair. En revanche dans le cas du dégoût cet avantage adaptatif est plus important si cette expression est dirigée vers des congénères consommant des aliments au même endroit, comme peut le faire l'expression faciale liée au mécanisme de régurgitation[24] et de blocage des odeurs déplaisantes, ce qui pourrait expliquer la prépondérance des expressions faciales dans l'expression du dégoût
- Toutefois, l'observation à un niveau de granularité plus fin des résultats obtenus sur l'expression du dégoût en condition audio seul vs. audiovisuel suggère une variation importante de la capacité à percevoir le dégoût à partir de seuls indices acoustiques d'un auditeur à l'autre. En effet, parmi les 24 auditeurs ayant reconnu de façon non équivoque au moins 4 des 5 expressions de dégoût présentées en condition AV, 2 ont été aussi performants en condition AS, tandis que 6 n'ont pas reconnu plus d'une expression de dégoût parmi les 5 présentées en condition AS. Une telle variabilité interindividuelle peut être rapprochée de la notion d'intelligence émotionnelle, en particulier dans sa composante de perception des émotions d'autrui (voir par exemple Mayer & Salovey (1997)). Bien que la variabilité des compétences pour la perception de l'expression vocale du dégoût n'ait, à notre connaissance, pas été étudiée, ce résultat est à mettre en parallèle avec l'étude de Rozin *et al.* (2005). A partir d'expressions faciales et gestuelles, ils ont observé une variabilité plus importante pour la perception du dégoût que pour celle d'autres émotions négatives, avec néanmoins une cohérence importante entre la perception du dégoût à partir de la face et de la gestualité. Les résultats que nous obtenons pour la perception du

[24] Le terme utilisé par les auteurs de « régurgitation », qui désigne en physiologie un mécanisme passif de reflux lié à des causes autres que le rejet d'aliments impropres à la consommation (Pocock & Richards, 2004), semble ici moins adapté que les mécanismes de haut-le-cœur et de vomissement décrits en détails par Borison *et al.* (1953).

dégoût dans ce test perceptif ne nous permettent pas de tirer de conclusions générales, mais suggèrent à l'inverse que la capacité à percevoir le dégoût pourrait être dissociée selon les modalités.

8.4.4 Regroupement de catégories par classification hiérarchique

L'analyse qualitative des données recueillies dans les matrices de confusion pour chacune des conditions de présentation laisse apparaître d'importantes confusions mutuelles entre les étiquettes proposées aux sujets, qui constituent une source d'information riche, comme le souligne Scherer (2003, p.236) :

« […] les confusions que font les auditeurs peuvent être considérées comme encore plus intéressantes [que la précision de décodage] car les erreurs ne sont pas distribuées au hasard et les patterns de ces erreurs d'identification fournissent d'importantes informations sur le processus de décodage. » [25]

Ainsi Johnstone & Scherer (2000) indiquent que les confusions sont plus larges à l'intérieur d'une même famille d'émotion, en s'appuyant notamment sur l'exemple de la colère chaude et de la colère froide, et proposent de mettre à profit ces confusions entre émotions comme une mesure de proximité entre émotions ou classes d'émotions. Ces conclusions prolongent les travaux de Banse & Scherer (1996), qui suggèrent que les patterns de confusion entre émotions peuvent être largement expliqués selon les trois dimensions de qualité de l'émotion, d'intensité et de valence. Néanmoins ils pointent également du doigt la dissymétrie des patterns de confusions observés dans leur étude, qui indique que ces trois dimensions ne permettent pas d'expliquer l'ensemble des confusions.

Afin de pouvoir étudier des tendances plus générales, notamment dans l'optique d'expérimentations ultérieures sur des stimuli présentés en condition audio, nous avons donc voulu prendre en compte les principales confusions entre étiquettes dans cette condition en définissant des classes plus larges, ce qui n'exclut pas pour autant d'étudier les patterns de confusion entre ces classes plus larges. Nous avons pour cela eu recours à une classification hiérarchique fondée sur les distances entre colonnes de la matrice de confusion (c'est-à-dire les catégories d'expressions émotionnelles telles qu'elles sont utilisées par les sujets pour donner leurs réponses) en condition audio seul, dont le

[25] Traduction personnelle de la citation originale *"[…] the confusions listeners make are arguably even more interesting [than decoding accuracy] as errors are not randomly distributed and as the patterns of misidentification provide important information on the judgment process."*

résultat est présenté Figure 30. Etant donné que nous considérons ici les colonnes de la matrice de confusion, toutes ces colonnes n'ont pas un poids égal. Afin de prendre en compte pour le calcul des distances dans la classification hiérarchique la forme de la distribution et non ces variations de poids des colonnes, chaque colonne a été normalisée par la somme de ses éléments.

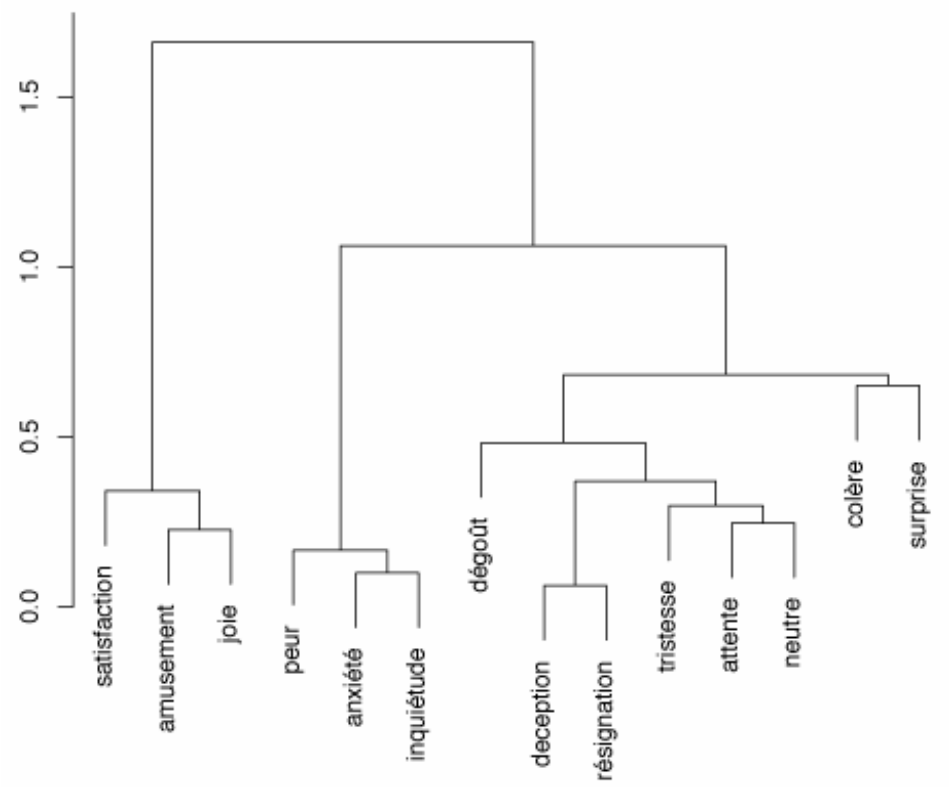


Figure 30 : Dendrogramme issu de la classification hiérarchique (critère d'agrégation de Ward) opérée sur les colonnes normalisées de la matrice de confusion correspondant à l'évaluation de productions actées sélectionnées du locuteur M2 en condition audio seul.

A partir de ces confusions, nous proposons de définir les regroupements suivants pour l'analyse des réponses données par les sujets en condition AS :

- Les étiquettes *anxiété*, *inquiétude* et *peur* sont regroupées. Ces étiquettes renvoient à des affects susceptibles d'apparaître en réaction à une menace directe ou indirecte.
- Les étiquettes *joie*, *satisfaction* et *amusement* sont regroupées au sein d'une même catégorie. Elles ont en commun une valence positive, qui les oppose à l'ensemble des autres étiquettes représentées dans cet ensemble.

- Les étiquettes *déception*, *résignation* et *tristesse* sont également regroupées. Le choix de regrouper *déception* et *résignation*, qui sont les deux premières étiquettes regroupées (cf. Figure 30), semble logique en dépit de la faiblesse des reports de *résignation* vers *déception*. Toutefois le regroupement de ces deux étiquettes avec *tristesse* est moins évident, sauf à considérer un regroupement global incluant également *attente* et *neutre* : en effet ces cinq étiquettes sont regroupées dans l'analyse hiérarchique à un niveau proche de celui correspondant au regroupement de *joie* et *satisfaction*. Plutôt que de conserver pour les analyses ultérieures une telle classe comprenant 5 étiquettes qui renvoient à des manifestations affectives par trop variables, nous avons choisi de la scinder en deux ensembles. En effet, un tel regroupement aurait conduit à considérer cette classe comme une catégorie par défaut sans que des informations pertinentes puissent être tirées du choix de l'un des membres de cette classe par les sujets. Les étiquettes *tristesse, déception* et *résignation* correspondant à des réactions passives du sujet face aux événements évalués comme négatifs pour lui, nous avons jugé préférable d'opérer un regroupement entre ces 3 catégories. Bien que, contrairement aux autres regroupements effectués, les liens entre *tristesse*, *déception* et *résignation* soient moins directs, nous considérons que la qualité des expressions de la tristesse actée est à mettre en cause plutôt que la proximité cognitive de ces étiquettes. L'acteur, peut-être car il n'avait pas ressenti ou exprimé dans la partie spontanée du corpus de la tristesse qui aurait pu servir de référence pour son expression actée, s'est efforcé de la produire avec une activation importante. Ceci peut expliquer la faiblesse des confusions avec *déception* ou *résignation*. De plus, une partie des expressions de tristesse produites par ce locuteur l'a été avec un niveau de tremblement vocal (*vocal tremor*) élevé, d'où l'importance relative des reports de *tristesse* vers *amusement*, cette caractéristique acoustique pouvant également évoquer le rire.
- Les étiquettes *neutre* et *attente* situées par le locuteur dans les passages du scénario, où il ne ressentait pas d'émotions mais était concentré sur la tâche, ont été regroupées. En effet, ces deux expressions transmettent une valence nulle et une activation faible. En ce sens, et bien qu'elles contribuent à l'expressivité du locuteur, il est possible de s'interroger sur la mesure dans laquelle ces expressions d'attente constituent ou non des expressions émotionnelles en tant que telle. En effet la plupart des théories des émotions, qu'elles proposent une approche simplifiée centrée sur les phénomènes de haut niveau dans une optique computationnelle comme le modèle OCC d'Ortony *et al.* (2000), ou aussi élaborées que le modèle des processus composants de Scherer (2001), considèrent l'évaluation de la valence d'un stimulus comme central dans la réaction émotionnelle.

- Enfin, nous avons conservé les étiquettes *colère*, *dégoût* et *surprise* comme des catégories séparées. Ce sont les dernières agrégées par la classification hiérarchique. De plus, elles correspondent à des expressions d'émotions dites « basiques », auxquelles il est délicat de rattacher les affects moins prototypiques rejoués par l'acteur d'après ceux ressentis durant l'induction, qui ne peuvent être considérés comme relevant de la même famille d'émotions.

Afin de comparer les principales confusions entre étiquettes en condition AS et AV, une classification hiérarchique ascendante a également été effectuée sur les colonnes de la matrice de confusion en condition AV, normalisées par la somme des intensités attribuées à chaque colonne. Le dendrogramme résultant de cette classification hiérarchique est présenté en Figure 31. De même qu'en condition AS, les confusions mutuelles importantes nous permettent d'aboutir à un premier regroupement entre *joie, satisfaction et amusement*, à un second entre *anxiété*, *inquiétude* et *peur* et un dernier entre *déception* et *résignation*. Les étiquettes *colère*, *surprise* et *dégoût* demeurent quant à elles bien séparées, et *attente* et *neutre* restent relativement proches. En revanche, la répartition est très différente pour la *tristesse*, qui est dans cette condition bien séparée des autres étiquettes.

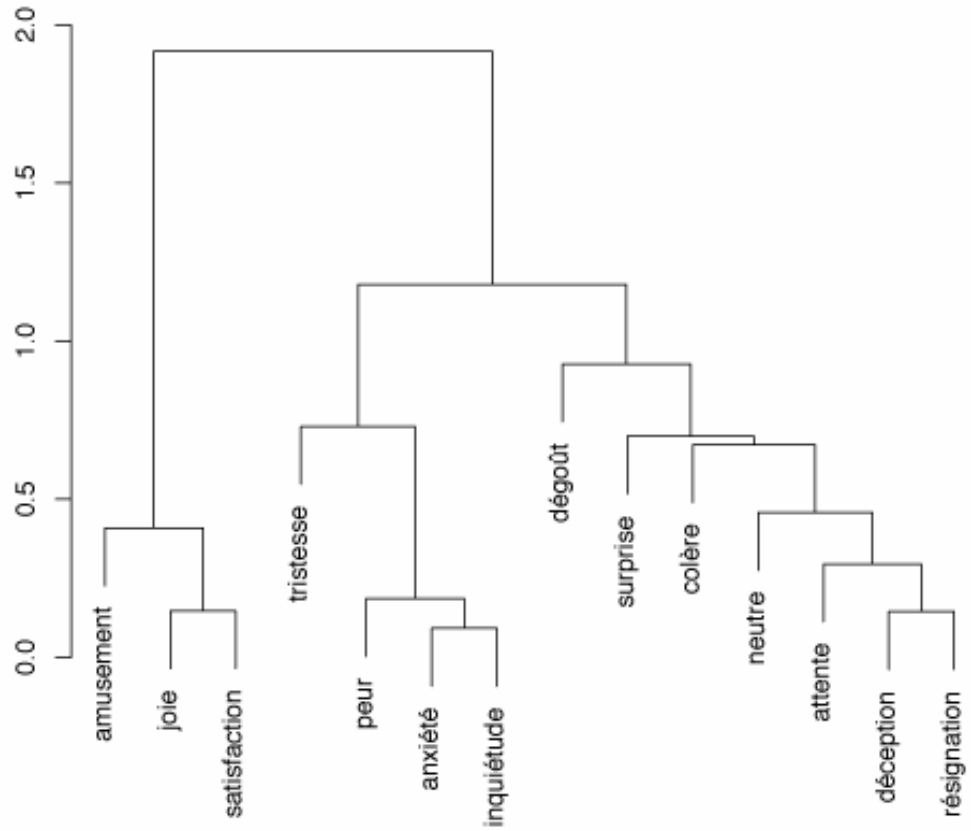


Figure 31 : Dendrogramme issu de la classification hiérarchique (critère d'agrégation de Ward) opérée sur les colonnes normalisées de la matrice de confusion correspondant à l'évaluation de productions actées sélectionnées du locuteur M2 en condition audiovisuelle.

8.4.5 Analyse statistique après regroupement

Nous avons choisi ici de concentrer l'analyse statistique sur les scores de reconnaissance plutôt que sur l'intensité des expressions émotionnelles exprimées par le locuteur. En effet, comme nous l'avons mentionné ci-dessus, la complexité de la tâche donnée aux juges a conduit une partie d'entre eux à utiliser les curseurs destinés à noter l'intensité perçue comme un moyen d'indiquer des proportions relatives plutôt que des intensités directement comparables d'un stimulus à l'autre.

En considérant les regroupements de catégories définis ci-dessus, les données ont été recodées sous forme dichotomique afin de pouvoir évaluer statistiquement les taux de reconnaissance de chacune des émotions présentées par rapport au niveau du hasard ainsi que les effets éventuels du type d'énoncé présenté. Pour chaque stimulus présenté, une réponse correcte est codée 1 (et 0 dans le cas contraire) si l'étiquette portant l'intensité la plus importante fait partie du groupe considéré. Ainsi, une expression de joie est considérée comme reconnue si l'intensité la plus élevée correspond à l'étiquette *joie*, *satisfaction* ou *amusement*. Ce recodage, malgré une inévitable perte d'information, nous a semblé plus pertinent que la solution alternative consistant à considérer la réponse comme correcte lorsqu'au moins l'un des curseurs choisis par le sujet appartient à la catégorie correspondant au stimulus, sans considération de l'intensité, ce qui aurait conduit à une surestimation des taux de catégorisation correcte.

Le protocole retenu pour le test perceptif laisse la possibilité aux juges de sélectionner plusieurs étiquettes sans imposer qu'une seule puisse être choisie avec l'intensité maximale. Par conséquent, le choix par les sujets d'attribuer à plusieurs étiquettes l'intensité maximale pour la réponse à un même stimulus peut également résulter en une surestimation des taux de catégorisation correcte. Ainsi, le cas extrême du choix systématique de l'ensemble des étiquettes (ou d'au moins une par catégorie regroupée) en leur associant la même intensité produirait artificiellement un taux de catégorisation correcte de 100% pour chacun des stimuli présentés. Le nombre moyen d'étiquettes par réponse auxquelles la plus forte intensité a été attribuée est ici de 1,185 en condition AS et de 1.133 en condition AV. Afin de corriger le biais positif sur les valeurs de reconnaissance et de confusions résultant de ces choix multiples, nous avons normalisé les matrices de confusion résultantes pour chacune des conditions de présentation en divisant chaque ligne par le nombre moyen d'étiquettes sélectionnées avec l'intensité maximale lors de la présentation d'expressions de l'émotion correspondante. Ces matrices de confusion normalisées, présentées en Table 6 pour la condition AS et en Table 7 pour la condition AV, reflètent donc les proportions de choix de chacune des étiquettes, fusionnées en fonction des regroupements définis ci-dessus. Les taux de

reconnaissance moyens obtenus après regroupement sont de 59% en condition AS et de 78% en condition AV.

	amu/joie/sat	anx/inq/peu	att/neu	rés/déc/tri	colère	dégoût	surprise
amusement	**99%**	0%	0%	1%	0%	0%	0%
joie	**47%**	1%	46%	2%	1%	1%	1%
satisfaction	**79%**	1%	5%	5%	1%	1%	8%
anxiété	2%	**63%**	13%	9%	8%	1%	3%
inquiétude	0%	**37%**	25%	20%	5%	1%	13%
peur	0%	**55%**	16%	20%	3%	5%	2%
attente	0%	17%	**52%**	21%	8%	0%	1%
neutre	0%	3%	**52%**	38%	3%	3%	0%
déception	0%	0%	13%	**76%**	5%	6%	0%
résignation	0%	4%	38%	**24%**	31%	1%	1%
tristesse	21%	29%	17%	**32%**	0%	0%	0%
colère	1%	2%	8%	2%	**88%**	0%	1%
dégoût	10%	1%	9%	38%	1%	**39%**	1%
surprise	4%	5%	14%	16%	2%	1%	**59%**

Table 6 : Matrice de confusion correspondant à l'évaluation en condition audio seule des productions sélectionnées du locuteur M2, normalisée après regroupement

Bien que les regroupements de catégories aient été effectués en tenant compte des principales confusions en condition AS entre choix de réponses proposés aux sujets, d'importantes confusions peuvent encore être observées après regroupement dans cette condition. Ainsi, tandis que les expressions d'amusement et de satisfaction sont reconnues avec très peu de confusions, celles de joie, produites par ce locuteur avec une activation moindre, sont largement confondues avec l'attente et l'expression neutre. Des tendances similaires sont observées pour les expressions de peur, d'anxiété et d'inquiétude, avec des reports portant majoritairement sur l'attente et l'expression neutre, ainsi que sur la résignation, la déception et la tristesse avec toutefois une moindre amplitude dans le cas de l'anxiété. L'attente et l'expression neutre sont majoritairement confondues avec déception, résignation et tristesse, avec des reports relativement importants vers anxiété, inquiétude et peur dans le cas de l'attente. Tandis que les expressions de déception montrent peu de confusions, les expressions de résignation sont largement confondues avec l'attente et l'expression neutre, et dans une moindre mesure avec la colère. La tristesse est principalement confondue avec anxiété, inquiétude et peur, ainsi qu'avec les émotions positives, ce qui s'explique probablement par la présence d'un important jitter dans certaines de ces expressions comme nous l'avons mentionné ci-dessus. Les expressions de colère sont très peu confondues, tandis que le dégoût l'est majoritairement avec résignation, déception et tristesse. Enfin, la surprise est

faiblement confondue avec attente et neutre ainsi qu'avec résignation, déception et tristesse.

Bien que l'ensemble d'émotions étudiées soit distinct de celui que nous avons retenu ici, ce qui ne permet pas une comparaison systématique des tendances de confusion, une partie de nos résultats en condition AS peut être mise en parallèle avec ceux obtenus par Banse & Scherer (1996). Ainsi, la reconnaissance et les confusions qu'ils obtiennent dans le cas d'expressions de peur panique et d'anxiété sont proches de celles que nous observons, avec d'importantes confusions mutuelles entre ces émotions. En revanche les autres tendances de confusion que nous obtenons divergent des résultats de Banse & Scherer (1996) : en effet les expressions de tristesse qu'ils ont évaluées, sensiblement mieux reconnues que celles produites par notre acteur, n'ont été que très marginalement confondues avec des expressions d'anxiété ou de joie, mais l'ont été plus fréquemment avec les expressions d'ennui que l'on peut rapprocher des expressions de résignation présentes dans notre corpus. Enfin, alors que Banse & Scherer (1996) indiquent que les expressions de dégoût, mal reconnues, ont été confondues de façon équilibrée entre les différentes émotions négatives proposées aux sujets comme choix de réponse, le dégoût en condition AS a été confondu de façon très majoritaire avec les expressions de résignation, déception et tristesse.

	amu/joie/sat	anx/inq/peu	att/neu	rés/déc/tri	colère	dégoût	surprise
amusement	**99%**	0%	0%	0%	0%	0%	1%
joie	**88%**	0%	8%	1%	0%	0%	2%
satisfaction	**96%**	0%	1%	0%	1%	0%	3%
anxiété	0%	**74%**	18%	3%	3%	0%	2%
inquiétude	0%	**67%**	10%	7%	2%	1%	13%
peur	0%	**71%**	12%	11%	2%	0%	4%
attente	0%	26%	**49%**	5%	1%	1%	18%
neutre	1%	3%	**66%**	27%	2%	1%	0%
déception	0%	0%	8%	**84%**	5%	3%	0%
résignation	0%	2%	24%	**52%**	19%	3%	0%
tristesse	0%	36%	9%	**54%**	1%	0%	1%
colère	0%	0%	3%	2%	**95%**	0%	0%
dégoût	0%	0%	1%	11%	0%	**88%**	0%
surprise	1%	3%	10%	8%	2%	0%	**77%**

Table 7 : Matrice de confusion correspondant à l'évaluation en condition audiovisuelle des productions sélectionnées du locuteur M2, normalisée après regroupement.

Comparativement aux confusions observées en condition AS, la matrice de confusion en condition AV indique que les informations données par la face et la gestualité (même limitée au buste) permettent principalement de désambigüiser les expressions de joie,

d'inquiétude, de résignation et celles de tristesse et de dégoût. En particulier, les confusions entre la tristesse et les expressions positives, vraisemblablement liées au tremblement vocal, disparaissent totalement lorsque l'information faciale et gestuelle est présente. De même, en dépit de la faible activation des expressions de joie, le sourire présent dans les expressions faciales semble suffisant pour permettre une reconnaissance de ces expressions par la majorité des sujets.

Les confusions observées en condition AV peuvent plus difficilement être mises en parallèle avec des résultats d'études similaires. En effet, outre le fait que peu d'études de catégorisation d'expressions émotionnelles produites par des acteurs indiquent les matrices de confusion et non uniquement les taux de reconnaissance (Johnstone & Scherer, 2000), le plus grand nombre porte sur des expressions faciales statiques et dans une moindre mesure sur des expressions vocales. Parmi les rares études portant sur la catégorisation par des juges d'expressions émotionnelles multimodales produites par des acteurs, des corpus d'expressions émotionnelles spontanées ont été annotés à partir du signal vidéo (voir par exemple Douglas-Cowie *et al.* (2005)), et quelques systèmes dédiés à la reconnaissance automatique d'expressions émotionnelles multimodales ont été construits à partir de données recueillies dans les modalités audio et visuelle (voir par exemple Busso *et al.* (2004)). Nous n'en avons trouvé aucune qui puisse être utilisée comme référence pour l'analyse des confusions observées ici.

Les taux de catégorisation correcte après regroupement, qui apparaissent en gras dans les matrices de confusion ci-dessus, ont été comparés aux niveaux théoriques du hasard résultant des regroupements de catégories pour chacune des émotions présentées et chaque condition de présentation, au moyen de tests t bilatéraux pour échantillon unique. En pratique, la distribution des taux de catégorisation correcte non normalisés à été comparée au niveau théorique du hasard multiplié par le terme correctif appliqué à la ligne correspondante lors de la normalisation des matrices de confusion. Les résultats de ces tests sont présentés dans la Table 8, qui fait également apparaître à titre indicatif les valeurs du taux de reconnaissance sans biais Hu (Wagner, 1993) extraits des matrices de confusion après regroupement présentées dans la Table 6 et la Table 7.

Ainsi, en considérant les regroupements d'étiquettes que nous avons définis, toutes les expressions émotionnelles présentées apparaissent comme reconnues par les sujets avec un taux significativement supérieur au hasard, à l'exception de la résignation. Néanmoins cet avantage est peu significatif dans le cas de la tristesse et de l'inquiétude. En condition audiovisuelle, toutes les expressions présentées sont en revanche reconnues à un taux très significativement supérieur au hasard.

Bien que l'écart entre le niveau du hasard et le taux de reconnaissance sans biais Hu (Wagner, 1993) calculé sur les proportions globales ne puisse être évalué statistiquement, les valeurs de cet indice ne sont dans la plupart des cas que légèrement inférieures aux taux de reconnaissance normalisés. La comparaison de ces valeurs permet de conclure que les taux de reconnaissance des expressions autres que celles de résignation et de tristesse en condition AS sont également significativement supérieurs au niveau du hasard.

En dépit des différences méthodologiques résultant notamment du nombre et de la nature des émotions et des alternatives de réponses proposées, les taux de reconnaissance ainsi extraits pour la condition AS peuvent être comparés qualitativement à ceux relevés dans la littérature pour les études de catégorisation d'expressions vocales d'émotions. Ainsi Scherer (1989), à partir d'une revue de 28 études de décodage d'expressions vocales d'émotions majoritairement fondées sur des productions d'acteurs, observe un taux moyen de reconnaissance de 55%, proche des 59% que nous obtenons, avec néanmoins une très importante variabilité entre les différentes études (les taux de reconnaissance moyens observés dans ces 28 études et corrigés pour tenir compte du hasard varient en effet de 8% à 96%). Toutefois, cette revue ne mentionne pas les taux de reconnaissance obtenus pour chacune des émotions étudiés.

Juslin & Laukka (2003) ont réalisé la revue de littérature la plus large à ce jour sur les expressions vocales des émotions, afin d'extraire les taux de reconnaissance des émotions étudiées ainsi que les caractéristiques acoustiques les plus fréquemment associées à ces émotions. Bien que cette revue ait également porté sur des études interculturelles ainsi que sur la communication d'émotions dans la performance musicale, nous ne considérerons ici que les études portant sur le décodage des expressions émotionnelles par des auditeurs de même langue que les locuteurs les ayant produites. Ce sous-ensemble représente 38 études portant sur 473 locuteurs. Afin de pouvoir comparer, indépendamment du nombre de choix possibles, les taux de reconnaissance obtenus dans ces 38 études pour les expressions de colère, peur, joie, tristesse et tendresse, Juslin & Laukka (2003) les ont convertis en utilisant l'indice π de taille d'effet pour choix multiples proposé par Rosenthal & Rubin (1989). Cet indice, qui prend la valeur 0.5 lorsqu'une catégorie est reconnue au niveau du hasard (c'est-à-dire P=1/k) et la valeur 1 lorsqu'elle est reconnue sans confusions, est calculé comme :

$\pi = \frac{P(k-1)}{1+P(k-2)}$, où k est le nombre réponses proposées et P le taux de reconnaissance brut.

Nous avons donc converti les taux de reconnaissance en valeurs de l'indice π, en considérant l'inverse du niveau théorique du hasard comme valeur de k afin de tenir compte des regroupements de catégories effectués. Ces valeurs sont présentées dans la Table 8. Parmi les émotions comparables à celles que nous évaluons ici, Juslin & Laukka (2003) indiquent pour le décodage des expressions vocales des valeurs moyennes de π de 0.93 pour la colère, 0.88 pour la peur, 0.87 pour la joie et 0.93 pour la tristesse. Le taux de reconnaissance correcte que nous obtenons en condition AS est donc supérieur à la moyenne dans le cas de la colère, mais inférieur pour la peur, la joie et la tristesse, cette différence étant particulièrement nette dans le cas de la tristesse. Néanmoins, si l'on considère l'anxiété comme relevant de la famille d'émotions de la peur, et la déception comme faisant partie de celle de la tristesse, les valeurs de π que nous obtenons en condition AS pour la reconnaissance de ces émotions est très proche des valeurs moyennes relevées par Juslin & Laukka (2003) pour respectivement la peur et la tristesse. En considérant de même la satisfaction et l'amusement comme appartenant à la famille de la joie, les valeurs de π que nous obtenons sont même supérieures aux valeurs moyennes relevées par ces auteurs pour les expressions de joie.

Condition	Emotion	Reconnaissance	π	Hasard	t	p	Significativité
AS	amusement	99% (94%)	1.00	21%	68.99	0.0000	***
	anxiété	63% (59%)	0.86	21%	9.02	0.0000	***
	attente	52% (44%)	0.87	14%	9.61	0.0000	***
	colère	88% (83%)	0.99	7%	31.31	0.0000	***
	déception	76% (65%)	0.92	21%	14.89	0.0000	***
	dégoût	39% (38%)	0.89	7%	7.69	0.0000	***
	inquiétude	37% (35%)	0.68	21%	2.71	0.0077	**
	joie	47% (44%)	0.76	21%	5.71	0.0000	***
	neutre	52% (44%)	0.87	14%	9.81	0.0000	***
	peur	55% (52%)	0.82	21%	6.85	0.0000	***
	résignation	24% (21%)	0.54	21%	0.07	0.9478	n.s.
	satisfaction	79% (75%)	0.93	21%	15.87	0.0000	***
	surprise	59% (57%)	0.95	7%	12.22	0.0000	***
	tristesse	32% (27%)	0.63	21%	2.41	0.0174	*
AV	amusement	99% (99%)	1.00	21%	97.41	0.0000	***
	anxiété	74% (69%)	0.91	21%	12.49	0.0000	***
	attente	49% (45%)	0.85	14%	8.76	0.0000	***
	colère	95% (93%)	1.00	7%	52.01	0.0000	***
	déception	84% (79%)	0.95	21%	20.34	0.0000	***
	dégoût	88% (88%)	0.99	7%	35.89	0.0000	***
	inquiétude	67% (63%)	0.88	21%	10.54	0.0000	***
	joie	88% (88%)	0.96	21%	21.92	0.0000	***
	neutre	66% (61%)	0.92	14%	12.51	0.0000	***
	peur	71% (67%)	0.90	21%	12.19	0.0000	***
	résignation	52% (48%)	0.80	21%	7.10	0.0000	***
	satisfaction	96% (96%)	0.99	21%	47.76	0.0000	***
	surprise	77% (75%)	0.98	7%	19.92	0.0000	***
	tristesse	54% (50%)	0.81	21%	8.28	0.0000	***

Table 8 : Taux de reconnaissance moyen normalisé par émotion (et taux de reconnaissance sans biais Hu (Wagner, 1993) entre parenthèses) correspondant à l'évaluation des productions sélectionnées du locuteur M2 en conditions de présentation audio seule et audiovisuelle, valeurs de l'indice π de taille d'effet pour choix multiples (Rosenthal & Rubin, 1989), et écart au niveau théorique du hasard (tests t bilatéraux pour échantillon unique à 129 degrés de liberté sur les taux de reconnaissance non corrigés avec la valeur théorique du hasard corrigée comme valeur de test). Le niveau théorique du hasard est indiqué dans la colonne *Hasard*. Niveaux de significativité : *** = $p < 0.001$, ** = $p<0.01$; * = $p<0.05$; n.s. = non significatif.

8.4.6 Influence du type d'énoncé sur le taux de reconnaissance

Afin de tester les effets principaux des facteurs ayant gouverné le choix des stimuli présenté, une ANOVA à mesures répétées a été effectuée avec la reconnaissance correcte après regroupement comme variable dépendante et la condition, l'émotion et le type d'énoncé comme facteurs fixés. Les résultats de cette ANOVA confirment le gain de

reconnaissance de la condition AV (78% de reconnaissance en moyenne après regroupement) par rapport à la condition AS (59%), avec un effet significatif de la condition ($F(1,25)=185.1$; $p<0.001$). Une ANOVA à mesure répétées sur ces taux de reconnaissance a également été effectuée pour chacune des conditions de présentation, avec l'émotion et le type d'énoncé comme facteurs fixés. Les résultats de ces trois ANOVAs sont récapitulés dans la Table 9 : l'émotion présentée, le type d'énoncé et l'interaction entre ces deux facteurs ont un effet significatif globalement et quelle que soit la condition de présentation.

Effet	Global	AS	AV
émotion	$F(9.5, 236.4)= 23.8$ ***	$F(10.4, 261.0)=25.6$ ***	$F(12.0, 298.7)=29.1$ ***
type d'énoncé	$F(4, 100)= 2.5$ ***	$F(4, 100)=18.8$ ***	$F(4, 100)=3.1$ *
émotion*énoncé	$F(41.8, 1044.6)= 6.5$ ***	$F(43.8, 1094.4)=5.8$ ***	$F(43.4, 1085.1)= 3.9$ ***

Table 9 : Principaux résultats des ANOVAs à mesures répétées sur les taux de reconnaissance correspondants à l'évaluation des productions sélectionnées du locuteur M2 après regroupement, pour l'ensemble des données (colonne *Global*), la condition audio seule et la condition audiovisuelle. Les degrés de libertés sont corrigés par la méthode de Huynh-Feldt (Huynh & Feldt, 1976) pour compenser d'éventuelles violations de la condition de sphéricité. Niveaux de significativité : *** = $p<0.001$; * = $p<0.05$.

Dans l'optique d'expérimentations fondées sur les stimuli extraits de la partie spontanée du corpus, une donnée particulièrement intéressante est l'influence éventuelle du type d'énoncé présenté sur les taux de reconnaissance obtenus. En effet le corpus spontané se compose d'occurrences de *page suivante* et d'énoncés monosyllabiques (*brique*, *vert*, *rouge*, *jaune*, *sable*). Il est donc essentiel de s'assurer que les expressions émotionnelles portées par de tels stimuli peuvent être reconnues à un niveau supérieur au hasard, de même que dans le cas de phrases porteuses de longueur plus importante.

Les taux moyens de reconnaissance pour chaque type d'énoncé et chaque condition sont présentés dans la Table 10. Afin d'évaluer les différences entre ces taux de reconnaissance au-delà de la significativité de l'effet du type d'énoncé dans les ANOVAs à mesures répétées présentées ci-dessus, nous avons également testé des contrastes spécifiques entre les taux de reconnaissance recueillis pour les énoncés monosyllabiques (respectivement les occurrences de *page suivante*) et la moyenne des phrases porteuses de 3 à 7 syllabes.

Ces contrastes révèlent qu'en condition AV, seules les émotions exprimées sur *page suivante* sont significativement moins bien reconnues que celles exprimées sur des phrases porteuses plus longues ($F(1,25)=13.2$; $p<0.01$), tandis que les émotions exprimées sur des énoncés monosyllabiques sont aussi bien reconnues que sur les phrases porteuses ($F(1,25)=2.4$; $p=0.132$ n.s.).

Type d'énoncé	AS	AV
monosyllabe	55%	76%
page suivante	45%	72%
phrase 3 syllabes	64%	83%
phrase 5 syllabes	63%	78%
phrase 7 syllabes	66%	80%
Moyenne	59%	78%

Table 10 : Taux de reconnaissance moyen correspondants à l'évaluation des productions sélectionnées du locuteur M2 après regroupement pour chaque type d'énoncé présenté et chaque condition de présentation.

En revanche, en condition AS, ces contrastes indiquent que les émotions exprimées tant sur les énoncés monosyllabiques ($F(1,25)=20.9$; $p<0.001$) que sur les occurrences de *page suivante* ($F(1,25)=86.2$; $p<.001$) sont significativement moins bien reconnues que celles exprimées sur les phrases plus longues. Dans cette condition, les expressions d'émotions sur *page suivante* sont de plus significativement moins bien reconnues que celles sur les énoncés monosyllabiques ($F(1,25)= 14.4$; $p<0.001$).

Néanmoins, un test t bilatéral pour échantillon indépendant indique que les émotions exprimées sur des énoncés monosyllabiques ne sont pas significativement moins bien reconnues que la moyenne des expressions présentées en condition audio seul ($t(363)=-1,47$; $p=0.142$ n.s.). Ainsi, dans l'optique d'expérimentations s'appuyant sur des énoncés monosyllabiques, qui permettent de geler la variabilité de la prosodie linguistique, nous pouvons considérer que les résultats généraux pour l'ensemble des stimuli présentés en condition AS restent valables pour de tels énoncés minimaux.

9. Dimensionnalité, localisation et pertinence perceptive des indices acoustiques

Dans cette section, nous nous efforçons d'évaluer l'influence perceptive relative des différentes dimensions prosodiques ainsi du décours temporel dans le décodage d'expressions émotionnelles sélectionnées pour les études présentées dans le chapitre précédent.

Nous présentons tout d'abord les résultats de deux études d'évaluation perceptive fondées sur une projection séparée de dimensions prosodiques d'énoncés actés (à l'aide de TD-PSOLA et de l'algorithme ARX-LF développé par Orange-labs), que nous mettons en relation avec les caractéristiques acoustiques des énoncés évalués, et comparons les résultats obtenus dans ces deux études. Les principaux résultats de ces évaluations sont que, pour les énoncés testés, les contours de fréquence fondamentale véhiculent l'essentiel de l'information affective des expressions positives, tandis que la qualité de voix et la durée du stimulus véhiculent l'essentiel de l'information affective pour les expressions négatives, sans qu'aucune dimension ne soit en mesure de véhiculer seule l'ensemble de l'information affective.

Nous présentons ensuite une étude de dévoilement progressif d'énoncés actés exprimant les mêmes émotions, qui permet d'évaluer le décours temporel du décodage de ces expressions. Les résultats de cette étude montrent une progression globalement linéaire du décodage de l'information affective de la plupart des énoncés présentés, qui pourrait être interprétée comme révélatrice d'un décodage en contours plutôt qu'au moyen d'indices saillants.

9.1. Introduction générale

L'objectif général que nous poursuivons dans la série d'expériences présentée dans cette section est d'étudier finement des phénomènes prosodiques particuliers à travers la perception qu'en ont des auditeurs naïfs, en séparant l'information selon des critères dimensionnels ou temporels. Afin de ne prendre en compte que les variations prosodiques induites par les émotions exprimées par le locuteur et non celles liées aux valeurs prises par les fonctions linguistiques de la prosodie, nous nous sommes focalisé sur les énoncés monosyllabiques, dans lesquels les variations prosodiques sont supposées restreintes aux affects exprimés par le locuteur.

Les stimuli utilisés comme référence dans cette section ont été choisis parmi les productions actées du locuteur M2, dont nous avons précédemment évalué un sous-ensemble en condition audiovisuelle. Etant donnée l'impossibilité de séparer dans le signal vidéo les indices faciaux liés aux différentes dimensions prosodiques, nous avons choisi de ne retenir que les stimuli audio produits par ce locuteur.

Le choix de conserver les productions actées de ce locuteur pour la série d'expériences présentées dans cette section s'est fait principalement en raison de la plus grande clarté des expressions actées recueillies, comparativement aux expressions spontanées (cf. section 5 pour des expériences sur les différences entre expressions actées et spontanées). Ce choix serait problématique si nous nous attachions à tester les hypothèses du modèle d'Aubergé (2002) relatives à la distinction temporelle entre contrôle volontaire et contrôle involontaire. Cependant cela n'est pas notre objectif dans cette partie de nos travaux, et nous ne nous attaquons pas à la dimensionnalité des expressions spontanées. En revanche, les résultats de la dernière expérience présentée dans cette section, dans laquelle les affects perçus à partir d'énoncés présentés graduellement selon le paradigme du dévoilement progressif sont évalués ne seront, si nous retenons cette hypothèse, pas généralisables à des expressions spontanées.

L'auto-annotation produite par ce locuteur laisse apparaître cinq grandes tendances d'expression émotionnelle rejouées par le locuteur en condition actée à la suite de l'induction par le scénario Sound Teacher (anxiété, déception, inquiétude, résignation et satisfaction). Nous avons donc retenu les expressions vocales correspondantes dans la partie actée de son corpus. Trois des « émotions de base » (joie, dégoût et tristesse) ont également été retenues pour compléter cette sélection. La joie a été sélectionnée afin de rééquilibrer l'ensemble d'émotions positives retenues, dont la faible représentation dans la plupart des tâches de catégorisation peut être problématique (cf. section 8). Le dégoût a été retenu en raison de l'intérêt porté aux conséquences acoustiques et perceptives de son expression, considérée comme un possible antagoniste du sourire, au sein de

l'équipe de recherche dans laquelle nos travaux se sont intégrés (Rebreyend, 2006), ainsi que dans le but de compléter l'ensemble d'émotions retenues. La tristesse a été retenue pour cette raison. En revanche la surprise, qui possède un statut ambivalent d'émotion et d'attitude, a été éliminée de l'ensemble des émotions retenues. La peur, dont les expressions en conditions audio seul ont dans l'expérience précédente de catégorisation été très largement confondue avec celles d'anxiété et d'inquiétude, a également été éliminée. Enfin, les expressions de colère, dont un nombre important se sont avérées saturées, n'ont pas non plus été retenues. Cet ensemble a été complété par les expressions neutres produites par ce locuteur, correspondant à l'étiquetage « rien » dans sa partie spontanée et dont les contours de fréquence fondamentale sont également remarquablement plats (cf. le début de la section), ont été retenues comme expressions de référence.

Nous présentons tout d'abord les deux étapes successives d'une évaluation par projection dimensionnelle des poids relatifs des dimensions prosodiques dans l'expression de l'ensemble d'émotions retenu, puis une évaluation par dévoilement progressif de l'information affective portée par des unités inférieures à la syllabe.

9.2. Etude des poids relatifs des dimensions prosodiques

9.2.1 Introduction

Nous souhaitons identifier la part de chaque dimension prosodique dans la morphologie vocale des émotions, et donc dans la perception de leurs expressions. Pour cela, il est nécessaire de pouvoir évaluer les émotions perçues par des auditeurs naïfs à partir de stimuli dans lesquels les paramètres prosodiques varient séparément, toutes choses égales par ailleurs. Toutefois, étant donné que la variation de ces paramètres résulte d'un contrôle global du conduit vocal et non de contrôles séparés, il paraît impossible de collecter de tels stimuli produits naturellement, quand bien même ces expressions émotionnelles seraient produites par des acteurs entraînés spécifiquement pour cette tâche. Une démarche semblable à celle adoptée par Ekman & Friesen (1978) pour l'étude des paramètres musculaires des expressions faciales des émotions, dans laquelle les auteurs ont demandé à des acteurs entraînés d'actionner séparément chacun des muscles ou groupes de muscles responsables de ces expressions, semble en conséquence exclue.

En revanche, les avancées récentes de la synthèse vocale permettent de mettre en œuvre de façon de plus en plus fine une autre approche, qui est celle de l'analyse par synthèse. Le principe général de cette méthode est de synthétiser des stimuli en leur appliquant tout ou partie des variations des paramètres déterminées par l'analyse acoustique de

stimuli de référence, avant d'évaluer la pertinence perceptive de ces paramètres au travers des stimuli resynthétisés.

Dans le cadre de l'étude des expressions émotionnelles dans la parole, certaines études se sont ainsi appuyées sur des stimuli resynthétisés à partir de variations prosodiques extraites de stimuli naturels. Le but était d'évaluer la pertinence de ces mesures multiparamétriques pour la perception des expressions émotionnelles. L'expérience pionnière en ce domaine est celle menée par Lieberman & Michaels (1962), qui ont extrait les variations de fréquence fondamentale et d'intensité d'énoncés produits par trois locuteurs simulant un ensemble d'états émotionnels, et les ont appliquées séparément et de façon conjointe à des voyelles tenues synthétisées. Les résultats de l'évaluation perceptive de ces voyelles modifiées ont montré que la fréquence fondamentale permettait globalement d'identifier les affects exprimés bien qu'à un niveau de reconnaissance nettement inférieur à celui obtenu par les énoncés naturels. De plus l'ajout de l'information d'intensité à la fréquence fondamentale améliorait légèrement cette reconnaissance, tandis que le lissage des contours de fréquence fondamentale la faisait baisser. Enfin, l'utilisation des seules variations d'intensité ne permettait pas de reconnaitre les affects exprimés.

Bien que directement orienté vers l'application en synthèse vocale expressive plutôt qu'à visée théorique, le système HAMLET développé par Murray & Arnott (1995), qui applique à des énoncés neutres des règles de transformation de la fréquence fondamentale, de la qualité de voix et du débit de parole définies à partir d'une revue de littérature, s'inscrit également dans ce cadre. Bänziger *et al.* (2003) ont quant à eux appliqué à des énoncés produits par un système de synthèse à partir du texte des contours de fréquence fondamentale analysés à partir d'expressions actées de tristesse, désespoir, joie, exaltation, anxiété, panique, irritation et rage extraites du corpus de Munich dont une partie a été exploitée par Banse & Scherer (1996). Ils concluent de l'évaluation perceptive de ces stimuli synthétiques que la manipulation des contours de fréquence fondamentale est insuffisante pour produire des expressions émotionnelles synthétiques.

Une autre approche de l'analyse par synthèse est d'appliquer une variation systématique des paramètres prosodiques aux stimuli synthétisés avant qu'ils ne soient évalués perceptivement. C'est par exemple la démarche adoptée par Ladd *et al.* (1985), qui ont manipulé des stimuli naturels pour faire varier la qualité de voix (modale ou *harsh*), la forme générale des contours de fréquence fondamentale (montant ou descendant) et l'étendue de la fréquence fondamentale, concluant à un effet net de l'étendue de F0 sur l'activation perçue par les auditeurs, l'influence de la qualité de voix et de la forme du

contour étant moins évidente. Gobl & Ní Chasaide (2003a) ont quant à eux évalué le rôle de la qualité de voix dans les expressions émotionnelles à partir de stimuli resynthétisés avec une onde de débit glottique modifiée par rapport à celle d'un stimulus de référence modal pour correspondre aux différents modes de phonation décrits en section 7 L'évaluation perceptive des émotions ou attitudes évoquées chez des juges naïfs par ces stimuli les a conduits à conclure qu'à l'exception de la colère transmise avec une relative efficacité par la qualité de voix tendue (*tense voice*), la seule variation de la qualité de voix, ici assimilée à la source glottique, n'est à même de transmettre que des états affectifs de faible intensité. Suivant une approche plus semblable à celle que nous nous proposons d'appliquer ici, les mêmes stimuli ont été utilisés par Yanushevskaya *et al.* (2005) en appliquant les contours de F0 proposés par Mozziconacci (1998). Ces contours sont considérés comme étant prototypiques d'un ensemble d'émotions et d'attitudes afin de confronter les influences relatives de ces dimensions sur la perception des affects véhiculés par les expressions synthétisées, la qualité de voix modale et les contours de fréquence fondamentale associés à l'expression neutre étant considérés comme référence. Les auteurs ont conclu de l'évaluation perceptive de ces stimuli que la qualité de voix permettait d'expliquer une part plus importante des attributions émotionnelles.

Bulut *et al.* (2002) ont quant à eux utilisé un système de synthèse par concaténation dans lequel la base de diphones était extraite d'expressions de colère, de joie et de tristesse ainsi que d'expressions neutres produites par des acteurs anglophones, et les règles de modification de la fréquence fondamentale et de l'intensité déduites de l'analyse des mêmes expressions. L'évaluation perceptive de stimuli générés en appliquant des modifications de fréquence fondamentale et d'intensité non congruentes avec les jeux de diphones utilisés a conduit les auteurs à conclure que les expressions de colère étaient majoritairement véhiculées par la qualité de voix portée par les diphones, la tristesse par les variations de fréquence fondamentale et d'intensité, tandis qu'aucune conclusion claire ne pouvait être donnée concernant les expressions de joie.

Dans l'optique de l'évaluation des poids relatifs des différentes dimensions prosodiques pour l'expression d'états émotionnels, plutôt que d'adopter une méthode telle que celle de Bulut *et al.* (2002) dans laquelle l'information portée par les différentes dimensions est mise en opposition au sein de stimuli émotionnellement incongruents afin de déterminer pour chaque valeur d'affect une dimension « gagnante », nous opterons pour une méthode dans laquelle la part d'information affective portée par chacune des dimensions est comparée à une valeur de référence. Ainsi, ces poids relatifs pourront être évalués à partir de stimuli synthétisés dans lesquels seules les variations d'une partie des paramètres prosodiques sont issues de l'analyse d'un stimulus exprimant un état émotionnel donné, les autres étant extraites d'une expression « neutre » sur le même

énoncé. Nous considérons en effet que cette expression « neutre » correspond à l'absence d'expression d'un état émotionnel par le locuteur, ce qui a été intuitivement transcrit par l'étiquette « rien » par le locuteur M2 que nous avons sélectionné dans sa tâche d'auto-annotation de la partie spontanée du corpus (cf. section 8.1.). La méthode que nous retenons consiste donc en une projection partielle des variations prosodiques extraites des stimuli porteurs d'expressions émotionnelles sur ces stimuli neutres de référence. Nous rejoignons en cela l'approche retenue par Gendrot (2004), qui a évalué le rôle joué par la qualité de voix dans la perception d'expressions actées de colère, tristesse, joie et surprise en appliquant en resynthèse les variations de fréquence fondamentale, d'intensité et de durée extraites d'expressions neutres.

Nous réalisons ainsi une partition de l'information prosodique, en partant du postulat que les dimensions prosodiques gelées en attribuant la valeur de l'expression neutre ne sont pas porteuses d'information affective. Cette méthode implique néanmoins le risque de générer des artefacts : le gel d'une partie des dimensions prosodiques produit un son de parole qui constitue une possible chimère, ne correspondant pas à une expression écologique et susceptible de par son incongruité de perturber le traitement cognitif de l'information et le décodage de l'émotion exprimée. Il est en effet bien établi que certains aspects de la qualité de voix, et notamment le quotient ouvert (Fant, 1997) ou d'autres mesures supposées capturer la dimension tendu-relâché comme la différence entre les amplitudes des deux premiers harmoniques (Swerts & Veldhuis, 2004) covarient avec la fréquence fondamentale. De plus la fréquence fondamentale est connue pour être perçue différemment en fonction de l'intensité (voir par exemple Houtsma (1995)), et bien que les corrélats acoustiques de la composante de la qualité de voix qu'est l'effort vocal soient encore mal connus (d'Alessandro, 2006), il semble raisonnable de supposer que ce dernier covarie avec l'intensité.

Une autre possibilité est que de ces combinaisons de dimensions gelées avec d'autres présentant des variations induites par l'expression d'une émotion, résultent des objets reconnus comme expressions émotionnelles existantes mais correspondant à une émotion autre que celle attendue. Ceci est à rapprocher des résultats obtenus par Rilliard (2000), qui a montré qu'un énoncé sur lequel était transplanté un contour prosodique incongru pouvait être perçu comme réalisant une fonction autre que celle visée.

9.2.2 Sélection de stimuli de référence

Notre objectif dans l'étude présente est de faire varier systématiquement les dimensions prosodiques issues de stimuli expressifs dont l'information est incorporée dans les stimuli synthétisés en prenant en compte les combinaisons de ces dimensions. La

combinatoire pour chacun des stimuli naturels ainsi évalués peut être importante quand bien même toutes les combinaisons possibles ne seraient pas retenues. Or, les stimuli synthétiques générés sont destinés à être évalués perceptivement par des sujets naïfs, ce qui impose de restreindre le nombre de stimuli de référence afin que le nombre de stimuli synthétiques à évaluer reste raisonnable. Toutes les expressions actées du locuteur sélectionné, produites sur des énoncés monosyllabiques, n'étant pas de qualité équivalente, nous avons choisi pour chacun des états émotionnels retenus un exemplaire que nous avons jugé suffisamment représentatif de l'expression considérée en condition audio seul. En raison de la nécessité de disposer pour chacun des stimuli expressifs retenus d'un stimulus de référence correspondant à l'expression neutre sur le même énoncé, nous avons jugé préférable de ne pas prendre en compte dans notre sélection les expressions sur l'énoncé [bʁik]. En effet, la voyelle [i] dans l'expression neutre produite sur cet énoncé par notre locuteur étant en partie dévoisée, le risque que la projection du contour de fréquence fondamentale correspondant soit peu fidèle à l'original aurait été plus élevé.

Comme ces critères de sélection ne nous permettent pas de retenir des stimuli représentatifs des 8 états émotionnels retenus sur un même énoncé, nous avons finalement sélectionné une combinaison d'expressions sur les énoncés [ʁuʒ] et [sabl], ce que nous avons considéré comme le meilleur compromis possible. En effet la durée moyenne de la voyelle [a] dans l'ensemble du corpus du locuteur Y. (166 ms) est la plus proche de celle de la voyelle [u] (191 ms). Ainsi, les expressions d'anxiété, d'inquiétude et de satisfaction sur l'énoncé [ʁuʒ] ont été retenues, complétées par les expressions de déception, résignation, dégoût et tristesse sur l'énoncé [sabl].

Bien que correspondant à un sous-ensemble des expressions émotionnelles produites par un même locuteur, les stimuli ainsi sélectionnés ne correspondent pas directement à ceux dont les valeurs d'affect ont été précédemment évaluées à partir du signal audio seul et audiovisuel (section 8.4.), à l'exception de l'expression de joie. Il est donc particulièrement important d'inclure dans l'évaluation perceptive une condition de contrôle correspondant aux stimuli naturels, à laquelle les stimuli générés en n'utilisant qu'une partie de l'information prosodique extraite des stimuli porteurs d'expressions émotionnelles pourront être comparés.

9.2.3 Propriétés acoustiques des stimuli retenus

9.2.3.1 Fréquence fondamentale

Nous avons tout d'abord extrait les valeurs de fréquence fondamentale pour l'ensemble des portions voisées de chaque stimulus avec Praat (Boersma &Weenink, 1992-2008) de façon semi-automatique. Pour cela, nous avons adapté un script développé par Yi Xu (*_TimeNormalizeF0*[26] version 2.5.1), qui permet de corriger manuellement les marques de pulsations glottiques détectées automatiquement par Praat avant d'extraire les contours de fréquence fondamentale bruts et lissés, que nous avons adapté pour prendre en compte l'étiquetage des frontières phonémiques réalisé auparavant et ne pas extrapoler aux portions non voisées les valeurs initiale et finale du contour. Les valeurs extraites ont été converties en demi-tons, le niveau de référence étant le niveau moyen de fréquence fondamentale du locuteur Y. dans l'ensemble du corpus (98,6 Hz), selon la formule suivante :

$F0_{demi-tons} = \frac{12}{\ln(2)} \cdot \ln(\frac{F0_{Hz}}{F0ref_{Hz}})$, où F0 est la fréquence fondamentale et F0ref la valeur de référence, toutes deux exprimées en hertz. Les contours de fréquence fondamentale des voyelles des huit stimuli expressifs retenus ainsi que des deux expressions neutres sont représentés Figure 32.

Nous avons également extrait les caractéristiques statiques générales de la fréquence fondamentale des voyelles de ces stimuli. Ces contours sont ici décrits par le niveau moyen, l'attaque, la déclinaison (différence entre la finale et l'attaque) et la dynamique (différence entre les valeurs maximale et minimale de F0). En raison du lien direct entre ces paramètres et d'autres mesures plus classiques des caractéristiques générales de la distribution que sont l'écart-type et un certain nombre de quantiles dont les plus fréquemment utilisés sont la médiane et les 25ème et 75ème centiles (voir par exemple Banse & Scherer (1996)), nous n'avons pas retenu ces derniers paramètres.

[26] La version la plus récente du script original *_TimeNormalizeF0* peut être téléchargée depuis la page Web de Yi Xu : http://www.phon.ucl.ac.uk/home/yi/downloads.html

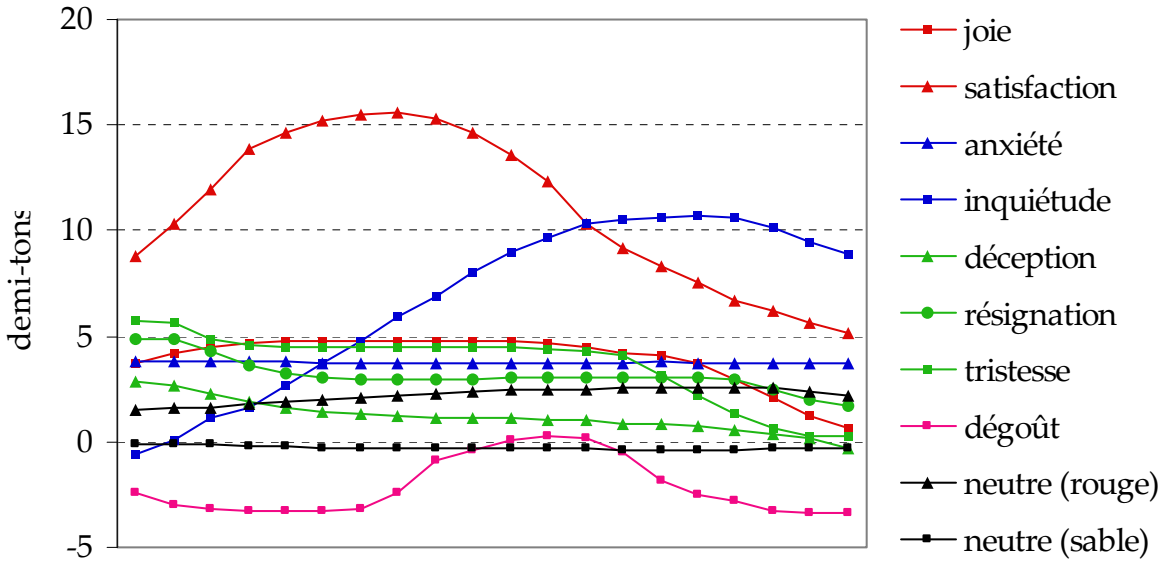


Figure 32 : Contours de fréquence fondamentale lissés et normalisés sur 20 points des voyelles des stimuli sélectionnés comme référence pour les expériences de projection dimensionnelle.

9.2.3.2 Durée

Nous avons normalisé la durée de la voyelle de chaque stimulus, extraite de l'étiquetage phonétique, par rapport à la durée de la même voyelle dans l'expression neutre correspondante considérée comme référence, afin de l'exprimer sous forme de proportion de cette durée de référence. En outre la durée totale du stimulus, ainsi que la proportion de la durée de la voyelle dans la durée totale ont également été calculées. Dans le cas des stimuli comprenant un schwa final, la durée de ce schwa également été extraite afin de calculerla proportion de la durée du schwa dans la durée totale du stimulus.

9.2.3.3 Intensité

Bien que nous ayons observé de fréquents mouvements de buste de la part des locuteurs piégés par Sound Teacher, induisant une variation de la distance de la bouche du locuteur au microphone, le locuteur M2 dont nous avons sélectionné les productions semble, d'après les enregistrements vidéo, avoir peu fait varier cette distance dans ses productions actées. L'intensité semble donc pouvoir fournir des informations exploitables.

Nous avons extrait à l'aide de Praat les valeurs moyennes ainsi que l'écart-type de l'intensité exprimées en décibels, selon les formules :

$\mu(E)_{dB} = 10.\log(\frac{\int_{t1}^{t2} 10^{\frac{x(t)}{10}}.dt}{t2-t1})$, et $\sigma(E)_{dB} = \sqrt{\frac{\int_{t1}^{t2} (10^{\frac{x(t)}{10}} - \mu_{dB}).dt}{t2-t1}}$, où t1 est l'indice de la trame la plus proche de l'instant correspondant au début de la portion de signal analysée, t2 la trame correspondant à la fin de cette portion, et x(t) l'intensité en dB à la tramet t.

Pour chaque stimulus considéré, ces valeurs ont été extraites pour l'ensemble de l'énoncé ainsi que pour la portion de signal restreinte à la voyelle. De plus, dans les cas où un schwa final a été ajouté, son énergie moyenne a été extraite afin de calculer la différence d'intensité entre la voyelle et le schwa, exprimée en dB.

9.2.3.4 Qualité de voix

Comme nous l'avons exposé en section 7, le lien entre la plupart des mesures acoustiques supposées liées à la qualité de voix et les régimes vibratoires de la glotte n'est pas clairement établi, pas plus que ne l'est celui entre ces mesures et des émotions ou familles d'émotions particulières. Néanmoins certaines de ces mesures permettent de capturer une partie des variations de qualité de voix perçues. N'étant pas en mesure d'extraire une estimation de NAQ adaptée à des mesures fines, nous avons choisi de restreindre l'extraction de paramètres acoustiques que nous avons réalisée aux deux paramètres spectraux fréquemment utilisés dans la littérature et dont le lien avec des phénomènes articulatoires et perceptifs liés à la qualité de voix est le plus clairement établi.

Nous avons donc choisi de limiter les mesures spectrales de qualité de voix à la pente spectrale au dessus de 1kHz, supposée rendre compte de la vitesse de fermeture de la glotte, et à la différence entre les amplitudes des deux premiers harmoniques supposée rendre compte à coefficient d'asymétrie constant du continuum voix pressée-voix relâchée (Doval *et al.*, 2006 ; d'Alessandro, 2006), corrigée de l'influence des fréquences et des largeurs de bande des formants (Hanson, 1995 ; Iseli & Alwan, 2004).

La pente spectrale au dessus de 1kHz (que nous noterons par la suite DO1000 selon la terminologie adoptée par Banse & Scherer (1996)) a été extraite au moyen de la fonction de Praat prévue à cet effet. Les mesures spectrales nécessaires au calcul de H1*-H2* ont été calculées sur une portion de 80% de la voyelle, centrée sur le milieu de cette voyelle tel que défini par l'étiquetage des frontières phonémiques. Nous avons extrait la différence entre les amplitudes des deux premiers harmoniques (H2-H1) ainsi que les trois premiers formants et leur largeur de bande à l'aide d'un script Praat, en nous fondant sur la valeur moyenne de F0 étiquetée manuellement dans la région considérée pour déterminer l'emplacement des deux premières harmoniques. L'extraction des

formants a été réalisée de façon semi-automatique, en considérant comme valeurs de référence celles indiquées par Tubach (1989) pour les voyelles du français dans le cas d'un locuteur masculin afin de guider l'extraction. Les valeurs extraites sont validées par inspection visuelle du spectrogramme. La pente spectrale corrigée de l'influence des formants H1*-H2* a été calculée à partir des valeurs de H1 et H2 mesurées, selon l'équation de Iseli & Alwan (2004) limitée aux deux premiers formants. Comme nous l'avons souligné en section 7, une fréquence fondamentale proche de la valeur du premier formant peut se révéler problématique pour l'extraction de la largeur de bande B1 du premier formant. Cependant, la fréquence fondamentale moyenne du locuteur que nous avons sélectionné (96.8 Hz) est suffisamment basse pour que ce problème ne se pose pas de façon cruciale.

Nous avons également extrait avec Praat à partir de l'étiquetage rectifié manuellement des pulsations deux mesures de la perturbation de la fréquence fondamentale supposées caractériser en partie la phonation craquée. Le taux de jitter a ainsi été extrait, calculé comme la différence moyenne entre les durées de deux périodes consécutives divisée par la durée moyenne d'une période, ce qui correspond à la définition du jitter considérée par Fourcin & Abberton (2008) comme la plus commune. Le shimmer a été extrait à l'aide de Praat de façon similaire, comme la différence moyenne entre les amplitudes de deux périodes consécutives divisée par l'amplitude moyenne. En raison de la variabilité inter-voyelle de ces mesures de perturbations (Atif Kiliç *et al.*, 2004), nous avons également normalisé ces mesures en les exprimant comme une proportion de la même mesure calculée sur l'expression neutre correspondant au même énoncé.

Nous avons enfin extrait le ratio harmonicité sur bruit (HNR) par autocorrélation, selon la méthode de Boersma (1993) implémentée dans Praat, sur l'énoncé entier et la portion centrale de la voyelle. Afin de compenser les potentielles variations d'harmonicité intrinsèques au contenu phonétique de l'énoncé, nous avons normalisé ces valeurs de HNR en leur retranchant la valeur de HNR mesurée sur l'expression neutre correspondante, considérée comme référence.

Les valeurs les plus représentatives des paramètres acoustiques extraits des stimuli sélectionnés, normalisées par rapport au stimulus neutre correspondant selon les principes exposés ci-dessus, sont récapitulées dans la Table 11.

Emotion	Enoncé	Durée	F0 (demi-tons)				Int.	Qualité de voix			
			moy.	att.	décl.	dyn.		Jitt.	Shim.	H2*-H1*	HNR
joie	[sabl]	-30%	3.6	5.7	-5.4	5.4	-0.5	200%	271%	-4.9	-0.2
satisfaction	[ʁuʒ]	-10%	6.7	-0.6	9.5	11.3	-8.5	60%	175%	-5.3	-0.6
anxiété	[ʁuʒ]	-10%	3.9	3.7	-3.1	4.2	5.1	80%	225%	-0.9	-3.1
inquiétude	[ʁuʒ]	24%	11	8.8	-3.7	10.4	8.2	60%	100%	-3.7	2.5
déception	[sabl]	32%	1.2	2.9	-3.2	3.2	-1.7	25%	114%	-6.5	7.7
résignation	[sabl]	-17%	2.2	1.6	0.6	1	-2.7	100%	71%	-16.4	12.8
tristesse	[sabl]	-33%	3.6	-0.3	-0.1	-0.2	-2.9	75%	143%	-13.4	2
dégoût	[sabl]	136%	-2.1	3.2	4.9	-3.2	3.1	100%	157%	-2.2	-0.2

Table 11 : Principales caractéristiques acoustiques des voyelles des stimuli expressifs sélectionnés parmi les productions du locuteur M2. Les énoncés sur lesquels ces expressions sont produites sont indiqués dans la seconde colonne. Les durées et mesures relatives à la qualité de voix sont normalisées par rapport à l'expression neutre sur le même énoncé. Les mesures d'énergie sont exprimées sous forme de différence en dB, les autres mesures en proportion relaitve. *Int.* : Intensité acoustique ; *Jitt.* : Jitter ; *Shim.* : Shimmer.

9.2.4 Projection de contours de fréquence fondamentale et d'intensité

9.2.4.1 Choix des combinaisons de dimensions traitées

Dans un premier temps, nous avons généré les stimuli synthétiques à l'aide du module de manipulation du logiciel Praat (Boersma &Weenink, 1992-2008). Ce module permet de modifier les contours de fréquence fondamentale ainsi que la durée de tout ou partie d'énoncés naturels en s'appuyant sur l'algorithme TD-PSOLA, ainsi que de modifier leurs contours d'intensité avant de générer les fichiers sons ainsi modifiés. En revanche il ne permet pas d'appliquer en synthèse des modifications de qualité de voix. De ce fait, les variations de qualité de voix ne peuvent être prises en compte directement comme l'ont fait par exemple Gobl & Ní Chasaide (2003). Il reste toutefois possible d'évaluer l'influence de la qualité de voix sur la perception des expressions émotionnelles « par élimination » : en appliquant à un stimulus représentatif d'un état émotionnel donné les contours de fréquence fondamentale et d'intensité ainsi que les variations de durée du stimulus de référence correspondant (en l'occurrence l'expression neutre produite sur le même énoncé), on peut obtenir, nonobstant d'éventuels artefacts d'analyse/resynthèse, un stimulus resynthétisé dans lequel seule la qualité de voix est caractéristique de l'expression de l'émotion considérée.

La manipulation de la durée est cependant plus problématique que celle des contours de fréquence fondamentale et d'intensité. En effet, au contraire de ces deux paramètres, la durée ne constitue pas une quantité qui varie au cours du temps, mais l'ensemble des phénomènes temporels de la prosodie, pour l'analyse desquels un modèle de durée (ne serait-ce qu'au niveau du choix des unités minimales à prendre en compte) est indispensable. En conséquence, et bien que l'algorithme TD-PSOLA implanté dans le module de modification de Praat propose un outil pour allonger ou rétrécir certaines portions d'un signal de parole (au prix de l'apparition fréquente d'artefacts de synthèse très perceptibles dans le cas de modifications importantes de durée), il nous est impossible de manipuler la durée de façon fiable dans nos stimuli.

Nous avons donc exclu de notre analyse les phénomènes de durée. Etant donné que la méthode d'analyse/synthèse que nous avons adopté ne permet de manipuler directement ni la qualité de voix ni la durée, nous ne pouvons analyser que les variations conjointes de la qualité de voix et de la durée, sans pouvoir étudier leurs influences relatives.

Ainsi, nous avons choisi de générer pour chacun des stimuli originaux porteurs d'une expression émotionnelle, 5 stimuli distincts correspondants chacun à une condition de resynthèse :

- Un stimulus de contrôle étiqueté *resynthèse complète*, construit en appliquant à lui-même les contours stylisés de F0 et d'intensité du stimulus source et destiné à évaluer d'éventuels artefacts dus à la stylisation de ces contours ou au processus de resynthèse lui-même.
- Un stimulus étiqueté *F0 seule*, construit en appliquant le contour stylisé de F0 du stimulus source au stimulus neutre correspondant au même énoncé.
- Un stimulus étiqueté *intensité seule* obtenu en appliquant le contour d'intensité du stimulus source à l'expression neutre correspondante.
- Un stimulus étiqueté *F0 et intensité* construit en appliquant les contours de F0 et d'intensité à l'expression neutre correspondante.
- Un stimulus étiqueté *qualité de voix et durée*. Cette dernière condition a été obtenue en appliquant les contours de F0 et d'intensité de l'expression neutre au stimulus source. Ainsi seuls les phénomènes de durée et de qualité de voix du stimulus source subsistent, tandis que ses variations spécifiques de F0 et d'intensité sont neutralisées.

En complément des 40 stimuli ainsi générés à partir des 8 expressions émotionnelles sélectionnées, nous avons choisi de générer également un stimulus en condition

resynthèse complète pour chacune des 2 expressions neutres, pour un total de 42 stimuli synthétiques.

9.2.4.2 Description de la méthode de projection

Les stimuli synthétiques ont été générés à l'aide de Praat (Boersma & Weenink, 2008), au moyen d'un script que nous avons développé pour rendre ce processus semi-automatique. Pour chacune des paires de stimuli traités (un stimulus source porteur d'une expression émotionnelle et un stimulus cible *neutre*), nous avons dans un premier temps appliqué le contour de fréquence fondamentale du stimulus source à la cible. A la suite de la copie automatique du contour de F0 de la source à la place de celui de la cible, avec un changement d'échelle fonction des durées relatives des phonèmes de chacun des stimuli, nous avons ajusté lorsque nous l'avons jugé nécessaire d'après une inspection visuelle des contours source et cible la répartition des points du contour ainsi copié. Pour cela le contour copié à été stylisé à l'aide de la fonction de Praat prévue à cet effet, tout en s'assurant que cette stylisation ne neutralise pas de points saillants du contour.

Le principe de l'algorithme itératif de stylisation des contours de Praat est de considérer les ensembles de trois points consécutifs en partant du début du contour, et de supprimer le point central si son écart au segment reliant le premier au troisième point est supérieur à la résolution spécifiée comme argument. L'opération est répétée en conservant le premier point du contour comme point initial de la série de trois points considérée, jusqu'à ce que le point central soit conservé. Ce point central devient alors le premier point de la série considérée, jusqu'à ce que l'ensemble du contour ait été traité. La résolution par défaut est de 2 demi-tons. Nous avons opté pour une résolution plus conservative d'un quart de ton, inférieure au seuil de perception du glissando (établi par Rossi (1971) pour les glissandos montants à 19 Hz pour une fréquence fondamentale de référence de 135 Hz, soit une variation de 2,3 demi-tons, et confirmé par Rossi (1978) pour les glissandos descendants), afin de ne pas éliminer de points saillants du contour. L'exemple de l'application du contour de fréquence de l'expression de tristesse au stimulus neutre correspondant est illustré par la Figure 33.

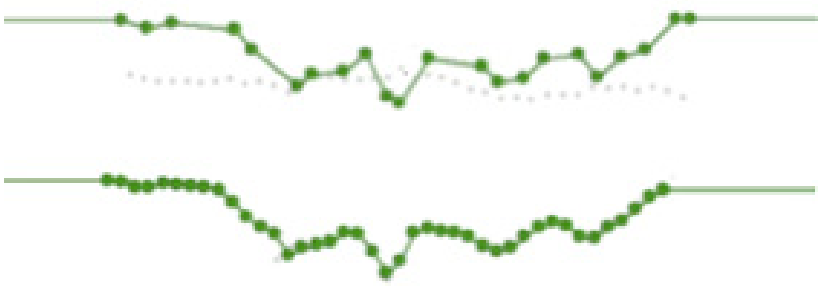


Figure 33 : Application du contour de F0 stylisé avec une résolution d'un quart de ton de l'expression de tristesse au stimulus neutre (haut) et contour de F0 résultant après resynthèse (bas). Le contour original de F0 de l'expression neutre apparaît en gris dans la partie supérieure de la figure. Les stimuli originaux utilisés correspondent à des productions du locuteur M2 sur l'énoncé [sabl].

La transplantation des contours d'intensité a été réalisée de manière similaire, le contour d'intensité du stimulus source étant appliqué au stimulus cible après un changement d'échelle fonction des durées phonémiques relatives, avant correction éventuelle de la répartition des points d'intensité. Néanmoins, Praat ne proposant pas pour les contours d'intensité de fonctionnalité de stylisation similaire à celle utilisée pour traiter les contours de fréquence fondamentale, la stylisation des contours d'intensité a été réalisée manuellement en supprimant les points paraissant les moins pertinents pour la paramétrisation du contour. Cette transplantation n'est pas suffisante car Praat ne traite que des contours d'intensité relatifs : la méthode décrite ci-dessus permet ainsi de modifier la forme globale du contour, mais pas le niveau moyen d'intensité qui participe également de la perception de l'intensité. Le signal acoustique de chaque stimulus généré a donc été rééchelonné afin que l'intensité moyenne en décibels des stimuli source et cible soient égales, en multipliant l'ensemble de ses échantillons par un facteur α défini comme $\alpha = 10^{(IC-IS)/20}$, où IC est l'intensité moyenne exprimée en dB du stimulus cible et IS celle du stimulus source.

La Figure 34 illustre l'exemple du contour d'intensité de l'expression d'inquiétude appliqué à l'expression neutre correspondante.

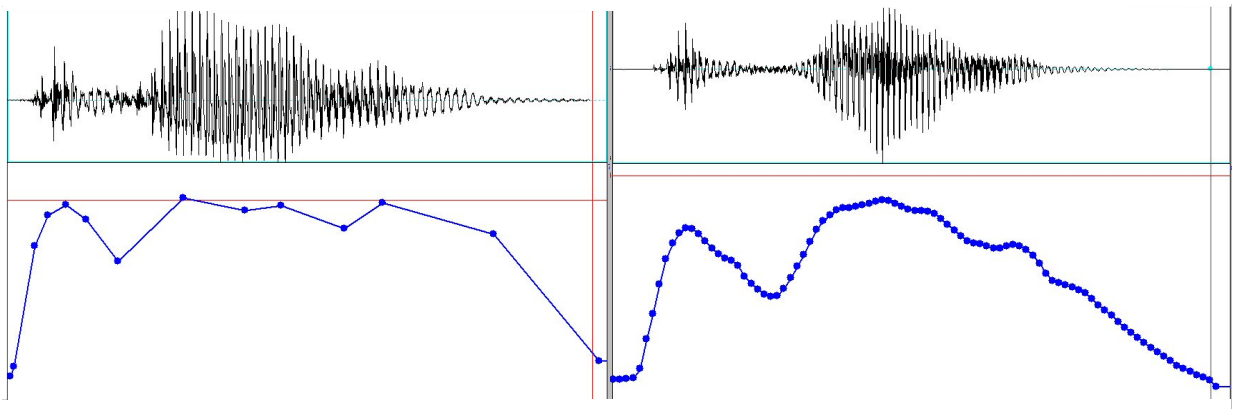


Figure 34 : Stylisation du contour d'intensité de l'inquiétude (gauche) et contour d'intensité résultant après application à l'expression neutre et resynthèse (droite). Le stimulus original utilisé a été produit par le locuteur M2 sur l'énoncé [ʁuʒ].

Dans le cas de la condition de synthèse *qualité de voix et durée*, les stimuli source et cible ont été intervertis, les contours de fréquence fondamentale et d'intensité de l'expression neutre étant appliqués au stimulus expressif, afin de neutraliser les variations de fréquence fondamentale et de durée de ce stimulus et n'en conserver que les caractéristiques de qualité de voix et de durée.

Dans le cas de la condition resynthèse complète, les modifications appliquées aux contours de fréquence fondamentale et d'intensité avant resynthèse se sont limitées à la stylisation de ces contours. Une écoute attentive ne nous pas permis de déceler de différences perceptibles entre les stimuli naturels d'origine et ceux générés dans cette condition. Néanmoins nous avons pris le parti d'inclure dans l'évaluation perceptive ces stimuli générés en condition de contrôle plutôt que les stimuli naturels d'origine, afin que d'éventuels artefacts de synthèse liés à la stylisation des contours et que nous n'aurions pas perçus soient pris en compte dans la comparaison des résultats de l'évaluation perceptive de cette condition de contrôle et des conditions de resynthèse partielle.

9.2.4.3 Evaluation perceptive des stimuli générés

La présentation des stimuli et l'enregistrement des réponses ont été réalisés à l'aide d'une interface dont la page de réponse est présentée Figure 35, que nous avons développée avec le logiciel *Revolution Runtime* sur le modèle de celle utilisée pour la première évaluation perceptive (cf. section 8.4.), mais de façon à ce que les sujets n'aient pas la possibilité de sélectionner plusieurs étiquettes pour indiquer leur réponse. Les sujets avaient pour instruction de sélectionner l'une des 8 étiquettes émotionnelles proposées (*Anxiété*, *Déception*, *Dégoût*, *Inquiétude*, *Joie*, *Résignation*, *Satisfaction* ou *Tristesse*) ou l'étiquette *Neutre* lorsque le stimulus était perçu comme n'exprimant aucune émotion.

De plus il leur était demandé de noter l'intensité émotionnelle perçue sur une échelle de 1 à 10 à l'aide du curseur choisi pour indiquer l'émotion perçue.Ils n'avaient la possibilité de valider leur réponse et de passer au stimulus suivant qu'après avoir sélectionné l'une des étiquettes proposées.

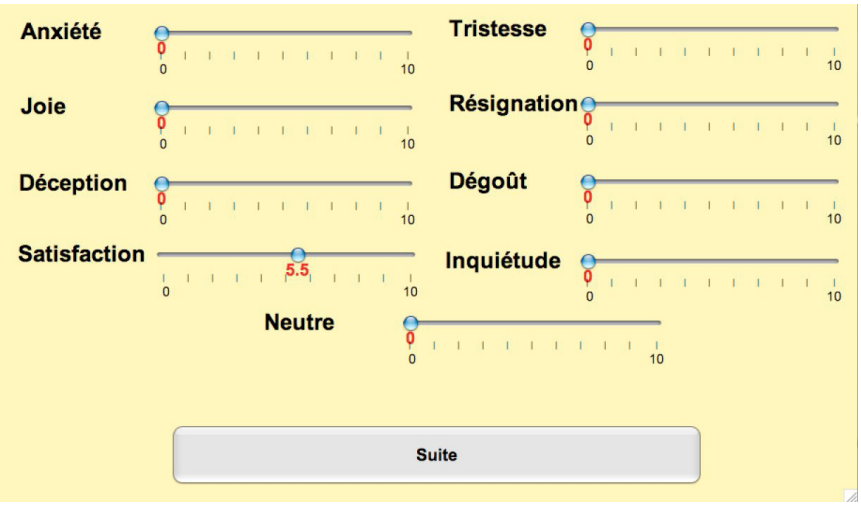


Figure 35 : Interface graphique utilisée pour l'évaluation perceptive des stimuli générés par projection dimensionnelle à l'aide du module TD-PSOLA de Praat de productions du locuteur M2.

De même que lors de la première évaluation perceptive, un curseur permettant de noter l'intensité perçue a été associé à l'étiquette *Neutre*. Néanmoins, ayant constaté lors de cette première évaluation que cela pouvait être source de perturbation pour les sujets, nous les avons informé que les intensités attribuées à cette étiquette ne seraient pas traitées et que la perception d'aucun expression émotionnelle pouvait en conséquence être indiquée par n'importe quelle valeur non nulle.

Les 42 stimuli générés ont été notés par 40 auditeurs de langue maternelle française sans problèmes d'audition connus (6 hommes et 34 femmes, d'âge moyen 23,3 ans) dans un environnement calme, les stimuli étant présentés au moyen d'un casque fermé de bonne qualité. Chaque stimulus a été présenté 3 fois à chaque auditeur, les stimuli étaient présentés dans un ordre aléatoire différent pour chacun, contrôlé pour éviter 2 présentations successives d'un même stimulus.

9.2.4.4 Matrices de confusion et regroupement d'étiquettes

Une matrice de confusion a été extraite des réponses données par les 40 auditeurs pour chacune des 5 conditions de resynthèse. La matrice de confusion correspondant à la condition de contrôle est présentée dans la Table 12. Les valeurs indiquées pour l'expression neutre correspondent à la moyenne des valeurs obtenues pour les expressions neutres sur les énoncés [ʁuʒ] et [sabl], présentées uniquement en condition de contrôle.

	joie	anxiété	résign.	tristesse	inquiét.	décep.	dégoût	satisf.	neutre
joie	**26%**	7%	6%	2%	4%	1%	0%	45%	10%
anxiété	0%	**58%**	11%	0%	20%	8%	4%	0%	0%
résignation	3%	15%	**30%**	0%	10%	13%	1%	13%	15%
tristesse	20%	17%	14%	**14%**	8%	16%	3%	7%	1%
inquiétude	15%	11%	1%	0%	**31%**	8%	9%	23%	3%
déception	0%	0%	39%	6%	0%	**35%**	19%	1%	0%
dégoût	1%	0%	2%	8%	1%	10%	**62%**	17%	0%
satisfaction	76%	0%	0%	0%	0%	0%	0%	**24%**	0%
neutre	1%	4%	19%	10%	9%	19%	5%	1%	**32%**

Table 12 : Matrice de confusion correspondant à la projection dimensionnelle à l'aide du module TD-PSOLA de Praat à partir de productions du locuteur M2, en condition de contrôle (utilisation de toutes les dimensions analysées). Les taux de reconnaissance sont indiqués en gras.

Les regroupements d'étiquettes définis à partir des principales confusions observées en condition audio lors de la première évaluation perceptive ont ensuite été appliqués à l'analyse des réponses recueillies afin de pouvoir effectuer des traitements statistiques sur les taux de reconnaissance en tenant compte de ces confusions et faire apparaître de façon plus claire les principales tendances. En effet, comme l'illustre le dendrogramme présenté Figure 36, qui résulte de la classification hiérarchique opérée sur les colonnes normalisées par la somme des éléments de la matrice de confusion en condition de contrôle, les confusions entre étiquettes correspondant aux stimuli sélectionnés sont très similaires à celles observées précédemment.

Ainsi, les réponses pourront être analysées en considérant les regroupements d'étiquettes *résignation*, *déception* et *tristesse*, ainsi que *joie* et *satisfaction* et *anxiété* et *inquiétude*. En revanche, les réponses attribuées aux étiquettes *neutre* et *dégoût* restent analysées de façon isolée.

La matrice de confusion en condition de contrôle après regroupement d'étiquettes est présentée dans la Table 13 avec le niveau théorique du hasard correspondant au nombre d'étiquettes regroupées. Les matrices de confusion brutes et après regroupement d'étiquettes pour chacune des conditions de resynthèse partielles sont présentées en Annexe III :.

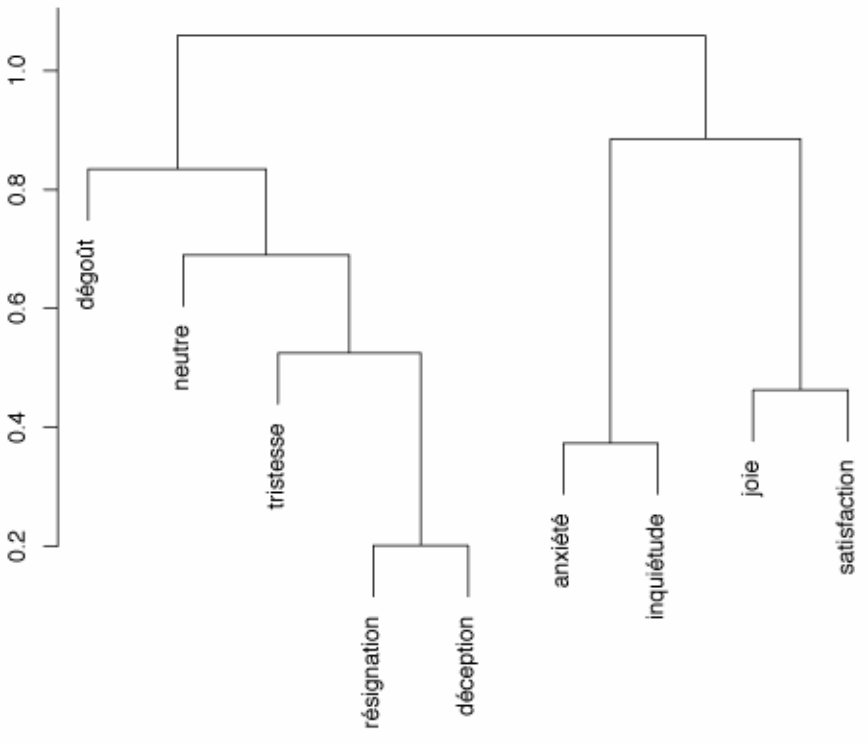


Figure 36 : Dendrogramme issu de la classification hiérarchique (critère d'agrégation de Ward) opérée sur les colonnes normalisées de la matrice de confusion correspondant à la projection dimensionnelle à l'aide du module TD-PSOLA de Praat à partir de productions du locuteur M2, en condition de contrôle (utilisation de toutes les dimensions analysées).

	joie/satisf.	rés./déc./trist.	anx./inq.	Dégoût	neutre
hasard	*22%*	*33%*	*22%*	*11%*	*11%*
joie	**71%**	8%	11%	0%	10%
satisfaction	**100%**	0%	0%	0%	0%
résignation	17%	**43%**	25%	1%	15%
tristesse	27%	**44%**	25%	3%	1%
déception	1%	**80%**	0%	19%	0%
anxiété	0%	18%	**78%**	4%	0%
inquiétude	38%	8%	**42%**	9%	3%
dégoût	18%	20%	1%	**62%**	0%
neutre	3%	48%	13%	5%	**32%**

Table 13 : Matrice de confusion correspondant à la projection dimensionnelle à l'aide du module TD-PSOLA de Praat à partir de productions du locuteur M2, en condition de contrôle (utilisation de toutes les dimensions analysées) après regroupement d'étiquettes. Les taux de reconnaissance sont indiqués en gras. Le niveau théorique du hasard, fonction du nombre d'étiquettes dans chaque regroupement, est indiqué sur la 1ère ligne.

9.2.4.5 Analyse statistique

En fonction des ces regroupements, chacune des réponses indiquées par les auditeurs a été convertie en réponse juste (1) ou fausse (0) en fonction de l'étiquette choisie : une réponse est ainsi considérée comme juste si l'une des étiquettes de la catégorie après regroupement a été sélectionnée. Chacun des stimuli générés ayant été présenté 3 fois, nous obtenons ainsi pour chaque émotion, chaque condition de resynthèse et chaque juge une valeur comprise entre 0 et 3.

L'intensité moyenne de l'émotion a également été extraite pour chaque réponse. Etant donné les instructions données aux sujets pour l'utilisation de l'étiquette *neutre*, nous n'avons pas pris en compte les intensités attribuées à cette étiquette : le choix par les auditeurs de l'étiquette *neutre* est donc considéré comme associé à une intensité d'émotion perçue nulle.

Des tests t pour échantillon unique indiquent que dans la condition de contrôle, les taux de reconnaissance après regroupement (diagonales de la matrice) des expressions de joie, satisfaction, déception, anxiété, inquiétude et dégoût sont significativement supérieurs au hasard, de même que ceux des expressions neutres ($p<0.001$ pour chaque comparaison à l'exception de l'inquiétude et de l'expression neutre sur l'énoncé [ʁuʒ], $p<0.01$). En revanche les expressions de résignation et de tristesse n'obtiennent pas des taux de reconnaissance supérieurs au hasard. Les résultats des tests t pour échantillon unique comparant le taux de reconnaissance de chaque émotion au niveau du hasard dans chaque condition de resynthèse sont présentés en Annexe III :.

Les taux moyens de reconnaissance obtenus dans les différentes conditions de resynthèse pour chacune des émotions présentées sont représentés Figure 37. Les intensités d'émotion perçues moyennes correspondantes sont représentées Figure 38.

Nous avons effectué une ANOVA à mesures répétées avec le nombre de réponses correctes comme variable dépendante et la condition de resynthèse et l'émotion présentée comme facteurs fixés. Afin de conserver un plan expérimental factoriel les résultats relatifs à l'évaluation des expressions neutres sur les énoncés [ʁuʒ] et [sabl], présentés uniquement en condition de contrôle, ont été exclus de l'analyse. Les résultats de cette ANOVA à mesures répétées indiquent un effet significatif de la condition ($F(3.8, 147.3)=137.5$; $p<0.001$), de l'émotion présentée ($F(6.6, 257.4)=35.5$; $p<0.001$), ainsi que de l'interaction émotion*condition ($F(22.6, 882.2)=21.8$; $p<0.001$). Des contrastes spécifiques indiquent que les émotions exprimées par les stimuli resynthétisés en condition de contrôle sont globalement mieux reconnues que dans toutes les autres conditions de resynthèse ($p<0.001$ pour tous les contrastes).

Une seconde ANOVA à mesures répétées avec les mêmes facteurs fixés et l'intensité d'émotion perçue comme variable dépendante montre également un effet significatif de la condition ($F(3.0, 116.4)=204.7$; $p<0.001$), de l'émotion présentée ($F(5.5, 213.4)=63.0$; $p<0.001$) et de l'interaction émotion*condition ($F(19.0, 740.2)=34.6$; $p<0.001$). Les stimuli resynthétisés en condition de contrôle sont également jugés avec une intensité d'émotion plus élevée que dans toutes les autres conditions de resynthèse ($p<0.001$ pour tous les contrastes).

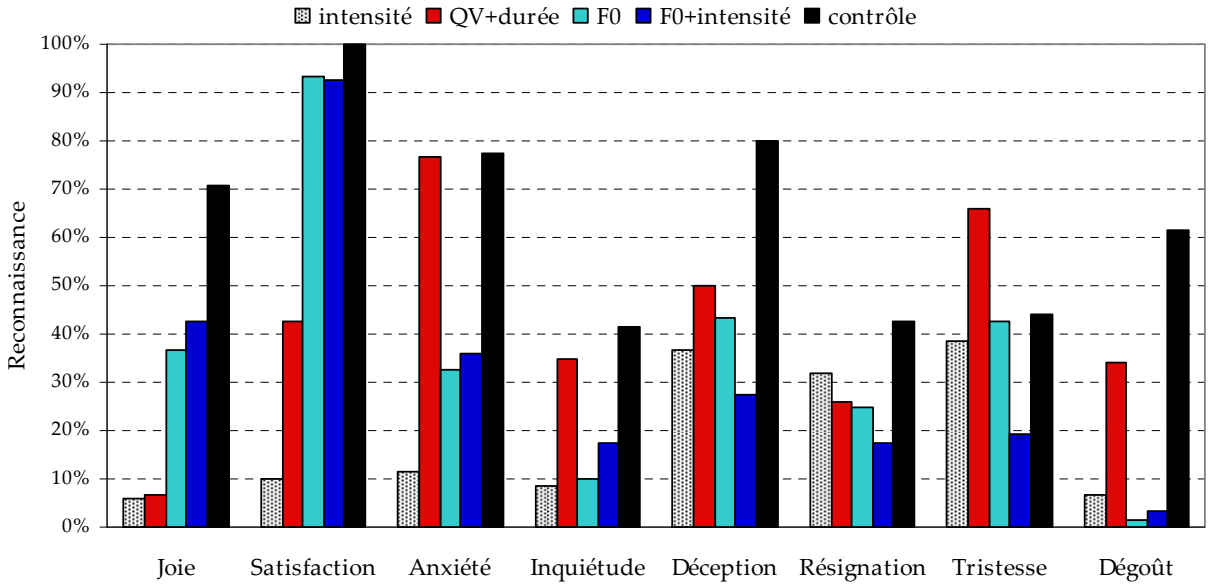


Figure 37 : Taux moyen de reconnaissance pour chaque émotion présentée dans chacune des conditions de resynthèse réalisée à l'aide du module TD-PSOLA de Praat à partir de productions du locuteur M2.

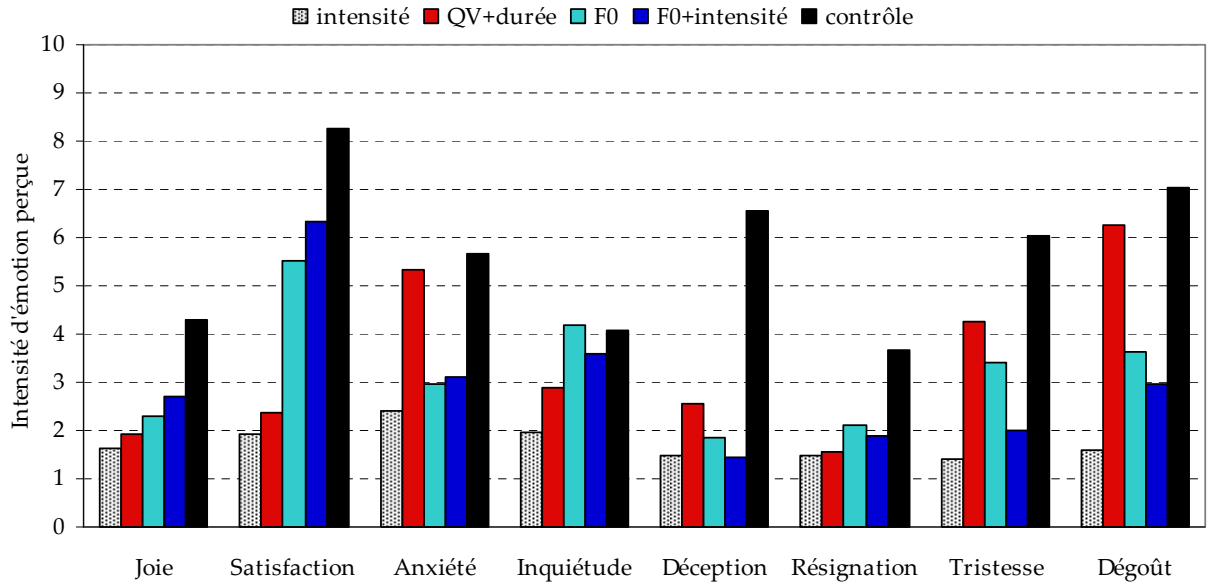


Figure 38 : Intensité moyenne de l'émotion perçue pour chaque émotion présentée dans chacune des conditions de resynthèse générées à l'aide du module TD-PSOLA de Praat à partir de productions du locuteur M2.

On peut également observer qu'à l'exception de l'expression de tristesse, aucune des émotions évaluées n'obtient de scores de reconnaissance aussi élevés dans l'une des conditions manipulées qu'en condition de contrôle. Cette différence n'est toutefois pas significative[27] pour les expressions d'anxiété et d'inquiétude resynthétisées en condition *qualité de voix et durée* (respectivement p=0.905 et p=0.305, n.s.), indiquant que les dimensions de qualité de voix et de durée présentées de façon conjointe transmettent une part suffisante de l'information affective de ces deux émotions proches l'une de l'autre pour permettre d'identifier la qualité de l'émotion correspondante. De plus les jugements d'intensité de l'émotion reçus par l'expression d'anxiété resynthétisée en condition *qualité de voix et durée* ne sont pas significativement plus faibles que ceux reçus par cette expression en condition de contrôle (p=0.407 n.s.). Ce résultat s'explique aisément par la comparaison des contours de fréquence fondamentale des expressions d'anxiété et d'inquiétude avec celui de l'expression neutre sur l'énoncé [ʁuʒ]. En effet, les stimuli générés en condition *qualité de voix et durée* l'ont été en remplaçant les contours de fréquence fondamentale et d'intensité originaux par celui de l'expression neutre correspondante. Or, le contour de fréquence fondamentale de l'expression d'anxiété a une forme très similaire à ceux des expressions neutres, n'en différant sur l'énoncé [ʁuʒ] que par un ancrage plus élevé (1.5 demi-tons d'écart en moyenne). Etant donné que les intensités de ces deux stimuli sont également proches et que la projection du contour d'intensité de l'expression ne semble pas produire d'artefact important, il n'est pas surprenant que l'expression d'anxiété projetée en condition *qualité de voix et durée* soit évaluée de façon similaire à l'expression originale d'anxiété.

En revanche, les jugements d'intensité de l'émotion reçus par l'expression d'inquiétude dans cette condition *qualité de voix et durée* sont plus faibles que ceux attribués en condition de contrôle ($p<0.01$), ce qui s'explique vraisemblablement par la projection du contour plat de fréquence fondamentale correspondant à l'expression neutre à la place du contour de F0 original de l'expression d'inquiétude, dont résulte un bruit additionnel qui rend le stimulus généré moins semblable à une production naturelle.

Bien que le pic de fréquence fondamentale de l'expression d'inquiétude soit décalé vers la fin de la voyelle comparativement au contour de fréquence fondamentale des expressions de joie et de satisfaction, la forme montante puis descendante du contour de

[27] Les niveaux de significativité des comparaisons par paire entre conditions de resynthèse ont été corrigés avec la procédure de Bonferroni modifiée de Benjamini & Hochberg (1995). Les valeurs de p indiquées dans les cas où l'hypothèse nulle ne peut être rejetée sont celles correspondant à un test t isolé.

F0 de cette expression d'inquiétude semble responsable du taux élevé de reports vers les étiquettes joie et satisfaction en condition de contrôle. En effet, les projections en conditions *F0 seule* et *F0 et intensité*, qui ne sont pas reconnues comme anxiété ou inquiétude à un niveau supérieur au hasard, sont très largement reconnues comme des expressions de joie ou de satisfaction (par 75% des auditeurs en condition *F0 seule*, et 68% en condition *F0 et intensité*).

Les expressions de satisfaction resynthétisées en conditions *F0 seule* et *F0 et intensité* ne sont pas significativement moins bien reconnues qu'en condition de contrôle (respectivement $p=0.131$ n.s. et $p=0.089$ n.s.). Néanmoins la resynthèse de cette expression de satisfaction limitée aux contours de fréquence fondamentale et le cas échéant d'intensité induit des jugements d'intensité de l'émotion plus faibles qu'en condition de contrôle ($p<0.001$ dans les deux cas). Cela indique que si le contour de fréquence fondamentale extrait du stimulus original est suffisant pour identifier l'émotion exprimée, il ne permet pas de transmettre l'ensemble de l'information affective. On peut remarquer que l'expression de satisfaction est la seule pour laquelle la transplantation des contours d'intensité en plus des contours de fréquence fondamentale induit une augmentation significative de l'intensité de l'émotion perçue ($p<0.05$). L'expression de joie en conditions *F0 seule* et *F0 et intensité* n'est pas aussi bien reconnue qu'en condition de contrôle ($p<0.001$ pour les deux comparaisons), et induit des jugements d'intensité d'émotion plus faibles ($p<0.001$ dans les deux cas). Cependant la transplantation du contour de fréquence fondamentale (et éventuellement de celui d'intensité, qui en première analyse ne semble pas ajouter d'information affective) est dans le cas de l'expression de joie la seule projection partielle permettant une reconnaissance supérieure au niveau du hasard.

L'expression de déception n'est reconnue à un niveau significativement supérieur au hasard qu'en condition *qualité de voix et durée*. Néanmoins elle semble portée dans des proportions comparables par la qualité de voix et la durée d'une part, et les contours de fréquence fondamentale d'autre part. La différence entre les taux de reconnaissance en condition *qualité de voix et durée* et en condition *F0* est en effet non significative ($p=0.357$ n.s.), les jugements d'intensité de l'émotion étant par ailleurs aussi faibles dans chacune de ces conditions ($p=0.048$, n.s. avec correction de Benjamini & Hochberg, 1995).

Les expressions de résignation et de tristesse n'étant pas reconnues à un niveau supérieur au hasard, il est plus difficile d'interpréter les taux de reconnaissance obtenus lorsque seule une partie de l'information est présentée. Cependant le taux de reconnaissance de l'expression de tristesse en condition *qualité de voix et durée*, plus élevé qu'en condition de contrôle ($p<0.01$) et supérieur au hasard tout en induisant des

jugements d'intensité de l'émotion plus faibles ($p<0.001$), mérite d'être commenté. Ce résultat peut s'expliquer par le tremblement vocal porté par l'expression originale de tristesse et que nous supposons être à l'origine des reports vers *joie* et *satisfaction* en condition de contrôle par confusion avec le rire. La transplantation du contour de F0 extrait de l'expression neutre sur l'énoncé [sabl] a pour effet de neutraliser ce jitter et par ricochet l'essentiel des reports vers *joie* et *satisfaction*, tandis que la forme générale du contour de F0 de l'expression neutre est insuffisamment distincte de celle du contour de F0 de l'expression de tristesse pour modifier de façon décisive la catégorisation effectuée par les auditeurs, ne se traduisant que par une intensité d'émotion perçue moindre.

Les taux de reconnaissance obtenus par les expressions de déception, résignation et tristesse resynthétisés en condition *intensité seule* ne sont pas significativement différents de ceux obtenus en condition *F0 seule* (respectivement pour ces trois émotions $p=0.357$, $p=0.313$ et $p=0.568$, n.s.) et ne sont pas supérieurs au hasard. De fait, les trois stimuli originaux correspondants sont caractérisés par un contour de fréquence fondamentale de faible dynamique, et sont en cela proches de l'expression neutre sur l'énoncé [sabl] utilisé comme référence pour la projection de ces énoncés. Les expressions de déception et de résignation reçoivent ainsi des jugements d'intensité de l'émotion aussi faibles en condition *intensité seule* qu'en condition *F0 seule* (respectivement $p=0.356$ et $p=0.077$, n.s.), non significativement supérieurs au niveau du hasard. Toutefois l'expression de tristesse, dont le contour de fréquence fondamentale présente une excursion légèrement plus importante, se voit attribuer des jugements d'intensité d'émotion plus élevés en condition *F0 seule* qu'en condition *intensité seule* ($p<0.001$).

Dans le cas du dégoût, seules la qualité de voix et la durée de l'expression originale permettent une reconnaissance à un niveau significativement supérieur au hasard ($p<0.001$), les contours de fréquence fondamentale et d'intensité ne semblant pas transmettre une part quelconque de l'information affective de cette expression. Cependant le taux de reconnaissance dans la condition de resynthèse *qualité de voix et durée* reste nettement en deca de celui obtenu en condition de contrôle ($p<0.001$), tandis que le jugement d'intensité de l'émotion n'est que légèrement inférieur à celui attribué en condition de contrôle ($p=0.048<0.05$). L'explication la plus vraisemblable de cette incapacité du stimulus resynthétisé en condition *qualité de voix et durée* à véhiculer l'ensemble de l'information affective du stimulus d'origine réside dans l'écart important entre la durée de ce stimulus et celle de l'expression neutre sur l'énoncé [sabl] : la transplantation du contour plat de fréquence fondamentale du stimulus neutre vers l'expression originale de dégoût pour générer le stimulus en condition *qualité de voix et*

durée a en effet pour résultat une perte de naturalité très perceptible, le stimulus synthétique pouvant difficilement être assimilé à une production naturelle.

9.2.4.6 Un artefact de projection des contours d'intensité

Il est surprenant de constater que le taux global de reconnaissance induit par la transplantation des contours de fréquence fondamentale et d'intensité est plus faible que lorsque seuls les contours de fréquence fondamentale sont transplantés (30% vs. 36%), de même que l'intensité d'émotion perçue moyenne (3.0 vs. 3.2).

Bien que ces différences ne soient pas significatives ($p=0.196$ n.s. pour la reconnaissance, $p=0.157$ n.s. pour l'intensité), il aurait été plus vraisemblable que nous observions la tendance inverse, à savoir que la reconnaissance et l'intensité d'émotion perçue soient supérieures lorsque le contour d'intensité est copié en complément des variations de fréquence fondamentale. L'examen des contours d'intensité résultant de l'application du contour relatif extrait d'un stimulus cible suivi d'un rééchelonnage du signal révèle que la méthode que nous avons appliquée, bien que préservant la forme générale du contour d'intensité d'origine, a dans la plupart des cas pour effet d'amplifier les écarts des minima et maxima locaux par rapport à l'intensité moyenne. Cette tendance est particulièrement sensible au niveau des minima locaux, avec pour effet d'atténuer fortement l'attaque et la coda des signaux générés. Afin de quantifier l'impact acoustique de la transplantation des contours relatifs d'intensité, nous avons extrait à l'aide d'un script développé avec Praat l'intensité de l'attaque, du noyau et de la coda, ainsi que l'intensité de l'ensemble de l'énoncé, pour les stimuli naturels et issus de la projection des contours d'intensité. Les valeurs moyennes ainsi que les écarts-types des différences exprimées en décibels entre les valeurs d'intensité mesurées sur les stimuli générés en conditions *F0 et intensité* et *qualité de voix et durée* et celles extraites des stimuli naturels correspondants, sont présentées dans la Table 14. Les valeurs sont présentées en séparant le cas de la projection de l'intensité des stimuli expressifs correspondant aux conditions *intensité seule* et *F0 et intensité* de celui de la projection des stimuli neutres correspondant à la condition *qualité de voix et durée*.

La première observation est que, alors que la méthode à laquelle nous avons recouru pour projeter les contours d'intensité est supposée ajuster le niveau global d'intensité du stimulus généré à la valeur mesurée sur le stimulus naturel, les différences sur l'ensemble de l'énoncé entre l'intensité mesurée sur le stimulus naturel et elle mesurée sur le stimulus projeté sont conséquentes et globalement significatives (test t pour échantillon unique avec la valeur de test 0, $t(15)=-6.9$, $p<0.001$). Ces différences sont d'un niveau équivalent entre les conditions *F0 et intensité* et *qualité de voix et durée* ($t(14)=1.2$, $p=0.248$ n.s.), en revanche elles sont plus importantes sur l'énoncé [sabl] que sur l'énoncé

[ʁuʒ](t(14)=-3.2, p<0.01). Ces écarts inattendus s'expliquent par le fait que nous avons contrôlé le niveau global d'intensité lors de la génération de stimuli en prenant en compte le stimulus dans son ensemble, les segments de longueurs inégales étiquetés comme silencieux étant pris en compte dans le calcul.

condition		F0 et intensité				qualité de voix et durée			
énoncé	émotion	attaque	noyau	coda	énoncé	attaque	noyau	coda	énoncé
[ʁuʒ]	anxiété	6.0	-0.2	0.7	2.4	-5.8	-0.9	-1.5	-3,0
	inquiétude	-9.1	1.5	-9.0	-5.4	-1.0	0.9	-3.7	-2,2
	satisfaction	-11.5	0.5	-12.0	-8.6	-6.5	0.0	-9.4	-4,1
[sabl]	déception	-8.0	3.1	-5.8	-4.9	-18.4	-5.7	-17.1	-13,1
	dégoût	-4.8	4.8	-4.3	-6.2	-19.2	-4.7	-21.6	-10,3
	joie	-10.7	1.8	-4.2	-5.6	-8.0	-1.4	-19.2	-10,4
	résignation	-11.6	1.4	-8.1	-7.3	-8.2	-0.7	-15.6	-9,5
	tristesse	-9.7	0.6	-10.1	-7.6	-7.3	0.4	-13.5	-8,5
moyenne		-7,4	1.7	-6.6	-5.4	-9.3	-1.5	-12.7	-7.6

Table 14 : Différences d'intensité exprimées en décibels entre le stimulus issu de la projection des contours d'intensité et de fréquence fondamentale et le stimulus naturel au niveau de l'attaque consonantique, du noyau vocalique et de la coda, ainsi que pour l'ensemble de l'énoncé, dans le cas de la projection à l'aide du module TD-PSOLA de Praat des contours extraits du stimulus expressif (condition *F0 et intensité*) et de ceux extraits du stimulus neutre (condition *qualité de voix et durée*), pour chacune des émotions traitées. Tous les stimuli sont extraits des productions du locuteur M2.

A un niveau de granularité plus fin, l'application de contours relatifs d'intensité a globalement pour effet en condition *F0 et intensité* une légère amplification de l'intensité du noyau vocalique, porteur du maximum d'énergie des stimuli originaux, significativement différente de l'atténuation observée en condition *qualité de voix et durée* (t(14)=3.1, p<0.01). Bien que cette tendance soit plus marquée dans le cas de la projection sur l'énoncé [sabl] que sur l'énoncé [ʁuʒ], les différences entre énoncés ne sont significatives dans aucune de ces conditions (t(6)=1.7, p=0.144 n.s. en condition *F0 et intensité*, t(5.3)=-1.8, p=0.115 n.s. avec correction de Welsh en condition *qualité de voix et durée*). L'intensité de l'attaque et de la coda des stimuli projetés sont atténuées dans des proportions plus importantes, sans que les différences entre conditions de resynthèse ne soient significatives (t(14)=0.6, p=0.550 n.s. pour l'attaque, t(14)=2.1, p=0.056 n.s. pour la coda). En particulier, les schwas de détente finaux sont fortement atténués, ce qui tend à les rendre imperceptibles, tout particulièrement dans le cas de la projection des contours neutres. Cette atténuation est également plus marquée sur l'énoncé [sabl] que sur l'énoncé [ʁuʒ], la différence étant significative pour l'attaque (t(14)=-2.2, p<0.01) mais pas pour la coda (t(14)=-2.0, p=0.065 n.s.).

Dans le cas de la projection de contours d'intensité extraits de l'expression de satisfaction pour la génération des stimuli projetés en conditions *intensité seule* et *F0 et intensité*, l'atténuation de l'attaque et de la coda sont particulièrement marquées. L'information affective portée par le contour de fréquence fondamentale de l'expression de satisfaction semble suffisamment robuste pour que le stimulus généré en condition *F0 et intensité* soit reconnu comme exprimant de la joie ou de la satisfaction à un niveau comparable à celui généré en condition *F0 seule*, et jugé comme reflétant une intensité d'émotion comparable. Bien que l'atténuation de la coda soit plus limitée pour l'expression de joie, un schéma similaire peut être observé pour l'expression de joie. Ceci ne permet toutefois pas de conclure que l'intensité de ces expressions ne porte pas d'information affective. En effet il est possible qu'avec une répartition de l'intensité dans l'énoncé plus semblable à celle de l'expression naturelle, ces expressions resynthétisées en condition *F0 et intensité* obtiennent un taux de reconnaissance et une intensité d'émotion plus proche de ceux obtenus en condition de contrôle.

9.2.5 Projection à partir d'une modélisation ARX-LF

9.2.5.1 Méthode d'analyse/resynthèse

Comme nous l'avons souligné ci-dessus, la méthode d'analyse/resynthèse retenue pour la première expérience de projection dimensionnelle ne permet que la transplantation de contours de fréquence et d'intensité, et nous a donc contraints à laisser la qualité de voix hors du champ d'investigation en nous contentant d'évaluer le poids des variations conjointes de qualité de voix et de durée. Afin de pouvoir étudier séparément l'influence de ces dimensions prosodiques sur les valeurs affectives véhiculées et plus généralement recentrer l'analyse sur les phénomènes de qualité de voix, nous avons procédé en collaboration avec Damien Vincent et Olivier Rosec (Audibert, Vincent, Aubergé, Rilliard &Rosec, 2006) à l'évaluation perceptive de stimuli modélisés et projetés à l'aide d'un modèle ARX-LF, suivant un protocole similaire à celui de l'étude présentée ci-dessus.

Vincent *et al.* (2005) proposent d'adapter le processus autorégressif exogène (ARX) proposé par Ding *et al.* (1995) pour l'estimation conjointe des paramètres de la source glottique et du conduit vocal à une modélisation de la source par le modèle LF (Fant *et al.*, 1985). Selon cette approche, tout son voisé s(n) peut être modélisé par le processus ARX défini comme :

$$s(n) = -\sum_{k=1}^{p} a_k(n).s(n-k) + b_0.u(n) + e(n)$$

Dans cette équation, la composante u(n) désigne la partie de la source modélisée par les paramètres de LF (cf. section 7), le coefficient b_0 étant l'amplitude de l'onde de débit

glottique dérivée. La partie de la source glottique non modélisée par LF est représenté par le résidu stochastique e(n). Ce résidu correspond notamment aux bruits d'aspiration et de friction, ainsi qu'à des variations de la source non pris en compte par le modèle LF, comme par exemple le phénomène d'oscillation de l'onde de débit glottique dérivée ou *ripple* (Vincent, 2007). Enfin les coefficients $a_k(n)$ sont ceux du filtre d'ordre p qui modélise les résonances du conduit vocal.

Dans le cadre de la modélisation d'un son voisé par un modèle ARX excité par une source LF, l'analyse revient à estimer les trois paramètres du modèle LF décrivant la forme de l'onde de débit glottique, le résidu stochastique, les coefficients du filtre correspondant au conduit vocal, ainsi que la fréquence fondamentale et l'énergie. Etant donné que les paramètres du filtre et du résidu peuvent être estimés par la méthode des moindres carrés une fois les paramètres LF connus, une méthode fondée sur une recherche exhaustive dans un espace d'ondes LF quantifiées a été proposée pour l'estimation de ces paramètres (Vincent *et al.*, 2005 ; Vincent, 2007). La fréquence fondamentale est estimée par une méthode temporelle adaptée de l'algorithme YIN (de Cheveigné & Kawahara, 2002), auquel ont été ajouté des contraintes spécifiques à la parole : une désambiguïsation entre la fréquence fondamentale et les fréquences des harmoniques ainsi qu'une contrainte de continuité des contours de fréquence fondamentale extraits (Vincent, 2007).

Dans ce cadre, le principe du processus d'analyse/resynthèse de signaux naturels appliqué à la projection dimensionnelle est de remplacer certains paramètres issus de l'analyse du stimulus neutre (source) par les valeurs de ces paramètres obtenues par l'analyse du stimulus expressif (cible). Ce processus comprend une procédure d'alignement ainsi qu'un algorithme de synthèse. L'alignement nécessite que les stimuli source et cible aient le même contenu phonétique : après avoir apparié les frontières phonémiques des deux stimuli, les points analysés dans un même phonème sont reliés par interpolation linéaire, dont le résultat doit être contrôlé afin de corriger d'éventuelles erreurs dues à la non congruence des informations de voisement des stimuli source et cible. L'algorithme de synthèse, similaire à ceux utilisés pour les modifications prosodiques fondées sur TD-PSOLA (Moulines & Laroche, 1995), détermine les instants de synthèse et génère pour chacun de ces instants une paire de trames issues respectivement des stimuli source et cible, permettant ainsi d'obtenir le stimulus synthétisé.

Avec cette modélisation ARX-LF, la qualité de voix est ainsi modélisée par trois composantes : les paramètres LF de la source, les coefficients du filtre correspondant à la modélisation du conduit vocal, et le résidu stochastique correspondant aux variations de

la source non modélisées par LF. Au-delà du recentrage sur les phénomènes de qualité de voix de l'évaluation des poids relatifs des dimensions prosodiques dans les expressions d'affects, l'étude de la perception par des auditeurs naïfs de stimuli projetés à l'aide de cette méthode a également pour objectif d'évaluer la pertinence de la modélisation ARX-LF pour des échantillons de parole expressive.

9.2.5.2 Choix des combinaisons traitées

L'application de cette méthode à une paire de stimuli constituée d'un stimulus expressif et d'un stimulus neutre permet donc d'attribuer à chaque stimulus synthétisé les valeurs extraites de l'un ou l'autre des stimuli naturels analysés pour chacun des six paramètres ou ensembles de paramètres suivants : source glottique, coefficients du filtre, résidu stochastique, durées phonémiques, contour de fréquence fondamentale et intensité. Ainsi, 64 combinaisons différentes peuvent être générées à partir de chacun des stimuli expressifs, incluant l'analyse/resynthèse à l'identique du stimulus expressif et du stimulus neutre de référence.

Les huit stimuli expressifs naturels ainsi que les stimuli neutres de référence utilisés pour la resynthèse à partir de Praat ont ainsi été analysés et resynthétisés partiellement par Damien Vincent, les 64 combinaisons possibles étant générées pour chaque stimulus expressif traité. Cependant, la méthode de synthèse décrite ci-dessus n'a pas permis de générer des stimuli synthétiques de qualité suffisante à partir de l'expression de satisfaction.

Pour chacune des sept émotions restantes, à savoir joie, anxiété, inquiétude, déception, résignation, tristesse et dégoût, nous avons sélectionné pour l'évaluation perceptive les sept combinaisons suivantes :

- Le stimulus de contrôle construit en appliquant à lui-même les paramètres de source glottique, le résidu et le filtre ainsi que les variations de fréquence fondamentale et d'intensité analysées à partir du stimulus source, étiqueté *contrôle*.
- Le stimulus construit en appliquant les variations de F0 et d'intensité extraites du stimulus expressif, les autres paramètres provenant de l'analyse de l'expression neutre correspondante, étiqueté *F0 et intensité*.
- Le stimulus construit à partir des paramètres de la source LF, du résidu, des paramètres du filtre et des durées phonémiques extraites du stimulus expressif, étiqueté *qualité de voix et durée*.
- Le stimulus construit à partir des paramètres de la source LF, du résidu et des paramètres du filtre extraits du stimulus expressif, étiqueté *qualité de voix*.
- Le stimulus construit à partir des paramètres de la source LF et du résidu extraits du stimulus expressif, étiqueté *source et résidu*.

- Le stimulus construit à partir des paramètres de la source LF extraits du stimulus expressif, étiqueté *source seule*.
- Le stimulus construit à partir des durées phonémiques extraites du stimulus expressif, étiqueté *durée seule*.

Les conditions *contrôle*, *F0 et intensité* et *qualité de voix et durée* permettent de comparer les résultats de l'évaluation perceptive avec ceux obtenus précédemment lors de l'évaluation des stimuli générés avec Praat. De même que lors de l'évaluation perceptive des stimuli resynthétisés avec Praat, les expressions neutres sur les énoncés « rouge » et « sable » resynthétisées en condition de contrôle ont été incluses dans l'ensemble des stimuli évalués par les auditeurs.

9.2.5.3 Evaluation perceptive

Les stimuli ainsi générés ont été évalués selon un protocole très semblable à celui utilisé dans l'évaluation perceptive des stimuli resynthétisés à l'aide de Praat présentée ci-dessus. En raison de l'impossibilité de générer des stimuli synthétiques de qualité suffisante pour les expressions de satisfaction, l'étiquette *satisfaction* a été retirée de l'ensemble des choix proposés aux auditeurs.

Les 51 stimuli générés ont été évalués par 25 auditeurs francophones sans problème d'audition connu (7 hommes, 18 femmes, âgés de 25.7 ans en moyenne), dans une pièce calme avec un casque fermé de haute qualité. De même que dans la précédente évaluation, chaque stimulus a été présenté trois fois, et les stimuli ont été présentés dans un ordre aléatoire différent pour chacun des auditeurs et contrôlé pour éviter deux présentations successives d'un même stimulus.

9.2.5.4 Matrices de confusion et regroupement d'étiquettes

Une matrice de confusion a été extraite pour chacune des sept conditions de resynthèse. La matrice de confusion en condition de contrôle est présentée dans la Table 15. Les matrices de confusions pour les six autres conditions sont présentées en Annexe IV :.

La classification hiérarchique opérée sur les colonnes de la matrice de confusion en condition de contrôle normalisées par la somme de leurs éléments, et illustrée par le dendrogramme présenté Figure 39, indique que la plupart des confusions en condition de contrôle sont similaires à celles observées précédemment. Néanmoins l'expression de tristesse, pour laquelle les reports en condition de contrôle vers les étiquettes *résignation* et *déception*, et dans une moindre mesure *anxiété*, sont plus faibles que dans l'évaluation précédente, apparaît ici comme une catégorie isolée plutôt que confondue avec déception et résignation.

	joie	anxiété	résignation	tristesse	inquiétude	déception	dégoût	neutre
joie	**77%**	7%	1%	1%	5%	0%	1%	7%
anxiété	0%	**64%**	3%	3%	27%	1%	1%	1%
résignation	11%	33%	**19%**	0%	21%	9%	3%	4%
tristesse	23%	12%	3%	**43%**	9%	11%	0%	0%
inquiétude	51%	8%	5%	0%	**36%**	0%	0%	0%
déception	0%	0%	51%	3%	1%	**31%**	15%	0%
dégoût	16%	0%	5%	8%	0%	0%	**71%**	0%
neutre	5%	5%	7%	7%	6%	16%	1%	**53%**

Table 15 : Matrice de confusion correspondant à l'évaluation en condition de contrôle de la resynthèse réalisée à l'aide de l'algorithme ARX-LF à partir de productions du locuteur M2. Les taux de reconnaissance sont indiqués en gras.

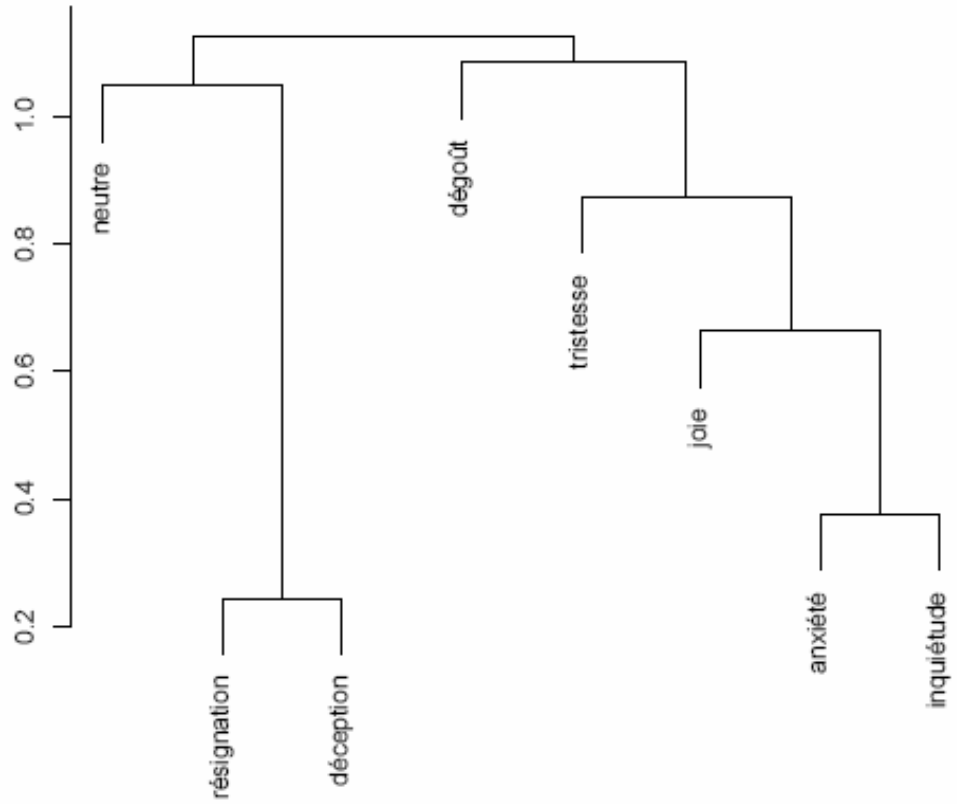


Figure 39 : Dendrogramme issu de la classification hiérarchique (critère d'agrégation de Ward) opérée sur les colonnes normalisées de la matrice de confusion confusion correspondant à l'évaluation en condition de contrôle de la resynthèse réalisée à l'aide de l'algorithme ARX-LF à partir de productions du locuteur M2.

Afin de pouvoir comparer les résultats de cette évaluation perceptive à ceux obtenus lors de l'évaluation des stimuli générés avec Praat, nous avons néanmoins considéré les mêmes regroupements de catégories. La matrice de confusion obtenue suite à ces regroupements est présentée Table 16.

	joie/satisf.	anx./inq.	rés./déc./trist.	dégoût	neutre
hasard	*25%*	*25%*	*38%*	*13%*	*13%*
joie	**77%**	12%	3%	1%	7%
anxiété	0%	**91%**	7%	1%	1%
inquiétude	51%	**44%**	5%	0%	0%
résignation	11%	55%	**28%**	3%	4%
déception	0%	1%	**84%**	15%	0%
tristesse	23%	21%	**56%**	0%	0%
dégoût	16%	0%	13%	**71%**	0%
neutre	5%	11%	31%	1%	**53%**

Table 16 : Matrice de confusion confusion correspondant à l'évaluation en condition de contrôle de la resynthèse réalisée à l'aide de l'algorithme ARX-LF à partir de productions du locuteur M2, après regroupement d'étiquettes. Les taux de reconnaissance sont indiqués en gras. Le niveau théorique du hasard, fonction du nombre d'étiquettes dans chaque regroupement, est indiqué sur la 1ère ligne.

9.2.5.5 Analyse statistique

Selon la même procédure que celle adoptée pour l'analyse des résultats de l'évaluation des stimuli générés avec Praat, les réponses des sujets ont été converties en réponses justes ou fausses afin que pour chaque émotion, chaque condition de resynthèse et chaque juge une valeur comprise entre 0 et 3. Nous avons également extrait l'intensité perçue de l'émotion, en considérant cette intensité comme nulle lorsque l'étiquette *neutre* avait été sélectionnée.

Des tests t pour échantillon unique indiquent que dans la condition de contrôle, les taux de reconnaissance après regroupement des expressions de joie, satisfaction, déception, anxiété, inquiétude et dégoût ainsi que ceux des expressions neutres sont significativement supérieurs au hasard ($p<0.001$ pour chaque comparaison à l'exception de l'inquiétude et de la tristesse, $p<0.05$). En revanche le taux de reconnaissance de l'expression de résignation n'est pas meilleur que le hasard. Les résultats des tests t pour échantillon unique comparant le taux de reconnaissance de chaque émotion dans chaque condition de resynthèse au niveau du hasard sont présentés en Annexe III :.

Les taux moyens de reconnaissance obtenus dans les différentes conditions de resynthèse pour chacune des émotions présentées sont représentés Figure 40. Les intensités moyennes d'émotion perçues correspondantes sont représentées Figure 41.

Nous avons effectué une ANOVA à mesures répétées avec le nombre de réponses correctes comme variable dépendante et la condition de resynthèse et l'émotion présentée comme facteurs fixés, les expressions neutres en condition de contrôle étant exclues de l'analyse pour conserver un plan d'expérience factoriel. Cette ANOVA indique un effet significatif de la condition ($F(5.2, 123.9)=42.7$; $p<0.001$), de l'émotion

($F(4.2, 102.0)=15.5$; $p<0.001$), ainsi que de l'interaction émotion*condition ($F(25.2, 604.4)=10.8$; $p<0.001$). Des contrastes spécifiques indiquent que les émotions exprimées par les stimuli resynthétisés en condition de contrôle sont globalement mieux reconnues que dans toutes les autres conditions de resynthèse ($p<0.001$ pour tous les contrastes). Une ANOVA à mesures répétées avec les mêmes facteurs fixés et l'intensité d'émotion perçue comme variable dépendante montre également un effet significatif de la condition ($F(4.2, 101.9)=100.8$; $p<0.001$), de l'émotion présentée ($F(5.4, 130.2)=23.1$; $p<0.001$) et de l'interaction émotion*condition ($F(27.1, 649.3)=11.9$; $p<0.001$). Les stimuli resynthétisés en condition de contrôle sont également jugés avec une intensité d'émotion plus élevée que dans toutes les autres conditions de resynthèse ($p<0.001$ pour tous les contrastes).

Une première indication sur la pertinence d'une modélisation ARX-LF appliquée à ces expressions émotionnelles est donnée par la comparaison des taux moyens de reconnaissance en conditions source seule (reconnu à 26% et jugé avec une intensité moyenne de 2.3), source et résidu (30%, 2.5) et qualité de voix (35%, 3.0). Bien que le gain lié à la prise en compte du résidu soit globalement non significatif ($p=0.230$ n.s. pour la comparaison des conditions *source* et *source et résidu*), de même que celui induit par la modélisation des résonances du conduit vocal ($p=0.249$ n.s. entre les conditions *source et résidu* et *qualité de voix*), le taux de reconnaissance en prenant en compte l'ensemble des informations relatives à la qualité de voix est plus élevé que celui obtenu avec seulement les paramètres LF de la source ($p<0.05$ entre les conditions *source seule* et *qualité de voix*). De même, les jugements d'intensité de l'émotion sont globalement plus élevés en condition *qualité de voix* qu'en conditions *source seule* ($p<0.01$) et *source et résidu* ($p<0.05$), tandis que le gain lié à la prise en considération du résidu est non significatif ($p=0.417$ n.s. entre les conditions *source seule* et *source et résidu*).

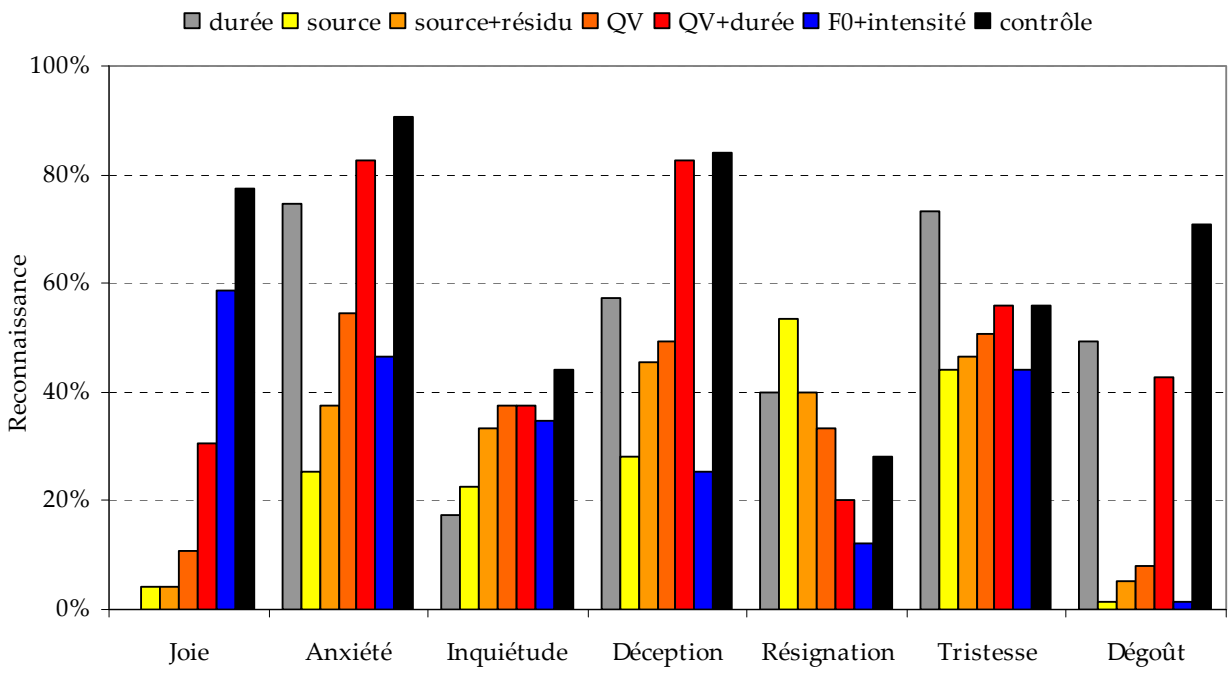


Figure 40 : Taux moyen de reconnaissance pour chaque émotion présentée et chaque condition de resynthèse générée avec l'algorithme ARX-LF à partir de productions du locuteur M2.

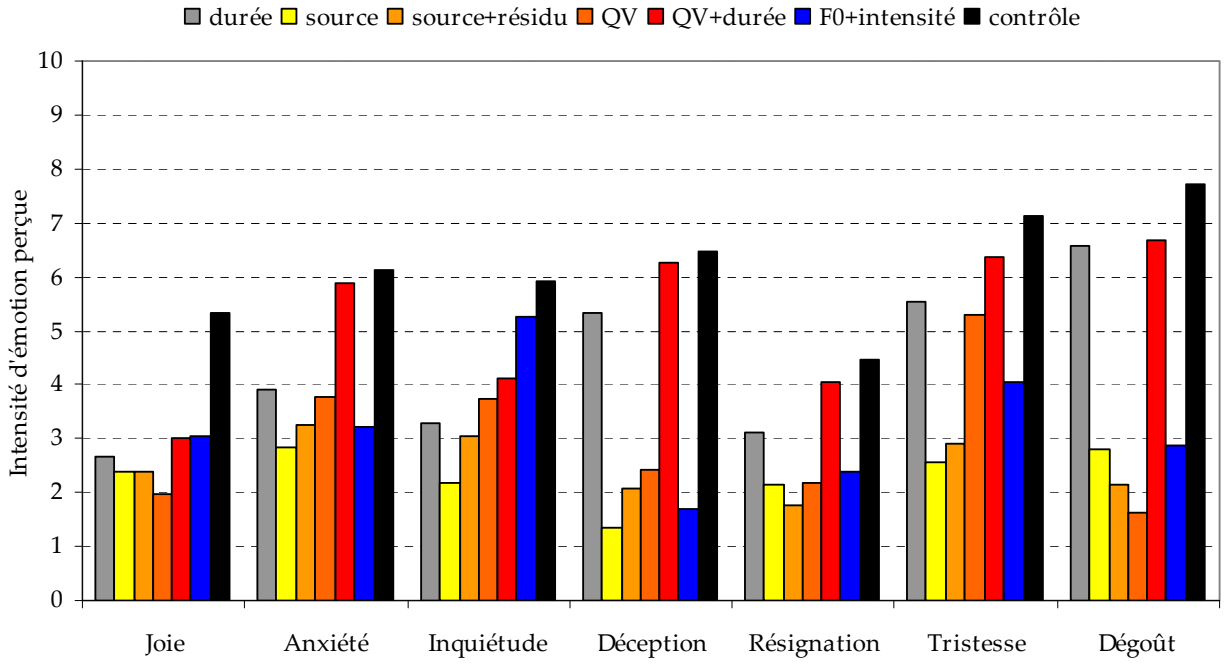


Figure 41 : Intensité moyenne perçue pour chaque émotion présentée dans chacune des conditions de resynthèse générée avec l'algorithme ARX-LF à partir de productions du locuteur M2.

Les comparaisons multiples entre conditions pour chacune des émotions indiquent que l'expression d'anxiété n'est pas significativement mieux reconnue en condition de contrôle qu'en conditions *qualité de voix et durée* (p=0.401 n.s.) ni *durée seule* (p=0.094 n.s.). Si la durée phonémique et notamment la brièveté de la voyelle [u] apparaît comme le

principal déterminant de l'information affective portée par cette expression, ces variations de durée ne semblent pas suffisantes pour la transmettre dans son ensemble. En effet, tandis que les jugements d'intensité de l'émotion reçus par l'expression d'anxiété resynthétisée en condition *qualité de voix et durée* ne sont pas significativement plus faibles que ceux reçus par cette expression en condition de contrôle ($p=0.649$ n.s.), les jugements d'intensité reçus en condition *durée seule* sont plus faibles qu'en condition de contrôle ($p<0.01$). De plus l'expression d'anxiété resynthétisée en condition *qualité de voix* est reconnue à un niveau supérieur au hasard ($p<0.01$), de même qu'en condition*F0 et intensité* ($p<0.01$).

Bien que n'étant pas significativement moins bien reconnue en condition *F0 et intensité* qu'en condition de contrôle ($p=0.271$ n.s.), l'expression d'inquiétude n'est reconnue à un taux significativement supérieur au hasard ($p<0.05$) qu'en condition de contrôle. Il est donc difficile de tirer des conclusions de la répartition des réponses obtenues pour cette expression. Notons toutefois que, de manière similaire à ce que nous avons pu observer dans le cas de la projection avec Praat, la projection en condition *F0 et intensité* de cette expression d'inquiétude est très largement confondue avec l'expression de joie (60%).

De même, l'expression de résignation, reconnue à un niveau proche du hasard en condition de contrôle ($p=0.195$ n.s.), n'est reconnue à un niveau supérieur au hasard qu'en condition *source seule* ($p<0.01$), ne permettant pas de formuler des interprétations exploitables de la répartition des réponses en conditions projetées. Bien que les différences de taux de reconnaissance entre les conditions *source seule*, *source et résidu* et *qualité de voix* soient toutes non significatives, elles présentent ici un schéma inverse de celui observé pour la plupart des autres émotions évaluées, pour lesquelles le taux de reconnaissance augmente avec l'ajout de composantes relatives à la qualité de voix du stimulus expressif. Cette répartition suggère que le résidu stochastique et la modélisation des résonances du conduit vocal issues de l'analyse de l'expression neutre permettent de produire un stimulus plus proche de l'idée que se font les auditeurs d'une expression de résignation que lorsque ces caractéristiques sont également extraites du stimulus expressif. Ce résultat n'est pas surprenant étant donnée la proximité acoustique apparente de ces stimuli.

L'expression de déception resynthétisée en condition *qualité de voix et durée* n'est pas significativement moins bien reconnue qu'en condition de contrôle ($p=0.874$ n.s.), de plus cette expression reçoit dans cette condition des jugements d'intensité de l'émotion aussi élevés qu'en condition de contrôle ($p=0.717$ n.s.). Les durées phonémiques ne portent pas ici une part de l'information significativement plus importante que celle portée par la qualité de voix ($p=0.342$ n.s. pour la comparaison entre les conditions *qualité de voix* et

durée seule), bien que l'expression de déception resynthétisée en condition *qualité de voix* ne soit pas reconnue à un niveau significativement plus élevé que le hasard (p=0.127 n.s.) contrairement à la resynthèse de cette expression en condition *durée seule* (p<0.01).

Pour l'expression de tristesse aucune condition de projection partielle ne montre de différence avec la condition de contrôle (les comparaisons par paires sont non significatives). La resynthèse en condition *durée seule* est mieux reconnue qu'en conditions *F0 et intensité*, *source seule*, *source et résidu* et *qualité de voix* (p<0.05 pour toutes les comparaisons), suggérant que la durée phonémique serait le déterminant principal de l'information affective véhiculée par l'expression de tristesse. Néanmoins l'intensité attribuée aux projections partielles de l'expression de tristesse reste inférieure à celle de la condition de contrôle (p<0.001 pour les conditions *F0 et intensité*, *source seule* et *source et résidu*, p<0.01 pour les conditions *durée seule* et *qualité de voix*), à l'exception de la condition *qualité de voix et durée* (p=0.171 n.s.).

L'expression de dégoût n'est jamais aussi bien reconnue qu'en condition de contrôle (p<0.001 pour toutes les comparaisons sauf avec la condition *durée seule*, p<0.01). Outre la condition de contrôle (p<0.001), les projections partielles de dégoût ne sont reconnues à un niveau supérieur au hasard qu'en conditions *durée seule* et *qualité de voix et durée* (p<0.001). Bien que les jugements d'intensité de l'émotion attribués à ces deux projections partielles soient plus élevés que ceux des autres projections partielles (p<0.001 pour toutes les comparaisons), ils restent inférieurs à ceux attribués à l'expression de dégoût en condition de contrôle (p<0.05 dans les deux cas).

L'expression de joie est reconnue à un niveau supérieur au hasard en condition *F0 et intensité*(p<0.001), avec néanmoins un taux de reconnaissance plus faible qu'en condition de contrôle (p<0.05). De plus les jugements d'intensité de l'émotion attribués à cette expression resynthétisée en condition *F0 et intensité*, plus faibles qu'en condition de contrôle (p<0.001), ne sont pas plus élevés que ceux attribués dans les autres conditions manipulées à l'exception de la condition *qualité de voix* (p<0.05). Cette expression de joie est également reconnue à un niveau supérieur au hasard en condition *qualité de voix et durée* (31%, p<0.05), alors que la resynthèse en condition *durée seule* n'est reconnue par aucun auditeur et que la resynthèse en condition *qualité de voix* ne dépasse pas le niveau du hasard (11%, p=0.626 n.s.). L'examen des signaux générés correspondants nous a permis de mettre en évidence la présence d'un artefact issu du traitement préalable au processus d'analyse/synthèse, consistant en un bruit de bouche copié dans l'expression générée en condition *qualité de voix et durée* depuis une portion de signal étiquetée comme silencieuse, et susceptible d'avoir été interprétée comme un coup de glotte annonciateur d'un rire (Figure 42), relevant ainsi artificiellement le taux de

reconnaissance de ce stimulus synthétisé. Ce bruit de bouche d'une durée de 9 millisecondes et d'une intensité de 48 dB, bien que minime en termes de quantité brute d'information, pourrait ainsi avoir eu un impact important sur l'attribution émotionnelle du stimulus synthétisé.

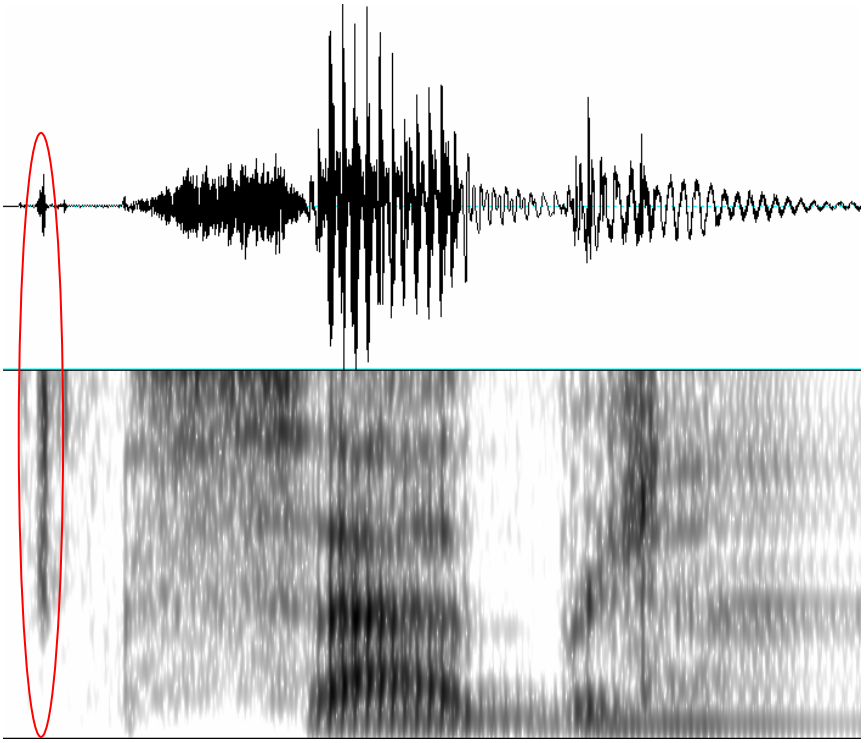


Figure 42 : Signal acoustique et spectrogramme de l'expression de joie sur l'énoncé [sabl] resynthétisée par ARX-LF en condition *qualité de voix et durée*. L'ellipse rouge indique le bruit de bouche copié depuis une portion du signal étiquetée comme silence.

9.2.6 Comparaison des résultats obtenus dans les deux expériences

En raison de l'absence de l'expression de satisfaction parmi les stimuli projetés au moyen d'une modélisation ARX-LF, ce qui nous a conduit a éliminer l'étiquette correspondante de l'ensemble des choix proposés aux auditeurs, les taux de reconnaissance obtenus lors des deux évaluations perceptives présentées ci-dessus ne peuvent être directement comparés. Pour mener à bien cette comparaison, nous pouvons néanmoins calculer un indice correspondant au taux de reconnaissance de chaque combinaison émotion x condition et prenant en compte le niveau du hasard. L'indice π proposé par Rosenthal & Rubin (1989) corrige les taux de reconnaissance bruts de l'influence du nombre de choix k proposés aux auditeurs, tout en restant directement interprétable contrairement à l'indice h de Cohen (1977) calculé comme la différence entre les transformations arcsinus du taux de reconnaissance brut et du niveau du hasard. L'indice π n'est pas directement transposable à l'analyse des taux de reconnaissance après regroupement d'étiquettes,

nous pouvons cependant le généraliser simplement en considérant l'inverse du niveau du hasard après regroupement en lieu et place du nombre d'étiquettes proposées dans la tâche de catégorisation. Nous nous appuierons donc pour comparer les résultats des deux évaluations perceptives sur l'indice π' défini comme :

$\pi' = \frac{P.(\frac{k}{q} - 1)}{1 + P.(\frac{k}{q} - 2)}$, où P est le taux de reconnaissance après regroupement, q le nombre d'étiquettes regroupées dans la catégorie considérée, et k le nombre d'étiquettes proposées aux auditeurs. Ainsi, π' suit exactement la distribution deπ lorsque q prend la valeur 1. De même que l'indice π de Rosenthal & Rubin (1989), π' prend la valeur 0.5 lorsque le stimulus est reconnu au niveau du hasard, et la valeur 1 lorsqu'il est reconnu sans confusions.

Il faut cependant souligner que lorsque le nombre de catégories considérées est élevé, cet indice est peu adapté à la comparaison de taux de reconnaissance proches du niveau du hasard et à plus forte raison inférieurs à ce niveau. En effet comme l'illustre la Figure 43 pour les six niveaux du hasard considérés dans les deux évaluations perceptives successives, les valeurs prises par indice en fonction du taux de reconnaissance brut montrent une tendance asymptotique à proximité de l'origine d'autant plus marquée que le niveau du hasard q/k est faible.

Ainsi, une variation de 10% du taux de reconnaissance centrée autour du niveau du hasard induit une variation de 0.108 de la valeur de π' dans le cas du regroupement de trois étiquettes parmi neuf (ce qui correspond au regroupement des expressions de déception, résignation et tristesse dans la première évaluation perceptive). A l'opposé, une variation de même amplitude dans le cas de l'analyse sans regroupement des réponses attribuées à une catégorie parmi neuf (ce qui correspond à l'expression de dégoût ainsi qu'aux expressions neutres dans la première évaluation perceptive) induit une variation de 0.266. Il convient donc de rester prudents dans notre comparaison des valeurs de π' obtenues pour les catégories non regroupées, en particulier lorsque le taux de reconnaissance brut est proche du niveau du hasard ou inférieur à celui-ci.

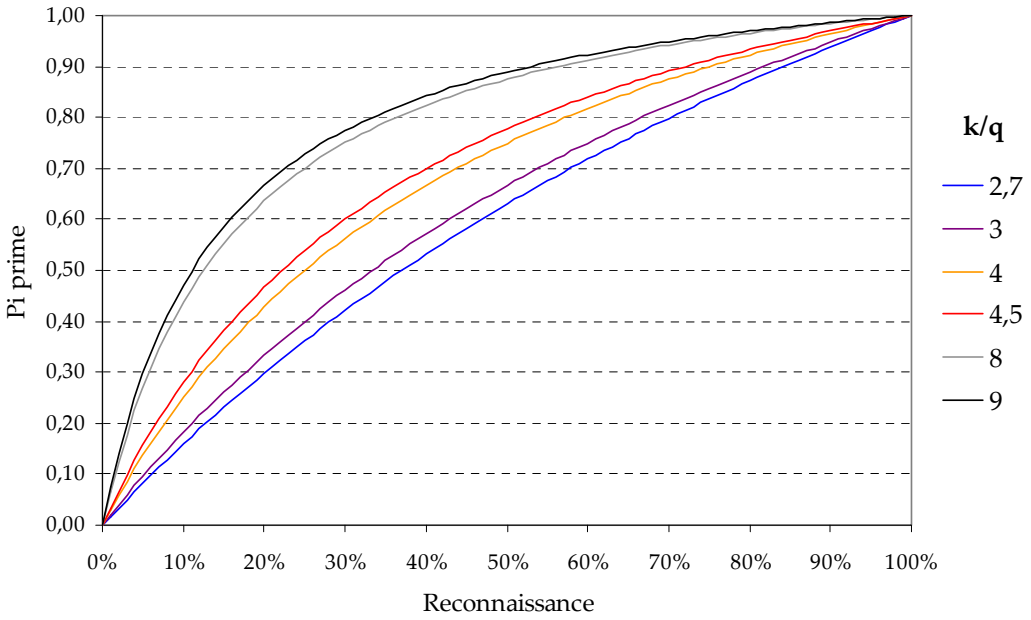


Figure 43 : Valeurs prises par π' en fonction du taux de reconnaissance brut, pour les six valeurs de k/q (inverse du taux de reconnaissance au hasard) représentées dans les résultats après regroupement des deux évaluations perceptives de stimuli générés par projection dimensionnelle partielle de productions du locuteur M2. La valeur π'=0.5 correspond au niveau de reconnaissance au hasard.

Les valeurs de π' pour chaque émotion et chaque condition de resynthèse sont présentées dans la Table 17. La première observation qui peut être tirée de ces résultats est que les valeurs obtenues en condition de contrôle pour chacune des émotions lors des deux évaluations perceptives successives sont proches, avec néanmoins une valeur de π' et une intensité légèrement supérieures pour l'évaluation des stimuli générés avec ARX-LF (π'=0.82 vs. 0.80 avec Praat, avec des intensités moyennes respectives de 6.2 et 5.7), tandis que des variations plus importantes peuvent être observées entre les résultats des deux évaluations perceptives en conditions *F0 et intensité* et *qualité de voix et durée*.

L'indice π permet non seulement une interprétation directe relativement aisée, mais également d'évaluer statistiquement les différences observées dans le cadre de méta-analyses. Rosnow & Rosenthal (2003) soulignent ainsi que cet indice est particulièrement adapté à la comparaison de performances de catégorisation lorsque le nombre de choix proposés aux sujets est variable.

Ces tests statistiques s'appuient sur l'évaluation de l'erreur standard associée à chaque π_i calculé. Cette erreur standard est définie par Rosenthal & Rubin (1989) comme :

$$SE_{(\pi)i} = \frac{\pi_i.(1-\pi_i)}{\sqrt{N.P_i.(1-P_i)}}$$

Il est alors possible d'évaluer la significativité de la différence globale entre les valeurs moyennes de π dans m études indépendantes au moyen d'un test du χ^2 défini comme :

$$\chi^2(m-1) = \sum_{i=1}^{m} (\frac{\pi_i - \overline{\pi}}{SE_{(\pi)i}})^2 \text{, où } \overline{\pi} = \frac{\sum w_i \pi_i}{\sum w_i} \text{, et } w_i = \frac{1}{(SE_{(\pi)i})^2}.$$

émotion		contrôle	F0+int	QV+dur	F0	int	durée	source	src+rés	QV
joie	P	.89 (4.3)	.72 (2.7)	.20 (1.9)	.67 (2.3)	.18 (1.6)				
	A	.96 (5.3)	.91 (3.1)	.76 (3.0)			.00 (2.7)	.23 (2.4)	.23 (2.4)	.46 (2.0)
satisfaction	P	1 (8.3)	.98 (6.3)	.72 (2.4)	.98 (5.5)	.28 (1.9)				
	A									
anxiété	P	.92 (5.7)	.66 (3.1)	.92 (5.3)	.63 (3.0)	.32 (2.4)				
	A	.97 (6.1)	.72 (3.2)	.93 (5.9)			.90 (3.9)	.50 (2.9)	.64 (3.2)	.78 (3.8)
inquiétude	P	.71 (4.1)	.43 (3.6)	.65 (2.9)	.28 (4.2)	.24 (2.0)				
	A	.70 (5.9)	.61 (5.3)	.64 (4.1)			.39 (3.3)	.47 (2.2)	.60 (3.1)	.64 (3.7)
déception	P	.89 (6.6)	.43 (1.4)	.67 (2.6)	.60 (1.8)	.54 (1.5)				
	A	.90 (6.5)	.36 (1.7)	.89 (6.3)			.69 (5.3)	.39 (1.3)	.58 (2.1)	.62 (2.4)
résignation	P	.60 (3.7)	.30 (1.9)	.41 (1.6)	.40 (2.1)	.48 (1.5)				
	A	.39 (4.5)	.19 (2.4)	.29 (4.1)			.53 (3.1)	.66 (2.1)	.53 (1.8)	.45 (2.2)
tristesse	P	.61 (6.0)	.32 (2.0)	.79 (4.3)	.60 (3.4)	.55 (1.4)				
	A	.68 (7.1)	.57 (4.1)	.68 (6.4)			.82 (5.5)	.57 (2.5)	.59 (2.9)	.63 (5.3)
dégoût	P	.93 (7.1)	.22 (3.0)	.81 (6.2)	.12 (3.6)	.36 (1.6)				
	A	.94 (7.7)	.09 (2.9)	.84 (6.7)			.87 (6.6)	.09 (2.8)	.28 (2.1)	.38 (1.6)
neutre	P	.79								
	A	.89								

Table 17 : Valeurs de π' issues des taux de reconnaissance bruts (et jugements moyens d'intensité entre parenthèses) pour chaque émotion et condition de resynthèse, dans les deux évaluations perceptives successives (stimuli générés à partir de productions du locuteur M2, respectivement avec le module TD-PSOLA de Praat et à l'aide de l'algorithme ARX-LF). Les données non disponibles sont indiquées par des cases grisées. Conditions : F0+int=*F0 et intensité* ; QV+dur=*qualité de voix et durée* ; F0=*F0 seule* ; int=*intensité seule* ; durée=*durée seule* ; source=*source seule* ; src+rés=*source et résidu* ; QV=*qualité de voix*. Méthode d'analyse/synthèse utilisée : P=*Praat* ; A=*ARX-LF*.

Afin de tester la significativité de tout contraste entre valeurs de π obtenues dans une même étude ou provenant d'études distinctes, Rosenthal & Rubin (1989) établissent de plus l'équation permettant de calculer la valeur de la distribution normale centrée réduite qui correspond à ce contraste :

$$Z = \frac{\sum \lambda_i . \pi_i}{\sqrt{\sum \lambda_i{}^2 . SE_{(\pi)i}{}^2}}$$, où les coefficients λ_i dont la somme est nulle définissent le contraste entre les π_i.

La comparaison, restreinte aux conditions et émotions directement comparables (c'est-à-dire les conditions *F0 et intensité* et *qualité de voix et durée* et la condition de contrôle, pour toutes les émotions à l'exception de la satisfaction), des valeurs moyennes de π' correspondant à la reconnaissance des stimuli générés avec Praat (0.66) vs. ARX-LF (0.68)

ne permet pas de conclure à une différence significative entre ces deux évaluations ($\chi^2(1)=0.380$; $p=0.537$ n.s.).

En revanche, un test pour échantillons indépendants limités aux mêmes conditions et émotions à l'exception de l'expression neutre à laquelle nous considérons que sont associés des jugements d'intensité nuls, révèle que les jugements moyens d'intensité d'émotion attribués aux stimuli générés avec ARX-LF sont plus élevés que ceux attribués aux stimuli générés avec Praat (4.9 vs. 3.8 en moyenne, $t(1363)=-7.6$; $p<0.001$). Cette différence pourrait s'expliquer par l'impression de plus grand naturel donnée par les stimuli générés avec ARX-LF en conditions de projection partielle, comparés à ceux générés avec Praat. Toutefois, si tel était le cas, les jugements d'intensité en condition de contrôle devraient être d'un niveau comparable. Or, en condition de contrôle, ces jugements d'intensité sont également plus élevés pour les stimuli générés avec ARX-LF (6.2 en moyenne vs. 5.3 pour ceux générés avec Praat ; $t(453)=-3.8$; $p<0.001$).

Nous n'avons pas évalué de façon systématique auprès d'auditeurs naïfs la possibilité de discriminer les stimuli naturels de ceux resynthétisés en condition de contrôle. Toutefois une écoute attentive et répétée ne nous a pas permis de déceler la moindre différence perceptible entre les stimuli naturels et ceux générés en condition de contrôle quel que soit le système de synthèse utilisé. En conséquence, il semble raisonnable de considérer l'influence sur les jugements d'intensité des différences acoustiques entre stimuli naturels et resynthétisés en condition de contrôle comme très limitée sinon nulle.

Il paraît donc raisonnable supposer les différences de jugements d'intensité plus largement dues à l'absence de l'expression de satisfaction parmi les stimuli générés par ARX-LF évalués, notamment en condition de contrôle. En effet, les sujets auxquels il est demandé de donner à l'aide d'une échelle de type Likert des jugements subjectifs relatifs à un ensemble de stimuli ont tendance à répartir les scores attribués relativement aux valeurs extrêmes de l'échelle, quelle que soit la valeur intrinsèque des stimuli présentés (voir par exemple Cohen (1937)). Il est donc probable que les valeurs les plus élevées des jugements d'intensité attribués l'aient été en fonction des expressions les plus intenses parmi celles présentées, et donc que l'absence de l'expression de satisfaction, à laquelle les intensités les plus élevées ont été attribuées en condition de contrôle et en condition *F0 et intensité* lors de l'évaluation des stimuli générés avec Praat, ait largement contribué à relever les intensités attribuées aux stimuli générés avec ARX-LF.

Au sein de l'ensemble des stimuli présents dans chacune des évaluations perceptives, nous avons comparé les valeurs de π' obtenues pour une émotion et une condition données, selon la méthode de calcul des contrastes proposée par Rosenthal & Rubin (1989). Les résultats de ces comparaisons par paire sont présentés dans la Table 18. Nous

avons également comparé pour les mêmes paires les jugements moyens d'intensité attribués au moyen de tests t pour échantillons indépendants (63 degrés de liberté). Les résultats de ces comparaisons sont présentés dans la Table 19.

émotion	contrôle			F0 et intensité			qualité de voix et durée		
	Z	p	Sig.	Z	p	Sig.	Z	p	Sig.
joie	-1.73	0.0417	*	-2.58	0.0049	**	-4.30	0.0000	***
anxiété	-1.25	0.1053	n.s.	-0.57	0.2828	n.s.	-0.35	0.3646	n.s.
inquiétude	0.11	0.4547	n.s.	-1.32	0.0933	n.s.	0.10	0.4602	n.s.
déception	-0.13	0.4465	n.s.	0.51	0.3045	n.s.	-2.53	0.0058	**
résignation	1.55	0.0607	n.s.	0.89	0.1880	n.s.	0.86	0.1954	n.s.
tristesse	-0.58	0.2817	n.s.	-1.86	0.0317	*	1.11	0.1341	n.s.
dégoût	-0.51	0.3062	n.s.	0.64	0.2613	n.s.	-0.44	0.3308	n.s.
neutre	-1.41	0.0786	n.s.						

Table 18 : Comparaison des valeurs de π' correspondant aux taux de reconnaissance après regroupement de catégories des stimuli générés avec Praat vs. ARX-LF, pour chaque émotion dans les trois conditions de resynthèse présentes dans les deux évaluations. Les niveaux de significativité sont indiqués dans la colonne Sig. : ***=p<0.001, **=p<0.01, *=p<0.05, n.s.=non significatif.

émotion	contrôle			F0 et intensité			qualité de voix et durée		
	t(63)	p	Sig.	t(63)	p	Sig.	t(63)	p	Sig.
joie	-1.90	0.0618	n.s.	-0.70	0.4846	n.s.	-2.73	0.0082	**
anxiété	-0.92	0.3621	n.s.	-0.26	0.7962	n.s.	-0.99	0.3279	n.s.
inquiétude	-4.15	0.0001	***	-3.48	0.0009	***	-2.50	0.0149	*
déception	0.16	0.8766	n.s.	-0.79	0.4319	n.s.	-6.75	0.0000	***
résignation	-1.77	0.0822	n.s.	-1.21	0.2312	n.s.	-5.71	0.0000	***
tristesse	-2.22	0.0298	*	-4.73	0.0000	***	-4.00	0.0002	***
dégoût	-1.24	0.2178	n.s.	0.18	0.8559	n.s.	-0.84	0.4050	n.s.

Table 19 : Comparaison des jugements moyens d'intensité d'émotion attribués aux stimuli générés avec Praat vs. ARX-LF, pour chaque émotion dans les trois conditions de resynthèse présentes dans les deux évaluations. Les niveaux de significativité sont indiqués dans la colonne Sig. : ***=p<0.001, **=p<0.01, *=p<0.05, n.s.=non significatif.

En condition de contrôle, seule l'expression de joie est mieux reconnue (p<0.05) dans l'évaluation des stimuli générés avec ARX-LF que dans celle des stimuli générés avec Praat. Les projections partielles de cette expression de joie en conditions *F0 et intensité* et *qualité de voix et durée* sont également mieux reconnues lorsqu'elles sont générées avec ARX-LF qu'avec Praat (respectivement p<0.01 et p<0.001). Parmi ces stimuli, seule l'expression de joie resynthétisée en condition *qualité de voix et durée* se voit attribuer des jugements d'intensité plus élevés lorsque cette expression est générée par ARX-LF que par Praat (p<0.01).

En condition *qualité de voix et durée*, l'expression de déception resynthétisée par ARX-LF est également mieux reconnue qu'avec Praat (p<0.01), avec des jugements d'intensité

plus élevés ($p<0.001$). Enfin en condition *F0 et intensité*, l'expression de tristesse resynthétisée par ARX-LF est significativement mieux reconnue que lorsqu'elle est générée avec Praat ($p<0.01$). Rappelons néanmoins qu'aucun de ces deux stimuli n'est reconnu à un niveau supérieur au hasard.

Une explication possible de la différence de valeurs de π' pour la reconnaissance des expressions de joie en condition de contrôle pourrait être liée aux propriétés acoustiques des stimuli correspondants, l'expression de joie étant la seule pour laquelle la resynthèse par ARX-LF induit une variation de l'amplitude des trois premiers formants supérieure à 1 dB. Etant donné que ces variations semblent avoir un impact perceptif très limité voire nul, il paraît peu vraisemblable qu'elles induisent une telle variation des valeurs de π'. Un autre facteur de variation non contrôlée possible pourrait être l'échantillonnage des groupes de sujets recrutés, pour lesquels nous n'avons pas procédé à une évaluation par questionnaires de traits de personnalité pouvant être liés à la notion de quotient émotionnel. Toutefois, à moins de postuler une variabilité interindividuelle spécifique à la perception d'expressions de joie ou d'expressions positives, une telle hypothèse impliquerait également une variabilité des valeurs de π' correspondant à la reconnaissance des autres émotions.

Bien que le recodage des taux de reconnaissance bruts en valeurs de π' soit supposé compenser les variations du nombre d'alternatives de réponses présentées aux auditeurs, il est donc vraisemblable que la différence entre les niveaux de reconnaissance de l'expression de joie en condition de contrôle découle de l'absence de la satisfaction parmi les expressions évaluées et parmi les alternatives de réponses proposées.

En effet, l'expression de satisfaction resynthétisée en condition de contrôle avec Praat étant mieux reconnue que l'expression de joie, et avec une intensité supérieure, on pourrait supposer que le contraste entre cette expression de satisfaction et l'expression de joie produite de façon moins intense par le locuteur que nous avons sélectionné induise une reconnaissance moindre de l'expression de joie. Bien que nous ne les ayons pas comparés statistiquement, les patterns de confusion très similaires observés pour l'expression de joie entre les conditions de contrôle des deux évaluations ne vont toutefois pas dans le sens de cette interprétation.

Une autre interprétation possible, qui serait compatible avec les taux de reconnaissance bruts de l'expression de joie en condition de contrôle, est que les émotions reconnues seraient d'abord discriminées de façon binaire par les auditeurs en fonction de leur valence avant l'attribution d'une catégorie émotionnelle plus fine. Si tel était le cas, le niveau du hasard pour le choix d'une étiquette quelconque parmi les réponses positives proposées (c'est-à-dire l'étiquette « joie » ou « satisfaction » dans le cas de l'évaluation

des stimuli générés avec Praat, et l'étiquette « joie » pour l'évaluation de ceux générés avec ARX-LF) devrait être fixé dans tous les cas à 50% comme le suggère Ekman (1994) pour les expressions faciales, ou à 33% si l'on considère qu'un choix au hasard se fait de manière équiprobable entre émotion positive, émotion négative et expression neutre, plutôt que de considérer comme nous le faisons le niveau du hasard comme directement lié au nombre d'alternatives de réponses proposées aux auditeurs. Rappelons cependant que, comme le relèvent Scherer *et al.* (2003), les confusions fréquentes dans la littérature entre expressions vocales de valence opposée, et notamment entre expressions de tristesse et de joie, plaident pour une interprétation moins tranchée dans le cas des expressions vocales. Il convient de surcroît de souligner que cette proposition d'Ekman (1994) a été formulée en réponse aux critiques de Russel (1994) quant au caractère universel des expressions faciales, avec l'objectif de montrer que, quand bien même le niveau théorique du hasard est fixé au niveau le plus élevé pouvant être considéré, le taux de catégorisation correcte d'expression faciales par des sujets naïfs reste nettement supérieur à ce niveau quelle que soit leur culture.

Au-delà de cette question non tranchée de la définition du niveau théorique du hasard en tenant compte ou non de la valence des alternatives de réponse proposées, les différences de reconnaissance entre les versions resynthétisées par Praat et ARX-LF de l'expression de joie en condition *qualité de voix et durée* semblent en majeure partie dues à la présence, discutée ci-dessus avec l'analyse des résultats de la seconde évaluation, d'un bruit de bouche en amorce du stimulus généré par ARX-LF.

La reconnaissance ainsi que l'intensité moyenne attribuée supérieures en condition *F0 et intensité* pour le stimulus généré avec ARX-LF dans le cas de l'expression de tristesse semblent pouvoir s'expliquer principalement par la méthode de transplantation des contours de F0 et d'intensité adoptée. En effet, tandis que les contours de F0 projetés avec Praat sont stylisés au préalable, les valeurs de F0 extraites pour chaque trame sont directement appliquées au stimulus cible au moyen d'une interpolation linéaire lorsque les stimuli synthétiques sont générés avec ARX-LF. L'expression de tristesse présente une variation de basse fréquence de la fréquence fondamentale, responsable de la perception d'une voix « tremblante ». Bien que cette variation ne soit pas à proprement parler périodique, elle se rapproche de la définition du tremblement vocal (*vocal tremor*) comme modulation de la fréquence fondamentale de fréquence inférieure à 12 Hz (Kreiman *et al.*, 2003), et se traduit par une valeur élevée du jitter (6.5%).

Bien que la stylisation effectuée préalablement à la projection par Praat préserve ces variations de fréquence fondamentale de basse fréquence, ne neutralisant qu'une partie des variations cycle-à-cycle, le processus de resynthèse par Praat semble lui-même

responsable d'un lissage plus important qui a pour effet de neutraliser en partie le tremblement vocal du stimulus naturel. Ainsi, le jitter mesuré sur l'expression de tristesse projetée par Praat en condition *F0 seule* présente une valeur de jitter de 3.2%, plus faible que celle du stimulus original. En revanche l'interpolation linéaire des variations de fréquence fondamentale mesurées sur le stimulus original effectuée par l'algorithme de synthèse par ARX-LF reproduit plus fidèlement les variations de fréquence fondamentale et permet de préserver la majeure partie de l'effet perceptif de tremblement vocal, ce qui se traduit par un jitter de 5.3%.

L'atténuation de l'effet de tremblement vocal dans le stimulus généré par Praat en condition *F0 et intensité* semble cependant plus largement lié à la méthode retenue pour la projection des contours d'intensité. En effet, dans le cas de cette expression de tristesse, l'application du contour relatif d'intensité dans le processus de resynthèse par Praat en condition *F0 et intensité* a pour effet de ne reproduire que partiellement le voisement présent sur l'expression originale de tristesse, contrairement à la resynthèse en condition *F0 seule*. En effet, le contour d'intensité de l'expression d'origine présente un minimum d'intensité à 60 dB au niveau des deux derniers tiers de la consonne [b]. Or, l'application du contour d'intensité relatif suivi d'un rééchelonnage de l'amplitude afin de conserver un niveau global d'intensité équivalent à celui du stimulus d'origine a pour effet de produire dans cette région un minimum d'énergie à 46 dB, ne permettant pas de percevoir le voisement initialement présent et donc le tremblement vocal supposé avoir été projeté. Cette disparition d'une partie voisée porteuse de tremblement vocal a pour effet d'atténuer davantage l'effet de tremblement vocal perçu, avec pour résultat un jitter mesuré de 2.9%, comparable à celui mesuré pour l'expression neutre sur l'énoncé [sabl] (2.3%).

Les contours bruts de fréquence fondamentale de l'expression de tristesse naturelle sur l'énoncé [sabl] et des projections en condition F0 et intensité par Praat et ARX-LF, ainsi que celui de la projection par Praat en condition F0 seule, sont illustrés Figure 44.

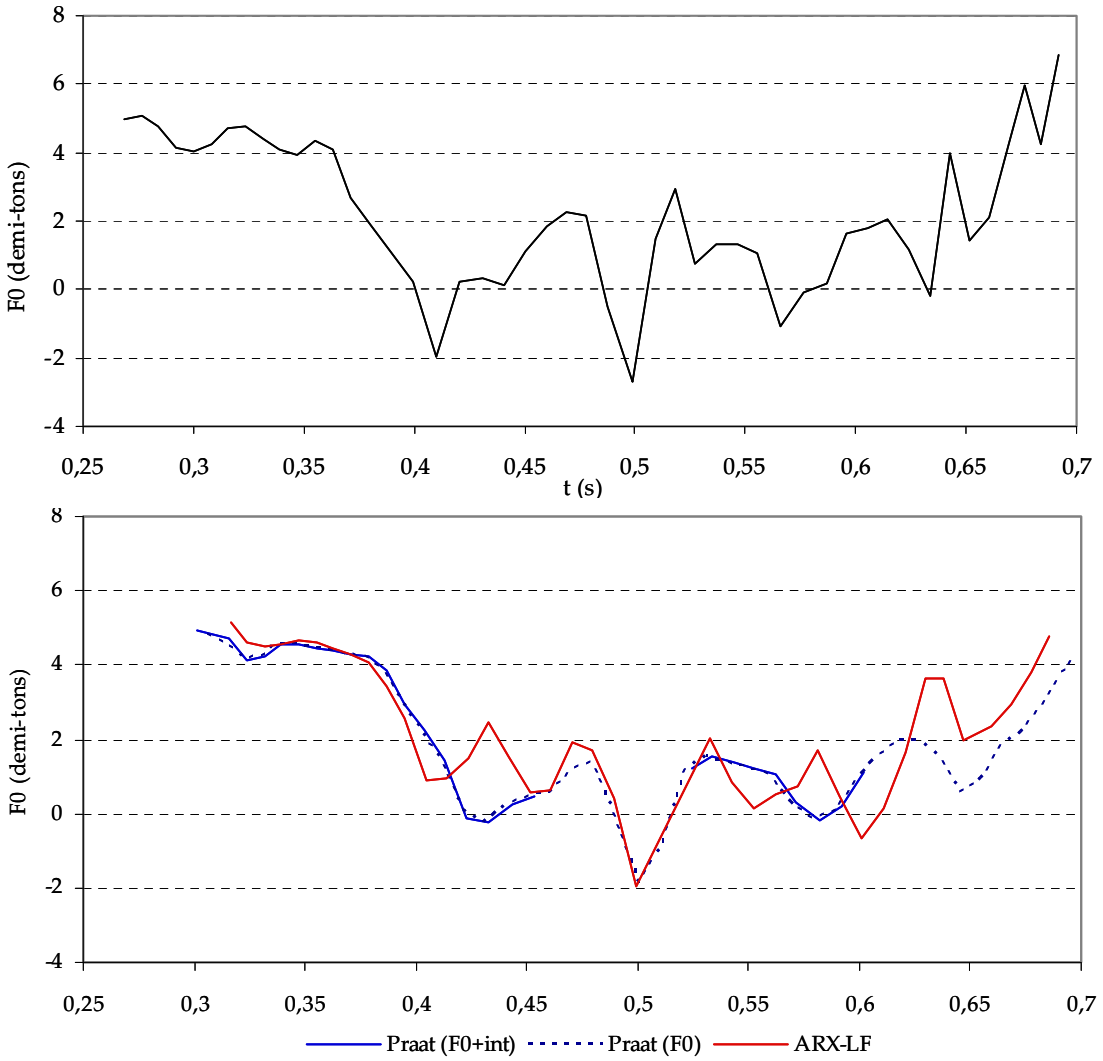


Figure 44 : Contour brut de fréquence fondamentale en demi-tons de l'expression de tristesse par le locuteur M2 sur l'énoncé [sabl] (partie supérieure) et résultat de la projection de ce contour sur l'expression neutre produite par le locuteur M2 sur le même énoncé, par Praat et ARX-LF (partie inférieure).

Au-delà du cas particulier de l'expression de tristesse détaillée ci-dessus, les différences observées entre les deux évaluations perceptives en conditions *F0 et intensité* et *qualité de voix et durée* sont vraisemblablement liées pour partie à la méthode utilisée pour la transplantation de contours d'intensité avec Praat. En conséquence, et bien que l'intensité ne semble jouer un rôle dans l'expression d'affects que lorsqu'elle est combinée avec d'autres dimensions, ce qui rejoindrait les résultats obtenus par Lieberman & Michaels (1962), on ne peut exclure que ces observations soient biaisées par les artefacts issus de la méthode de transplantation utilisée. Etant donné que les stimuli générés en condition de contrôle sont reconnus à un niveau comparable en condition de contrôle, une possibilité néanmoins pour évaluer le rôle de l'intensité combinée à la fréquence fondamentale est de comparer les valeurs de π' correspondant à la

reconnaissance des stimuli générés avec Praat en condition *F0 seule*, et celles correspondant à la reconnaissance des stimuli générés avec ARX-LF en condition *F0 et intensité*. Les résultats de ces comparaisons sont présentés dans la Table 20.

émotion	π' F0 seule (Praat)	π' F0 + intensité (ARX-LF)	Z	p	Sig.
anxiété	0.63	0.72	-0.86	0.1952	n.s
déception	0.60	0.36	1.86	0.0312	*
dégoût	0.12	0.09	0.17	0.4308	n.s
inquiétude	0.28	0.61	-2.29	0.0109	*
joie	0.67	0.91	-2.99	0.0014	**
résignation	0.40	0.19	1.68	0.0463	*
tristesse	0.60	0.57	0.24	0.4070	n.s

Table 20 : Comparaison des valeurs de π' dérivées des taux de reconnaissance des stimuli générés avec Praat en condition *F0 seule* et de celles correspondant à la reconnaissance des stimuli générés avec ARX-LF en condition *F0 et intensité*.

Etant donné que ces résultats correspondent à l'évaluation par deux groupes d'auditeurs différents de stimuli resynthétisés selon deux méthodes distinctes, ils doivent être interprétés avec prudence. Cependant quelques tendances semblent se dessiner. L'ajout de l'information d'intensité paraît ainsi améliorer la reconnaissance de l'expression d'intensité, de même que de celle de joie. Le gain observé dans le cas de l'expression de joie doit cependant être nuancé, du fait de l'hypothèse qui ne peut être exclue d'emblée que les expressions de joie et de satisfaction soient en premier lieu discriminées des expressions d'émotions négatives. Si l'on suppose, à la suite d'Ekman (1994), que les expressions vocales d'émotions positives et négatives sont discriminées de façon binaire par des indices aussi forts que l'action du zygomatique majeur pour les expressions faciales, ce qui peut porter à controverse, le gain de reconnaissance lié à l'ajout de l'information d'intensité n'est alors plus significatif. En effet les valeurs de π' alors obtenues sont de 0.37 en condition *F0 seule* pour la projection avec Praat et de 0.59 en condition *F0 et intensité* pour la projection avec ARX-LF (Z=-0.19, p=0.424 n.s.). La baisse du taux de reconnaissance lors de l'ajout de l'information d'intensité dans les cas des expressions de déception et de résignation, à laquelle nous n'avons trouvé d'explication satisfaisante, est en revanche plus surprenante.

9.2.7 Discussion

Les stimuli que nous avons générés en faisant varier un par un les paramètres prosodiques d'expressions émotionnelles sélectionnées, et fait évaluer perceptivement par des auditeurs « naïfs » afin de déterminer les poids relatifs des dimensions

manipulées dans l'expression d'affects, constituent des artefacts. En effet, les expressions émotionnelles naturelles utilisées comme référence pour la projection dimensionnelle résultent d'un contrôle global du conduit vocal. Il est difficile d'imaginer produire naturellement des expressions dans lesquelles seules certaines dimensions acoustiques varient en fonction de l'émotion exprimée tandis que les autres dimensions correspondent à une prosodie neutre.

Les auditeurs se sont toutefois montrés capables d'identifier des traits saillants de cette prosodie artefactuelle, en parvenant à reconnaître à un niveau supérieur au hasard bon nombre des stimuli générés par projection sur une expression neutre les variations liées à l'expression émotionnelle. Si aucune dimension n'apparaît comme dédiée spécifiquement à l'expression d'émotions, une tendance générale peut être extraite de ces résultats : comparativement aux expressions neutres que nous considérons comme référence, les émotions positives semblent plus largement portées par les variations de fréquence fondamentale, tandis que les émotions négatives le sont plutôt par la qualité de voix et la répartition des phénomènes de durée.

Ainsi, les auditeurs sont largement parvenus à identifier les expressions de joie et de satisfaction présentées à partir des seules variations de fréquence fondamentale, caractérisées pour ces expressions par des contours montants puis descendants. De plus l'expression d'inquiétude qui présente un contour de fréquence fondamentale de forme similaire, est reconnue comme anxiété ou inquiétude en condition de contrôle à un niveau supérieur au hasard mais également largement confondue avec joie et satisfaction, et reconnue de façon non équivoque comme joie ou satisfaction lorsque ce contour est combiné avec les propriétés de qualité de voix, de durée et d'intensité de l'expression neutre. Les contours de fréquence fondamentale, et particulièrement les contours montants/descendants, semblent donc jouer un rôle prépondérant dans l'expression des émotions positives. Remarquons de plus que, si ces contours de même forme générale mais présentant des valeurs d'ancrage variables permettent tous trois de générer des expressions reconnues comme joie ou satisfaction à un niveau supérieur au hasard, le taux de reconnaissance ainsi que l'intensité perçue de l'émotion augmentent avec l'amplitude du contour.

Les émotions négatives que nous avons évaluées, et notamment les expressions d'anxiété, de déception et de tristesse, sont quant à elles transmises de façon conjointe par la qualité de voix et la répartition des phénomènes de durée, avec une prépondérance de l'influence de la durée, ici considérée comme équivalente aux durées phonémiques, dans le cas de l'expression d'anxiété. La répartition entre l'influence de la qualité de voix et celle de la durée est en revanche moins clairement tranchée dans le cas

de l'expression de déception et de celle de tristesse. Si les variations de fréquence fondamentale influent également sur la reconnaissance de la tristesse, notons néanmoins que cette influence n'est pas liée à la forme générale du contour de fréquence fondamentale mais à un phénomène de tremblement vocal, traditionnellement rattaché aux phénomènes de qualité de voix en tant que variation du second ordre de la fréquence fondamentale. L'expression de dégoût semble pour sa part transmise par un allongement important de la voyelle, les autres dimensions ne permettant pas une reconnaissance même faible de cette expression, mais apparaît également comme moins robuste que les autres expressions évaluées aux manipulations par projection dimensionnelle, avec un taux de reconnaissance moindre qu'en condition de contrôle lorsque seul cet allongement vocalique est réalisé. Cette apparente fragilité de l'expression du dégoût pourrait être en partie liée au fait que le dégoût soit considéré par divers auteurs (voir notamment Rozin (1987, 2000)) comme une réaction adaptative de rejet de nourriture impropre à la consommation, son expression étant en conséquence destinée à être perçue dans un périmètre restreint (Johnstone & Scherer (2000), cf. section 8.4. pour une discussion plus détaillée). Ainsi les expressions de dégoût sont généralement considérées dans la littérature comme mal identifiées en modalité audio seule (voir par exemple Juslin&Laukka, 2003). Il est en revanche plus difficile de tirer des conclusions claires des résultats de l'évaluation perceptive des projections dimensionnelles des expressions d'inquiétude et de résignation, plus faiblement reconnues en condition de contrôle.

Bien que certaines dimensions apparaissent comme prépondérantes pour l'expression de certaines émotions parmi celles que nous avons évaluées, comme c'est le cas de la fréquence fondamentale pour l'expression de la joie et de la satisfaction, et de l'allongement de la voyelle pour l'expression du dégoût et dans une moindre mesure de l'anxiété (en l'occurrence respectivement un allongement et un raccourcissement, en particulier du noyau vocalique), l'incapacité de ces dimensions utilisées seules à reproduire l'intégralité de l'information affective doit être soulignée. En effet, si la conjonction de la qualité de voix et de la durée permet de reconnaître l'expression d'anxiété et celle de déception à un niveau équivalent à celui des expressions originales tout en leur attribuant une intensité perçue comparable, ce qui s'explique par des variations de fréquence fondamentale et un niveau d'intensité proches des caractéristiques de l'expression neutre, aucune des dimensions manipulées seules ne permet d'induire un taux de reconnaissance et une intensité d'émotion équivalents au stimulus en condition de contrôle. Ce résultat est particulièrement important dans l'optique d'une modélisation prosodique pour la synthèse d'expressions émotionnelles : il met en effet l'accent sur la nécessité de prendre en considération l'ensemble des

dimensions acoustiques, y compris des dimensions pouvant paraître secondaires pour l'expression de l'émotion visée, afin de générer des expressions synthétiques qui ne soient plus des chimères comme celles que nous avons construites et évaluées dans cette expérience mais puissent être assimilées à des productions naturelles.

En raison d'artefacts liés à la méthode utilisée pour la transplantation des contours d'intensité avec Praat dans la première expérience, nous n'avons pas pu clairement établir le rôle joué par l'intensité dans l'expression des émotions étudiées. Cependant, les données dont nous disposons nous permettent d'émettre l'hypothèse que l'intensité seule est insuffisante pour permettre la reconnaissance des émotions exprimées, mais qu'elle est en mesure de renforcer l'identification en complément de l'information apportée par les autres dimensions prosodiques.

L'artefact de synthèse que nous avons mentionné ci-dessus, bien qu'ayant pour conséquence de limiter la portée de notre étude, a le mérite de mettre l'accent sur l'importance de la répartition de l'information au-delà de valeurs globales extraites de l'ensemble de l'énoncé. De plus l'artefact issu du prétraitement de l'algorithme ARX-LF, dont nous avons pu observer l'impact perceptif, souligne l'importance de la prise en compte de phénomène locaux et mineurs en termes de quantité d'information au sens de Shannon, qui au regard de leur taille peuvent avoir d'importantes conséquences sur la perception des affects exprimés.

Enfin, cette étude nous a permis de valider la pertinence de la modélisation ARX-LF de la qualité de voix appliquée à des échantillons de parole expressive. En effet, bien que ces différences soient de faible ampleur, les projections partielles ont été mieux reconnues et jugées comme exprimant une émotion plus intense lorsque les informations du filtre supposé modéliser les résonances du conduit vocal et le résidu permettant de reproduire les caractéristiques de la source non modélisées par LF ont également été prises en compte, comparativement aux cas dans lesquels seule la paramétrisation LF de la source glottique est considérée.

9.3. Quel timing pour le décodage des expressions émotionnelles ?

9.3.1 Introduction

Comme nous l'avons exposé brièvement en section 4.5. , le modèle prosodique d'Aubergé (1991, 2002, voir aussi Aubergé *et al.* (2006)) émet l'hypothèse que la prosodie expressive est véhiculée par des contours gradients multiparamétriques. La forme des contours (c'est-à-dire la dynamique des paramètres pertinents pour l'expression d'affects par la prosodie au sens large) serait ainsi porteuse d'information affective, au-delà d'une simple information de gradience. Autrement dit, les contours prosodiques transmettraient des informations sur la nature de l'affect, en plus des informations globales comme la valeur moyenne de la fréquence fondamentale ou son étendue. En cela, l'hypothèse d'Aubergé peut être considérée comme une extension du concept de « clichés mélodiques » proposé par Fónagy (1983) pour la fréquence fondamentale.

Si l'on considère que les expressions émotionnelles sont décodées au moyen d'un traitement gradient, il serait intéressant d'évaluer dans quelle mesure ce traitement permet une identification anticipée des valeurs émotionnelles, et le cas échéant de déterminer les indices gradients dans l'énoncé.

Afin d'avoir un accès direct au décours temporel du traitement cognitif de l'information portée par des sons de parole, la méthode la plus fiable consiste en la mesure des signaux neuraux impliqués dans ce traitement. De telles mesures peuvent être effectuées avec une résolution temporelle élevée par des méthodes comme l'électroencéphalographie (EEG). Cependant, outre la complexité de la mise en œuvre de telles méthodes, la mesure directe de la reconnaissance des valeurs d'affects exprimées par des sons de parole nécessiterait une modélisation beaucoup plus fine des mécanismes neuraux impliqués dans la reconnaissance d'expressions d'affects que ne le permettent les connaissances actuelles dans ce domaine.

Une approche possible de l'évaluation du décours temporel du décodage de l'information affective pourrait être de s'appuyer sur un paradigme dans lequel le temps de réaction des auditeurs est mesuré. Dans une telle tâche, il est demandé à ces derniers de répondre le plus rapidement possible, sans attendre la présentation de l'énoncé dans son intégralité. Dans le domaine des expressions émotionnelles, le temps de réaction des sujets a par exemple été appliqué par de Gelder & Vroomen (2000) dans leur étude du décodage d'expressions incongruentes entre face et prosodie, ici limitée aux contours de fréquence fondamentale. Un tel paradigme est adapté aux études dans lesquelles l'objectif est de déterminer quelles émotions parmi celles présentées sont reconnues le

plus rapidement. Cependant dans le cas qui nous intéresse ici, étant donné le temps de latence pour la réponse motrice du sujet il ne nous permettrait pas de localiser les indices saillants responsables de l'identification ni de suivre l'évolution au cours du temps du décodage. De plus l'application de ce paradigme nécessite pour obtenir des résultats exploitables que les sujets indiquent leurs réponses par un geste moteur simple, ce qui est difficilement applicable à une tâche de catégorisation dans laquelle un nombre relativement élevé d'alternatives est proposé.

9.3.2 Le paradigme du dévoilement progressif

Le paradigme du dévoilement progressif ou *gating* (Grosjean, 1980) consiste en une présentation graduelle de stimuli audio dans une tâche de jugement perceptif. Les stimuli originaux sont coupés à des points définis en fonction de durées absolues, généralement incrémentées selon un pas fixe, ou relatives aux unités linguistiques composant l'énoncé. Ce découpage définit ainsi un ensemble de fenêtres de dévoilement ou *gates*. L'analyse des réponses données par les auditeurs permet d'évaluer la quantité minimale d'information nécessaire pour prédire correctement la valeur de l'énoncé considéré dans son ensemble. L'usage classique de ce paradigme n'implique pas de contraindre les auditeurs à répondre le plus rapidement possible. En conséquence, des biais de réponses pourraient être observés, liés au fait que les réponses ne sont pas indiquées au moment où l'auditeur perçoit l'énoncé. Cependant, Tyler & Wessels (1985) concluent de l'analyse comparative des résultats obtenus pour les mêmes stimuli présentés graduellement avec l'instruction donnée ou non aux auditeurs de répondre le plus rapidement possible que le dévoilement progressif dans son usage courant constitue bien une tâche en ligne. Une autre critique d'ordre méthodologique de la tâche de dévoilement progressif pourrait être un possible effet d'entraînement des auditeurs dû au fait que, à l'exception de la première fenêtre de dévoilement, les stimuli présentés aux auditeurs sont une version « allongée » de stimuli déjà présentés au préalable. Afin d'évaluer ce possible effet, Cotton & Grosjean (1984) ont comparé les résultats obtenus dans une tâche classique de dévoilement progressif à ceux obtenus en présentant individuellement les stimuli tronqués. Ils concluent à une absence d'effet de la répétition des stimuli sur les formes reconnues, le seul effet de la répétition étant une légère augmentation des scores de confiance recueillis en parallèle.

Le dévoilement progressif a été le plus souvent utilisé dans le domaine de la psycholinguistique pour évaluer les compétences de prédiction des auditeurs dans des tâches d'accès lexical (voir notamment Grosjean (1996) pour un état de l'art), et dans une moindre mesure à la prédiction de valeurs linguistiques en fonction de contours intonatifs. Thorsen (1980) l'a ainsi appliqué à des énoncés danois afin d'évaluer la

prédictibilité de leur caractère déclaratif, tandis que Vion & Colas (2006) ont testé l'identification précoce de questions fermées en français. Quelques études dans lesquelles le dévoilement progressif a été appliqué à la prédiction de la valeur expressive de l'énoncé peuvent également être mentionnées.

Ainsi Aubergé *et al.* (1997) ont proposé une tâche de catégorisation de six attitudes prosodiques du français enregistrées et étudiées par Morlec (1997) exprimées sur des énoncés de deux à cinq syllabes, présentées graduellement selon un pas défini en fonction de la syllabe. L'évaluation des résultats perceptifs a montré une identification précoce à la 2e syllabe (taux de reconnaissance moyen de 51% à comparer au niveau du hasard de 1/6) de l'ensemble des attitudes et modalités évaluées, que les auteurs ont interprétée comme liée à l'identification d'indices saillants dans l'énoncé.

Plus récemment, Shochi (2008) a effectué une série d'expériences de dévoilement progressif d'attitudes et modalités japonaises exprimées sur des énoncés sémantiquement neutres longs de deux, cinq et huit mores, dans lesquelles la more (unité rythmique du japonais) était considérée comme unité de base. L'analyse de la catégorisation effectuée par les auditeurs lui a permis de montrer que les attitudes d'autorité et d'exclamation de surprise étaient identifiées dès la première more, ainsi que les deux plus formelles des trois types d'expressions de politesse codifiées par la langue japonaise en dépit de confusions mutuelles entre types de politesse. Il a interprété ce résultat comme en majeure partie lié aux qualités de voix et variations d'intensité suffisamment distinctes de celles des autres attitudes et rapidement perçues par les auditeurs. A l'opposé, l'attitude d'irritation et la question simple qu'il a évaluées n'ont été reconnues que grâce aux indices acoustiques localisés en fin d'énoncé, tandis que les autres attitudes et modalités ont été décodées de façon linéaire en fonction de la quantité d'information présentée.

Simultanément à l'étude que nous avons menée et qui est présentée dans cette section, Grichkovtsova *et al.* (2007) ont proposé à des auditeurs francophones une tâche de catégorisation d'expressions actées de joie, de colère, de tristesse, d'évidence, d'ironie et de doute comprenant des indices lexicaux localisés en fin d'énoncé, présentées graduellement par pas fixe de 200 millisecondes. Les auteurs ont conclu de l'analyse des réponses que les expressions émotionnelles évaluées, à l'exception de la joie, étaient reconnues plus rapidement que les expressions attitudinales.

Le paradigme du dévoilement progressif n'a pas à notre connaissance été utilisé jusqu'alors pour évaluer ce que perçoivent des auditeurs naïfs lorsque l'information minimale présentée est inférieure à la syllabe. Nous nous proposons ici de l'appliquer à l'ensemble des expressions émotionnelles que nous avons sélectionné sur des énoncés

monosyllabiques, afin d'évaluer si des traits saillants ou un traitement gradient peuvent également mener à une identification anticipée des valeurs affectives sur des unités inférieures à la syllabe. Kohler (1987, 1991) a montré sur l'allemand en faisant varier la position du pic de F0 qu'un même stimulus pouvait véhiculer diverses valeurs communicatives pragmatiques. Il a associé le pic précoce à la valeur « connaître », définie comme le point final de l'exposé d'un argument, le pic médial à la valeur « réaliser », définie comme le commencement d'un nouvel argument, et le pic tardif à la valeur « réaliser en contraste avec les attentes de l'interlocuteur ». Bien que ces valeurs pragmatiques ne soient pas directement transposables à la catégorisation d'expressions d'affects en français hors d'un contexte d'interaction bien défini, un premier cadre pour l'étude des valeurs émotionnelles perçues sur de petites unités peut donc être de découper les voyelles en trois parts égales et de comparer l'information véhiculée par ces différentes parts.

9.3.3 Sélection des stimuli de référence

Afin de ne traiter que des énoncés entièrement voisés, nous avons décidé de retenir comme données de référence pour cette expérience de dévoilement progressif les expressions par le locuteur M2 sur les énoncés [ʒon], [ʁuʒ] et [vɛʁ] des huit émotions sélectionnées (anxiété, déception, dégoût, inquiétude, joie, résignation, tristesse et satisfaction), complétées par une expression neutre sur chacun de ces énoncés.

La durée totale de ces stimuli varie de 370 millisecondes pour l'expression de résignation sur l'énoncé [ʁuʒ] à 941 millisecondes pour l'expression de dégoût sur l'énoncé [ʁuʒ] (587 ms en moyenne). La durée de l'attaque consonantique, du noyau vocalique et de la coda pour chacun des 27 stimuli sélectionnés ainsi que leur durée totale sont représentées par la Figure 45.

Les variations des durées en fonction de l'énoncé, testées au moyen d'analyses de variance à un facteur, sont non significatives : $F(2,24)=0.0$, $p=0.985$ n.s. pour l'énoncé dans son ensemble, $F(2,24)=1.1$, $p=0.362$ n.s. pour l'attaque consonantique, $F(2,24)=1.2$, $p=0.315$ n.s. pour le noyau vocalique et $F(2,24)=1.4$, $p=0.276$ n.s. pour la coda.

En revanche, à l'exception de la durée de l'attaque consonantique ($F(8,18)=1.1$, $p=0.387$ n.s.), les durées varient significativement en fonction de l'émotion : $F(8,18)=10.1$, $p<0.001$ pour l'énoncé dans son ensemble, $F(8,18)=12.2$, $p<0.001$ pour le noyau vocalique et $F(8,18)=3.7$, $p<0.05$ pour la coda.

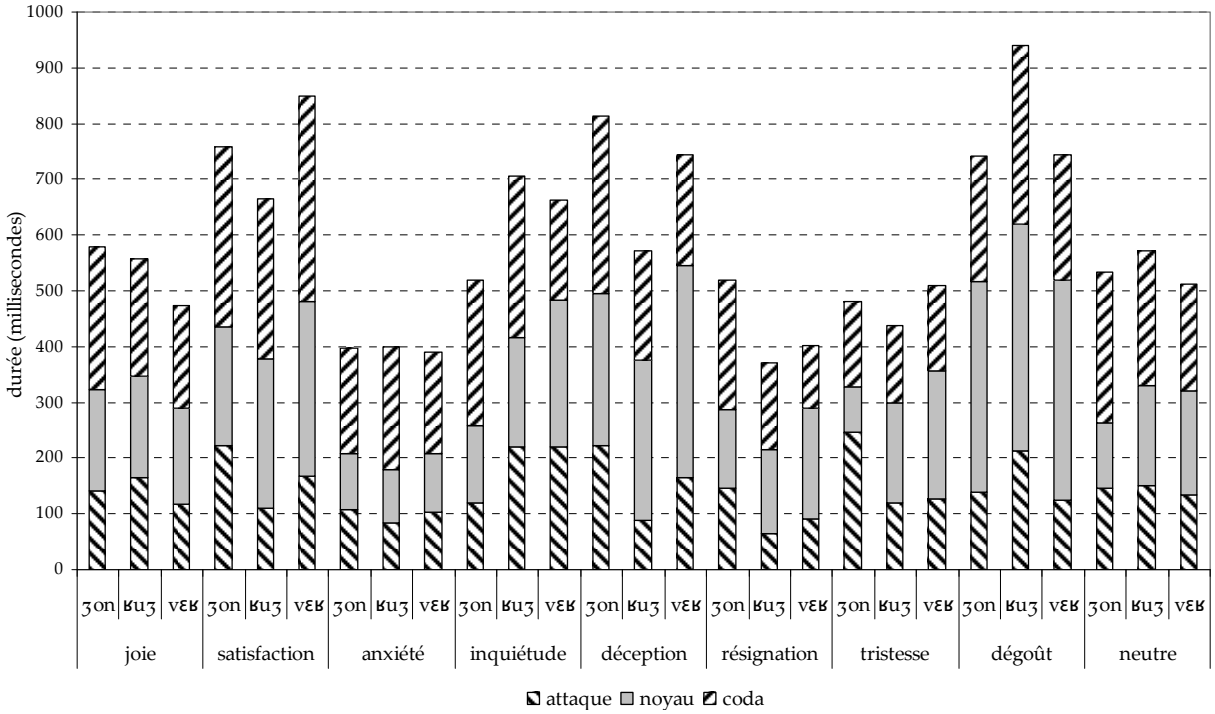


Figure 45 : Durée de l'attaque consonantique, du noyau vocalique et de la coda de chacun des stimuli sélectionnés parmi les productions du locuteur M2 pour l'expérience de dévoilement progressif. La durée totale de chaque stimulus est indiquée par les barres cumulées.

9.3.4 Découpage des stimuli pour le dévoilement progressif

Nous avons défini six fenêtres de dévoilement pour chaque stimulus, relativement aux frontières phonémiques étiquetées manuellement.

- Fenêtre 1 : du début du stimulus au premier tiers du noyau vocalique.
- Fenêtre 2 : du début du stimulus au deuxième tiers du noyau vocalique.
- Fenêtre 3 : du début du stimulus à la fin du noyau vocalique.
- Fenêtre 4 : du début du stimulus au premier tiers de la coda.
- Fenêtre 5 : du début du stimulus au deuxième tiers de la coda.
- Fenêtre 6 : du début du stimulus à la fin de la coda.

Les stimuli ont été tronqués et complétés par un bruit blanc de durée variable afin que tous les stimuli atteignent une durée totale de 1.25 ms et ainsi compenser un possible effet de la durée des stimuli, à l'aide du logiciel Praat (Boersma & Weenink, 1992-2008).

9.3.5 Evaluation perceptive

La présentation des stimuli ainsi que l'enregistrement des réponses ont été réalisés au moyen d'une interface automatisée développée à l'aide du logiciel *Runtime Revolution* et

présentée en Figure 46. Les juges devaient choisir une étiquette émotionnelle parmi les 8 proposées (anxiété, déception, dégoût, inquiétude, joie, résignation, tristesse ou satisfaction) ou opter pour l'étiquette *aucune émotion*. Contrairement aux méthodes utilisées auparavant pour l'évaluation des stimuli naturels en conditions audio et audiovisuel et pour l'évaluation des projections dimensionnelles, nous n'avons pas collecté de jugements d'intensité de l'émotion dans cette expérience. L'évaluation des capacités de prédiction par les auditeurs de l'intensité de l'émotion relative à l'ensemble de l'énoncé à partir de parties tronquées de cet énoncé pourrait fournir des informations intéressantes sur le décours temporel du décodage de l'information affective. Cependant, la catégorisation de la valeur affective portée par des stimuli tronqués, dans lesquels la durée de l'information utile présentée peut descendre jusqu'à 11.5 ms dans le cas de la première fenêtre de dévoilement des expressions d'anxiété et de résignation sur l'énoncé [ʁuʒ], nous a semblé constituer une tâche suffisamment complexe pour ne pas l'alourdir davantage. En conséquence l'interface de présentation des stimuli et d'enregistrement des réponses que nous avons élaborée ne propose qu'un choix multiple à l'aide de cases à cocher.

Les 162 stimuli générés ont été évalués perceptivement au laboratoire dans un environnement calme avec un casque fermé de bonne qualité par 20 juges francophones naïfs (13 femmes, 7 hommes, âgés de 30.6 ans en moyenne). Les stimuli étaient présentés par fenêtre de dévoilement de durée croissante afin d'éviter l'identification par des indices locaux insuffisants pour la reconnaissance des valeurs affectives d'expressions déjà présentées sous une forme moins réduite. Pour chaque durée de fenêtre, les stimuli étaient présentés dans un ordre aléatoire différent pour chaque juge et contrôlé afin qu'un même énoncé ne soit jamais présenté deux fois consécutives.

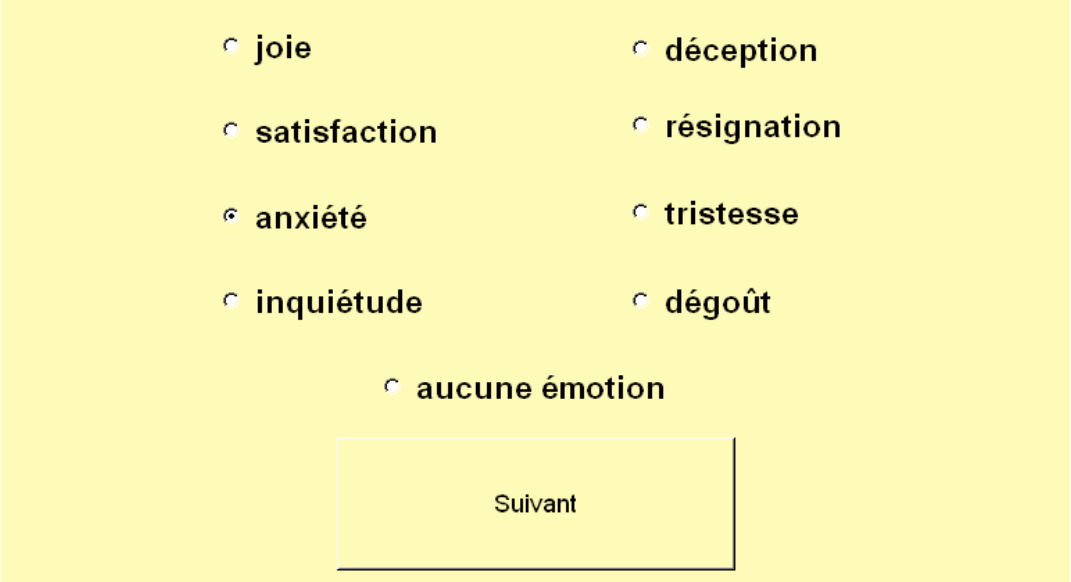


Figure 46 : Interface graphique utilisée pour l'évaluation des stimuli tronqués construits à partir de productions du locuteur M2 pour l'expériecne de dévoilement progressif.

9.3.6 Analyse des résultats

9.3.6.1 Matrices de confusion et regroupement d'étiquettes

Le premier niveau d'analyse a été l'extraction des matrices de confusion correspondant à chaque fenêtre de dévoilement. La matrice de confusion à la dernière fenêtre, qui correspond aux confusions résultant de la présentation des stimuli complets, est présentée dans la Table 21.

Afin de prendre en compte les principales confusions entre étiquettes présentées, nous avons effectué une classification hiérarchique ascendante sur les colonnes de la matrice de confusion à la dernière fenêtre de dévoilement. Les valeurs des colonnes sont normalisées par la somme de leurs éléments. Cette analyse est illustrée par le dendrogramme présenté Figure 47. Cette classification hiérarchique a été réalisée en considérant pour chaque émotion et chacune des colonnes de la matrice de confusion présentée en Table 21 la moyenne sur les énoncés [ʒon], [ʁuʒ] et [vɛʁ].

Le dendrogramme présenté Figure 47 indique que les principales confusions entre étiquettes à la dernière fenêtre de dévoilement sont semblables à celles observées précédemment. En conséquence, nous pouvons considérer les mêmes regroupements d'étiquettes. *Anxiété* et *inquiétude* seront donc considérées comme relevant d'une même classe d'émotions, *résignation*, *déception* et *tristesse* seront également regroupées, de même que *joie* et *satisfaction*, tandis que *dégoût* et *neutre* resteront des catégories séparées.

émotion	énoncé	joie	satisf.	anxiété	inq.	décep.	résign.	trist.	dégoût	neutre
joie	[ʒon]	**25%**	45%	0%	0%	0%	10%	5%	0%	15%
	[ʁuʒ]	**30%**	45%	5%	10%	0%	0%	0%	0%	10%
	[vɛʁ]	**15%**	20%	15%	25%	5%	5%	5%	0%	10%
satisfaction	[ʒon]	70%	**25%**	0%	0%	5%	0%	0%	0%	0%
	[ʁuʒ]	65%	**35%**	0%	0%	0%	0%	0%	0%	0%
	[vɛʁ]	60%	**40%**	0%	0%	0%	0%	0%	0%	0%
anxiété	[ʒon]	0%	5%	**35%**	35%	10%	5%	0%	0%	10%
	[ʁuʒ]	0%	0%	**55%**	15%	0%	10%	0%	5%	15%
	[vɛʁ]	5%	0%	**40%**	40%	5%	0%	5%	0%	5%
inquiétude	[ʒon]	0%	15%	20%	**25%**	0%	5%	10%	0%	25%
	[ʁuʒ]	0%	5%	10%	**65%**	10%	0%	10%	0%	0%
	[vɛʁ]	20%	5%	35%	**40%**	0%	0%	0%	0%	0%
déception	[ʒon]	0%	5%	0%	0%	**40%**	25%	20%	5%	5%
	[ʁuʒ]	0%	0%	0%	0%	**55%**	30%	10%	5%	0%
	[vɛʁ]	0%	0%	0%	0%	**35%**	25%	25%	15%	0%
résignation	[ʒon]	0%	5%	20%	10%	10%	**15%**	0%	0%	40%
	[ʁuʒ]	0%	10%	15%	5%	25%	**15%**	0%	0%	30%
	[vɛʁ]	0%	10%	30%	5%	10%	**25%**	5%	5%	10%
tristesse	[ʒon]	50%	15%	5%	10%	5%	0%	**15%**	0%	0%
	[ʁuʒ]	15%	45%	10%	10%	0%	0%	**15%**	0%	5%
	[vɛʁ]	50%	0%	30%	5%	0%	0%	**10%**	0%	5%
dégoût	[ʒon]	0%	25%	0%	0%	25%	15%	5%	**30%**	0%
	[ʁuʒ]	0%	15%	0%	0%	15%	25%	0%	**45%**	0%
	[vɛʁ]	0%	5%	0%	0%	10%	5%	0%	**80%**	0%
neutre	[ʒon]	0%	0%	0%	0%	20%	10%	25%	0%	**45%**
	[ʁuʒ]	0%	0%	10%	10%	20%	15%	10%	0%	**35%**
	[vɛʁ]	0%	5%	0%	5%	15%	15%	20%	0%	**40%**

Table 21 : Matrice de confusion à la dernière fenêtre de dévoilement (stimuli complets) dans l'expérience de dévoilement progressif d'expressions sélectionnées parmi les productions du locuteur M2. Les taux d'identification correcte apparaissent en gras.

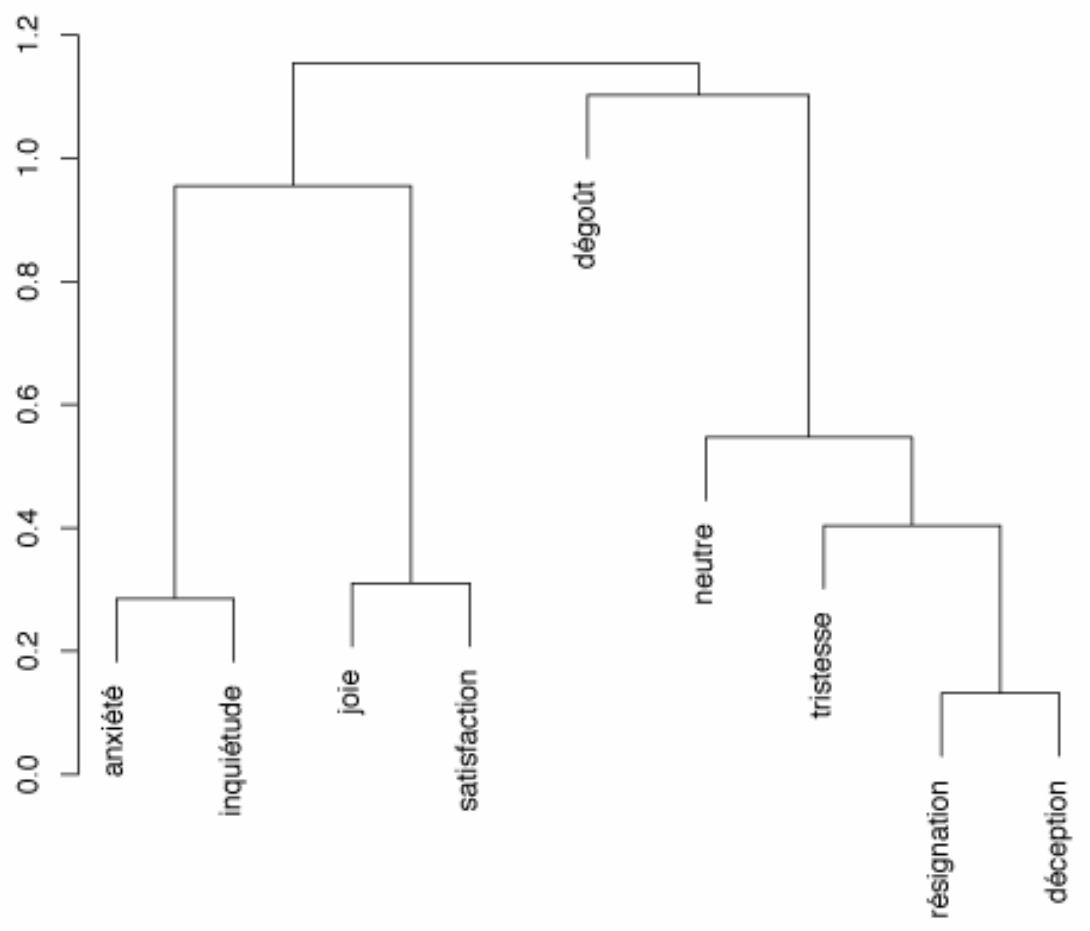


Figure 47 : Dendrogramme issu de la classification hiérarchique (critère d'agrégation de Ward) opérée sur les colonnes normalisées de la matrice de confusion correspondant à la dernière fenêtre de présentation (productions originales du locuteur M2 tronquées à la fin de la coda) dans l'expérience de dévoilement progressif.

La Table 22 présente la matrice de confusion à la dernière fenêtre de dévoilement, après regroupement. Bien que les distributions des résultats de ces deux études ne puissent être directement comparées par des tests statistiques standards, les scores d'identification et les confusions entre étiquettes sont proches de ceux observés dans l'étude de projection dimensionnelle présentée en section 9.2. Le nombre plus important d'exemplaires utilisés pour chaque émotion dans l'étude présente et le fait que les deux ensembles ne se recouvrent que partiellement peuvent expliquer les quelques divergences qui ont pu être observées. En dépit de la similarité prosodique apparente de la plupart des expressions actées d'une émotion donnée produites par le locuteur M2, qui constitue l'une des raisons qui nous ont conduites à sélectionner les productions de ce locuteur, certaines expressions présentent en effet une morphologie prosodique plus variable d'un énoncé à l'autre.

émotion	énoncé	joie-satisf.	anx.-inq.	déc.-res.-tri.	dégoût	neutre
hasard		*22%*	*33%*	*22%*	*11%*	*11%*
joie	[ʒon]	**70%**	0%	15%	0%	15%
	[ʁuʒ]	**75%**	15%	0%	0%	10%
	[vɛʁ]	**35%**	40%	15%	0%	10%
satisfaction	[ʒon]	**95%**	0%	5%	0%	0%
	[ʁuʒ]	**100%**	0%	0%	0%	0%
	[vɛʁ]	**100%**	0%	0%	0%	0%
anxiété	[ʒon]	5%	**70%**	15%	0%	10%
	[ʁuʒ]	0%	**70%**	10%	5%	15%
	[vɛʁ]	5%	**80%**	10%	0%	5%
inquiétude	[ʒon]	15%	**45%**	15%	0%	25%
	[ʁuʒ]	5%	**75%**	20%	0%	0%
	[vɛʁ]	25%	**75%**	0%	0%	0%
déception	[ʒon]	5%	0%	**85%**	5%	5%
	[ʁuʒ]	0%	0%	**95%**	5%	0%
	[vɛʁ]	0%	0%	**85%**	15%	0%
résignation	[ʒon]	5%	30%	**25%**	0%	40%
	[ʁuʒ]	10%	20%	**40%**	0%	30%
	[vɛʁ]	10%	35%	**40%**	5%	10%
tristesse	[ʒon]	65%	15%	**20%**	0%	0%
	[ʁuʒ]	60%	20%	**15%**	0%	5%
	[vɛʁ]	50%	35%	**10%**	0%	5%
dégoût	[ʒon]	25%	0%	45%	**30%**	0%
	[ʁuʒ]	15%	0%	40%	**45%**	0%
	[vɛʁ]	5%	0%	15%	**80%**	0%
neutre	[ʒon]	0%	0%	55%	0%	**45%**
	[ʁuʒ]	0%	20%	45%	0%	**35%**
	[vɛʁ]	5%	5%	50%	0%	**40%**

Table 22 : Matrice de confusion à la dernière fenêtre de dévoilement (stimuli complets) dans l'expérience de dévoilement progressif d'expressions sélectionnées parmi les productions du locuteur M2. Les taux d'identification correcte apparaissent en gras. Le niveau du hasard est indiqué en première ligne.

Ainsi, les expressions de tristesse sont porteuses d'un tremblement vocal plus important que sur l'expression [sabl] testée précédemment, susceptible d'avoir été confondu avec du rire dans une plus large mesure, ce qui expliquerait les importants reports vers la joie

et satisfaction. De même l'expression de joie sur l'énoncé [vɛʁ], présente un contour de fréquence fondamentale plus plat que sur les énoncés [ʒon] et [ʁuʒ]. Etant donné les résultats obtenus dans l'expérience de projection dimensionnelle sur les expressions de joie et de satisfaction produites par le même locuteur sur les énoncés [ʁuʒ] et [sabl], ceci semble pouvoir expliquer le taux de reconnaissance moindre sur ces expressions.

Bien que les tendances de confusions restent proches, le taux d'identification à la dernière gate des expressions du dégoût sur les énoncés [ʒon] et [ʁuʒ] est plus faible que celui observé précédemment sur l'énoncé [sabl], tandis que l'expression de dégoût sur l'énoncé [vɛʁ] est beaucoup mieux reconnue. Ceci est à rapprocher du fait que plusieurs auditeurs ont déclaré à la suite du test avoir éprouvé plus de difficultés à considérer comme naturelles les expressions de dégoût sur les énoncés [ʒon] et [ʁuʒ] que sur l'énoncé [vɛʁ]. Une explication possible est que, en dépit de l'allongement important du noyau vocalique commun à toutes les expressions de dégoût produites par cet acteur sur des énoncés monosyllabiques, et qui apparaît comme le principal déterminant du dégoût exprimé sur l'énoncé [sabl] dans l'expérience de projection dimensionnelle, une telle expression serait plus écologique sur des voyelles mi-ouvertes voire ouvertes que fermées en raison de la configuration de la bouche associée à une expression faciale de dégoût. De plus, et bien que cela puisse difficilement être attesté par des mesures acoustiques, les expressions de dégoût semblent de façon plus nette être réalisées avec une voix craquée par le locuteur M2 sur les voyelles [a] et [ɛ] que sur les voyelles [o] et [u]. Nous développons en section 13.1. une perspective sur les types de voyelles et la qualité de voix qui pourraient être associés de façon privilégiée aux expressions de dégoût, en regard des explications évolutionnistes de l'expression de dégoût proposées notamment par Rozin (1987, 2000).

9.3.7 Analyse statistique

Nous avons comparé les taux de catégorisation correcte après regroupement d'étiquettes au niveau du hasard correspondant à la catégorie considérée, pour chacun des stimuli présentés et pour chaque fenêtre de dévoilement, au moyen de tests t pour échantillon unique avec comme valeur de test le niveau du hasard. Les résultats de ces tests sont présentés en Annexe V :.

Tous les taux d'identification correcte à la dernière fenêtre de dévoilement sont significativement supérieurs au niveau du hasard, à l'exception de ceux des expressions de tristesse et de résignation ainsi que de ceux des expressions de dégoût et d'inquiétude sur l'énoncé [ʒon] et de l'expression de joie sur l'énoncé [vɛʁ].

Le nombre d'auditeurs ayant reconnu correctement le stimulus à la dernière fenêtre pour chaque émotion et chaque énoncé est présenté dans la Table 23.

	[ʒɔn]	[ʁuʒ]	[vɛʁ]
joie	14	15	7
satisfaction	19	20	20
anxiété	14	14	16
inquiétude	9	15	15
déception	17	19	17
résignation	5	8	8
tristesse	4	3	2
dégoût	6	9	16
neutre	9	7	8

Table 23 : Nombre d'auditeurs ayant identifié correctement l'émotion présentée à la dernière fenêtre dans l'expérience de dévoilement progressif d'expressions sélectionnées parmi les productions du locuteur M2, pour chaque émotion et chaque énoncé.

Nous avons défini le point d'identification stable pour chaque paire juge*stimulus comme l'indice de la fenêtre à partir de laquelle l'une des étiquettes correspondant à l'émotion présentée est choisie, sans changement de réponse aux fenêtres suivantes. Etant donné que tous les juges ne sont pas parvenus à correctement identifier à la dernière fenêtre certains des stimuli présentés, les données correspondantes ont été codées comme reconnues à une hypothétique septième fenêtre.

Nous avons évalué l'effet de l'énoncé et de l'émotion sur le point d'identification stable ainsi défini au moyen d'une analyse de variance à mesures répétées. Les données correspondant à la présentation des expressions de tristesse et de résignation, non reconnues à un niveau supérieur au hasard à la dernière fenêtre, n'ont pas été prises en compte dans l'analyse. En revanche, afin de conserver un plan d'expérience factoriel, des expressions de dégoût et d'inquiétude sur l'énoncé [ʒɔn] et de l'expression de joie sur l'énoncé [vɛʁ] l'ont été. Cette ANOVA à mesures répétées révèle un effet de l'émotion ($F(5.5,105.2)=20.7$, $p<0.001$) ainsi que de l'interaction émotion*énoncé ($F(10.4,198.5=4.7$, $p<0.001$) sur le point d'identification stable. En revanche elle ne permet pas de montrer un effet de l'énoncé ($F(2.0,38.0)=1.9$, $p=0.162$, n.s.). Ainsi, si l'émotion présentée influe sur la rapidité du décodage de l'information affective, l'homogénéité des morphologies prosodiques des différents énoncés semble confirmée du point de vue de cette vitesse de décodage.

Toutefois, étant donné que nous avons pris en compte des fenêtres de dévoilement dont le pas est défini selon des critères syllabiques et non en fonction de durées absolues, nous

ne pouvons exclure *a priori* que l'effet de l'émotion sur la rapidité du décodage de l'information affective soit dû en majeure partie aux différences de durées mesurées entre émotions. Nous avons donc également extrait les durées correspondant aux points d'identification stables pour chaque auditeur et chacun des stimuli présentés. Les valeurs moyennes de ces durées par stimulus et par énoncé, plus proches de la notion de point d'isolation que Grosjean (1980) définit comme la durée moyenne à laquelle la valeur de l'énoncé complet est identifiée avec une confiance supérieure à 50%, sont récapitulées dans la Figure 48. Ces valeurs moyennes sont calculées pour chaque stimulus uniquement sur les réponses données par les auditeurs qui sont parvenus à identifier correctement l'émotion présentée à la dernière fenêtre. Elles correspondent donc à un nombre variable d'auditeurs (cf. Table 23).

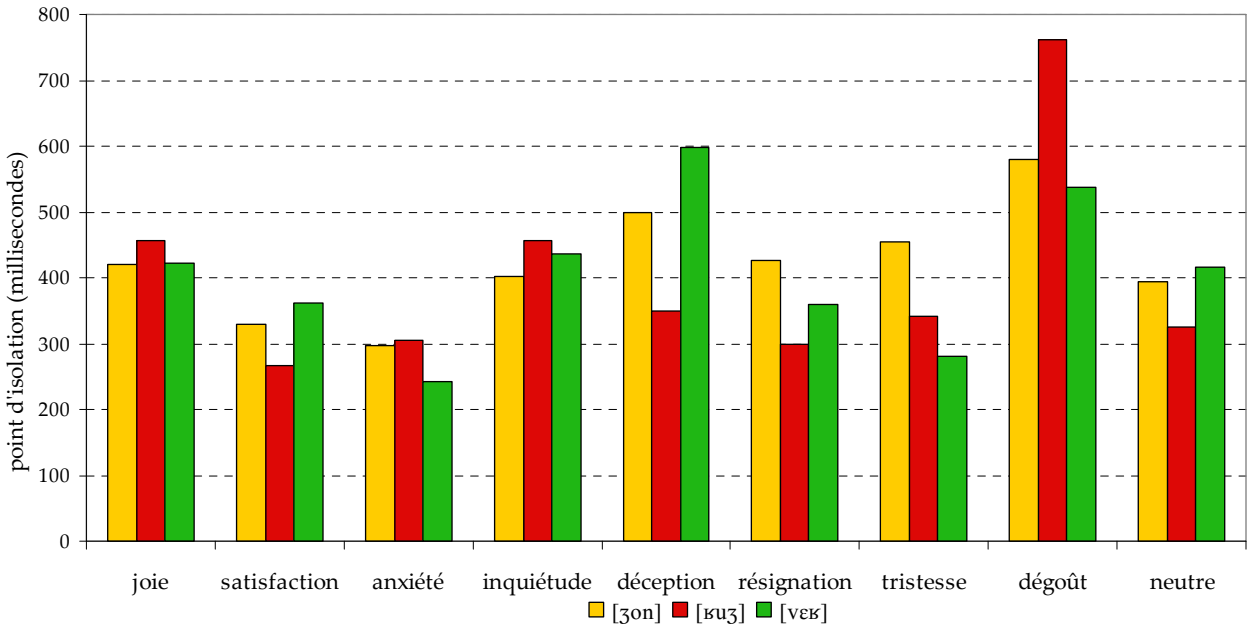


Figure 48 : Durée moyenne par émotion et par énoncé du point d'identification stable moyen obtenu dans l'expérience de dévoilement progressif de productions du locuteur M2, exprimée en millisecondes.

L'évolution au cours du dévoilement des taux d'identification correcte de chacun des stimuli (à l'exception de ceux n'étant pas reconnus au dessus du niveau du hasard à la dernière fenêtre) est présentée en Figure 49. Toutes les expressions suivent un schéma de progression linéaire, avec un taux d'identification à la gate 1 significativement en dessous du niveau du hasard, à l'exception des expressions de satisfaction. En effet ces dernières, identifiées dès la première gate à 60.7% en moyenne, suivent un schéma d'identification précoce, tandis qu'aucune des autres n'est identifiée au dessus du niveau du hasard avant la troisième gate.

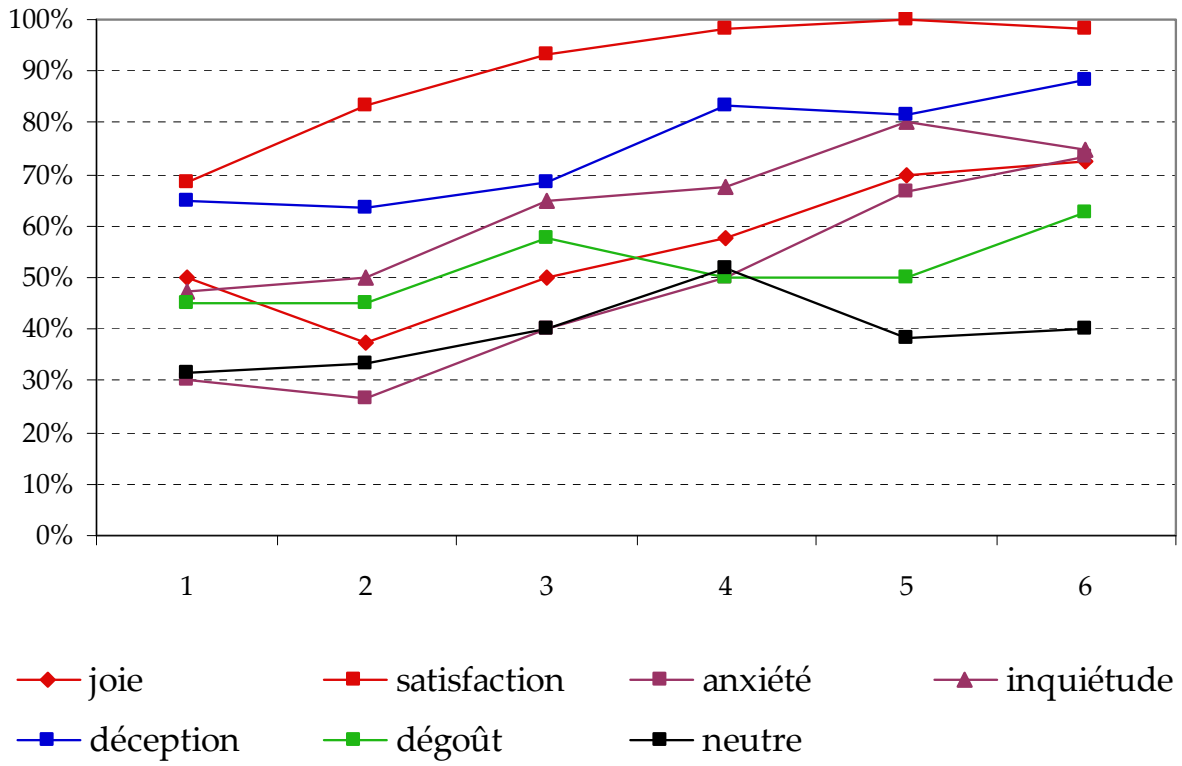


Figure 49 : Evolution des scores de reconnaissance au fil des fenêtres de dévoilement successives des productions du locuteur M2.

9.3.8 Discussion

Tandis que qualité de voix et durée jouent un rôle prépondérant dans le cas des émotions négatives, l'information affective des expressions de joie et de satisfaction est principalement portée par les contours de F0 (cf. section 9.2.). Les contours de F0 de ces expressions peuvent en conséquence être considérés comme de bons indicateurs de l'information affective présentée aux gates successives. Etant donné que l'expression de la joie sur [vɛʁ] a été exclue, l'analyse comparative des contours de F0 a été restreinte aux expressions de joie et satisfaction sur les énoncés [ʒon] et [ʁuʒ]. Les valeurs de F0 ont été converties en demi-tons, la valeur de référence (0 demi-tons) étant définie comme la valeur moyenne de F0 du locuteur dans l'ensemble du corpus (96.8 Hz).

L'analyse comparative de ces contours révèle que les expressions de joie et de satisfaction partagent la même forme de contour, mais avec des valeurs d'ancrage très différentes. Les contours correspondant à la satisfaction présentent une valeur moyenne et une variabilité plus élevées que dans le cas de la joie. Sur l'énoncé [ʁuʒ] la valeur moyenne de F0 est de 3.8 demi-tons pour la joie contre 7.8 pour la satisfaction; et l'étendue de F0 de 6.1 demi-tons pour la joie contre 14.9 pour la satisfaction. Sur l'énoncé [ʒon] la valeur moyenne de F0 est de 3.9 demi-tons pour la joie contre 7.0 pour la satisfaction; et l'étendue de F0 de 8 demi-tons pour la joie contre 16.5 pour la satisfaction.

La Figure 50 présente les contours de F0 des expressions de la joie et de la satisfaction sur l'énoncé [ʁuʒ], normalisés à la même durée, avec la localisation des 4 premières gates.

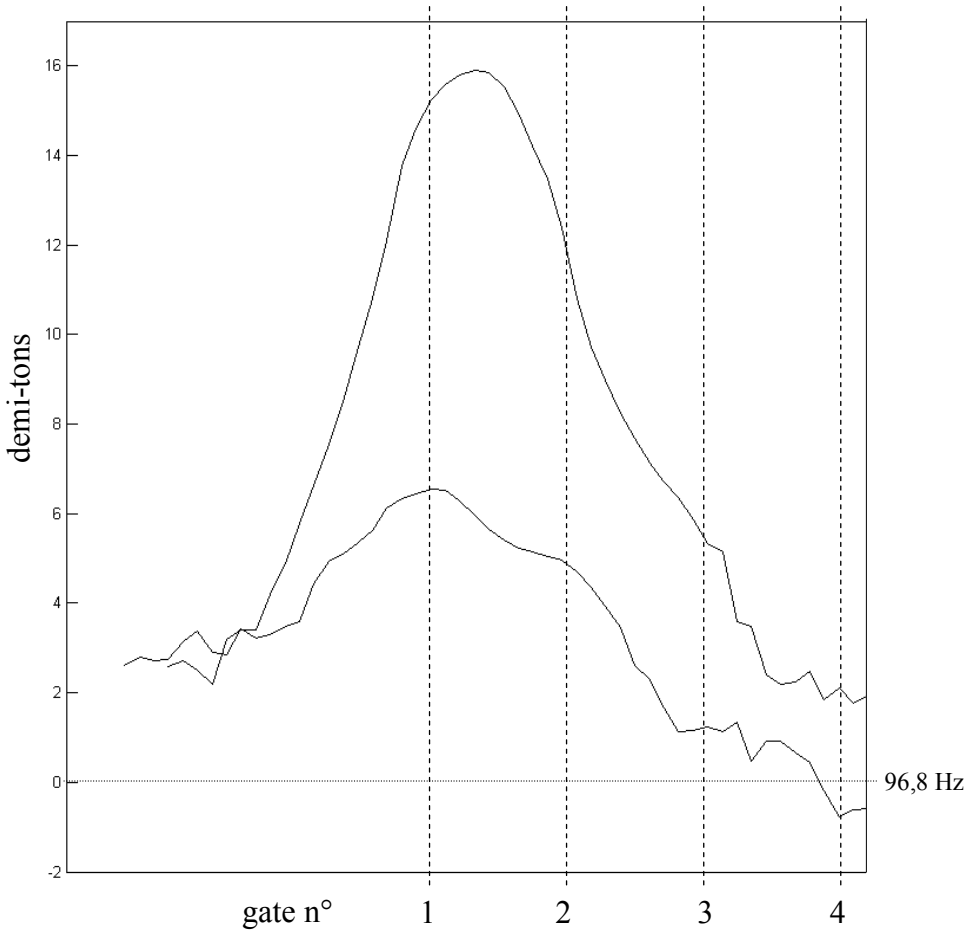


Figure 50 : Contours de F0 des expressions de la joie (*courbe du bas*) et de la satisfaction (*courbe du haut*) produites par le locuteur M2 sur l'énoncé [ʁuʒ], normalisés à la même durée et emplacement des 4 premières fenêtres de dévoilement (*lignes verticales pointillées* : depuis le 1er tiers de la voyelle jusqu'au 1er tiers de la consonne finale). La F0 moyenne du locuteur (96.8 Hz) est définie comme valeur de référence (0 demi-tons).

Les expressions de joie et de satisfaction semblent ainsi correspondre à des stratégies différentes de réalisation d'une même valeur affective. La comparaison des contours de F0 montre que les valeurs gradientes des contours de la satisfaction sont reconnues dès la 1ère gate, permettant une identification anticipée. À l'opposé les valeurs d'ancrage (notamment la valeur moyenne et la dynamique) du contour de fréquence fondamentale correspondant à la joie sont insuffisamment différentes de celles des contours des autres expressions testées pour permettre l'identification de l'expression cible avant que la forme générale du contour ne soit connue.

L'hypothèse de Bänziger & Scherer (2005) est que la valeur moyenne et l'étendue de F0 varient fortement avec l'activation des expressions émotionnelles et expliquent l'essentiel

des variations mesurables perceptivement, tandis que la forme des contours ne porterait que peu d'information. Etant donné que l'activation, bien que n'ayant pas été évaluée de façon systématique, est clairement supérieure pour les expressions de satisfaction que de joie, cette hypothèse reste compatible avec nos résultats. Une hypothèse alternative serait que la tendance à une évolution linéaire du taux de reconnaissance au fil des fenêtres successives de dévoilement serait révélatrice d'un décodage selon la forme de leurs contours multiparamétriques plutôt que par les valeurs gradientes de ces contours ou des évènements locaux saillants. Cependant les résultats que nous obtenons ne nous permettent pas de trancher de façon nette pour ou contre cette hypothèse. La portée des résultats obtenus dans cette étude doit de plus être relativisée. En effet les stimuli sur lesquels s'appuie cette étude ont été produits par un acteur, et il n'est donc pas possible, si nous retenons les hypothèses formulées par le modèle prosodique d'Aubergé (1991, 2000) d'interpréter ces résultats en termes d'affects sociaux vs. involontaires.

Les temps de décodage avant d'aboutir à un taux de reconnaissance supérieur au hasard doivent être comparés aux résultats obtenus par Paulmann *et al.* (2008) à partir d'une étude de potentiels évoqués, qui montrent que la prosodie émotionnelle actée est reconnue avec un délai d'environ 200 millisecondes, ce qui correspond environ à la durée d'une syllabe en parole continue. Dans notre expérience, une durée de 200 ms correspond pour la plupart des stimuli testés à une durée correspondant environ à la 3^{e} fenêtre de dévoilement progressif.

10. Parole émotionnelle actée vs. spontanée

Nous présentons tout d'abord dans cette section les résultats d'une série d'évaluations perceptives portant sur les déterminants de la typicalité d'expressions émotionnelles actées et spontanées d'affects appartenant aux classes d'émotions de la peur, de la colère et de la joie, produites par 6 des 7 acteurs enregistrés. Les expressions actées évaluées sont jugées comme exprimant une émotion plus intense que les expressions spontanées, confirmant les quelques résultats comparables de la littérature. Le principal résultat concernant les déterminants de la typicalité est que les expressions vocales des émotions semblent mieux décrites comme des catégories dérivées du but au sens de Barsalou que comme des catégories taxonomiques, sans pour autant que la distinction ne soit très marquée, ce qui confirme les résultats obtenus auparavant par Horstmann sur les expressions faciales.

Nous présentons ensuite les résultats d'une étude de discrimination d'expressions actées et spontanées produites par les mêmes locuteurs et relevant des mêmes classes d'émotions, présentées en conditions audio, visuelle et audiovisuelle. Cette étude montre que des sujets naïfs sont capables de discriminer la plupart des paires présentées sans différences notables entre classes d'émotions, et ce d'autant plus en condition audiovisuelle que lorsqu'une seule modalité est présentée, avec un fort effet inter-juge qui souligne la difficulté à proposer un protocole fiable de la similarité des expressions actées et spontanées. Les différences d'intensité perçue de l'émotion au sein des paires n'expliquent que partiellement les scores de discrimination, indiquant que d'autres indices audio et visuels permettraient de distinguer les expressions actées des expressions spontanées.

Remarque : Dans l'ensemble de cette section, le terme d'intensité désigne l'intensité perçue de l'émotion et non l'intensité acoustique.

10.1. Evaluation de la typicalité d'expressions actées et spontanées

10.1.1 Catégorisation taxonomique vs. dérivée du but

Les recherches sur la catégorisation en psychologie cognitive, majoritairement fondées sur la catégorisation sémantique d'unités lexicales, ont montré que tous les exemplaires d'une catégorie donnée n'étaient pas considérés comme des membres aussi bon de cette catégorie tout en étant reconnu comme en faisant partie (voir par exemple Smith *et al.* (1974) et Rosch & Mervis (1975)). Deux grands types de catégories dont les propriétés caractéristiques diffèrent ont ainsi été isolées (Barsalou, 1985, 1987) :

- Les catégories taxonomiques communes, qui généralement existent en tant que telles dans l'environnement et dont les membres partagent des propriétés rarement observées en dehors de la catégorie. C'est le cas de la catégorie des mammifères. Ces catégories sont supposées refléter la structure corrélationnelle de l'environnement, c'est-à-dire être représentatives des cooccurrences les plus fréquemment observées de certaines propriétés.
- Les catégories dérivées du but, qui impliquent une activité humaine et qui, selon Barsalou (1985), sont généralement dérivées de catégories *ad hoc* créées pour répondre à un but comportemental précis comme la catégorie des « endroits agréables pour dîner en ville ». Elles sont en conséquence moins saillantes que les catégories taxonomiques : en dehors du but considéré, les exemplaires de ces catégories peuvent apparaître comme très différents, et de telles catégories sont moins susceptibles d'être largement partagées d'un individu à l'autre.

Si les exemples canoniques donnés par Barsalou (1985, 1987) peuvent être considérés sans ambigüité comme des instances de catégories taxonomiques ou dérivées du but, il indique également que la distinction n'est pas nécessairement aussi évidente et que des catégories présentant à la fois des caractéristiques de catégories taxonomiques et dérivées du but peuvent exister. Barsalou (1987) estime ainsi que la question centrale pour déterminer ce qui fait qu'un exemplaire est considéré comme un membre plus ou moins bon d'une catégorie est celle de sa typicalité, la représentativité des exemplaires formant un continuum du membre le plus typique de la catégorie jusqu'au plus atypique. Parmi les déterminants de la typicalité d'un exemplaire au sein d'une catégorie, il en propose ainsi trois principaux capables de rendre compte de la structuration de la catégorie étudiée :

- La similarité de l'exemplaire aux idéaux, les idéaux étant vus comme les buts associés à cette catégorie. L'exemplaire idéal peut alors ne correspondre à aucun exemplaire

existant dans la réalité. Barsalou (1985) donne ainsi l'exemple d'une catégorie « aliments adaptés pour un régime », dont l'exemplaire idéal serait un hypothétique aliment sans la moindre calorie. Dans le cas d'une catégorie purement taxonomique, la question de la similarité aux idéaux peut néanmoins s'avérer difficile à formuler. Ainsi, pour reprendre l'un des exemples donnés par Barsalou, on peine à imaginer quel peut être le but poursuivi par la catégorie des mammifères.

- Sa fréquence d'instanciation, c'est-à-dire la fréquence avec laquelle cet exemplaire peut être observé comme membre de la catégorie considérée
- Sa similarité à la tendance centrale de la catégorie, c'est-à-dire à un exemplaire « central » existant ou hypothétique, souvent défini comme la moyenne des propriétés mesurables de l'ensemble des exemplaires de la catégorie.

Dans une série d'expériences, Barsalou (1985) a montré que la typicalité des exemplaires d'une catégorie taxonomique était déterminée de façon prédominante par leur fréquence d'instanciation en tant que membre de cette catégorie et par leur similarité à la tendance centrale de la catégorie. À l'inverse, la typicalité des exemplaires de catégories dérivées du but est plutôt déterminée par leur similarité aux membres idéaux de la catégorie, dont les caractéristiques en font les plus à même de remplir le but de façon appropriée mais qui sont généralement rares et éloignés de la tendance centrale.

10.1.2 Le cas des expressions émotionnelles

La catégorisation d'expressions faciales d'émotions a été étudiée par Horstmann (2002) à partir de dessins stylisés d'expressions de colère, dégoût, joie, tristesse et surprise d'intensité variable construits par interpolation entre expressions neutres et expressions prototypiques proposées par Ekman & Friesen (1978). Les jugements subjectifs de typicalité, d'idéalité, de fréquence d'instanciation et de similarité par paires, pris comme mesure de la tendance centrale, indiquent que la typicalité des expressions faciales est majoritairement prédite par la similarité aux idéaux, définis comme les expressions faciales exprimant le plus efficacement l'émotion considérée. Horstmann (2002) en conclut que les expressions faciales sont structurées comme des catégories dérivées du but. Néanmoins il souligne également leur spécificité d'un point de vue théorique : bien qu'apparaissant comme structurées plutôt selon leur efficacité qu'en fonction de la structure corrélationnelle des expressions faciales, ces expressions possèdent en effet certaines caractéristiques relevées par Barsalou (1985) comme classiques parmi les catégories taxonomiques communes. Ainsi, les expressions faciales constituent des catégories très largement partagées, voire universelles (voir par exemple Ekman (1994)), et différents exemplaires relevant d'une même catégorie émotionnelle partagent

indéniablement certaines propriétés. Horstmann (2002) suppose de plus que l'utilisation des expressions faciales pour la catégorisation est courante.

Considérant que les expressions vocales des émotions remplissent également un but communicationnel, nous faisons l'hypothèse que les expressions vocales sont également des catégories dérivées du but. Cette hypothèse est cohérente avec le fait que les expressions prototypiques d'émotions vocales sont rarement observées dans les interactions de la vie de tous les jours (Campbell, 2005; Cowie & Cornelius, 2003). Néanmoins, malgré l'importante variabilité observée, différentes expressions vocales d'une même émotion semblent partager des propriétés acoustiques communes quelque soit la culture (voir par exemple Scherer *et al.*, 2003 ; Juslin & Laukka, 2003), ce qui pourrait être compatible avec une structuration en catégories taxonomiques.

En collaboration avec Petri Laukka (Laukka, Audibert & Aubergé, 2007, 2008), nous avons donc évalué selon une méthodologie similaire à celle utilisée par Horstmann (2002) les déterminants de la typicalité d'expressions vocales d'émotions extraites du corpus Sound Teacher/E-Wiz. Afin de pouvoir comparer ces déterminants dans le cas d'expressions actées ou spontanées, nous avons utilisé des stimuli produits par 6 sujets acteurs, incluant des expressions de peur, de colère et de joie vues comme des catégories larges. Les stimuli ont été à la fois produits spontanément par les locuteurs, puis reproduites en s'appuyant sur le ressenti de l'expérience et enfin actées afin d'exprimer les expressions des *Big Six* ajoutées au corpus spontané.

10.1.3 Présélection de stimuli audio

Les productions de six sujets acteurs (trois hommes, trois femmes) parmi les sept sujets disponibles ont été sélectionnées. Les productions de la septième actrice, pour laquelle la grille d'auto-annotation n'avait pu être récupérée et qui de plus n'avait fourni que peu d'informations sur son ressenti à la suite du déroulement du scénario Sound Teacher, n'ont pas été retenues.

Une première étape a consisté en la sélection 'un sous-ensemble équilibré des productions actées et spontanées présentant autant de variabilité que possible dans l'intensité des expressions et comprenant les trois conditions de production de ces expressions : spontanée, actée en rejouant des émotions ressenties au cours du test (que nous désignerons dans cette section par le terme *acté non prototypique*) et actée correspondant aux émotions ajoutées (expressions des *Big Six*, que nous désignerons par le terme *acté prototypique*). Les énoncés correspondent à des noms de couleurs monosyllabiques et aux occurrences de « page suivante ». Afin d'obtenir des catégories émotionnelles comparables entre les conditions, trois classes d'émotions ont été

considérées : joie, peur et colère. Ces émotions ont été considérées comme des classes larges, l'étiquette *colère* pouvant ainsi correspondre à l'expression d'une légère irritation ou à l'inverse d'une rage intense.

Une présélection des 246 stimuli (dont 100 stimuli actés) correspondant aux trois classes émotionnelles retenues a tout d'abord été effectuée de façon concertée par les trois expérimentateurs, avec l'objectif de retenir des stimuli présentant autant de variation que possible dans leur activation et les stratégies d'expressivité prosodique utilisées par les locuteurs. Les 146 stimuli spontanés ont été choisis parmi les énoncés spontanés pour lesquels l'auto-annotation correspondante était claire et ne faisait pas référence à un mélange d'états émotionnels. Ainsi les expressions spontanées et actées non-prototypiques de satisfaction et d'amusement sont-ils rattachés à la classe *joie*, les expressions de stress et d'anxiété à la classe *peur*, et celles d'irritation et d'agacement à la classe *colère*.

Une première évaluation perceptive de ces 246 stimuli a alors été effectuée auprès de 15 auditeurs francophones naïfs (8 hommes, 7 femmes, âge moyen : 29,8 ans), à l'aide d'une interface développée avec le logiciel Revolution Runtime permettant de présenter les stimuli dans un ordre aléatoire différent pour chaque auditeur et de collecter automatiquement les réponses. Les juges, qui pouvaient écouter les stimuli autant de fois que nécessaire, avaient pour consigne de catégoriser les stimuli présentés entre les classes d'émotion *joie*, *colère* et *peur*, ou de leur attribuer l'étiquette *autre émotion*. Les consignes préalables au test indiquaient explicitement que les 3 émotions considérées devaient être vues comme des classes larges, en illustrant d'exemples pour chacune des classes. De plus, les juges devaient également indiquer l'intensité avec laquelle l'émotion était perçue au moyen d'une échelle variant de 0 (intensité très faible) à 10 (intensité très forte), de manière similaire au protocole utilisé par Juslin & Laukka (2001). L'interface du test perceptif est présentée Figure 51. Les instructions données aux sujets préalablement au test sont reproduites en Annexe VI :.

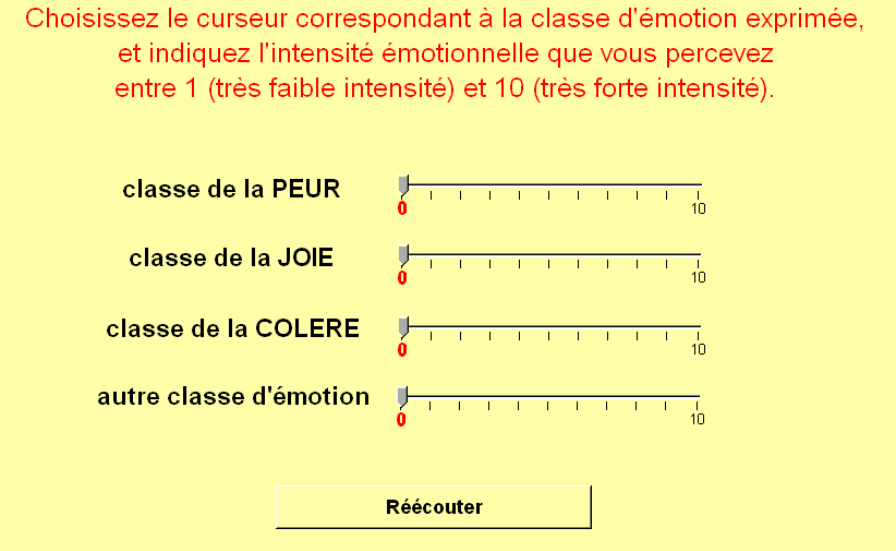


Figure 51 : Interface développée pour l'évaluation conjointe de la valeur des affects exprimés et de l'intensité de l'émotion dans le pré-test de l'étude de la typicalité.

Parmi les 193 stimuli reconnus avec une précision supérieure au hasard sans qu'une autre classe émotionnelle n'ait été sélectionnée plus fréquemment, nous avons sélectionné 12 stimuli pour chaque sous-catégorie de classe d'émotion et chaque condition d'enregistrement (spontané, acté prototypique ou acté non-prototypique), pour un total de 108 stimuli (3 classes d'émotion x 3 conditions x 12 stimuli). Chacun des 3 auteurs a proposé indépendamment une telle sélection, avec comme critères prioritaires une variabilité importante dans les scores d'identification obtenus et l'intensité émotionnelle notée par les 15 participants du pré-test, avant de s'accorder sur le choix d'une sélection commune. Etant donnée la distribution des stimuli correspondant à ces critères, un équilibre strict entre groupes quant aux locuteurs et énoncés représentés n'a pu être assuré. Néanmoins nous nous sommes efforcé de maintenir autant que possible l'équilibre entre hommes et femmes dans les différents sous-groupes. Le tableau présenté en Annexe VI : récapitule les caractéristiques principales de la distribution des stimuli retenus.

10.1.4 Evaluation perceptive de la typicalité et de ses déterminants

Nous avons effectué une série d'évaluations perceptives, dont l'objectif était d'obtenir de la part de groupes distincts d'auditeurs naïfs des jugements de typicalité, de similarité aux idéaux ou de fréquence d'instanciation pour chacun des 108 stimuli retenus. Chacun des trois groupes d'auditeurs a donc passé un test perceptif organisé de façon identique mais avec une consigne différente :

- Typicalité : *À quel point le mot que vous venez d'entendre vous semble-t-il typique d'une expression de la colère/peur/joie ? (pas du tout typique 0-10 très typique)*

- Similarité aux idéaux : *Si quelqu'un tentait d'exprimer de la colère/peur/joie, avec quelle efficacité la manière de prononcer que vous venez d'entendre exprimerait-elle cette émotion ? (pas du tout efficacement 0-10 très efficacement)*
- Fréquence d'instanciation : *Avec quelle fréquence pourriez-vous entendre cette manière de prononcer chez d'autres personnes, lorsque ces personnes ressentent de la colère/peur/joie ? (pas du tout fréquemment 0-10 très fréquemment)*

Dans les instructions données aux auditeurs préalablement au test, nous avons insisté sur le fait que les trois classes d'émotions considérées devaient être vues comme des classes larges. Les instructions écrites, que nous avons complétées par des commentaires donnés oralement, sont reproduites en Annexe VI :.

Les stimuli ont été présentés par classe d'émotion, dans un ordre aléatoire à l'intérieur de chaque classe différent pour chaque sujet, à l'aide d'une interface développée avec *Revolution Runtime*, dans laquelle seule la question posée variait en fonction du groupe. Un exemple de la page de réponse de cette interface dans le cas de la présentation d'une expression de la classe de la peur dans le test de typicalité est illustré Figure 52. L'ordre de présentation des classes émotionnelles a de plus été systématiquement alterné pour les différents sujets de chaque groupe.

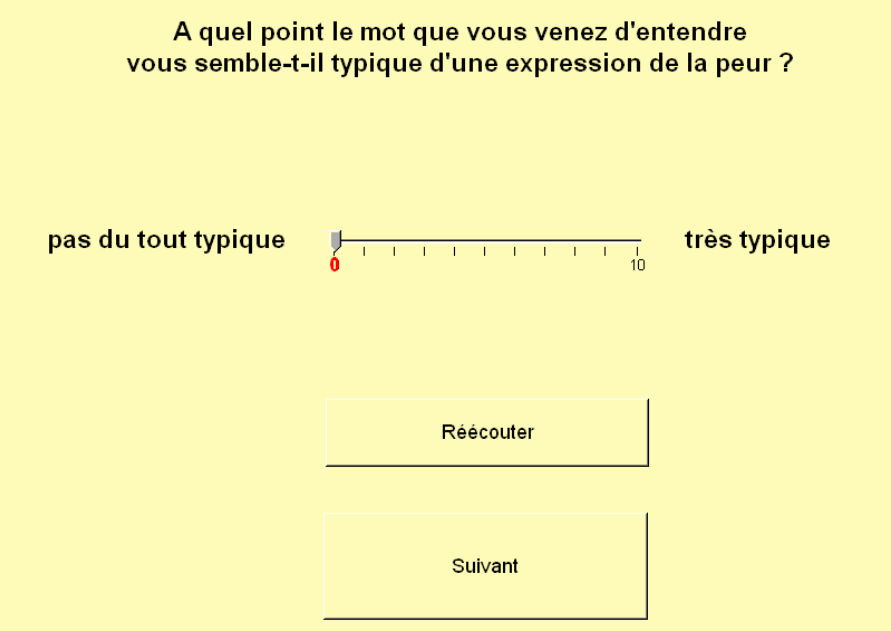


Figure 52 : Exemple d'une page de réponses dédiée à l'évaluation de la typicalité de l'interface développée pour la deuxième série de tests.

Chacun des trois tests perceptifs a été effectué par 16 sujets naïfs francophones (8 hommes et 8 femmes dans le groupe *typicalité*, âge moyen 27.0 ans; 8 hommes et 8 femmes dans le groupe *similarité aux idéaux*, âge moyen 27.9 ans; 7 hommes et 9 femmes

dans le groupe *fréquence d'instanciation*, âge moyen 27.9 ans; âge moyen des 48 sujets = 27.6 ans).

Une approximation de la similarité à la tendance centrale a été obtenue en calculant un score de ressemblance pour chaque paire de stimuli au sein d'une classe d'émotion prédéfinie, à partir de leur répartition dans l'espace allant du stimulus le mieux reconnu au stimulus le moins bien reconnu. D'après Barsalou (1985), la similarité d'un exemplaire aux autres exemplaires de la même catégorie peut être assimilée à sa similarité à la tendance centrale. En conséquence, nous avons calculé pour chaque stimulus la différence D en valeur absolue entre son taux d'identification correcte obtenu dans le pré-test et exprimé comme une proportion et le taux moyen d'identification correcte dans la catégorie considérée. La mesure que nous avons retenue comme estimation de la similarité S à la tendance centrale est donnée par la formule S=1-D, de sorte qu'une valeur élevée représente une similarité à la tendance centrale supérieure.

Si cette estimation présente l'avantage d'être aisée à calculer, soulignons qu'elle ne représente qu'une similarité entre expressions relevant d'une même classe d'émotion pour l'identification de cette classe. Sauf à estimer que ces scores d'identification sont directement liés à la cooccurrence de propriétés acoustiques, cette estimation ne représente pas directement la tendance centrale telle que la définit Barsalou (1985, 1987) et nous devrons donc l'utiliser avec circonspection dans nos conclusions.

10.1.5 Analyse des résultats

Une première analyse des résultats (Laukka, Audibert & Aubergé, 2007) a donné des résultats inattendus quant au rapport entre similarité aux idéaux et fréquence d'instanciation comparativement aux résultats obtenus dans des études précédentes et notamment par Horstmann (2002) pour les expressions faciales. En effet la corrélation entre les jugements de similarité aux idéaux et ceux de fréquence d'instanciation s'est avérée sensiblement plus élevée que celle observée par Horstmann (2002).

Nous avons donc voulu explorer l'impact sur les résultats d'une reformulation de la question relative à la fréquence d'instanciation. En effet, l'interview a posteriori de certains participants du test nous ont fait soupçonner que la notion de fréquence d'instanciation ait été mal comprise des sujets et confondue avec une notion plus proche de la similarité aux idéaux, d'autant plus que contrairement à nos attentes les jugements de fréquence d'instanciation se sont avérés très significativement plus élevés pour les expressions actées que spontanées ($t(106)=5.0$; $p<0.001$).

Nous avons donc reformulé cette question de la façon suivante : « *Avec quelle fréquence pourriez-vous entendre cette manière de prononcer chez d'autres personnes, lorsque ces personnes*

ressentent de la colère/peur/joie ? Notez bien que nous cherchons à évaluer la fréquence d'apparition de différentes expressions, et non la clarté avec laquelle elles expriment des émotions particulières. Il est possible qu'une expression fréquente exprime très clairement une émotion particulière, mais une expression fréquente peut aussi ne pas être facilement reconnue (pas du tout fréquemment 0-10 très fréquemment). ». 16 sujets naïfs (8 hommes, 8 femmes, âge moyen 28,9 ans) n'ayant pas participé aux tests précédents ont évalué les stimuli. La méthodologie utilisée était la même que précédemment à l'exception de la formulation de la question.

La comparaison des valeurs des jugements de fréquence d'instanciation recueillies avec les deux formulations de la question conforte l'hypothèse d'une mauvaise interprétation des sujets lors du premier test. En effet ces jugements de fréquence d'instanciation sont globalement significativement plus élevés que ceux recueillis précédemment dans le cas des expressions spontanées ($t(107)=-5.5$; $p<0.001$), mais cela n'est pas le cas en considérant uniquement les expressions actées. De plus, si les expressions actées reçoivent toujours des jugements de fréquence d'instanciation plus élevés que les expressions spontanées, cette différence évaluée au moyen d'un test t n'est plus significative dans cette version reformulée. L'analyse que nous présentons ici s'appuie donc sur cette seconde mesure de la fréquence d'instanciation.

Les statistiques descriptives (moyenne et écart-type) des différents descripteurs évalués par les sujets, ainsi que le taux d'identification correcte et l'intensité émotionnelle perçue, évalués lors du pré-test, sont présentés pour chacun des groupes de stimuli dans la Table 24. Au-delà des différences relevées ci-dessus entre les deux versions successives de l'estimation de fréquence d'instanciation, notons que les expressions actées ont reçu des jugements de typicalité ($t(106)=4,14$; $p<.001$) et de similarité aux idéaux ($t(106)=3,85$; $p<.001$) significativement supérieurs à ceux attribués aux expressions spontanées. En revanche, toutes les différences de jugement entre expressions actées prototypiques et expressions actées non prototypiques, ainsi que les différences entre classes d'émotion, évaluées au moyen de tests t, se sont avérées non significatives.

	Global	Acté prototypique	Acté non prototypique	Spontané
Typicalité	3.30 (1.79)	3.61 (1.78)	3.93 (1.85)	2.36 (1.35)
Sim. idéaux	3.41 (1.66)	3.57 (1.53)	4.08 (1.73)	2.60 (1.37)
Fréq. inst. 1	3.62 (1.59)	3.98 (1.46)	4.23 (1.54)	2.63 (1.32)
Fréq. inst. 2	4,13 (1,16)	4,28 (1,14)	4,32 (1,21)	3,78 (1,07)
Tend. centrale	0.83 (0.10)	0.85 (0.07)	0.82 (0.11)	0.82 (0.11)
Identification	77,4% (20,0%)	82,8% (16,4%)	79,6% (21,6%)	70% (20%)
Intensité émo.	4.76 (1.76)	4.93 (1.77)	5.54 (1.75)	3.82 (1.31)

Table 24 : Moyennes et écarts-types généraux et par groupe de stimuli pour les jugements de : typicalité ; similarité aux idéaux ; fréquence d'instanciation avec la question sous sa forme initiale (Fréq. inst. 1) et révisée (Fréq. inst. 2) ; mesure de similarité à la tendance centrale estimée ; taux d'identification correcte lors du pré-test (hasard=25%) ; intensité émotionnelle perçue (pré-test).

Suivant la méthodologie proposée par Barsalou (1985) et appliquée aux expressions faciales des émotions par Horstmann (2002), nous avons ensuite comparé la contribution des différentes variables évaluées à la prédiction des jugements de typicalité, au moyen de corrélations partielles entre les jugements moyennés sur l'ensemble des sujets attribués à chaque stimulus. Les corrélations partielles permettent d'étudier la corrélation entre 2 variables en retranchant l'influence d'une ou plusieurs autres variables dites de contrôle[28].. En considérant la typicalité comme variable dépendante et les variables *similarité aux idéaux*, *fréquence d'instanciation* et *tendance centrale* comme prédicteurs, nous avons donc calculé les corrélations partielles entre :

(a) *typicalité* et *similarité aux idéaux*, avec *fréquence d'instanciation* et *tendance centrale* comme variables de contrôle,

(b) *typicalité* et *fréquence d'instanciation*, avec *similarité aux idéaux* et *tendance centrale* comme variables de contrôle,

(c) *typicalité* et *tendance centrale*, avec *similarité aux idéaux* et *fréquence d'instanciation* comme variables de contrôle.

Ces corrélations partielles, globales et par groupes de stimuli, sont présentés dans la Table 25.

[28] Lorsque l'une des 2 variables est considérée comme variable dépendante, le carré de la corrélation partielle est alors interprété comme la proportion de variance unique de la variable dépendante expliquée par la variable indépendante, indépendamment de l'influence éventuelle des variables de contrôle.

Afin de compléter ces résultats, nous avons effectué une régression linéaire multiple avec la typicalité comme variable dépendante et les 3 prédicteurs *similarité aux idéaux*, *fréquence d'instanciation* et *tendance centrale* comme variables indépendantes. Cette régression indique une corrélation multiple significative (R^2=.837 ; F(3, 104)=178,3 ; p<.001), avec des coefficients normalisés significatifs β=.805 (t(106)=13,30 ; p<.001) pour la variable *similarité aux idéaux* et β=.142 (t(106)=2,32 ; p<.05) pour la variable *fréquence d'instanciation* qui confirment la prépondérance de la similarité aux idéaux comme prédicteur de la typicalité.

	r_P (sim. idéaux)	r_P (freq. inst.)	r_P (tend. .centrale)
Global	.79 **	.22 *	-.04 ns
Acté prototypique	.80 **	.18 ns	-.31 ns
Acté non prototypique	.81 **	.19 ns	.12 ns
Spontané	.65 **	.42 *	-.04 ns

Table 25 : Corrélations partielles entre (col. 1) typicalité et similarité aux idéaux (col. 2) typicalité et fréquence d'instanciation, et (col. 3) typicalité et similarité à la tendance centrale. Les 2 variables non mentionnées sont utilisées comme variables de contrôle dans chacune des corrélations partielles. La version de la fréquence d'instanciation obtenue avec la question reformulée est utilisée. ** : p < .001 ; * : p < .05 ; ns = non significatif.

Ainsi, la similarité aux idéaux explique une proportion beaucoup plus importante (62%) de la variance unique des jugements de typicalité que ne le fait la fréquence d'instanciation (5%), ce qui est à rapprocher des résultats de Horstmann (2002) qui obtient des proportions de la variance expliquées de respectivement 92% et 25%. Néanmoins cette tendance est moins marquée dans le cas des expressions spontanées (42% de la variance unique expliquée par la similarité aux idéaux contre 18% expliquée par la fréquence d'instanciation).

Nous avons également mesuré les corrélations entre ces différents jugements obtenus sur les 108 stimuli retenus et les jugements d'intensité émotionnelle collectés lors du pré-test (corrélation de Pearson). Les corrélations les plus importantes sont obtenues, pour l'intensité émotionnelle perçue, avec la typicalité (r = .86 ; p < .001), suivie de la similarité aux idéaux (r = .81 ; p<.001), et de la fréquence d'instanciation (r = .61; p< .001), tandis que l'intensité perçue n'est pas significativement corrélée à l'estimation de la tendance centrale que nous avons utilisée. Ces résultats suggèrent que les expressions les plus typiques et jugées comme exprimant plus efficacement l'émotion considérée sont aussi celles qui expriment le plus intensément l'émotion.

Bien que la corrélation positive entre intensité émotionnelle perçue et fréquence d'instanciation soit significative, la corrélation partielle entre intensité perçue et fréquence d'instanciation avec la similarité aux idéaux comme variable de contrôle ne l'est pas ($r = -0,05$; n.s.), à l'inverse de la corrélation partielle entre intensité perçue et similarité aux idéaux avec la fréquence d'instanciation comme variable de contrôle explique 45% de la variance unique de l'intensité perçue ($r = .67$, $p<.001$). Ainsi, les émotions les plus typiques ne sont pas nécessairement les plus fréquentes.

10.1.6 Discussion

Il ressort des résultats présentés ci-dessus que la similarité aux idéaux est le meilleur prédicteur des jugements de typicalité des expressions vocales d'émotions testées : la typicalité des expressions vocales varie donc en fonction de leur capacité à remplir le but, c'est-à-dire dans le cas présent à véhiculer l'information émotionnelle aussi efficacement que possible. Par contraste, ces jugements de typicalité reflètent beaucoup moins la fréquence relative avec laquelle ces expressions peuvent être observées dans la vie de tous les jours, ou leur similarité à la tendance centrale des expressions. Ainsi, les expressions vocales des émotions correspondraient plutôt à des catégories dérivées du but qu'à des catégories taxonomiques communes. Ces résultats indiquent de plus que les expressions jugées typiques sont généralement perçues avec une intensité émotionnelle élevée, ce qui est cohérent avec l'hypothèse d'une structuration de la catégorie en fonction du but. En effet, bien que parmi les 108 stimuli l'intensité perçue ne soit pas significativement corrélée au taux d'identification[29] ($r = .16$, n.s.), l'intensité émotionnelle perçue pour les 247 stimuli inclus dans le pré-test est significativement corrélée au taux d'identification de ces stimuli ($r = .50$, $p<.001$). Plusieurs études ont en outre pu montrer que les émotions exprimées plus intensément étaient généralement mieux reconnues (voir par exemple Juslin & Laukka, 2001). On peut considérer que, dans le cas général, les expressions intenses remplissent mieux le but de communication de l'état émotionnel du locuteur que ne le font les expressions moins intenses. Néanmoins, étant donné la moindre fréquence des émotions intenses dans la vie quotidienne, il est naturel que les expressions d'émotions intenses soient également plus rares. Ainsi, la catégorisation des expressions vocales des émotions en fonction du but pourrait en partie expliquer l'apparente contradiction entre la rareté des expressions prototypiques et leur meilleure reconnaissance (voir par exemple Campbell, 2003).

[29] Ce qui peut aisément s'expliquer par le fait que cette sélection a précisément été choisie en faisant varier l' intensité perçue et le taux d'identification.

Ces résultats ont par ailleurs montré que la fréquence d'instanciation, bien qu'ayant un pouvoir prédictif des jugements de typicalité beaucoup plus faible que celui de la similarité aux idéaux, explique néanmoins une partie de la variance unique des jugements de typicalité. Ainsi, dans le cas des expressions vocales des émotions, la distinction entre structuration des catégories dérivées du but et des catégories taxonomiques communes est moins évidente que dans les exemples initiaux introduits par Barsalou (1985) : tout en correspondant plutôt à la structuration de catégories dérivées du but, les expressions vocales partageraient ainsi certaines propriétés des catégories taxonomiques. Barsalou (1987, revenant sur ses propositions initiales, a ainsi noté que la structuration des catégories est un phénomène extrêmement flexible, les frontières entre catégories dérivées du but et taxonomiques n'étant pas clairement marquées.

Il est intéressant de noter que la répartition de l'influence des prédicteurs de la typicalité que nous avons mesurés est proche de la distribution trouvée par Horstmann (2002) pour les expressions faciales. Cette considération l'a mené à considérer ces expressions comme des catégories dérivées du but mais partageant certaines propriétés des catégories taxonomiques. Horstmann (2002, p. 303) a ainsi émis l'hypothèse que

« *cette répartition pourrait être caractéristique de signaux évolués qui apparaissent naturellement et sont utilisés par les receveurs pour réguler leurs actions dans des domaines biologiquement pertinents* »[30].

Si l'on considère l'expression des émotions comme issu de réponses adaptatives à l'environnement, cela peut s'appliquer tant à leurs expressions faciales que vocales.

Ainsi les expressions vocales des émotions seraient plutôt structurées en fonction du but, ce qui pourrait en partie expliquer que les expressions considérées comme les plus typiques puissent être très rares, voire inexistantes dans la réalité.

10.1.7 Limites de l'étude de la typicalité

Une limite potentielle de cette étude est liée à la complexité des tâches proposées aux auditeurs, qui peuvent avoir surinterprété certaines des questions auxquelles il leur était demandé de répondre. En particulier, les jugements de fréquence d'instanciation semblent susceptibles d'avoir été partiellement confondus avec les jugements de

[30] Traduction personnelle de la citation originale : *"[this] pattern may be characteristic for evolved signals which are naturally occurring and are used by receivers to regulate their actions in biologically relevant domains"*

similarité aux idéaux, ici définis comme l'efficacité avec laquelle l'émotion est transmise. En effet, bien que la formulation de la question ait été révisée afin d'améliorer sa compréhension et que nous ayons insisté sur ce point lors des explications données avant le test, les sujets ont plus fréquemment fait état d'une difficulté à accomplir cette tâche d'évaluation que les autres. Par ailleurs, et bien que cela ait également été largement développé dans les explications préalables tant orales qu'écrites, plusieurs participants ont mentionné une difficulté à considérer conceptuellement des expressions peu activées comme relevant d'une classe large étiquetée joie, peur ou colère, étiquettes généralement associées à des émotions plus intenses.

De plus, contrairement à Barsalou (1985) et Horstmann (2002), nous n'avons pas recueilli de jugements de similarité par paire de stimuli mais nous avons eu recours à une mesure de similarité à la tendance centrale estimée à partir des scores d'identification relatifs au sein de la catégorie. En raison du nombre élevé (630) de paires possibles au sein de chaque classe d'émotion, il est peu réaliste de faire évaluer la similarité des paires correspondant aux 3 classes d'émotion par des auditeurs naïfs en une seule session. Néanmoins, l'évaluation des 630 paires au sein de l'une des classes représente une tâche d'environ 1h30 et reste envisageable. Nous avons ainsi mis en place un tel test auprès de 6 premiers auditeurs, mais n'avons pu exploiter les résultats en raison d'un dysfonctionnement du logiciel utilisé pour présenter les stimuli et collecter les réponses. Cette campagne d'évaluation perceptive de la similarité par paire reste toutefois planifiée, afin de compléter les résultats présentés dans cette section. L'idée d'une mesure de similarité à la tendance centrale fondée sur des distances acoustiques, qui permettrait de s'abstraire des jugements subjectifs émis par les auditeurs, peut paraître séduisante. Cependant une telle solution présente un biais méthodologique important : pour pouvoir définir une distance acoustique pertinente en terme de différence de stratégie d'expression d'une émotion donnée, il faudrait disposer d'une modélisation acoustique fine des expressions vocales des émotions, ce qui est précisément ce que nous cherchons à établir. Quand bien même une distance définie sur un ensemble de paramètres moyennés sur la longueur de l'énoncé serait considérée comme une approximation acceptable, la question du choix des paramètres pertinents et de leur poids relatif reste entière.

Enfin, cette étude est à notre connaissance la première à tenter de répondre à la question d'une structuration des catégories en fonction du but ou taxonomique en s'appuyant sur des expressions vocales d'émotions. Avant d'être tenté de généraliser les résultats obtenus à l'ensemble des expressions vocales des émotions, il convient donc de la répliquer avec des stimuli produits et perçus dans d'autres contextes ainsi qu'avec des classes d'émotions définies différemment. De plus, des informations complémentaires

sur des traits de personnalité des locuteurs et des auditeurs permettraient une étude plus fine de la structuration de telles catégories.

10.1.8 Caractéristiques des expressions actées vs. spontanées

En restreignant l'ensemble considéré aux 108 stimuli sélectionnés, l'intensité émotionnelle perçue est significativement plus importante pour les expressions actées vs. spontanées ($t(90.9)=4.67$ avec correction de Welch ; $p<.001$), en accord avec les résultats de Wilting *et al.* (2006) sur la parole émotionnelle actée et spontanée visuelle. Ce résultat reste valable en considérant l'intégralité des 246 stimuli évalués lors du pré-test, pendant lequel les expressions spontanées ont reçu un jugement moyen d'intensité de 3.4 contre 4.9 pour les expressions actées ($t(145.5)=7.38$ avec correction de Welch ; $p<.001$). De plus, comme noté ci-dessus, les jugements de typicalité et de similarité aux idéaux sont significativement plus élevés dans le cas des expressions actées, confortant le constat de Campbell (2003) quant à la plus grande prototypicalité des expressions émotionnelles actées.

Contrairement à nos attentes, les tests t n'ont révélé aucune différence significative quant aux jugements de typicalité, de similarité aux idéaux ni de fréquence d'instanciation entre les expressions des *Big Six* (acté prototypique) et les expressions rejouées des émotions ressenties par l'acteur au cours du déroulement du scénario Sound Teacher (acté non-prototypique). En effet, nous aurions pu supposer que la validité écologique supérieure des expressions rejouées immédiatement après avoir été ressenties aurait induit des jugements de typicalité moins élevés ainsi qu'une fréquence d'instanciation plus élevée que pour les expressions actées des *Big Six* du fait de leur rareté (Campbell, 2005).

10.2. Discrimination d'expressions actées vs. spontanées

10.2.1 Expressions actées vs. spontanées

La validité des corpus d'émotions produits par des acteurs pour la modélisation de phénomènes propres à la parole expressive spontanée, soulève de nombreux débats, exposés en section 5. Pourtant, très peu d'études, à notre connaissance, se sont focalisées, en production ou en perception, sur les différences entre parole actée et spontanée. Williams & Stevens (1972) ont ainsi comparé des enregistrements d'expressions émotionnelles spontanées capturées *in vivo* aux reproductions par des acteurs professionnels de ces énoncés à partir de la description de la situation. Ils concluent de l'inspection visuelle des spectrogrammes que les deux enregistrements apparaissent comme suffisamment similaires pour considérer comme satisfaisantes les performances

des acteurs. Cependant, Aubergé & Cathiard (2003) ont montré que l'amusement acté audiovisuel pouvait être discriminé de l'amusement spontané produit par les mêmes locuteurs selon le même dispositif expérimental, avec une grande variabilité des performances des sujets indépendamment de la compétence des acteurs. Récemment, Wilting *et al.* (2006) ont induit au moyen de la méthode d'induction de Velten (1968) des humeurs positives et négatives chez des locuteurs néerlandais sans compétences d'acteurs particulières, qui ont ensuite reproduit les mêmes énoncés en simulant les affects ressentis. Une évaluation perceptive a montré que les expressions actées étaient perçues comme plus intenses que les expressions spontanées. Ces résultats ont été confirmés par Shahid *et al.* (2008) qui ont répliqué la même expérience avec des locuteurs et auditeurs pakistanais.

Suivant l'hypothèse formulée par Aubergé (2002) d'une distinction cognitive des affects entre contrôle volontaire et involontaire, exposée en section 4.5. , nous postulons ici que la compétence de simulation, centrale dans la communication expressive, est également utilisée par les acteurs lorsque ceux-ci s'appuient sur des méthodes ayant pour but de produire des expressions les plus proches possibles de celles observées dans la vie de tous les jours et s'appuyant sur le souvenir d'épisodes émotionnels pour la production des affects.

La principale question posée ici est de savoir si des énoncés d'un même locuteur, exprimant des valeurs d'affects similaires, avec un contrôle volontaire vs. involontaire (dans notre cas actés vs. spontanés), peuvent être discriminés par des auditeurs naïfs. Nous chercherons aussi à savoir si tous les auditeurs ont des compétences similaires pour accéder à ces indices. Nous nous appuyons pour cela sur des productions spontanées extraites du corpus Sound Teacher, appariées par locuteur, par classe d'émotion exprimée et par intensité d'émotion perçue à des productions actées correspondant aux expressions des émotions ressenties au cours du déroulement du scénario et jouées immédiatement après l'expérience.

De nombreuses études ont montré que le décodage des expressions émotionnelles devait être considéré comme un processus multimodal. Ainsi Scherer & Ellgring (2007) ont étudié des expressions multimodales d'émotions produites par des acteurs professionnels en considérant conjointement l'expressivité véhiculée par la parole, la face et la gestualité. Par ailleurs, Aubergé & Cathiard (2003) ont montré que l'information acoustique peut être intégrée dans le décodage visuel de l'information affective, quand bien même les expressions faciales sont porteuses d'informations suffisantes pour permettre une identification presque parfaite. Nous considérons donc ici des expressions multimodales (transmises par la voix et la face), et cherchons à évaluer la part

d'information relative portée par ces modalités visuelle et audio, ainsi que par les deux modalités conjointes.

10.2.2 Sélection

Nous avons sélectionné un ensemble de 48 stimuli audiovisuels produits par 4 sujets acteurs, afin de former 24 paires comprenant chacune une expression actée et une expression spontanée. Quatre locuteurs parmi les six dont les productions actées et spontanées sont utilisés dans l'étude des déterminants de la typicalité présentée en section 10.1. , ont été retenus. Il s'agit de ceux qui nous ont semblé montrer le plus d'expressivité dans leurs productions spontanées : 2 hommes âgés de 21 et 25 ans que nous désignerons dans le suite comme M1 et M2, et 2 femmes de 19 et 35 ans que nous désignerons comme F1 et F2.

Les paires de stimuli sélectionnées, appariés par locuteur, type d'énoncé (mot monosyllabique de couleur ou « page suivante ») et émotion, correspondent à l'une des classes d'émotions utilisées pour l'étude des déterminants de la typicalité : joie, colère ou peur. Étant donné que les énoncés spontanés ne comprennent pas d'émotions intenses, les expressions retenues sont la satisfaction, l'irritation, l'anxiété et l'inquiétude. Les deux stimuli choisis dans la même paire véhiculent, autant que possible, un niveau d'intensité émotionnelle similaire.

Ainsi, 33 stimuli parmi les 48 sélectionnés ont été choisis sur la base des résultats du pré-test réalisé précédemment, avec comme critère prioritaire la similarité des jugements d'intensité émotionnelle perçue et du taux de catégorisation correcte parmi les classes émotionnelles considérées. Cependant, certaines expressions étant peu représentées parmi les stimuli du pré-test (en particulier les expressions de satisfaction des locuteurs F2 et M1), 15 stimuli ont dû être choisis en dehors de cet ensemble. En conséquence, il n'est possible de calculer la différence entre intensités émotionnelles perçues entre deux stimuli appariés que pour 11 des 24 paires. La Table 26 présente l'intensité émotionnelle perçue et le taux de catégorisation correcte mesurés lors du pré-test pour chacun des stimuli pour lesquels cette information est disponible. Les différences d'intensité perçue et de taux de confiance moyen au sein des paires sont également présentées en Table 26. Ainsi, pour les 11 paires comparables en fonction de l'intensité perçue et du taux moyen de catégorisation correcte, les stimuli actés ont été en moyenne jugés avec une intensité émotionnelle de 0.5 points supérieure à celle des stimuli spontanés. Ils ont été correctement catégorisés dans une proportion supérieure de 13% à celle des stimuli spontanés.

Locuteur		Anxiété		Irritation		Satisfaction	
		mono.	p. suiv.	mono.	p. suiv.	mono.	p. suiv.
F1	Spontané	3.3 (47%)		3.5 (73%)	4.3 (67%)		3.5 (80%)
	Acté	4.5 (93%)	3.2 (60%)	5.5 (87%)	4.3 (87%)	3.5 (80%)	5.7 (100%)
	Diff. A-S	1.2 (47%)		2.1 (13%)	0.0 (20%)		2.2 (20%)
F2	Spontané	4.8 (100%)	2.7 (67%)		5.7 (100%)		
	Acté	6.0 (100%)		3.0 (60%)	4.1 (80%)		3.0 (80%)
	Diff. A-S	1.2 (0%)			-1.7 (-20%)		
M1	Spontané	4.5 (93%)		2.9 (53%)	2.8 (60%)		3.3 (67%)
	Acté	3.5 (73%)	2.9 (60%)	3.3 (93%)	2.3 (80%)	3.9 (73%)	
	Diff. A-S	-1.1 (-20%)		0.5 (40%)	-0.5 (20%)		
M2	Spontané		4.8 (67%)	2.3 (40%)		7.3 (100%)	
	Acté	2.8 (67%)		6.7 (93%)		4.5 (67%)	2.2 (73%)
	Diff. A-S			4.4 (53%)		-2.8 (-33%)	

Table 26 : Intensité émotionnelle perçue et taux de catégorisation correcte (entre parenthèses) évalués en condition audio dans le pré-test (voir supra) pour chacun des stimuli retenus. La différence de mesure entre l'expression actée et l'expression spontanée est présentée pour chaque paire à la ligne étiquetée Diff. A-S. Les cases vides correspondent aux stimuli non évalués lors du pré-test. Les abréviations mono et p. suiv. correspondent aux types d'énoncés utilisés, respectivement monosyllabe et page suivante.

10.2.3 Evaluation perceptive

Les 24 paires de stimuli ont été présentées aux juges séquentiellement, avec une latence de 1.5 secondes, selon 3 conditions de présentation : audio seul (A), visuel seul (V) et audiovisuel (AV). Les stimuli ont été regroupés par condition et triés aléatoirement au sein de chaque condition. La moitié des sujets ont passé la condition A avant V et vice-versa, tandis que la condition AV était toujours présentée en dernier. Chaque paire a été présentée deux fois dans chaque condition, en inversant l'ordre acté/spontané au sein de la paire. Les sujets devaient donc évaluer 144 paires (4 locuteurs x 3 classes d'émotion x 2 types d'énoncé x 2 ordres dans la paire x 3 conditions) lors de l'expérience.

La tâche des sujets étaient d'indiquer pour chaque paire lequel des stimuli correspond à une expression spontanée. La réponse était donnée à l'aide d'un curseur allant de « certainement le premier » à « certainement le deuxième ». Le curseur, destiné à capturer à la fois la discrimination des expressions actées et la confiance du sujet en cette discrimination, était initialement placé en position centrale. Les sujets devaient obligatoirement déplacer ce curseur pour valider leur réponse.

La présentation des stimuli et l'enregistrement des réponses ont été automatisés au moyen d'une interface développée à l'aide du logiciel *Revolution Runtime*. Les instructions sur le contexte d'enregistrement des stimuli étaient communiquées aux

sujets par écrit et oralement au début du test. L'interface destinée à recueillir les réponses est présentée dans la Figure 53.

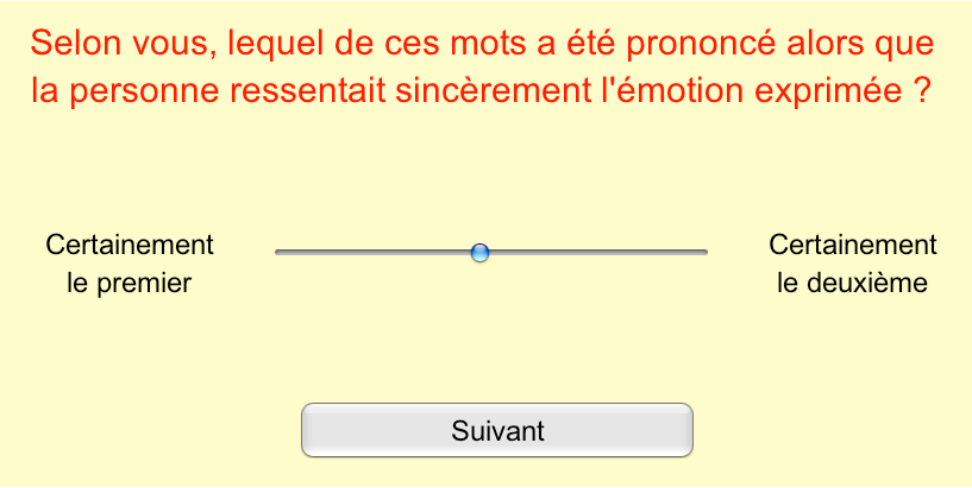


Figure 53 : Page de réponse de l'interface utilisée pour le test perceptif de discrimination des expressions actées vs. spontanées.

33 juges francophones sans problèmes particuliers d'audition ni trouble de la vision non corrigé (15 hommes, 18 femmes, âgés en moyenne de 33.1 ans) ont participé à cette évaluation perceptive, d'une durée moyenne de 25 minutes.

10.2.4 Analyse statistique

Considérer une mesure conjointe de la discrimination correcte et de la confiance accordée à la réponse donnerait un poids plus important aux réponses associées à un taux de confiance élevé. Ainsi, nous avons choisi d'analyser séparément ces deux variables et de centrer notre analyse sur les taux de discrimination des paires présentées.

Locuteur	A	V	AV	Total
F1	56% (60%)	57% (60%)	72% (68%)	62% (63%)
F2	60% (58%)	51% (51%)	60% (57%)	57% (56%)
M1	55% (59%)	53% (53%)	70% (62%)	59% (58%)
M2	73% (64%)	64% (63%)	77% (67%)	71% (65%)
Anxiété	55% (60%)	58 % (59%)	72% (65%)	Anxiété
Irritation	60% (60%)	63% (58%)	70% (63%)	Irritation
Satisfaction	67% (61%)	48% (53%)	68% (63%)	Satisfaction
Total	61% (60%)	56% (67%)	70% (64%)	62% (60%)

Table 27 : Valeurs moyennes de discrimination correcte et de taux de confiance (entre parenthèses) par locuteur et par condition.

Les positions du curseur ont ainsi été converties en scores de discrimination (bonne ou mauvaise réponse, codées respectivement par 1 et 0) en fonction du côté choisi. Le taux de confiance associé à chaque réponse a été calculé comme la distance entre la position

finale du curseur et sa position centrale initiale. Les valeurs moyennes de discrimination et de confiance pour chaque condition sont présentées pour chaque locuteur et chaque classe d'émotion dans la Table 27.

Les scores de discrimination ne sont que faiblement corrélés au taux de confiance quelque soit la condition, bien que ces corrélations soient significatives (condition A : $r=0.152$; V : $r=0.207$, AV : $r=0.152$; global : $r=0.176$; $p<0.001$ pour toutes les corrélations). Néanmoins un test t pour échantillons indépendants ($t(3260)=-11.8$; $p<0.001$, avec correction de Welch) révèle que les discriminations correctes l'ont été avec un degré de confiance moyen (64.8%) significativement plus élevé que les discriminations incorrectes (52.7%).

Les scores de discrimination et les taux de confiance ont été analysés au moyen d'une ANOVA à mesures répétées prenant le locuteur, la classe émotionnelle, la condition de présentation, la longueur de l'énoncé et l'ordre de présentation comme facteurs fixés, et le genre du juge comme variable inter-sujets. Les principaux résultats des deux ANOVAs à mesures répétées sont présentés dans la Table 28 pour les effets sur les scores de discrimination et dans la Table 29 pour les effets sur les taux de confiance. Les effets des interactions du facteur genre des sujets avec les différents effets et interactions intra-sujets testés, non-significatifs, ne sont pas reproduits dans ces tables.

Afin de tester les effets de ces facteurs pour chaque locuteur et chaque condition de présentation, quatre ANOVAs à mesures répétées ont été effectuées séparément sur les scores de discrimination correspondant aux productions de chaque locuteur, et trois ANOVAs à mesures répétées sur les scores de discrimination correspondant à chaque condition de présentation. Ces ANOVAs, dont les principaux résultats sont présentés dans les pages suivantes, ont été effectuées sur les mêmes facteurs que dans le cas de l'ANOVA à mesures répétées effectuées sur l'ensemble des données, à l'exception de celui utilisé comme critère de séparation.

En raison de l'impossibilité de spécifier le sujet lui-même comme variable inter-sujets dans une ANOVA à mesures répétées, deux analyses de variance univariées avec respectivement le score de discrimination et le taux de confiance comme variable dépendante et le juge comme facteur fixé ont été effectuées afin d'évaluer l'effet éventuel du juge sur ces variables.

Un test t bidirectionnel pour échantillon unique a été effectué pour chaque juge et chaque condition de présentation afin de comparer les scores de discrimination correspondants au niveau théorique du hasard (50%). Les résultats complets de ces tests t sont présentés en Annexe VII : avec les statistiques descriptives correspondantes.

Effet	Degrés de liberté	F	p	Significativité
émotion	(1.75, 54.20)	1.14	0.3205	n.s.
locuteur	***(3, 93)***	***16.04***	***0.0000***	***
condition	***(2, 62)***	***21.33***	***0.0000***	***
énoncé	***(1, 31)***	***6.33***	***0.0173***	*
ordre	***(1, 31)***	***8.32***	***0.0071***	**
émotion * locuteur	***(6, 186)***	***4.92***	***0.0001***	***
émotion * condition	***(3.74 , 116.07)***	***8.65***	***0.0000***	***
locuteur * condition	***(6, 186)***	***2.99***	***0.0082***	**
émotion * locuteur * condition	***(11.19, 347.04)***	***2.93***	***0.0009***	***
émotion * énoncé	***(2, 62)***	***5.61***	***0.0058***	**
locuteur * énoncé	(3, 93)	0.57	0.6342	n.s.
émotion * locuteur * énoncé	***(6, 186)***	***13.58***	***0.0000***	***
condition * énoncé	***(2, 62)***	***5.82***	***0.0048***	**
émotion * condition * énoncé	***(4, 124)***	***7.43***	***0.0000***	***
locuteur * condition * énoncé	(6, 186)	0.86	0.5219	n.s.
émotion * locuteur * condition * énoncé	***(12, 372)***	***3.05***	***0.0004***	***
émotion * ordre	(2, 62)	0.14	0.8663	n.s.
locuteur * ordre	(2.66, 82.40)	0.35	0.7658	n.s.
émotion * locuteur * ordre	(6, 186)	1.13	0.3458	n.s.
condition * ordre	(2, 62)	2.26	0.1131	n.s.
émotion * condition * ordre	(4, 124)	0.73	0.5756	n.s.
locuteur * condition * ordre	(6, 186)	1.59	0.1534	n.s.
émotion * locuteur * condition * ordre	(12, 372)	0.59	0.8519	n.s.
énoncé * ordre	(1, 31)	0.15	0.7021	n.s.
émotion * énoncé * ordre	***(1.76, 54.62)***	***5.57***	***0.0083***	**
locuteur * énoncé * ordre	(3, 93)	0.70	0.5560	n.s.
émotion * locuteur * énoncé * ordre	(6, 186)	1.16	0.3301	n.s.
condition * énoncé * ordre	(2, 62)	0.57	0.5707	n.s.
émotion * condition * énoncé * ordre	(4, 124)	1.63	0.1701	n.s.
locuteur * condition * énoncé * ordre	(6, 186)	1.34	0.2433	n.s.
émotion * locuteur * condition * énoncé * ordre	(12, 372)	1.07	0.3828	n.s.
sexe du sujet	(1, 31)	0.70	0.4084	n.s.

Table 28 : Résultats de l'ANOVA à mesures répétées avec le score de discrimination comme variable dépendante. Les degrés de libertés sont rectifiés à l'aide de la correction de Huynh-Feldt afin de compenser les éventuelles violations de la condition de sphéricité. Les notations ***, **, * et n.s. correspondent respectivement aux niveaux de significativité $p<0.001$, $p<0.01$, $p<0.05$ et non significatif. Les effets significatifs sont repérés par une notation en gras et italiques.

Effet	Degrés de liberté	F	p	Significativité
émotion	***(2, 62)***	***4.16***	***0.0202***	*
locuteur	***(2.16, 67.09)***	***26.44***	***0.0000***	***
condition	***(2, 62)***	***5.75***	***0.0051***	**
énoncé	***(1, 31)***	***13.90***	***0.0008***	***
ordre	***(1, 31)***	***13.62***	***0.0009***	***
émotion * locuteur	***(6, 186)***	***6.02***	***0.0000***	***
émotion * condition	***(4, 124)***	***3.58***	***0.0085***	**
locuteur * condition	***(6, 186)***	***3.91***	***0.0011***	**
émotion * locuteur * condition	***(10.98, 340.47)***	***3.61***	***0.0001***	***
émotion * énoncé	***(2, 62)***	***3.69***	***0.0305***	*
locuteur * énoncé	***(3, 93)***	***3.96***	***0.0105***	*
émotion * locuteur * énoncé	***(6, 186)***	***4.47***	***0.0003***	***
condition * énoncé	***(2, 62)***	***6.27***	***0.0033***	**
émotion * condition * énoncé	(4, 124)	1.94	0.1074	n.s.
locuteur * condition * énoncé	***(6, 186)***	***5.34***	***0.0000***	***
émotion * locuteur * condition * énoncé	(11.01, 341.15)	1.42	0.1598	n.s.
émotion * ordre	(2, 62)	0.94	0.3945	n.s.
locuteur * ordre	***(2.68, 83.14)***	***3.00***	***0.0407***	*
émotion * locuteur * ordre	(5.38, 166.66)	1.78	0.1150	n.s.
condition * ordre	(1.72, 53.33)	0.94	0.3855	n.s.
émotion * condition * ordre	(4, 124)	0.19	0.9407	n.s.
locuteur * condition * ordre	(6, 186)	1.19	0.3108	n.s.
*émotion * locuteur * condition * ordre*	*(12, 372)*	*2.14*	*0.0143*	*
énoncé * ordre	(1, 31)	0.43	0.5184	n.s.
émotion * énoncé * ordre	(2, 62)	2.42	0.0974	n.s.
locuteur * énoncé * ordre	***(3, 93)***	***4.34***	***0.0066***	**
émotion * locuteur * énoncé * ordre	(6, 186)	0.44	0.8541	n.s.
condition * énoncé * ordre	(2, 62)	0.74	0.4815	n.s.
émotion * condition * énoncé * ordre	(4, 124)	0.62	0.6497	n.s.
locuteur * condition * énoncé * ordre	(6, 186)	0.71	0.6421	n.s.
émotion * locuteur * condition * énoncé * ordre	(12, 372)	0.91	0.5418	n.s.
sexe du sujet	(1, 31)	1.85	0.1842	n.s.

Table 29 : Résultats de l'ANOVA à mesures répétées avec le taux de confiance comme variable dépendante. Les degrés de libertés sont rectifiés à l'aide de la correction de Huynh-Feldt afin de compenser les éventuelles violations de la condition de sphéricité. Les notations ***, **, * et n.s. correspondent respectivement aux niveaux de significativité p<0.001, p<0.01, p<0.05 et non significatif. Les effets significatifs sont repérés par une notation en gras et italiques.

Pour l'ensemble des comparaisons multiples présentées dans cette section, la significativité indiquée a été corrigée en utilisant la méthode séquentielle proposée par Benjamini & Hochberg (1995) pour le contrôle du faux rejet d'hypothèses nulles.

10.2.5 Un fort effet inter-juge

Le résultat le plus remarquable est le fort effet du juge tant sur les scores de discrimination ($F(32, 4719) = 10.2$; $p<0.001$) que sur les taux de confiance ($F(32, 4719) = 163.6$; $p<0.001$), qui confirme les résultats obtenus sur l'amusement par Aubergé &

Cathiard (2003) : les performances des auditeurs s'étagent en effet de 32.7% à 80.6% de paires correctement discriminées.

Les tests t présentés en Annexe VII : révèlent que 32% des juges globalement, et jusqu'à 64% en condition audiovisuelle, discriminent correctement plus de la moitié des paires présentées et réalisent donc une performance meilleure que le hasard, tandis que deux juges (6%) discriminent moins bien que le hasard en condition V ($p<0.05$). Cette distribution est récapitulée dans la Table 30.

Discrimination	A	V	AV	Total
Inférieure au hasard	0	2	0	2%
Non significativement différente du hasard	27	26	12	66%
Supérieure au hasard	6	5	21	32%

Table 30 : Nombre de juges par condition ayant correctement discriminé les paires présentées avec une fréquence respectivement inférieure au hasard, non significativement différente du hasard et supérieure au hasard.

La Figure 54 présente la distribution des scores de discrimination obtenus par les différents juges.

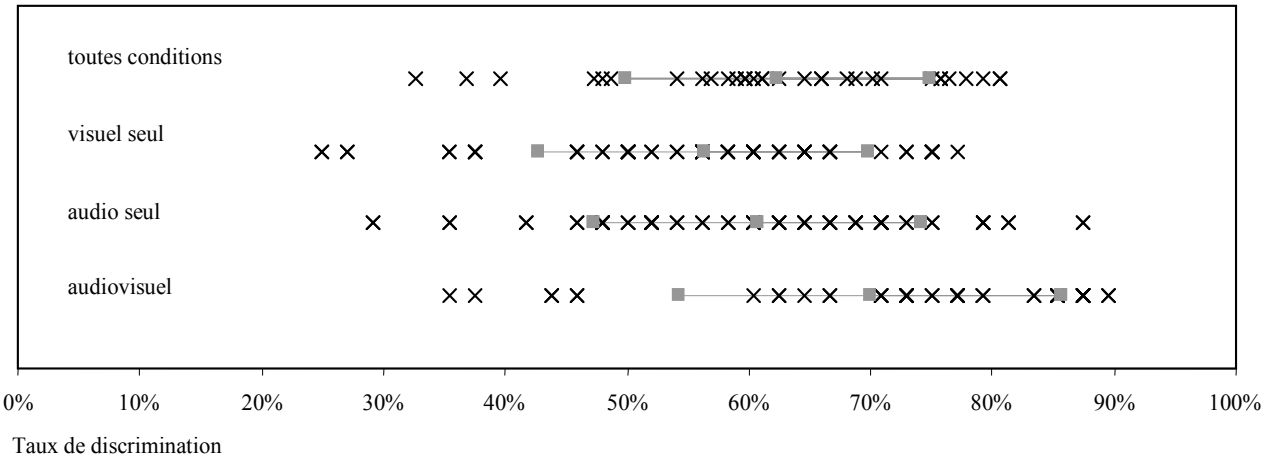


Figure 54 : Distribution des scores individuels de discrimination des expressions actées vs. spontanées produites par les locuteurs M1, M2, F1 et F2, par condition de présentation. Les lignes et carrés gris indiquent la moyenne et l'écart type pour chaque condition.

Malgré la grande variabilité des performances des juges, aucune préférence marquée pour un locuteur ne ressort, indépendamment de la qualité supposée de leurs performances d'acteurs. En effet, comme le montre la valeur élevée du coefficient alpha de Cronbach calculé entre juges sur les taux moyens de discrimination correspondants aux productions de chacun des locuteurs (α = 0.940), les performances des juges ont été consistantes d'un locuteur à l'autre.

Bien que les scores de discrimination et de confiance attribués par les juges soient légèrement plus élevés pour les femmes, particulièrement en conditions A (63% vs. 58%)

et AV (72% vs. 67%), un résultat surprenant est l'absence d'effet significatif du genre du juge tant sur le taux de discrimination correcte (p = 0.408 n.s.) que sur le taux de confiance (p = 0.184 n.s.). En effet de nombreuses études, pour la plupart centrées sur les expressions faciales mais comprenant, pour certaines des conditions audio et audiovisuelle, indiquent que les femmes perçoivent plus efficacement et plus rapidement les indices émotionnels que les hommes (voir Hall (1978, 1984) et Hall *et al.* (2000) pour trois états de l'art distincts). On aurait donc pu s'attendre à ce que les femmes soient également plus performantes que les hommes pour discriminer les expressions émotionnelles spontanées des expressions actées, ce qui ne semble pas être le cas ici. Ce résultat doit cependant être nuancé en raison de la faiblesse de la puissance de ces tests (13% dans le cas de la discrimination, 26% dans le cas de la confiance)[31], qui ne nous permet pas d'accepter l'hypothèse nulle H0 avec certitude. L'évaluation de telles différences inter-genre devrait donc faire l'objet d'une évaluation spécifique avec un nombre de sujets plus élevé.

Un résultat particulier mérite en outre d'être relevé : pour 2 juges masculins, S23 et S30 (cf. Annexe VII :), les performances en condition A sont sensiblement meilleures qu'en condition AV. Pour ces 2 juges, l'interprétation erronée de l'information visuelle semble ainsi avoir largement grevé leurs performances de discrimination en condition AV.

10.2.6 Effet de la condition de présentation et du locuteur

Un effet très significatif de la condition de présentation sur les scores de discrimination ($F(2, 62)=21.3$, $p<0.001$) a été observé, avec un gain significatif de discrimination pour la condition AV par rapport à A et V (respectivement $F(1, 31)=17.2$ et $F(1, 31)=65.7$, $p<0.001$ dans les deux cas). En revanche, la différence entre A et V n'est pas significative ($F(1, 31) =3.3$; $p=0.081$ n.s.). Cette nette amélioration des performances en condition AV montre l'important potentiel de synergie des informations de chaque modalité pour la discrimination des expressions actées et spontanées. Cela souligne l'intérêt d'autant plus important de s'appuyer sur des données spontanées lorsque l'objectif est de modéliser des expressions émotionnelles multimodales. Cependant cet avantage global pour la condition AV n'est pas consistant pour tous les locuteurs (cf. Table 27). Le gain entre A et

[31] Rietveld & van Hout (2005) indiquent que la puissance d'un effet doit être au minimum de 80% pour pouvoir accepter une hypothèse nulle. Dans les analyses de variance que nous avons effectuées dans le cadre des expériences présentées dans les sections précédentes, nous n'avons pas mesuré la puissance observée de l'effet. Il convient donc de nuancer la portée de certains des résultats que nous obtenons, comme par exemple l'absence d'effet de la longueur des énoncés dans l'évaluation perceptive présentée en section 8.4.

AV s'est avéré non significatif pour la locutrice F2 ($F(1, 31)=0.1$; $p=0.737$ n.s.) et le locuteur M2 ($F(1, 31)=2.2$; $p=0.148$ n.s.). La stratégie du locuteur importe donc pour l'utilisation des modalités.

Un effet très significatif du locuteur sur les scores de discrimination ($F(3, 93)=16.0$, $p<0.001$) et le taux de confiance ($F(2.2, 67.1)=26.44$, $p<.001$) a également été observé. Les contrastes par paires entre locuteurs sont présentés dans la Table 31 (cf. Table 27 pour les valeurs correspondantes). Seules les productions du locuteur M2 ont été significativement mieux discriminées que celles des autres locuteurs. Ce résultat indique des compétences moindres de cet acteur pour simuler des expressions émotionnelles similaires à ses productions spontanées. Un nombre important de juges a considéré la tâche de discrimination plus difficile pour les locuteurs F2 et M1. Ceci s'est traduit par des scores de confiance significativement plus faibles que ceux attribués aux productions des autres locuteurs. Toutefois, les paires produites par les locuteurs M1, F1 et F2 ont été discriminées avec des scores comparables.

		F2	M1	M2
F1	discrimination	F(1, 31) = 3.90 p = 0.0573 n.s.	F(1, 31) = 1.92 p = 0.1754 n.s.	F(1, 31) = 16.75 p = 0.0003 **
	confiance	F(1, 31) = 23.19 p = 0.0000 ***	F(1, 31) = 19.79 p = 0.0001 ***	F(1, 31) = 10.35 p = 0.0030 **
F2	discrimination		F(1, 31) = 0.64 p = 0.4313 n.s.	F(1, 31) = 53.14 p = 0.0000 ***
	confiance		F(1, 31) = 5.47 p = 0.0260 *	F(1, 31) = 40.26 p = 0.0000 ***
M1	discrimination			F(1, 31) = 32.87 p = 0.0000 ***
	confiance			F(1, 31) = 43.61 p = 0.0000 ***

Table 31 : Contrastes par paire des scores de discrimination et taux de confiance obtenus pour les productions des différents locuteurs. Les notations ***, **, * et n.s. correspondent respectivement aux niveaux de significativité $p<0.001$, $p<0.01$, $p<0.05$ et non significatif. Les niveaux de significativité indiqués sont corrigés avec la procédure de Bonferroni modifiée de Benjamini & Hochberg (1995).

10.2.7 Effet de l'émotion

Aucun effet global de la classe d'émotion sur le score de discrimination n'a été trouvé ($F(1.8, 54.2)=1.14$, $p=0.321$ n.s.). Les juges auraient donc des compétences similaires pour discriminer les productions actées vs. spontanées quelle que soit l'émotion. Cependant, la faible puissance de l'effet (36%) nous amène à relativiser cette conclusion. De plus, un

effet global de l'émotion sur le taux de confiance ($F(2, 62)=4.2$, $p<0.05$) est observé. En effet, les expressions d'anxiété sont discriminées avec une confiance légèrement supérieure, ce qui tendrait à indiquer qu'elles présentent des indices plus faciles à noter de façon consciente.

Les effets des interactions émotion*condition et émotion*locuteur sur les scores de discrimination ainsi que sur les taux de confiance sont significatifs ($p<0.001$ dans tous les cas sauf pour l'interaction émotion*condition sur la confiance, $p<0.01$, cf. Table 28 et Table 29).

Les valeurs moyennes de discrimination et de confiance par émotion et par condition sont présentées dans la Table 32, avec l'effet de l'émotion sur ces variables au sein de chaque condition de présentation (ANOVAs à mesures répétées). Le principal résultat concerne la discrimination des expressions de satisfaction. En effet, les paires de stimuli exprimant de la satisfaction ont été significativement mieux discriminés que les autres en condition A ($F(1,31)=6.66$ entre irritation et satisfaction, $p<0.05$), tandis que ces mêmes paires ont été significativement moins bien discriminées en condition V ($F(1,31)=4.49$ entre anxiété et satisfaction, $p<0.05$). Ces variations se compensent en condition AV. Bien que cette moins bonne discrimination en condition V soit associée à un taux de confiance moindre ($F(1,31)=13.77$ entre anxiété et satisfaction, $p<0.001$), ce n'est pas le cas en condition A avec une absence d'effet de l'émotion sur le taux de confiance.

Émotion	A	V	AV
Anxiété	55% (60%)	58 % (59%)	72% (65%)
Irritation	60% (60%)	63% (58%)	70% (63%)
Satisfaction	67% (61%)	48% (53%)	68% (63%)
Effet(émotion)	F(2,62) =8.07 *** (F(2,62) =0.35 n.s.)	F(2,62) =6.93 ** (F(2,62) =7.58 **)	F(2,62) =0.63 n.s. (F(2,62) =3.10 n.s.)

Table 32 : Valeurs moyennes de discrimination correcte et de taux de confiance (entre parenthèses) par émotion et par condition, et effet de l'émotion sur ces variables dans chaque condition (ANOVAs à mesures répétées).

Les valeurs moyennes de discrimination et de confiance par émotion et par locuteur sont présentées dans la Table 33. Si pour les locuteurs F2 et M1 aucun effet de l'émotion n'est observé ni sur les scores de discrimination ni sur les taux de confiance, les expressions de satisfaction produites par la locutrice F1 sont significativement moins bien discriminées ($F(1,31)=5.92$ entre satisfaction et irritation, $p<0.05$). Le locuteur M2 montre quant à lui une importante variabilité dans sa capacité à produire des expressions actées similaires à ses expressions spontanées selon l'émotion exprimée : ses expressions d'irritation sont mieux discriminées que ses expressions d'anxiété ($F(1,31)=22.88$; $p<0.001$), les

expressions de satisfaction étant également mieux discriminées que celles d'anxiété ($F(1,31)=4.34$; $p<0.05$)

Émotion	F1	F2	M1	M2
Anxiété	64% (66%)	57% (57%)	62% (60%)	64% (63%)
Irritation	66% (59%)	53% (55%)	58% (57%)	79% (69%)
Satisfaction	55% (62%)	61% (55%)	58% (57%)	71% (62%)
Effet(émotion)	F(2,62) =3.40 * (F(2,62) =8.75 ***)	F(2,62) =1.94 n.s. (F(2,62) =0.62 n.s.)	F(2,62) =1.22 n.s. F(2,62) =3.30 *)	F(2,62) =9.86 *** (F(2,62) =15.41 ***)

Table 33 : Valeurs moyennes de discrimination correcte et de taux de confiance (entre parenthèses) par émotion et par locuteur, et effet de l'émotion sur ces variables pour chaque locuteur (ANOVAs à mesures répétées).

10.2.8 Autres effets

L'effet de l'ordre de présentation des stimuli dans la paire (stimulus spontané en 1ère ou 2ème position) sur les scores de discrimination s'est révélé significatif ($p<0.001$). Ainsi, les paires sont mieux discriminées lorsque le stimulus spontané est présenté en premier (65.0%) que lorsque le stimulus acté est présenté en premier (59.7%). Quelques juges ayant rapporté avoir placé le curseur destiné à indiquer leurs réponses à une position très proche du centre dans les cas d'indécision, nous avons examiné un possible biais expérimental qui aurait consisté en un léger déplacement du curseur vers la gauche et la réponse « certainement le premier » plutôt que vers la droite en cas d'indécision. Cependant, un test t bilatéral pour échantillons indépendants montre que la proportion de déplacements vers la gauche pour les 11% de réponses associées à un déplacement de moins de 5% de la longueur totale du curseur (soit un taux de confiance associé à la réponse de 10%) n'est pas significativement différente de celle parmi les 89% restants ($t(4750)=0.535$ n.s.).

Une analyse plus approfondie montre que cette différence n'est significative qu'en condition V ($F(1, 31)=10.93$, $p<0.01$), et uniquement pour les deux locuteurs pour lesquels le gain entre les conditions A et AV est non significatif. Le locuteur le mieux discriminé est M2 ($F(1, 31)=5.68$, $p>0.05$) et la locutrice la moins bien discriminée est F2 ($F(1,31) = 4.93$, $p<0.05$). Bien que les expressions faciales et la gestualité des locuteurs n'aient pas fait l'objet d'une analyse objective systématique, l'amplitude des gestes de M2 et F2 est de toute évidence plus importante que celle des autres locuteurs. De plus, tous deux bougeaient quasi-systématiquement la tête vers le bas pendant leurs productions spontanées. Ainsi, la force informative du stimulus visuel présenté en premier pourrait expliquer ce résultat surprenant de prime abord.

L'effet du type d'énoncé est significatif ($p<0.05$) avec une discrimination légèrement meilleure sur les énoncés [paʒsɥivãt] (64.7%) que sur les monosyllabes (60.0%). Cet effet pourrait provenir de la quantité d'information présentée, supérieure pour ces énoncés de 3 syllabes que pour les énoncés monosyllabiques. Cependant, l'analyse par condition de présentation révèle que cette différence n'est significative qu'en condition AV ($F(1, 31) = 19.2$; $p<0.001$), tandis qu'aucun effet de l'énoncé n'est observé ni en condition A ($F(1,31) = 0.1$; $p =0.916$ n.s.) ni en condition V ($F(1, 31) = 0.8$; $p =0.385$ n.s.). De plus, cet effet n'est significatif que pour les productions du locuteur M1 ($F(1, 31) = 6.4$; $p<0.05$). On retiendra donc plutôt un effet dû aux caractéristiques des productions de ce locuteur sur les deux types d'énoncés proposés qu'un effet global de la quantité d'information présentée.

Les corrélations entre les différences de durée dans chaque paire (de -480 ms à 760 ms) d'une part, et les scores de discrimination et de confiance moyens par paire d'autre part ont été calculés afin d'étudier la possible influence de ces différences de durées sur les performances de discrimination. Cependant les très faibles valeurs de ces corrélations (respectivement $r=0.047$ et $r=0.037$, non significatives) suggèrent que cet indice n'a pas été directement utilisé par les juges.

10.2.9 Rôle de l'intensité émotionnelle perçue

Afin d'évaluer le rôle de l'intensité émotionnelle perçue dans la discrimination des expressions actées vs. spontanées, la corrélation entre la différence d'intensité perçue dans la paire et les scores de discrimination (respectivement de confiance) obtenus par chacune des 11 paires utilisées durant le pré-test a été calculée pour chaque condition de présentation. Ces corrélations, dont aucune n'est significative, sont présentées dans la Table 34.

Condition	A	V	AV	Global
Discrimination	r = 0.428	r = -0.050	r = 0.319	r= 0.245
Degré de confiance	r = 0.252	r = 0.246	r = 0.218	r= 0.213

Table 34 : Corrélations entre la différence d'intensité émotionnelle perçue et le score de discrimination (respectivement confiance). Toutes ces corrélations sont non-significatives ($p>.05$).

Les intensités perçues ayant été attribuées en condition audio seul, il n'est pas surprenant d'observer des corrélations avec les scores de discrimination plus élevées dans cette condition. Une inspection graphique du lien entre intensité émotionnelle perçue et discrimination (Figure 55) révèle néanmoins une relation qui semble quasi linéaire pour 10 des 11 paires considérées. En effet, le taux de discrimination correcte d'une paire est généralement plus élevé lorsque l'expression actée est perçue comme

plus intense que l'expression spontanée. Le point en haut à gauche du graphique qui ne suit pas cette tendance correspond, quant à lui, à la paire composée des expressions actée (jugée avec une intensité moyenne de 4,5) et spontanée (jugée avec une intensité moyenne de 7,3) de satisfaction du locuteur M2 sur des énoncés monosyllabiques. Cette paire est correctement discriminée par 82% des juges. En ne prenant pas cette paire en compte, la corrélation entre la différence d'intensité perçue dans les 10 paires restantes et les scores de discrimination est significative ($r = .889$; $p<.001$).

Bien que les données disponibles soient en quantité trop limitée pour établir solidement un tel résultat, la différence d'intensité de l'émotion perçue entre les 2 membres de la paire à discriminer semble ainsi à même d'expliquer une part importante des scores de discrimination. Cependant, les juges ne semblent pas s'être appuyés uniquement sur cet indice. En effet, outre le contre-exemple flagrant évoqué ci-dessus, toutes les paires pour lesquelles cette différence d'intensité émotionnelle perçue est faible ne sont pas discriminées à un niveau supérieur au hasard, et inversement. Ainsi, la paire d'expressions d'irritation produites par la locutrice F1 sur l'énoncé [paʒsɥivãt] est discriminée par 62% des juges en condition A. Ces deux paires ont été discriminées au dessus du niveau du hasard en condition A (tests t pour échantillon avec la valeur de test 50%, $t(65) = 6.65$; $p<0.05$ pour la 1ère paire, $t(65) = 2.02$; $p<0.05$ pour la seconde paire.

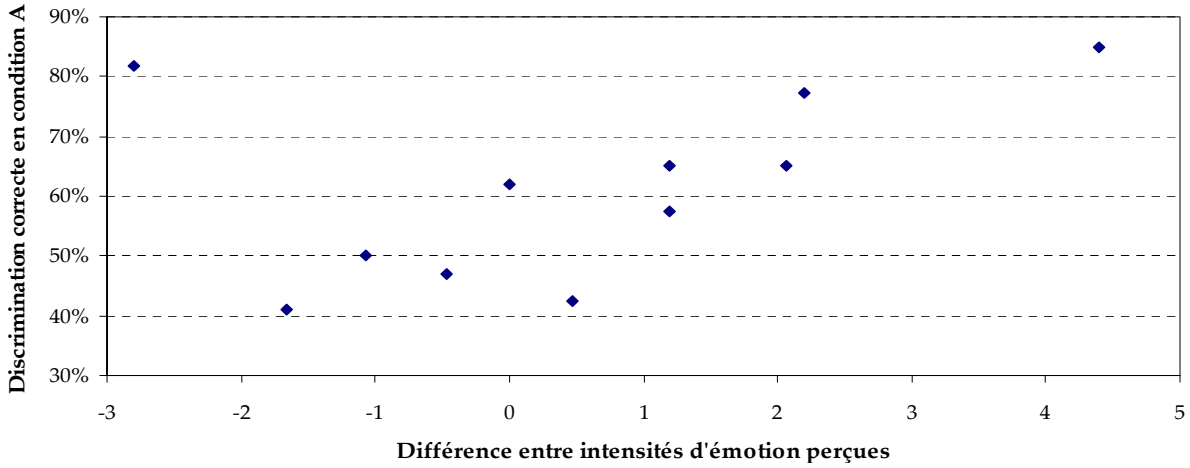


Figure 55 : Taux moyen de discrimination correcte en condition audio en fonction de la différence entre l'intensité émotionnelle perçue pour l'expression actée et celle perçue pour l'expression spontanée, pour les 11 paires évaluées durant le pré-test de l'étude de la typicalité.

La faible quantité de données disponibles ne permet pas ici de confronter les scores de discrimination obtenus par les paires présentées aux jugements de typicalité, de similarité aux idéaux et de similarité à la tendance centrale. En effet, seules 5 paires sont composées de stimuli tous deux évalués dans la seconde phase de l'étude portant sur les déterminants de la typicalité des expressions vocales présentée en section 10.1.

10.2.10 Discussion

Les résultats obtenus suggèrent une capacité globale des auditeurs naïfs à discriminer la nature actée ou spontanée d'expressions émotionnelles multimodales, sans effet global de l'émotion exprimée mais avec un fort effet inter-juge. Les différences d'intensité émotionnelle perçue, mesurée auparavant en condition audio sur les stimuli individuels (section 10.1.), suggèrent que cette différence entre intensités perçues pourrait expliquer une part importante de la capacité à discriminer, sans pour autant être le seul facteur explicatif. Afin de permettre une mise en correspondance plus systématique entre différence d'intensité émotionnelle perçue et discrimination des expressions actées vs. spontanées, nous prévoyons de prolonger cette expérience par une évaluation perceptive complémentaire. Cette dernière reprendrait le même dispositif expérimental avec un jugement de la différence d'intensité de l'émotion perçue pour chaque paire dans toutes les conditions de présentation. De plus, une analyse acoustique et visuelle des stimuli utilisés, fondée sur des hypothèses fortes quant à l'ancrage temporel des expressions volontaires vs. involontaires, pourrait permettre d'isoler certains des indices utilisées par les juges pour discriminer les expressions actées des expressions spontanées.

Afin de pouvoir étudier plus finement les différences entre expressions spontanées et actées au-delà des différences d'intensités d'émotion perçues, il serait souhaitable de pouvoir contrôler cette dimension afin de proposer des paires d'expressions véhiculant une intensité perçue similaire. En l'absence d'un modèle acoustique et facial du contrôle de l'intensité émotionnelle exprimée qui permettrait de produire en synthèse de tels énoncés, la seule approche réaliste pour ce faire semble de passer par une évaluation perceptive de cette intensité sur un grand nombre d'expressions actées et spontanées produites par les mêmes locuteurs afin d'isoler une quantité suffisante de stimuli comparables selon ce critère.

Bien que les acteurs enregistrés ne soient pas nécessairement les plus performants qui puissent être recrutés pour une telle tâche, tous ont été en mesure de « piéger » plus de la moitié des juges sur au moins l'une des paires présentées. Sans préjuger de la capacité d'autres acteurs à piéger un nombre plus important de juges, ou de l'existence de l'hypothétique « acteur parfait », l'utilisation de la parole actée comme référence pour la modélisation de phénomènes relatifs à l'expression d'émotions involontaires pourrait être reconsidérée en tenant compte de cette capacité de discrimination. Au-delà de la discrimination, la question de la variabilité des performances en identification de la simulation (qui est dans notre hypothèse à la base des expressions d'affects sociaux) reste ouverte, ainsi que celle de son possible lien avec la notion d'intelligence affective.

11. Quelques limites de l'étude réalisée

11.1. Protocole de recueil de la partie actée du corpus

Les acteurs recrutés pour l'enregistrement du corpus acté, et détournés pour participer au soi-disant test du logiciel Sound Teacher ont pour la plupart considéré que le protocole retenu, dans lequel il leur était demandé de reproduire les expressions ressenties et exprimées au cours du déroulement du scénario, leur permettait effectivement de se replacer dans un état émotionnel proche de celui ressenti pendant cette tâche. Certains d'entre eux ont cependant indiqué qu'ils avaient été perturbés par le fait de ne pas être libres de leurs mouvements et donc de ne pas pouvoir se placer dans une situation propice à l'élicitation des émotions qui ne correspondaient pas à celles ressenties au cours de l'expérience.

Une autre limite du protocole appliqué dans le cas des acteurs est que ces derniers, hormis les cas dans lesquels ils n'étaient pas satisfaits de leur performance, n'ont pas produit de répétitions du même énoncé pour chacune des émotions visées. Ceci nous a donc limité par la suite dans les sélections que nous avons eu à effectuer parmi ces stimuli actés, et ne nous a pas permis de tester statistiquement des hypothèses relatives aux paramètres acoustiques. Il aurait bien entendu été préférable de demander aux acteurs de produire un nombre élevé de répétitions, en faisant varier l'activation des expressions produites. Cependant, étant donné le caractère primordial dans l'optique d'une élicitation réussie de la session actée survenant immédiatement après l'expérience d'induction, et la durée totale de la tâche déjà élevée pour chacun des acteurs, il se serait avéré délicat de solliciter plus longuement les acteurs ayant accepté de participer à l'étude.

11.2. Limites techniques à l'analyse de l'intensité et de la fréquence fondamentale

Bien que nous nous soyons efforcés de mettre en place un protocole de recueil de données de la meilleure qualité possible, celui-ci peut souffrir de certaines limitations dues au choix du matériel utilisé. En particulier, l'utilisation d'un microphone fixe, bien adapté à l'enregistrement de corpus de parole lue par des locuteurs entraînés, est plus problématique dans le cas de l'enregistrement d'un corpus de relativement longue durée au cours duquel les états émotionnels des locuteurs varient, comme nous avons pu le constater d'après l'enregistrement vidéo. En effet, les locuteurs ont fréquemment effectué des mouvements du buste avec pour effet de faire varier la distance de la bouche au microphone, et par la même de modifier la mesure de l'intensité acoustique réalisée.

Nous avons en partie contourné cette difficulté en centrant une grande partie de notre analyse sur les productions d'un locuteur qui, d'après l'enregistrement vidéo, faisait relativement peu varier la distance de sa bouche au microphone par de tels mouvements. De plus, des indices convergents nous permettent de supposer que l'intensité acoustique n'est pas un élément prépondérant dans l'expression prosodique des affects. Cependant, il aurait été préférable afin de garder constante cette distance tout en laissant les locuteurs relativement libres de leurs mouvements de recourir à un micro-casque de qualité professionnelle similaire à ceux utilisés pour le chant sur les plateaux de télévision. Après avoir pris conscience lors des pré-tests effectués avec les deux premiers sujets féminins que ce problème pouvait se poser, nous avons commandé un micro-casque de ce type. Cependant, les délais de livraison de ce matériel ne nous ont pas permis de l'utiliser pour la campagne d'enregistrement de ce corpus.

Le choix d'étudier, pour nos analyses acoustiques, un nombre restreint de stimuli, en contrôlant méticuleusement les mesures réalisées (notamment en vérifiant manuellement la détection des pulsations réalisée de façon automatique), ainsi que le fait de ne pas avoir mesuré le signal électroglottographique pour l'ensemble des sujets ne semblent pas avoir de conséquence importante sur nos mesures de fréquence fondamentale. Dans l'optique d'une analyse réalisée à plus large échelle, une telle mesure serait souhaitable afin d'obtenir une mesure de référence fiable à la fois de la fréquence fondamentale et du quotient ouvert. Cependant, si un protocole incluant une partie actée tel que celui que nous avons élaboré est envisagé, la possibilité de réaliser cette mesure électroglottographique sans perturber le jeu d'acteur, et en particulier en le laissant libre de ses mouvements afin de faciliter l'élicitation reste à étudier. En effet, bien que cette instrumentation soit non invasive, il n'existe pas à notre connaissance d'électroglottographe permettant de transmettre le signal enregistré selon une technique similaire à celle utilisée par les microphones HF, ni d'électroglottographe portable qui permette de stocker provisoirement le signal sans nécessiter d'être directement relié à la station d'enregistrement. Cette question de la compatibilité avec un jeu d'acteur dans lequel les mouvements corporels ne sont pas contraints se pose d'ailleurs de la même façon pour les capteurs biophysiologiques.

12. Conclusion

La première étape de notre travail a été de développer une méthode et des outils pour produire des corpus combinant un degré élevé de contrôle phonétique et linguistique tout en guidant l'évolution de l'état émotionnel du locuteur, sans pour autant éliminer la spontanéité des processus émotionnels mis en œuvre pendant la production de ses énoncés. Afin de pouvoir varier les protocoles, nous avons développé une plate-forme logicielle dédiée, E-Wiz, ainsi que deux scénarios d'induction, Top Logic et Sound Teacher. Cette dernière application nous a permis de constituer un corpus d'expressions émotionnelles authentiques positives et négatives induites chez 17 sujets, dans un contexte dialogique minimal et non variant (gel des variantes attitudinales des sujets), et sur des énoncés phonétiques phonologiquement distribués pour vérifier la pertinence prosodique des analyses acoustiques ultérieures, notamment spectrales. Parmi ces 17 sujets, 7 acteurs avec une expérience de jeu d'acteur de type ligue d'improvisation ont rejoué les affects ressentis au cours de l'expérience d'induction. Un protocole original d'auto-annotation des états émotionnels exprimés préalable à leur validation perceptive par un panel d'auditeurs naïfs a été mis en place. Si ce protocole ne peut prétendre remplacer les protocoles plus classiques d'annotation fondés sur l'accord inter-annotateur sans intervention directe du locuteur, il peut en revanche permettre de s'affranchir de certains biais inhérent à l'annotation dite experte d'expressions émotionnelle.

L'analyse de contours de fréquence fondamentale de stimuli monosyllabiques actés et spontanés produits par l'un des locuteurs suggère que les formes de ces contours, au-delà des valeurs globales de paramétrisation de l'énoncé, seraient porteuses d'informations affectives, une forme de contour sans aucune valeur émotionnelle pouvant être isolée pour les énoncés spontanés et actés.

Une évaluation perceptive en conditions audio et audiovisuelle des productions actées d'un locuteur a montré que les affects exprimés étaient reconnus dans ces deux conditions, sans que ceux-ci ne soient mieux reconnus sur les énoncés de longueur plus importante que sur les énoncés monosyllabiques. Cette évaluation a également montré une variabilité inter-sujet pour la reconnaissance des expressions de dégoût en condition audio seule. Deux expériences perceptives de projection dimensionnelle de stimuli porteurs d'émotions sur des stimuli neutres, fondées sur l'algorithme classique de manipulation des dimensions de fréquence fondamentale et de durée TD-PSOLA puis sur un modèle d'analyse/resynthèse ARX-LF (Orange-labs) permettant également de modéliser et manipuler la qualité de voix, ont été menées. L'évaluation perceptive des énoncés générés dans ces deux expériences a montré une prépondérance des contours de

F0 pour la perception des expressions positives, et de la qualité de voix et de la durée pour les expressions négatives, sans qu'une dimension ne soit spécifiquement liée à une classe particulière d'émotions. Dans l'optique d'applications en synthèse de la parole expressive, l'extension de ces résultats à un panel plus large d'émotions et d'affects sociaux permettrait de guider les transformations prosodiques nécessaires ou en pointant les dimensions acoustiques à prendre en compte prioritairement pour la sélection d'unités dans la synthèse par corpus. Bien que les descriptions de la parole émotionnelle proposées par la psychologie expérimentale et fondées sur des mesures prosodiques globales puissent s'avérer suffisantes pour identifier certaines expressions, les résultats obtenus dans cette série d'analyses questionnent la validité d'une approche de modélisation ne prenant pas également en considération la dynamique des contours pour la génération de parole expressive.

La typicalité, la fréquence d'instanciation et la similarité aux idéaux d'expressions actées et spontanées relevant des classes de la joie, de la colère et de la peur produites par 6 locuteurs (3 hommes et 3 femmes) ont été évaluées perceptivement comme descripteurs de la catégorisation cognitive des expressions vocales d'émotions, confirmant le statut de catégorie liée au but au sens de Barsalou auquel avait conclu une précédente étude portant sur des expressions faciales prototypiques, avec toutefois une tendance moindre pour les expressions spontanées. L'intensité perçue de l'émotion a de plus été jugée supérieure pour les émotions actées que pour les émotions spontanées, en accord avec les quelques études précédentes s'étant attachées à la description comparative des expressions émotionnelles actées et spontanées.

Une tâche de discrimination audio, visuelle et audiovisuelle de paires de stimuli actés vs. spontanés a montré que des sujets naïfs étaient capables de discriminer les énoncés actés vs. spontanés sans différence notable entre classes d'émotion mais avec une importante variabilité interindividuelle, et sans que les différences entre intensités d'émotion perçues n'expliquent entièrement les scores de discrimination.

Ce dernier résultat souligne l'importance de prendre en compte des expressions spontanées pour l'analyse et la modélisation des expressions émotionnelles. En effet, l'ampleur de la variabilité inter-juge mesurée nous indique que, quand bien même certains acteurs seraient en mesure de simuler des expressions en tous points similaires à celles qu'ils produiraient spontanément, l'incapacité d'un panel de juges naïfs à évaluer ces expressions comme simulées ne préjugerait pas de la possibilité que d'autres juges parviennent à les discriminer d'expressions spontanées. Les performances limitées des systèmes automatiques d'identification ou catégorisation des expressions vocales d'émotions s'appuyant sur des modèles d'expressions actées pour reconnaître des

expressions spontanées ont déjà conduit au cours des dernières années à une forte orientation vers l'utilisation de corpus spontanés de la recherche dans ce domaine. Toutefois, ces résultats ont également des implications directes en synthèse de parole expressive, à la fois dans ses applications en recherche comme outil de simulation de la cognition des expressions et pour des applications industrielles, tout particulièrement pour la construction d'agents conversationnels animés réalistes. En effet, étant donnée cette capacité de l'humain à distinguer la parole émotionnelle actée de la parole émotionnelle spontanée et la variabilité interindividuelle observée, il est vraisemblable que la parole expressive produite par un agent conversationnel anthropomorphique reproduisant des expressions actées soit perçue comme actée par les utilisateurs de cet agent, sans que la perception de l'expressivité de l'agent comme réaliste ou non puisse être aisément prédite pour l'ensemble de ses utilisateurs potentiels à partir des réactions d'un sous-ensemble d'utilisateurs jugé représentatif. Dans l'optique du développement d'agents conversationnels animés crédibles, qui reproduisent les compétences communicationnelles de l'humain, il est donc nécessaire de modéliser l'expressivité de ces agents à partir de données spontanées afin que l'ensemble des utilisateurs puissent les percevoir comme réalistes quelles que soient leurs compétences d'identification de la simulation.

13. Perspectives

13.1. Une perspective sur l'expression vocale du dégoût

Dans l'expérience de dévoilement progressif que nous avons présentée en section 9.3. , nous avons observé un taux d'identification à la dernière fenêtre de dévoilement des expressions du dégoût sur les énoncés [ʒon] et [ʁuʒ] plus faible que celui observé précédemment sur l'énoncé [sabl], bien que les confusions observées restent proches. Une explication possible est que, en dépit des morphologies prosodiques très similaires des expressions de dégoût produites par le locuteur M2 sur les énoncés [ʒon], [ʁuʒ] et [vɛʁ], une telle expression serait moins contrainte et potentiellement plus écologique sur des voyelles ouvertes et antérieures. En effet, plusieurs auditeurs ont déclaré à la suite du test avoir éprouvé plus de difficultés à considérer comme naturelles les expressions de dégoût sur les énoncés [ʒon] et [ʁuʒ] que sur l'énoncé [vɛʁ].

A la suite de Darwin (1872), le dégoût a largement été considéré comme une réponse évolutive à la présence d'aliments impropres à la consommation, dérivée du geste de vomissement et dont l'expression faciale permettrait d'avertir les congénères situés à proximité. Curtis *et al.* (2004) ont ainsi montré que la présentation de stimuli visuels évoquant la possible transmission de maladies, tels que fluides corporels et parasites internes, induisait un dégoût plus important que la présentation de stimuli semblables mais n'évoquant pas de risque de contagion.

Schröder (2003, p. 103), revient sur la définition des bouffées d'affect (*affect bursts*) donnée par Scherer (1994), et en propose une simplification préalable à son étude expérimentale. Il considère ainsi les bouffées d'affects comme

« des expressions non-verbales courtes et émotionnelles, comprenant à la fois des sons non-verbaux clairs (comme le rire) et des interjections ayant une structure phonémique (par exemple « Wow ! »), mais excluant des interjections « verbales » pouvant apparaître comme des actes de parole différents (comme « Ciel », « Non ! », etc.). »[32]

Dans cette étude expérimentale, Schröder a demandé à six locuteurs germanophones dont quatre acteurs amateurs, de produire de telles expressions pour un ensemble de dix

[32] Traduction personnelle de la citation originale *« Affect bursts are short, emotional non-speech expressions, comprising both clear non-speech sounds (e.g. laughter) and interjections with a phonemic structure (e.g. ''Wow!''), but excluding ''verbal'' interjections that can occur as a different part of speech (like ''Heaven!'', ''No!'', etc.). »*

émotions. Il note que la bouffée d'affect la plus fréquemment réalisée (six occurrences sur un total de huit) par les acteurs germanophones exprimant le dégoût est [buæː] accompagnée d'une voix craquée.

Les interjections non-verbales, ou emblèmes affectifs, se situent à l'extrémité du continuum de classification des bouffées d'affect proposé par Scherer (1994) en fonction du degré de conventionnalisation par les langues et les cultures, l'autre extrémité étant occupée par les bouffées brutes d'affect comme le rire. Le rire et autre « grognements » (notion de *grunts* introduite par Campbell *et al.* (2005)) en rapport avec l'expression émotionnelle n'ont été étudiés en tant que tels que de façon très marginale jusqu'à une époque récente (voir notamment les actes du colloque *Laughter* organisé en août 2007 par Campbell). En revanche, les interjections l'ont été plus largement, notamment dans le domaine de la pragmatique.

Wierzbicka (1992) passe ainsi en revue les interjections liées au dégoût en anglais, allemand, polonais, russe et dans les langues scandinaves, et propose une taxonomie d'interjections de dégoût selon le type de répugnance et le geste de rejet associé. Elle considère ainsi en particulier que l'interjection anglaise *yuk* ([jʌk]), mise en avant par Wharton (2003) comme la plus typique de l'expression de dégoût en anglais, correspondrait à une réaction de « répugnance pharyngale », que nous pouvons lier au geste de vomissement. Selon cette taxonomie, les interjections constituées d'une consonne fricative bilabiale ou labiodentale sourde ou d'une plosive bilabiale suivie d'une voyelle fermée comme par exemple l'interjection anglaise *phew* ([fjuː]) correspondraient à une « répugnance buccale » et imiteraient un crachat destiné à se débarrasser d'un fluide dégoûtant tout en en laissant pénétrer la quantité la plus faible possible en minimisant l'aire aux lèvres.

Aucune étude à notre connaissance ne s'est attachée spécifiquement à dresser un inventaire des interjections du français liées au dégoût. Cependant l'édition 2006 du dictionnaire de langue française Le Petit Robert atteste « berk » ([bɛʁk]) et sa variante « beurk » ([bœʁk]) comme interjections liées à l'expression du dégoût, ainsi que « pouah » ([pwa]) qui impliquerait également une dimension de mépris, tandis que le Trésor de la Langue Française informatisé définit « pouah » ([pwa] ou [pwɑ]) comme une interjection servant à exprimer le dégoût physique, intellectuel ou moral du locuteur[33]. « Berk » et « beurk » semblent être les interjections liées le plus directement au dégoût. Pochon *et al.* (2006) ont ainsi recouru parmi d'autres stimuli visuels et audio, afin de comparer les

[33] http://www.cnrtl.fr/definition/pouah

performances en reconnaissance d'émotions d'enfants atteints de trisomie 21 à celles d'enfants de même âge développemental, à « *des « berk…» de dégoût* ». Les auteurs ne précisent néanmoins pas dans quelles conditions ces stimuli audio ont été enregistrés ni ne donnent d'indication sur leurs propriétés acoustiques. Par ailleurs Rose & dos Santos (2006), dans une étude longitudinale de l'harmonie consonantique et de la métathèse dans l'acquisition du français langue première, mentionnent la réalisation de « berk » comme [bɛk] par un enfant alors âgé de 1 an et 11 mois.

Comme nous l'avons mentionné ci-dessus, nous ne disposons pas de données expérimentales relatives à la production de bouffées d'affects liées au dégoût en français. Néanmoins il nous semble fréquent que les bouffées d'affect liées à une expression spontanée de dégoût en français soient réalisées comme la consonne bilabiale [b] suivie d'une voyelle antérieure semi-ouverte ([ɛ] ou [œ]) voire ouverte ([a]) allongée et réalisée avec une fréquence fondamentale basse, soit une forme phonétique proche de la bouffée d'affect de dégoût majoritairement recueillie par Schröder (2003) en allemand.

L'expression faciale qu'Ekman & Friesen (1975) considèrent comme prototypique du dégoût est caractérisée par un plissement du nez mais également par la rétractation de la lèvre supérieure et l'élévation du menton. Rozin *et al.* (1994) ont cependant constaté qu'une importante variabilité pouvait être observée entre expressions faciales de dégoût. Ils ont élaboré une tâche d'association dans laquelle les sujets devaient attribuer aux différentes configurations de la face présentées sur des photographies le stimulus qu'ils considéraient comme étant à l'origine du dégoût exprimé par la personne représentée. Ils en ont conclu que l'ouverture de la bouche codée dans le système FACS (Ekman & Friesen, 1978) par l'unité d'action n°26 (Figure 56), était principalement associée à l'ingestion d'aliment brûlants, irritants, avariés ou prémâchés par une autre personne. Il semble donc naturel que cette ouverture buccale favorise la production de voyelles mi-ouvertes telles que [ɛ], [œ], [æ] ou [ʌ], et donc qu'une expression de dégoût correspondant au rejet d'un aliment soit associée prioritairement à une voyelle mi-ouverte. Rebreyend (2006) a étudié l'influence de la forme du conduit vocal associée à des expressions de dégoût simulées afin de les comparer à l'effet du geste de sourire qui semble antagoniste. Il a ainsi enregistré deux locuteurs produisant l'ensemble des voyelles du français tout en simulant des expressions faciales de dégoût plus ou moins prononcées. Les confusions importantes observées entre voyelles mi-ouvertes lors de l'évaluation perceptive de ces productions est compatible avec notre hypothèse, cependant l'effet plafond pour la reconnaissance des autres voyelles quelque soit l'intensité du geste facial de dégoût ne nous permet pas d'en tirer de conclusions.

Figure 56 : Unité d'action n°26 du système FACS, dans laquelle les lèvres sont relâchées et séparées et la mandibule abaissée (d'après Ekman & Friesen (1978)).

L'interprétation de l'expression de dégoût comme dérivée du geste de vomissement pourrait conduire à supposer que la qualité de voix liée à son expression dans la parole serait celle correspondant à l'ouverture minimale de la glotte permettant la phonation. En effet, le geste de vomissement est décrit en physiologie comme l'expulsion à travers la glotte fermée avec une forte pression du contenu de l'estomac (Borison & Wang, 1953).

La qualité de voix qui pourrait ainsi être associée prioritairement à une telle configuration de la glotte serait la voix grognée (*growl voice*) décrite en voix chantée par Sakakibara *et al.* (2004), qui est caractérisée par une vibration des structures supérieures du larynx plis aryépiglottiques et dont l'usage le plus célèbre en voix chantée est celui qu'en fait Louis Armstrong. Il est également vraisemblable que les variations de fréquence fondamentale soient limitées, le conduit vocal étant déjà « occupé » à l'expulsion de l'aliment impropre à la consommation. Ces hypothèses restent cependant à vérifier d'un point de vue expérimental, ce qui, bien qu'un tel protocole semble en première approche très complexe à mettre en œuvre, pourrait être évalué au moyen d'une expérience d'induction d'expressions spontanées du dégoût combinée à des mesures directes du comportement du larynx.

13.2. Un plus large panel d'émotions et d'affects sociaux

Nos travaux apportent des éléments de réponse à certaines des nombreuses questions de recherche qui se posent dans l'étude des interactions entre parole et émotion et de poser des jalons méthodologiques. Outre celles qui s'inscrivent dans le prolongement direct des expériences que nous avons réalisé et dont certaines sont développées à la suite de ces expérience, et celle sur les expressions du dégoût développée dans la section précédente, les perspectives qui s'ouvrent à la suite des travaux que nous avons présentés dans cette thèse sont cependant multiples. Parmi les principales, nous pouvons ainsi mentionner :

- La problématique des affects sociaux, et en particulier la question du niveau de contrôle cognitif mis en œuvre lors de la production spontanée d'attitudes prosodiques comparativement à celui mis en œuvre lors de la production actée de ces attitudes. En effet, et bien qu'il soit possible de s'interroger sur le statut cognitif de

certaines des productions spontanées des locuteurs que nous avons capturés dans notre expérience de Magicien d'Oz, nous nous sommes efforcé, en plaçant les locuteurs face à une machine afin d'éliminer toute interaction sociale, de favoriser la production d'expressions émotionnelles plutôt que celle d'affect sociaux. Il serait donc intéressant de mettre en place un protocole, dont les modalités restent à définir, permettant le recueil d'expressions spontanées et actées de telles attitudes afin de pouvoir les comparer systématiquement.

- Le spectre des émotions capturées et étudiées, en s'attachant à l'étude de situations plus proches de la vie courante tout en conservant un degré élevé de contrôle expérimental. En effet la situation que nous avons proposée dans le cadre du scénario Sound Teacher constitue un cas spécifique d'interaction personne-machine dans lequel il nous a été possible d'induire un certain nombre d'affects, mais ne peut rendre compte de l'expression émotionnelle dans son ensemble, ce qui nécessite la collecte d'autres corpus. Le cadre général que nous avons défini de recueil par un protocole de Magicien d'Oz simulant une application fondée sur la reconnaissance vocale nous semble cependant suffisamment souple pour permettre d'imaginer et mettre en œuvre de tels scénarios, qui restent cependant eux aussi à définir.
- L'aspect développemental de la production et du décodage d'expressions d'affects chez l'enfant. Tant le recueil de données que l'analyse acoustique des signaux recueillis soient problématiques lorsqu'il s'agit de les étudier chez de jeunes enfants. Cependant, à condition d'imaginer des scénarios d'induction capables de captiver l'attention d'enfants, le cadre méthodologique que nous avons proposé devrait permettre de proposer de tels protocoles de recueil de données.

13.3. Multilinguisme et inter-culturalité

La question principale posée par la problématique de l'inter-culturalité, largement étudiée en ce qui concerne les expressions faciales (voir notamment Ekman, 1994), l'a été beaucoup moins pour les expressions vocales. En effet, comme nous l'avons évoqué en introduction de cette thèse, d'importants problèmes méthodologiques se posent pour l'étude d'éventuels universaux dans les expressions vocales des émotions. Le tour d'horizon le plus complet de ces problématiques est donné par Scherer *et al.* (2001). Ils proposent également l'étude la plus large et la plus complète à ce jour du décodage de mêmes expressions vocales actées d'émotions par des auditeurs de différentes langues et cultures, suggérant que des universaux existeraient bien pour les expressions vocales d'émotions, la précision du décodage étant cependant moins élevée pour les langues plus éloignées de la langue maternelle des locuteurs.

En collaboration avec le laboratoire SpeechLab de l'université de Budapest (Fék, Audibert, Szabo, Rilllard, Németh & Aubergé, 2008), nous avons contribué à l'enregistrement d'un corpus de parole expressive spontanée reprenant les principes de Sound Teacher en l'adaptant à la langue hongroise, qui a permis de capturer les productions de 8 locuteurs. Bien que le hongrois soit linguistiquement très éloigné du français, notamment au niveau en ce qui concerne sa prosodie (voir par exemple Fonagy (2003) pour une description comparative de la réalisation de certaines fonctions prosodiques), des valeurs expressives proches de celles recueillies dans le corpus français ont ainsi pu être collectées. 149 stimuli parmi les 962 ainsi recueillis ont été évalués perceptivement selon le même protocole que dans le pré-test présenté en section 10.1. Nous avons ainsi posé les fondements d'une évaluation perceptive d'expressions émotionnelles spontanée hongroise par des auditeurs francophones et vice-versa, que nous prévoyons de mener à bien ultérieurement.

En revanche, la tentative de réplication de ce protocole auprès de sujets japonais, menée dans le cadre du projet *Expressive Speech Processing*, s'est soldée par un échec, l'implication des sujets dans la tâche ne pouvant être assurée. Cet échec souligne ainsi la question de l'influence relative des faits de langue et de culture, souvent considérés conjointement, dans les réalisations expressives et les facteurs à l'origine de ces expressions.

Bibliographie

Abrilian S., Devillers L. & Martin J.-C. (2006). Annotation of Emotions in Real-Life Video Interviews: Variability between Coders. *5th International Conference on Language Resources and Evaluation (LREC 2006)*, Gênes, Italie, pp. 2004-2009.

Abrilian, S., Devillers, L. Buisine, S. & Martin, J-C. (2005). EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. *11th International Conference on Human-Computer Interaction (HCI International 2005)*, Las Vegas, USA.

D'Agostino R.B. (1972). A second look at analysis of variance on dichotomous data. *Journal of educational measurement*, 8(4), pp. 327-333.

Airas M. (2008). *Methods and studies of laryngeal voice quality analysis in speech production*. Thèse de doctorat, University of Technology, Helsinki.

Akif Kiliç M., Öüt F., Dursun G., Okur E., Yildirim I. & Midilli R. (2004). The effects of vowels on voice perturbation measures. *Journal of Voice*, 18 (3), pp. 318 - 324

Alku P. & Vilkman E. (1996). Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering. *Speech Communication*, 18 (2), pp. 131-138.

Alku P., Bäckström T. & Vilkman E. (2002). Normalized amplitude quotient for parametrization of the glottal flow. *Journal of the Acoustic Society of America*, 112 (2), pp. 701-710.

Alku P., Strik H. & Vilkman E. (1997). Parabolic spectral parameter – A new method for quantification of the glottal flow. *Speech Communication*, 22, pp. 67–79.

d'Alessandro C. (2006). Voice source parameters and and prosodic analysis. In *Method in Empirical Prosody Research*, Sudhoff S., Leternovà D., Meyer R., Pappert S., Augurzky P., Mleinek I., Richter N., Schliesser J., de GruyterW. (Eds), Berlin, New York, pp 63-87.

d'Alessandro C. & Doval B. (2003). Voice quality modification for emotional speech synthesis. *8th European Conference on Speech Communication and Technology* (*EUROSPEECH 2003)*, Geneva, Switzerland, pp. 1653-1656.

Amir N. & Ron S. (1998). Toward an Automatic Classification of Emotions in Speech. 5th *International conference on spoken language processing (ICSLP 1998)*, Sydney, Australie, pp. 29-33.

Ang J., Dhillon R., Krupski A., Shriberg E. & Stolcke A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, USA, pp. 2037-2040.

Arnal A. (2003). *Schéma de la table de mixage CREST/ESP*, document interne ICP.

Arnold M. B. (1960). *Emotion and personality*. NY, Columbia University Press.

Arroabarren I., & Carlosena A. (2003). Glottal spectrum based inverse filtering. *8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, Genève, Suisse, pp. 57–60.

Aubergé V. (1991). *La synthèse de la parole : des règles au lexique*. Thèse de doctorat, Université Pierre Mendès-France, Grenoble.

Aubergé V. (2002). A gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP. *1st International Conference Speech Prosody*, Aix-en-Provence, France, pp. 151-155.

Aubergé V. (2003). Expressions, attitudes et expressivité : une architecture cognitive distribuée pour les voies parlées des émotions. *Interfaces Prosodiques*, Nantes, France.

Aubergé V. & Cathiard M. (2003). Can we hear the prosody of smile ? *Emotional Speech*, *Speech Communication Review*, 40, pp. 87-97.

Aubergé V., Grepillat T. & Rilliard A. (1997). Can we perceive attitudes before the end of sentences? the gating paradigm for prosodic contours. 5th *European Conference on Speech Communication and Technology* (*EUROSPEECH 1997)*, Rhodes, Greece, pp. 871-874.

Aubergé V. & Rilliard A. (2006). More than pointing with the prosodic focus: the valence-intensity-domain (VID) model. *3rd International Conference on Speech Prosody*, Dresden, Allemagne (actes CD-ROM).

Averill J. R. (1980). A Constructivist View of Emotion. *Emotion : Theory, Research and Experience : Vol. I. Theories of Emotion* (R. Plutchik & H. Kellerman, Eds.). New York: Academic Press, pp. 305-339. (ré-édité dans : Halberstadt, A. G., & Ellyson, S. L., ed. (1990)).

Bailly G. (2001). A parametric harmonic + noise model. In E. Keller, G. Bailly, A. I. C. Monaghan, J. Terken and M. Huckvale (Eds.) *Improvements in Speech Synthesis*. Chichester, England, J. Wiley & Sons, Ltd, pp. 22-38.

Banse R. & Scherer K. R. (1996). Acoustic Profiles in Vocal Emotion Expression. *Journal of Personality and Social Psychology*, 70 (3), pp. 614-636.

Bänziger T. (2004). *Communication vocale des émotions, Perception de l'expression vocale et attributions émotionnelles*, Thèse de doctorat, spécialité Psychologie, Université de Genève.

Bänziger T., Morel M. & Scherer K. R. (2003). Is there an emotion signature in intonational patterns? And can it be used in synthesis? *8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, Genève, Suisse, pp. 1641-1644.

Bänziger T. & Scherer K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, 46, pp. 252-267.

Bänziger T. & Scherer K. R. (2007). Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus. *2nd International Conference on Affective Computing and Intelligent Interaction (ACII2007)*, Lisbonne, Portugal, pp. 476-487.

Barkhuysen P., Krahmer E. & Swertz M. (2007). Cross-modal perception of emotional speech. *16th International Congress of Phonetic Sciences*, Saarbrücken, Allemagne, pp. 2133-2136.

Barney A., De Stefano A. & Henrich N. (2007) The effect of glottal opening on the acoustic response of the vocal tract. *Acta Acustica united with Acustica*, 93 (6), pp. 1046-1056.

Barsalou L.W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 11, pp. 629-653.

Barsalou L.W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development. Ecological and intellectual factors in categorization*. Cambridge, Cambridge University Press, pp. 101-140.

Batliner A., Fischer K., Huber R., Spilker J. & Nöth E. (2003a). How to find trouble in communication. *Speech Communication*, 40, pp. 117-143.

Batliner A., Hacker C., Steidl S., Nöth E. & Haas J. (2003b). User states, user strategies, and system performance: how to match the one with the other. *ISCA Workshop on Error Handling in Spoken Dialogue Systems*, Château d'Oex, Vaud, Suisse, pp. 5-10.

Batliner A., Hacker C., Steidl S.,Nöth E.,.D'Arcy S., Russel M. & Wong M. (2004). 'You stupid tin box' – children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. *4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbonne, Portugal, vol.1, pp.171–174.

Beller G., Obin N. & Rodet X. (2008). Articulation degree as a prosodic dimension of expressive speech. *4th International Conference on Speech Prosody*, Campinas, Brésil, pp. 681-684.

Benjamini Y. & Hochberg Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), pp. 289-300.

Benjamini Y. & Hochberg Y. (1995). Controlling the False Disco-very Rate : a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B*, 57(1), pp. 289-300.

Benzecri J.P. (1973). *L'Analyse des Données*, 2 vols, Paris, Dunod.

Beskow J., Cerrato L., Granström B., House D., Nordstrand M. & Svanfeldt G. (2004). The Swedish PF-Star multimodal corpora. *International Workshop on Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, Lisbonne, Portugal, pp. 34-37.

Blomgren M. & Chen Y. (1998) Acoustic, aerodynamic, physiologic and perceptual properties of modal and vocal fry registers. *Accoustic Soc Am* 103(5), pp. 2649-2658.

Blomgren M., Chen Y., Ng M.L. & Gilbert H.R. (1998). Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. *Journal of the Acoustical Society of America*, 103 (5, Pt. 1), pp. 2649-2685.

Boë L.J., Gabioud B. & Perrier P. (1995) Speech MAPS Interactive Plant "SMIP". *XIIIth Int. Congr. of Phonetic Sciences*, Stockholm, Suède, 2, pp. 426-429.

Boë L.J., Gabioud B. & Perrier P. (1995) The SMIP: An interactive articulatory-acoustic software for speech production studies. *Bulletin de la Communication Parlée*, 3, pp: 137-154.

Boersma P. & Weenink D. (1992–2008). Praat, doing phonetics by computer. Institute of Phonetic Sciences, University of Amsterdam, Pays-Bas. http://www.fon.hum.uva.nl/praat

Boersma P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences* (17), pp. 97-110.

Bolinger D. L. (1949). Intonation and analysis. *Word* (5), pp. 248-254.

Bonastre J.-F., Scheffer N., Matrouf D., Fredouille C., Larcher A., Preti A., Pouchoulin G., Evans N., Fauve B. & Mason J. (2008). ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition. *Odyssey - The Speaker and Language Recognition Workshop*, Spier Wine Estate, Stellenbosch, Afrique du Sud.

Borison H. L. & Wang S. C. (1953). Physiology and pharmacology of vomiting. *Pharmacological Reviews*, 5, pp. 192-230.

Borkenau, P. (1990). Traits as ideal-based and goal-derived social categories. *Journal of Personality and Social Psychology*, 58, pp. 381-396.

Bottineau D. (2007). Language and enaction, Stewart, J., Gapenne, O. & Di Paolo, E. (eds),*Enaction: towards a new paradigm for cognitive science*, MIT.

Bougé K., Aubergé V., Vilain C. (2006). Cohérence des signaux de communication expressive. Esquisse d'une méthode originale. *2ème Workshop Francophone sur les Agents Conversationnels Animés*, pp. 207-211.

Bryan Loyall A. (1997). *Believable Agents: Building Interactive Personalitie*s. Carnegie Mellon University.

Brown R. & Kulik J. (1977). Flashbulbs memories, *Cognition*, 5, pp. 73-99.

Buisine S., Abrilian S., Niewiadomski R., Martin J.-C., Devillers L. & Pelachaud C (2006). Perception of Blended Emotions: from Video Corpus to Expressive Agent. *6th International Conference on Intelligent Virtual Agents (IVA'2006)*, Marina del Rey, USA, pp. 93-106.

Busso C. & Narayanan S.S. (2008). Recording audio-visual emotional databases from actors: a closer look. in Second International Workshop on Emotion: Corpora for Research on Emotion and Affect,

6th International conference on Language Resources and Evaluation (LREC 2008), Marrakech, Maroc. pp. 17-22.

Bulut M., Narayanan S.S. & Syrdal A. K. (2002). Expressive speech synthesis using a concatenative synthesiser. *7th International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, Colorado, USA.

Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W. & Weiss B. (2005). A Database of German Emotional Speech. *9th European Conference on Speech Communication and Technology (INTERSPEECH 2005)*, Libonne, Portugal, pp. 1517-1520.

Busso C. & Narayanan S. S. (2008). Recording audio-visual emotional databases from actors: a closer look. *5th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Maroc.

Busso C., Deng Z., Yildirim S., Bulut M., Lee C.M., Kazemzadeh A., Lee S., Neumann U. & Narayanan S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. *6th International Conference on Multimodal Interfaces (ICMI 2004)*, State College (PA), Etats-Unis, pp. 205–211.

Cahn J. (1990). The Generation of Affect in Synthesised Speech. *Journal of the I/O Voice American Society* 8, pp. 1-19.

Campbell N. (2000). Databases of Emotional Speech. *ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, pp. 34-38.

Campbell N. & Mokhtari (2003a). Voice Quality: the 4th Prosodic Dimension. *15th International Congress of Phonetic Sciences*, Barcelone, Espagne, pp. 2417-2420.

Campbell N. (2003b). The JST/CREST ESP Project – a Mid-Term Progress Report.

Campbell N (2004). Speech & Expression; the value of a longitudinal corpus, *4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbonne, Portugal.

Campbell N. (2005). Getting to the Heart of the Matter: Speech as the Expression of Affect; Rather than Just Text or Language? *Journal of Language Resources and Evaluation*, 39, pp. 109–118.

Campbell W.N., Black A. (1997). Prosody and the selection of source units for concatenative synthesis. In: van Santen J., Sproat R., Olive J., Hirschberg J. (Eds.), *Progress in Speech Synthesis. Springer Verlag*, NY, pp. 279-292.

Campbell N., Kashioka H. & Ohara R. (2005). No laughing matter. *9th European Conference on Speech Communication and Technology* (*INTERSPEECH 2005)*, pp. 465-468.

Campbell N., Devillers L., Douglas-Cowie E., Aubergé V., Batliner V. & Tao J. (2006). Resources for the Processing of Affect in Interactions (Panel). *5th international conference on Language Resources and Evaluation (LREC 2006)*, Gênes, Italie, pp. 25-28.

Campbell N., Sadanobu T., Imura M., Iwahashi N., Noriko S. & Douxchamps D. (2006). A Multimedia Database of Meetings and Informal Interactions for Tracking Participant Involvement and Discourse Flow. *5th international conference on Language Resources and Evaluation (LREC 2006)*, Gênes, Italie, pp. 391-394.

Cannon (1927). The James-Lange theory of emotions: a critical examination and an alternative. *American Journal of Psychology*, 39, pp. 106-124.

Carletta J. (2006) Announcing the AMI Meeting Corpus. *The ELRA Newsletter*, 11(1), pp. 3-5.

Carton F. (1999). L'épithèse vocalique en français contemporain : étude phonétique. *Faits de langues*, 13, pp. 35-45.

Cassell J. (2000). More than Just Another Pretty Face: Embodied Conversational Interface Agents. *Communications of the ACM*, 43(4), pp. 70-78.

Castellengo M. & Dubois, D. (2005). Timbre ou timbres ? Propriété du signal, de l'instrument, ou construction cognitive ?, *Actes du colloque interdisciplinaire de musicologie*. Montréal, Canda.

Chauny C. (1996). *Catégorisation et espace perceptif des sons de la parole : vers des cartes sensori-motrices*. Mémoire de DEA Sc. Cognitives, INPG.

Chen A., Gussenhoven C. & Rietveld T. (2004). Language specificity in the perception of paralinguistic intonational meaning. *Language and Speech*, 47, pp. 311-349.

Childers D. G. & Lee C. K. (1991). Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America*, 90(5), pp. 2394–2410.

Chung S. (2000). *L'expression et la Perception de l'Emotion dans la Parole Spontanée : Evidences du Coréen et de l'Anglais.* Thèse de doctorat, spécialité Sciences du Langage, Université de la Sorbonne Nouvelle (Paris III).

Clavel C. (2007). *Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales.* Thèse de doctorat, spécialité Signal et Images, Ecole Nationale Supérieure des Télécommunications.

Coadou M. (2006) Voice quality and variation: a pilot study of the Liverpool accent. *Speech Prosody*, Dresde, Allemagne, actes CD-ROM.

Cohen J. (1960), A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1), pp. 37-46.

Cohen J. (1977). S*tatistical power analysis for the behavioral sciences* (edition révisée). New York: Academic Press.

Cohen N. E. (1937). The Relativity of Absolute Judgments. *American Journal of Psychology*, 49, pp. 93-100

Cole R., Oshika B.T., Noel M., Lander T. & Fanty M. (1994). Labeler Agreement In Phonetic Labeling Of Continuous Speech. 3rd *International Conference on Spoken Language Processing (ICSLP 1994)*, Yokohama, Japon, pp. 2131-2134.

Colletta J. M. & Tcherkassof A. (2001). *Les émotions : Cognition, langage et développement*. Liège, Belgique : Mardaga.

Cornelius R. R. (2000). Theoretical Approaches to Emotion, *ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, pp. 3-10.

Cornelius R. R. (1996). *The science of emotion. Research and tradition in the psychology of emotion*. Upper Saddle River (NJ), USA : Prentice-Hall.

Cowie R. & Cornelius R. R. (2003). Describing the emotional states that are expressed in speech. *Spech Communication*, 40(1-2), pp. 5-32.

Cronbach L.J. (1951). Coefficient alpha and the internal structure of tests, *Psychometrika* 16, pp. 297-334.

Cummings K. E. & Clements M. A. (1995). Analysis of the glottal excitation of emotionally styled and stressed speech. *The Journal of the Acoustical Society of America*, 98(1), pp. 88-98.

Curtis C. E. Rao V. Y. & D'Esposito M. (2004). Maintenance of spatial and motor codes during oculomotor delayed response tasks. *The Journal of Neuroscience,* 24(16), pp. 3944-52.

d'Alessandro C. & Doval B. (2003). Voice quality modification for emotional speech synthesis, *8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, Genève, Suisse, pp. 1653-1656.

Damasio A. R. (1994). *Descartes' error. Emotion, reason, and the human brain*. A. Grosset/Putnam Books.

Damasio A. R. (2003). *Looking for Spinoza. Joy, Sorrow an the Feeling Brain*. Orlando (FL), USA : Harcourt.

Danes F. (1994). Involvement with language and in language. *Journal of Pragmatics*, 22, pp. 251-164.

Darwin C. (1872). *The expression of the emotion in Man and Animals.* New York, USA: Philosophical Library.

Dohen M., Lœvenbruck H., Cathiard M.-A. & Schwartz J.-L. (2004). Visual perception of contrastive focus in reiterant French speech. *Speech Communication*, 44, pp. 155-172.

de Cheveigné A. & Kawahra H. (2002). YIN, A fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4), pp. 1917-1930.

de Gelder B. & Vroomen J. (2000). The Perception of Emotions by Ear and by Eye. *Cognition and Emotion*, 14 (3), pp. 289-311

Delattre P. (1966). Les dix intonations de base du français, *The French Review*, 40 (3), pp. 1-14.

Dellaert F., Polzin T. & Wajbel A. (1996). Recognizing emotion in speech, *4th International Conference on Spoken Language Processing (ICSLP 1996)*, Philadelphia, pp. 1970-1973.

Descartes R. (1649). *Les passions de l'âme.* Paris : Gallimard, 1988.

Di Cristo A. (1975). *Soixante-dix Ans de Recherches en Prosodie.* Editions de l'Université de Provence, Aix en Provence.

Ding W., Kasuya H. & Adachi S. (1995). Simultaneous estimation of vocal tract and voice source parameters based on an ARX model. *IEICE Transactions on Information Systems*, E78-D(6), pp. 738-743.

Douglas-Cowie E., Cowie R. & Schröder M. (2000). A new emotion database: considerations, sources and scope, *ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, pp. 39-44.

Douglas-Cowie E., Devillers L., Martin J.-C., Cowie R., Savvidou S., Abrilian S. & Cox C. (2005). Multimodal databases of everyday emotion: facing up to complexity. *9th European Conference on Speech Communication and Technology (INTERSPEECH 2005)*, Lisbonne, Portugal, pp. 813-816.

Doval B. & d'Alessandro C. (1997). Spectral correlates of glottal waveform models: an analytic study. In *International Conference on Acoustics, Speech and SignalProcessing (ICASSP 97)*, pp. 446--452,

Doval B., d'Alessandro C. & Henrich N. (2006). The Spectrum of Glottal Flow Models. *Acta Acustica*, 92, pp. 1026-1046.

Doyon D., Iffenecker C. & Cabanis E.A (1998). *Nomenclature anatomique radiologique internationale*. Elsevier Masson.

Dutoit T. (1997). *An introduction to Text-To-Speech Synthesis*. Dordrecht : Kluwer Acamdic Publishers.

Dutoit T., Pagel V., Pierret N., Bataille F. & van der Vreken O. (1996). The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes, *4th International Conference on Spoken Language Processing (ICSLP 1996)*, Philadelphia, pp. 1393-1396.

Ekman P. (1999a). Facial Expressions. In T. Dalgleish & T. Power (Eds.) *The Handbook of Cognition and Emotion*, Sussex, R.U.: John Wiley & Sons, Ltd., pp. 301-320.

Ekman P. (1999b). Basic emotions. In T. Dalgleish and T. Power (Eds.) *The Handbook of Cognition and Emotion*, Sussex, R.U.: John Wiley & Sons, Ltd., pp. 45-60.

Ekman P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin,* 115*,* pp. 268-287.

Ekman P. (1989). L'Expression des Emotions. In *Les Emotions* (B. Rimé and K. R. Scherer, ed.). Neuchâtel ; Paris : Delachaux-Niestlé, pp.183-201.

Ekman P., Davidson R. J. & Freisen W. V. (1990). Duchenne's smile : emotional expression and brain physiology II. *Journal of Personality and Social Psychology* (58), pp. 342-353.

Ekman P. & Friesen W.V. (1975). *Unmasking the face : A guide to recognizing emotions from facial clues*. Englewood Cliffs, NJ: Prentice-Hall.

Ekman P. & Friesen W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement.* Palo Alto, California: Consulting Psychologists Press.

Elfenbein H. A., Mandal M. K., Ambady N., Harizuka S. & Kumar S. (2002). Cross-cultural patterns in emotional communication: Highlighting design and analytical techniques. *Emotion*, 2, pp. 75-84.

Enos F. & Hirschberg J. (2006). A Framework for Eliciting Emotional Speech: Capitalizing on the Actor's Process. *1st International Workshop on Corpora for Research on Emotion and Affect*, Gênes, Italie, pp. 6-10.

Fant G. (1960). *Acoustic Theory of Speech Production*. La Hague, Pays-Bas : Mouton.

Fant G. (1997). The voice source in connected speech. *Speech Communication*, 22, pp. 125-139.

Fant G., Liljencrants J. & Lin Q. (1985). A four parameter model of glottal flow. *Speech Transmission Laboratory Quarterly Progress Report* (4), pp. 1-13.

Feng G. (2003). *Détection de la fréquence fondamentale de la parole avec filtrage*. Mémoire de DEA en Signal, Image, Parole, Institut National Polytechnique de Grenoble.

Fernandez R. & Picard R. (2000). Modelling drivers' speech under stress. *ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, pp. 219-224.

Février F., Jamet E., Rouxel G., Dardier V., Breton G. (2006). Induction d'émotions pour la motion capture dans une situation de vidéo-conversation. *2ème Workshop sur les Agents Conversationnels Animés (WACA 2006)*, Toulouse, France, pp. 17-26.

Fonagy I. & Magdics K. (1963). Emotional patterns in intonation and music. Zeitschrift für Phonetik, 16, pp. 293–326.

Fonagy I. (1983). *La vive voix*. Paris, Payot.

Fonagy I. (2003). Des fonctions de l'intonation : essai de synthèse. *Flambeau*, 29, pp. 1-20.

Fourcin A. J. & Abberton E. (1971). First Applications of a New Laryngograph. *Medical and Biological Illustration*, 21, pp. 172-182.

Fourcin A. & Abberton E. (2008). Hearing and phonetic criteria in voice measurement: Clinical applications. *Logopedics Phoniatrics Vocology*, 33(1), pp. 35-48.

Fouard A. (2002). *Etablissement d'un corpus de parole émotionnelle spontanée basé sur le paradigme du Magicien d'Oz*. Mémoire de DEA Sciences du Langage, Université Stendhal – Grenoble III

Fraser N.M. & Gilbert G.N. (1991). Simulating speech systems. *Computer Speech and Language*.

Frick R.W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin*, 97, pp. 412–429.

Frijda N. H. (1987). Emotions, Cognitive structures and Action tendency. *Cognition and Emotion*, 1, pp. 115-143.

Fujisaki H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In Fujimura O. (Ed.) *Vocal Physiology: Voice Production, Mechanisms and Functions.* Raven, New York, pp. 347–355.

Garcia-Prieto P. & Scherer K. R. (2006). Connecting social identity theory and cognitive appraisal theory of emotions. In Brown R. & Capozza D. (Eds.*). Social Identities: Motivational, Emotional, Cultural Influences*. Hove, England: Psychology Press.

Gendrot C. (2002). *L'influence de la qualité de la voix sur la perception de quatre émotions simulées : Une étude acoustique, perceptive et laryngographique*. Mémoire de DEA de Phonétique, Université Paris 3.

Gendrot, C. (2004). Influence de la qualité de la voix sur la perception de quatre émotions simulées : une étude perceptive et physiologique. *Parole*, 13(1), pp. 1-18.

Gerrards-Hesse A., Spies K. & Hesse F. V. (1994). Experimental induction of emotional states and their affectiveness: a review. *British Journal of Psychology*, 85, pp. 55-78.

Gerratt B. R. & Kreiman J. (2001). Measuring vocal quality with speech synthesis. *The Journal of the Acoustical Society of America*, 110 (5 part 1), pp. 2560-2566.

Glass G.V., Peckham P.D., Sanders J.R. (1972). Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance. *Review of Educational Research*, 42(3), pp. 237-288.

Gobl C. & Ní Chasaide A. (2000). Testing affective correlates of voice quality through analysis and resynthesis. *ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, pp. 178-183.

Gobl C. & Ní Chasaide A. (2003). The role of the voice quality in communicating emotions, mood and attitude, *Speech Comm*. 40, pp. 189–212.

Grandjean D., Ducommun C., Bernard P.-J. & Scherer K.R. (2002). *Comparison of cerebral activation patterns in identifying affective prosody, semantic prosody, and phoneme differences* (résumé). International Organization of Psychophysiology, Montréal, Canada.

Greasley P., Setter J., Waterman M., Sherrard C., Roach P., Arnfield S., Horton D. (1995). Representation of prosodic and emotional features in a spoken language database. *13th International Congress of Phonetic Sciences*, Stockholm, Suède, Vol. 1, pp. 242–245.

Grichkovtsova I., Lacheret A., Morel M., Beaucousin V. & Tzourio-Mazoyer N. (2007). Affective speech gating. *16th International Congress of Phonetic Sciences (ICPhS 2007)*, Saarbrücken, Allemagne, pp. 805-808.

Grosjean F. (1996). Gating. *Language and Cognitive Processes*, 11, pp. 597-604.

Grosjean F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, 28, pp. 267-283.

Gussenhoven C. (2002). Intonation and interpretation: phonetics and phonology. *1st International on Speech Prosody,* Aix-en-Provence, France, pp. 47-57.

Hall J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin*, 85, pp. 845-57.

Hall J. A. (1984). *Nonverbal sex differences: Communication accuracy and expressive style*. Baltimore: Johns Hopkins University Press. (Paperback edition,1990).

Hall J. A., Carter J. D., & Horgan T. G. (2000). Gender differences in the nonverbal communication of emotion. In A. H. Fischer (Ed.), *Gender and emotion: Social psychological perspectives*. Paris: Cambridge University Press, pp. 97-117.

Hallahan W. I. (1996). DECtalk software: text-to-speech technology and implementation. http://www.speech.cs.cmu.edu/comp.speech/Section5/Synth/dectalk.html

Hanson H. M. (1995). *Glottal characteristics of female speakers*. Thèse de doctorat, Harvard University, Division of Applied Sciences.

Harris J. C. (2003). Social neuroscience, empathy, brain integration, and neurodevelopmental disorders. *Physiology and Behavior*, 79 (3), pp. 525-531(7)

Hartigan J. A. & Wong M. A. (1979). A K-means clustering algorithm. *Applied Statistics*, 28, pp. 100–108.

Hecker M. H., Stevens K. N., von Bismarck G. & Williams, C. E. (1968). Manifestation of Task-Induced Stress in the Acoustic Speech Signal, *Journal of the Acoustical Society of America*, 44(4), pp. 993-1001.

Helfrich, H., Standke, R., Scherer, K.R., (1984). Vocal indicators of psychoactive drug effects. *Speech Communication,* 3, pp. 245-252.

Henrich N. (2001). *Etude de la source glottique en voix parlée et chantée : modélisation et estimation, mesures acoustiques et électroglottographiques, perception*. Thèse de doctorat, Université Paris 6.

Henrich N. (2006). Mirroring the voice from Garcia to the present day: Some insights into singing voice registers. *Logopedics Phoniatrics Vocology*, 31, pp. 3-14.

Henrich N. (2008) The source-filter approach in singing and its implication on inverse-filtering techniques. *The 6th International Conference on Voice Physiology and Biomechanics (ICVPB 2008).*Tampere, Finland.

Healy H. & Williams J. M. G. (1999). Autobiographical memory. In T. Dalgleish & M. Power (Eds.), *Handbook of cognition and emotion,* pp. 229–242. New York: Wiley.

Hirst D. & di Cristo A. (1999). A survey of intonation system. In Hirst & Di Cristo (Eds.), *Intonation System, A Survey of twenty languages*

Holm S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, pp. 65-70.

Holmberg E., Hillman R., Perkell J., Guiod P. & Goldman S. (1995). Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice, *Journal of Speech and Hearing Research*, 38, pp. 1212-1223.

Horstmann G. (2002). Facial expressions of emotion: Does the prototype represent central tendency, frequency of instantiation, or an ideal? *Emotion*, 2(3), pp. 297-305.

Houtsma A. J. M. (1995). Pitch Perception in Hearing, *Handbook of Perception and Cognition*, 2nd ed., edited by B. C. J. Moore, Academic, San Diego, pp. 267–295.

Huynh H. & Feldt L.S. (1976). Estimation of the BOX correction for degrees of freedom sample data in randomized block and split-plot designs. *Journal of Educational Statistics*. 1, pp. 69-82.

Huynh H. & Mandeville G. K. (1979). Validity Conditions in Repeated Measures Design. *Psychological Bulletin*, 86 (5), pp. 964-973.

Iida A, Campbell N., Iga S., Higushi F & Yasumura M. (2000). A Speech Synthesis System with Emotion for Assisting Communication. *ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, pp. 167-172.

Iseli M. & Alwan A. (2004). An Improved Correction Formula for The Estimation of Harmonic Magnitudes and Its Application to Open Quotient Estimation. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, Montreal, Canada, pp. 669-672.

Izard C. E. (2000). Organizational and Motivational Functions of Discrete Emotions. In Lewis M. & Haviland J. (Eds.). *The Handbook of Emotions*. New York: Guilford, pp. 631-641.

James W. (1884). What is an emotion ? *Mind*, 9, pp. 188-205.

Johnstone T. & Scherer K. R. (1999). The effects of emotions on voice quality. *14th International Congress of Phonetic Sciences*, San Francisco, Etats-Unis, pp. 2029-2032.

Johnstone T. & Scherer K. R. (2000). Vocal communication of emotion. The In Lewis M. & Haviland J. (Eds.). *The Handbook of Emotions*. New York: Guilford, pp. 226–235.

Johnstone T., van Reekum C. M., Hird K., Kirsner K. & Scherer K. R. (2005). Affective Speech Elicited With a Computer Game. *Emotion*, 5(4), pp. 513–518.

Juslin P. N. & Laukka P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. Emotion, 1, pp. 381-412.

Juslin P. N. & Laukka P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129 (5), pp. 770-814.

Kaiser S. & Wehrle T. (1994). Emotion research and AI: some theoretical and technical issues. *Geneva Studies in Emotion and Communication*, 8 (2), pp. 1-16.

Kanellos I., Suciu I. & Moudenc T. (2007). Expressivité et synthèse vocale. Isotopies expressives, cohérence discursive et structures prosodiques. *Nouveaux Cahiers de Linguistique Française, N° spécial sur « Interfaces Discours-Prosodie »*, 28, pp. 199-206.

Kätsyri J. (2006). *Human recognition of basic emotions from posed and animated dynamic facial expressions*, Ph.D. thesis, Helsinki University of Technology, Finland.

Kehrein R. (2002). The Prosody of Authentic Emotions. *1st International Conference Speech Prosody*, Aix-en-Provence, pp. 423-426.

Kelley J.F. (1983). An empirical methodology for writing user-friendly natural language computer applications. *ACM SIG-CHI'83 Human Factors in Computing systems*, Boston, Etats-Unis, pp. 193-196.

Kittur A., Hummel J. E. & Holyoak K. J. (2006). Ideals Aren't Always Typical: Dissociating Goodness-of-Exemplar From Typicality Judgments. *Twenty Eighth Annual Meeting of the Cognitive Science Society*. Vancouver, Canada.

Klasmeyer G., Johnstone T., Bänziger T., Sappok C. & Scherer K. R. (2000). Emotional Voice Variability in Speaker Verification. *ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, pp. 213-218.

Klatt D. H. & Klatt L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87, pp. 820-857.

Kohler K. J. (1987) Categorical pitch perception. Proc. *11th Int. Conf. Phonetic Science*, Tallinn, 5, pp. 331-333.

Kohler K. J. (1991) Terminal intonation patterns in single-accent utterances of German: phonetics, phonology and semantics. Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK), 25, pp. 115-185.

Kreiman J., Gerratt B. R., Precoda K. & Berke G.S. (1992). Individual Differences in Voice Quality Perception. *Journal of Speech and Hearing Research*, 35, pp. 512-520.

Kreiman J., Gabelman B. & Gerratt B. R. (2003). Perception of vocal tremor. *Journal of Speech, Language, and Hearing Research*, 46, pp. 203-214.

Kreiman J., Gerratt B. R. & Antoñanzas-Barroso N. (2007). Measures of the glottal source spectrum. (2007). *Journal of speech, language, and hearing research*, 50 (3), pp. 595-610.

Kreiman J., Gerratt B. R. & Antoñanzas-Barroso N. (2007) Measures of the glottal source spectrum. *Journal of speech, language, and hearing research*, 50(3), pp. 595-610.

Kreiman J., Vanlancker-Sidtis D. & Gerratt B. R. (2003). Defining and measuring voice quality. *Voice Quality: Functions, Analysis (VOQUAL'03)*, Geneva, Switzerland, pp. 115-120.

de Krom G. (1995). Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of Speech and Hearing Research*, 38, pp. 794–811.

Kohler K.J. (2005). Timing and Communicative Functions of Pitch Contours. *Phonetica*, 62, pp. 88-105.

Labov W. (1997). Some further steps in narrative analysis. *Journal of Narrative & Life History*, 7 (4), pp. 395-415.

Labov W. (1970). The Study of Language in its Social Context. *Studium Generale*, 3, pp. 30–87.

Lacheret-Dujour A. & Beaugendre F. (1999). *La prosodie du français*, CNRS EDITIONS.

Ladd D.R., Silverman K., Tolkmitt F., Bergmann G. & Scherer K.R. (1985). Evidence for the independent function of intonation contour type, voice quality and f0 range in signalling speaker affect. *Journal of the Acoustic Society of America*, 78 (2). pp. 435-444.

Ladefoged P. & Maddieson I. (1996) *Sounds of the world's languages*. Oxford: Blackwells.

Landis J.R. & Koch G.G. (1977). The measurement of observer agreement for categorical data, *Biometrics*, 33, pp. 159–174.

Lange C. (1885). *Les émotions.* Kronar, Copenhague.

Laver J. (1980). *The phonetic description of voice quality*. Cambridge University Press, Cambridge.

Laver J. (2000). Phonetic evaluation of voice quality. In: Kent, RD; Ball, M.J. *Voice Quality Measurement*. San Diego: Singular Publishing Group.

Lazarus R.S. (1991). *Emotion and adapation*. New York: McGraw Hill.

LeDoux J. (1996). *The emotional brain: the mysterious underpinnings of emotional life*. Simon & Schuster, London.

Lea W. A., Medress M. F., Skinner T. E. (1975). A Prosodically Guided Speech Understanding Strategy. *IEEE transactions on Acoustics, Speech and Signal Processing*, 23 (1), pp. 30-39.

Léon P. R. (1971). *Essais de phonostylistique*. Ottawa: Didier.

Léon P. R. (1993). *Précis de phonostylistique : parole et expressivité*. Nathan université.

Levine L. J. & Pizarro D. A. (2004). Emotion and Memory Research: A Grumpy Overview. *Social Cognition*. 22, pp. 530-544.

Levine L. J. & Safer M. A. (2002). Sources of bias in memoty for emotions. *Current Directions in Psychological Science*, 11, pp. 169-173.

Lewin I., Becket R., Boye J., Carter D., Rayner M. & Wirén M. (1999). Language processing for spoken dialogue systems : is shallow parsing enough ? *ESCA ETRW Workshop*, Cambridge, Grande-Bretagne, pp. 37-42.

Lewis M. (2005). Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and Brain Sciences*, 28(02), pp. 169-194.

Liberman A. M. & Mattingly I. G. (1985). The motor theory of speech perception revised, *Cognition* 21, pp. 1–36.

Lieberman P. & Michaels S. B. (1962). Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech, *The Journal of the Acoustical Society of America*, 34 (7), pp. 922-927.

Loyau F. (2002). *Etablissement d'un corpus de parole émotionnelle "capturée" par un magicien d'Oz sur un simulateur de vol*. Mémoire de DEA de Sciences cognitives, Institut National Polytechnique de Grenoble.

Loyau F. (2007). *Expressions des états mentaux et émotionnels de l'humain en interaction : ébauches du "Feeling of Thinking"*. Thèse de Doctorat en Sciences Cognitives. Institut National Polytechnique de Grenoble.

Lunney G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement*, 7 (4), pp. 263-269.

MacNeil D. (1992) *Hand and Mind.* University of Chicago Press.

Maeda S. (1976). *A characterization of American English intonation*. Thèse de doctorat, Massachussets Institute of Technology.

Massaro D. W. (2000). Multimodal emotion perception: Analogous to speech processes. *ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, pp. 114-121.

Max L. & Onghena P. (1999). Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research*, 42, pp. 261–270.

Mayer J.D. & Salovey P. (1997). What is emotional intelligence?. In: P. Salovey and D. Sluyter, Editors, *Emotional development and emotional intelligence: Implications for educators*, Basic Books, New York, pp. 3–31.

Mertens P. & d'Alessandro C. (1995). Pitch contour stylization using a tonal perception model, *13th International Congress of Phonetic Sciences*, pp. 228-231.

Mishra R., Shriberg E., Upson S., Chen J., Weng F., Peters S., Cavedon L., Niekrasz J., Cheng H. & Bratt H. (2004). A Wizard of Oz Framework for Collecting Spoken Human-Computer Dialogs. *8th International Conference on Spoken Language Processing (INTERSPEECH 2004 – ICSLP),* Jeju Island, Korea, pp. 2269-2272.

Mokhtari P. & Campbell N. (2002). Automatic Detection of Acoustic Centres of Reliability for Tagging Paralinguistic Information in Expressive Speech. *3rd International Conference on Language Evaluation and Resources*, Las Palmas, Espagne, pp. 2015-2018.

Montero J. M., Gutiérrez-Arriola J., Colás J., Enríquez E. & Pardo J. M. (1999). Analysis and Modelling of Emotional Speech in Spanish, *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 1999)*, San Fransisco, USA, pp. 957-960.

Morel M., Lacheret-Dujour A., Lyche C. & Poiré F. (2006) Vous avez dit proéminence ? *15èmes Journées d'Etudes sur la Parole (JEP 2006)*, Dinard, France, pp. 183-186.

Morlec Y., Rilliard A., Bailly G. & Aubergé V. (1998). Evaluating the adequacy of synthetic prosody in signalling synthetic boundaries: methodology and first re-sults. *1st International Conference on Language resources and Evaluation*, Granada, Spain, 1, pp. 647-650.

Morlec Y. (1997). *Génération multiparamétrique de la prosodie du français par apprentissage automatique.* Thèse de doctorat, spécialité Sciences Cognitives, Institut National Polytechnique de Grenoble.

Morlec Y., Bailly G. & Aubergé V. (1999). Training an application-dependent prosodic model: corpus, model and evaluation. *European Conference on Speech Communication and Technology*, Budapest, Hongrie, pp. 1643-1646.

Moulines E. & Laroche J. (1995). Non-parametric techniques for pitch-scale and time-scale modifications of speech. *Speech Communication* (16), pp. 175–205.

Mozziconacci S. (1998). *Speech Variability and Emotion: Production and Perception.* These de doctorat, Eindhoven University, Pays-Bas.

Mozziconacci S. (2000). The Expression of Emotion Considered in the Framework of an Intonation Model. *ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, pp. 45-52.

Mozziconacci S. J. L. & Hermes D. J. (1999). Role of intonation patterns in conveying emotion in speech, *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 1999)*, San Fransisco, USA, pp. 2001-2004.

Murray I. R. & Arnott J. L. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication,* 16(4), pp. 369-390.

Ní Chasaide A. & Gobl C. (2000). Voice source variation. In W. J. Hardcastle and J. Laver (eds.). *The Handbook of Phonetic Sciences*. Oxford: Blackwell, pp. 427-461.

Ní Chasaide A., Gobl C. (2003). Voice quality and expressive speech. *1st International Workshop on Expressive Speech Processing (JST/CREST)*, Kobe, Japon, pp. 19-27.

Noiray A. (2003). Comparaison entre acteurs et non-acteurs en parole émotionnelle. Mémoire de DEA Sciences du Langage, Université Stendhal – Grenoble III.

Ohala J. J. (1983). Cross-language use of pitch: an ethological view. *Phonetica*, 40, pp. 1-18.

Ohala J. J. (1996). Ethological theory and the expression of emotion in the voice. *4th International Conference on Spoken Language Processing*, Philadelphia, USA. Vol. 3, pp. 1812-1815.

Pakosz M. (1983). Attitudinal judgments in intonation: Some evidence for a theory. *Journal of Psycholinguistic Research*, 12, pp. 311–326.

Pasdeloup V. (1992). A prosodic model for French text-to-speech synthesis: a psycholinguistic approac. in Bailly, G. & Benoît, C. (ed) *Talking Machines: theories, Models and designs*. Elsevier Science Publ. North Holland, pp. 335-348.

Pashler H. (1997). *The Psychology of Attention.* MIT Press Pub.

Paulmann S., Schmidt P., Pell M. & Kotz S. (2008). Rapid processing of emotional and voice information as evidenced by ERPs. *4th International Conference Speech Prosody*, Campinas, Brésil, pp. 205-209.

Pavlenko A. (2005). *Emotions and Multilingualism*. Cambridge, UK: Cambridge University Press.

Peterson R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Psychology*, 21, pp. 381-391.

Petrushin V. (2000). Emotion recognition in speech signal: experimental study, development, and application. *6th International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, China, vol. 2, pp. 222-225.

Pfitzinger H. (2006). Five dimensions of prosody : Intensity, intonation, timing, voice quality, and degree of reduction, *3rd International Conference Speech Prosody*, Dresde, Allemagne. Actes CD-ROM.

Philippot P. (1993). Inducing and Assessing differentiated Emotion-Feeling states in the Laboratory. *Cognition and Emotion*, 7 (2), pp.171-193.

Picard R. (1997). *Affective Computing*. MIT Press, Cambridge, Mass, London, Angleterre.

Pochon R., Brun P. & Mellier D.(2006) Développement de la reconnaissance des émotions chez l'enfant avec trisomie 21. *Psychologie Française* 51(4), pp. 381-390.

Pocock G. & Richards C.D. (2004). *Physiologie humaine : Les fondements de la médecine*. Trad. de Brun J.F., Caillaud C., Mercier J., Raynaud E., Paris : Masson.

Pourtois G., De Gelder B. & Vroomen J. (2000). The time-course of intermodal binding between seeing and hearing affective information. *Neuroreport*, 11 (6), pp. 1329-1333.

Protopapas A. & Lieberman P. (1997). Fundamental Frequency of Phonation and Perceived Emotional Stress. *The Journal of the Acoustical Society of America.* 101 (4), pp. 2267-2277.

Rabiner L. & Juang B. H. (1993). *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall.

Rao A. S. & Georgeff M. P. (1995). BDI Agents: From Theory to Practice, *1st International Conference on Multi-Agent Systems (ICMA 1995)*, San Francisco, USA. pp. 312-319.

Rebreyend J. (2002). *Réalisation d'une application client/serveur basée sur le paradigme du Magicien d'Oz*. Mémoire de DUT Informatique, Grenoble.

Rebreyend J. (2006). Le *dégoût et l'amusement dans la parole : Un « liptube » écologique*. Mémoire de Master en Modèles & Instruments en Médecine et Biologie, Université Joseph Fourier - Grenoble 1.

Reidsma D., Heylen D. & Ordelman R. (2006). Annotating Emotion in Meetings. In processedings 5th Int. Conf. on *Language Resources and Evaluation* (LREC 2006), Genova, Italy.

Rietveld T. & van Hout R. (2005). *Statistics in Language Research: Analysis of Variance*. Berlin : Walter de Gruyter GmbH & Co.

Rilliard A. (2000). *Vers une mesure de l'intelligibilité linguistique de la prosodie – évaluation diagnostique des prosodies synthétique et naturelle*. Thèse en Sciences Cognitives de l'INP Grenoble.

Rilliard A., Shochi T., Martin J.-C., Erickson D. & Aubergé V. (sous presse). Multimodal indices to Japanese and French prosodically expressed social affects. *Language and Speech*.

Roach P. (2000). Techniques for the phonetic description of emotional speech. *ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, pp. 53-59.

Rosch E. & Mervis C. B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology*, 7, pp.573-605.

Rose Y. & dos Santos C. (2006). Facteurs prosodiques et articulatoires dans l'harmonie consonantique et la métathèse en acquisition du français langue première. *Recherches linguistiques de Vincennes*, 35, pp. 77-102.

Rosec O. & Vincent D. (2008). Modification of a voice signal. Brevet d'invention consultable à l'URL : http://www.freepatentsonline.com/EP1944755.html

Rosenberg E.L. & Ekman P. (1994). Coherence between expressive and experiential systems in emotion. *Cognition & Emotion*, 8, pp. 201-229.

Rosenthal R. & Rubin D. B. (1989). Effect Size Estimation for One-Sample Multiple-Choice-Type Data: Design, Analysis, and Meta-Analysis. *Psychological Bulletin*, 106(2), pp. 332-337.

Rosnow R. L. & Rosenthal R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology*, 57(3), pp. 221-237.

Rossi M., (1971). L'intensité spécifique des voyelles. *Phonetica*, 24, pp. 129-161.

Rossi M. (1971). Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica,* 23, pp. 1-33.

Rossi M. (1978). La perception des glissando descendants dans les contours prosodiques. *Phonetica* 35, pp. 11–40.

Rossi M. (1993). A model for predicting the prosody of spontaneous speech (PPSS model). *Speech Communication*, 13, pp. 87-107.

Rossi M. (1999). *Intonation: past, present and future*. A. Botonis (ed). Elsevier Pub.

Rothenberg M. (1991). A Multichannel Electroglottograph. *Journal of Voice*, 6 (1), pp. 36-43.

Rozin P., Haldt J., McCauley C. R. & Imada S. (1997). Disgust Preadaptation and the Cultural Evolution of a Food-Based Emotion in *Food Preferences and Taste: Continuity and Change*, ed. Helen Macbeth. Providence, R.I.: Berghahn.

Rozin P., Taylor C., Ross L., Bennett G. & Hejmadi A. (2005). General and specific abilities to recognise negative emotions, especially disgust, as portrayed in the face and the body. *Cognition and Emotion* (19), pp. 397–412.

Russell J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin,* 115, pp. 102–141.

Sakakibara K-I., Fuks L., Imagawa H. & Tayama N. (2004). Growl voice ethnic and pop styles. *proceedings of the international symposium on musical acoustics*. Nara, Japan.

Salber D. & Coutaz J. (1993). Applying the Wizard of Oz Technique to the Study of Multimodal Systems. *East-West International Conference on Human-Computer Interaction*, Springer-Verlag, Berlin, Allemagne, pp. 219-230.

Sander, D. (2008). Basic tastes and basic emotions: basic problems, and perspectives for a nonbasic solution. *Behavioral and Brain Sciences*, 31(1), pp. 88.

Scassellati B. (2001) *Foundations for a Theory of Mind for a Humanoid Robot.* PhD Thesis, MIT Department of Computer Science and Electrical Engineering, Etats-Unis.

Scherer K. R. (1984). Emotion as multicomponent process : A model and some cross-cultural data. *Review of Personality and Social Psychology* (5), pp. 37-63.

Scherer K. R. (1989). Vocal measurement of emotion. In Plutchik R. & Kellerman H. (Eds.), *Emotion: Theory, research, and experience. Vol. 4. The measurement of emotion.* New York: Academic Press, pp. 233-260.

Scherer K. R. (1986). Vocal Affect Expression: A Review and a Model for Future Research. *Psychological Bulletin*, 99 (2), pp. 143-165.

Scherer, K. R. (1994). Affect bursts. *Emotions: Essays on emotion theory,* pp. 161–196.

Scherer K. R. (2000). Psychological models of emotion. In J. Borod (Ed.). *The neuropsychology of emotion*. Oxford/New York: Oxford University Press, pp. 137-166.

Scherer K. R. (2001). Appraisal considered as a process of multi-level sequential checking. In K Scherer, A Schorr, & T. Johnstone (Eds.). *Appraisal processes in emotion: Theory, Methods, Research*, Oxford University Press, pp. 92-120.

Scherer K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40, pp. 227-256.

Scherer K. R. (2004). Feelings integrate the central representation of appraisal-driven response organization in emotion. In Manstead A. S. R., Fridja N. & Fisher A. (Eds.). *Feelings and Emotions: The Amsterdam Symposium*, Cambridge: Cambridge University Press, pp. 136–157.

Scherer K. R., Ladd D. R. & Silverman (1984). Vocal cues to speaker affect: testing two models. *Journal of the Acoustic Society of America,* 76 (5), pp. 1346-1356.

Scherer K. R. & Ceschi G. (1997). Criteria for emotion recognition from verbal and nonverbal expression: Studying baggage loss in the airport. P*ersonality and Social Psychology Bullettin*, 26(3), pp. 327-339.

Scherer K. R., Banse R. & Wallbott H.G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, pp.76–92.

Scherer K. R., Johnston T. & Klasmeyer G. (2003). Vocal Expression of Emotion. In R.J. Davidson, K.R. Scherer, H.H. Goldsmith (Eds). *Handbook of Affective Sciences*, pp. 433-456.

Scherer K. R. & Ellgring H. (2007). Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns? *Emotion*, 7(1), pp. 158-171.

Schröder M. (1998). *L'expression vocale de l'amusement : premières expériences audiovisuelles.* Mémoire de maîtrise en Sciences du Langage, mention Industrie de la Langue, Université Stendhal – Grenoble III.

Schröder M. (2001). Emotional speech synthesis: a review. *7th European Conference on Speech Communication and Technology (EUROSPEECH 2001),* Aalborg, Denmark, pp. 561-564.

Schröder M. (2003). Experimental study of affect bursts. *Speech Communication* 40(1-2), pp. 99-116.

Searle J. R. (1969). *Speech acts*. Cambridge, England: Cambridge University Press.

Searle J. R. (1975). A taxonomy of illocutionary acts. In *Language, Mind and Knowledge*. University of Minnesota, USA, pp. 344-369.

Seppi D., Batliner A., Schuller B., Steidl S., Vogt T., Wagner J., Devillers L., Vidrascu L., Amir N. & Aharonson V. (2008). Patterns, Prototypes, Performance: Classifying Emotional user States. *12ème European Conference on Speech Communication and Technology (INTERSPEECH 2008)*, Brisbane, Australie, pp. 601-604.

Shochi T. (2008). *Prosodie des affects socioculturels en japonais, français et anglais : à la recherche des vrais et faux-amis pour le parcours de l'apprenant*. Thèse en Sciences du Langage, Université Stendhal – Grenoble III.

Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J. & Hirschberg J. (1992). TOBI: A Standard for Labeling English Prosody. *International Conference on Spoken Language Systems*, Banff, Canada, pp. 867-870.

Sinha R. (1996). Multivariate Response Patterning of Fear and Anger. *Cognition & Emotion*, 10 (2), pp. 173-198.

Smith E. E., Shoben, E. J. & Rips L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 81 (3), pp. 214-241.

Spinoza B. (1677). *L'Ethique*. Paris, Gallimard, 1994.

Stanislavski K. (1966). *La construction du personnage*. Pygmalion Gerard Watelet.

Shahid S., Krahmer E. & Swerts M. (2008). Real vs. acted emotional speech: Comparing South-Asian and Caucasian speakers and observer. *4th International Conference on Speech Prosody*, Campinas, Brésil, pp. 669-672.

Suciu I., Kanellos I. & Moudenc T. (2006). Formal expressive indiscernibility underlying a prosodic deformation model, *International Speech Communication Association (ISCA) : Tutorial and Research Workshop on Experimental Linguistics*, Athens, Greece.

Suciu I., Kanellos I. & Moudenc T. (2006). What about the text ? Modelling global expressiveness in speech synthesis. *IEEE International Conference on Information and Communication Technologies: from Theory to Applications (ICCTA'06)*, Damas, Syria.

Sundberg J. (2002). Estimating perceived phonatory pressedness in singing from flow glottograms, *31st Annual Symposium Care of the Professionnal Voice*, Philadelphia, Etats-Unis, pp. 89-96.

Swerts M. & Krahmer E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language.* 53:1, pp. 81-94.

Swerts M. & Veldhuis R. (2001). The effect of speech melody on voice quality. *Speech Communication*. 33(4), pp. 297-303.

Syrdal A. K. & Kim Y.-J. (2008). Dialog speech acts and prosody: Considerations for TTS, *4th International Conference on Speech Prosody*, Campinas, Brésil, pp. 661-665.

Talarico J. M., LaBar K. S. & Rubin D. C. (2004). Emotional intensity predicts autobiographical memory experience. *Memory & Cognition*, 32, pp. 1118-1132.

Talarico J. M. & Rubin D. C. (2003). Confidence, not consistency, characterizes flashbulb memories. *Psychological Science*, 14, pp. 455-461.

Tartter V. C. (1980). Happy talk: perceptual and acoustic effects of smiling on speech. *Perception & psychophysics*, 27 (1), pp. 24-27.

Tartter V. C. & Braun D. (1994). Hearing smiles and frowns in normal and whisper registers. In *Journal of the Acoustical Society of America*, 96 (4), pp. 2101-2107

Tcherkassof A. (1999). Les indices de préparation à l'action et la reconnaissance des expressions émotionnelles faciales. *Revue Européenne de Psychologie Appliquée*, 49 (2), pp. 99-105.

Terrier M. (2003). Stratégies expressives écrit versus oral : quels degrés de liberté de l'oral. Mémoire de DEA en Sciences Cognitives. Institut National Polytechnique de Grenoble.

Thomas D. & Diener E. (1990). Memory accuracy in the recall of emotions. *Journal of Personality and Social Psychology*, 59, pp. 291-297.

Thompson K. G., Hanes D. P., Bichot N. P. & Schall J. D. (1996). Perceptual and motor processing stages identified in the activity of macaque frontal eye field neurons during visual search. *Journal of Neurophysiology*, 76, pp. 4040-4055.

Thorsen N. (1980). A study of the perception of sentence intonation - evidence from Danish. *Journal of the Acoustical Society of America*, 67(3), pp. 1014-1030.

't Hart J., Collier R. & Cohen A. (1990). *A perceptual study of intonation.* Cambridge: Cambridge University Press.

Tolkmitt F. & Scherer K.R. (1986). Effet of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology: Human Perception and Performance*, 12, pp. 302-313.

Tomkins S. S. (1980). Affect as amplification: some modifications in theory. In Plutchick R. & Kellerman H. Eds. *Emotion, Theory, research and experience: theories of emotions*, NY: Academic Press, pp. 141-165.

Trouvain J. & Barry W. J. (2000). The prosody of excitement in horse race commentaries. *ISCA Workshop on Speech and Emotions*, Newcastle, Irlande du Nord, pp. 86-91.

Tubach J. P. (1989). La parole et son traitement automatique. Masson, Paris.

Tyler L & Wessels J. (1985). Is gating an on-line task? Evidence from naming latency data. *Perception and Psychophysics*, 38, pp. 217–222.

Vaissière J. (1995). Nasalité et Phonétique In Le voile du palais et la parole, *Colloque sur le voile pathologique*, Société Française de Phoniatrie et Groupe Francophone de la Communication Parlée, Lyon, France.

Vaissiere J. (1997). Phonological use of the larynx. *Larynx 97*, Marseille, pp. 115-125.

Velten E. (1968). A laboratory task for induction of mood states. *Behavior Research and Therapy*, 6, pp. 473-482.

Vanpé A. & Aubergé V. (2008). Static vs. dynamic Gestural Icons of "Feeling of Thinking". 2nd International Workshop on *Corpora for research on Emotion & Affect*, Marrakech, Maroc.

Vernet-Maury E., Collet C., Robin O., Delhomme G. & Dittmar A. (2001). Capter et mesurer la qualité de la réponse émotionnelle. In *L'Art, La Pensée & Les Emotions*. Y. Orlarey (ed.).

Vidrascu L. & Devillers L. (2005). Detection of real-life emotions in call centers. *9th European Conference on Speech Communication and Technology (INTERSPEECH 2005)*, Libonne, Portugal, pp. 1841-1844.

Vidrascu L. & Devillers L. (2005). Annotation and detection of blended emotions in real human-human dialogs recorded in a call center, *IEEE International Conference on Multimedia (ICME 2005),* Amsterdam, The Netherlands.

Viola I. C. & Madureira S. (2007). Voice quality and Speech Expressiveness. Proccedings of the *First International Workshop on Paralinguistic Speech07-* between models and data. Saarbrücken, Allemagne, pp. 65-70.

Vion M. & Colas A. (2006). Pitch cues for the recognition of Yes-No questions in French, *Journal of Psycholinguistic Research*, 35(5), pp. 427-445.

Vincent D. (2007). *Analyse et contrôle du signal glottique en synthèse de la parole*. Thèse de doctorat, spécialité Traitement du Signal et Télécommunications, Ecole Nationale Supérieure des Télécommunications de Bretagne.

Vincent D., Rosec O. & Chonavel T. (2005). Estimation of LF glottal source parameters based on ARX model. *9th European Conference on Speech Communication and Technology (INTERSPEECH 2005)*, Libonne, Portugal, pp. 333-336.

Wagner H.L (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior,* Springer Netherlands, vol 17, (1), pp. 3-28.

Wallbott H. & Scherer K.R. (1989). Assessing emotion by questionnaire. *Emotion: Theory, research, and experience, 4. The measurement of emotion*, pp. 55-82.

Wharton T. (2003). Interjections, language and the 'showing-saying' continuum. *Pragmatics & Cognition* (11), pp. 39-91.

Wierzbicka A. (1992). Talking about Emotions : Semantics, Culture, and Cognition, *Cognition and Emotion*, 6 (3/4), pp. 285-319.

Williams C.E. & Stevens K.N. (1972). Emotions and speech:some acoustical correlates. *Journal of the Acoustic Society of America*, 52, 4 (part 2), pp. 1238-1250.

Wilting J., Krahmer E. & Swerts M. (2006). Real vs. acted emotional speech. *10th European Conference on Speech Communication and Technology (INERSPEECH 2006)* (CD-ROM proceedings).

Yanushevskaya I., Gobl C. & Ní Chasaide A. (2005). Voice quality and f0 cues for affect expression: implications for synthesis. *9th European Conference on Speech Communication and Technology (INTERSPEECH 2005)*, Libonne, Portugal, pp. 1849-1852.

Yegnanarayana B., d'Alessandro C. & Darsinos V (1998). *An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. IEEE transaction on Speech and Audio Processing*, 6(1), pp. 1-11.

Yumoto E., Gould W. J. & Baer T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America*, 71, pp. 1544-1550.

Publications du candidat

Revue avec comité de lecture

Aubergé, V., **Audibert, N.** & Rilliard, A. (2006). De E-Wiz à E-Clone : recueil, modélisation et synthèse d'expressions authentiques. *Revue d'Intelligence Artificielle – Numéro spécial « Interactions émotionnelles »*, vol. 20 (4-5), pp. 499-528.

Chapitres d'ouvrages

Laukka, P., **Audibert, N.** & Aubergé, V. (2009). Exploring the graded structure of vocal emotion expressions. In Hancil, S. (Ed.), *The role of prosody in the expression of emotions in English and in French, Linguistic Insights 97*, Bern, Suisse: Peter Lang AG, pp. 241-258.

Boë, L.-J., Bessière, P., Ladjili, N., **Audibert, N.** (2008). Simple combinatorial considerations challenge the proof of Ruhlen's mother tongue theory. In Davis, B.L., Zajdo, K. (Eds.), *Syllable Development: The Frame/Content Theory and Beyond*. Mahwah, NJ: L. Erlbaum, pp. 63-9

Conférences et workshops internationaux avec actes publiés

Audibert, N., Aubergé, V. & Rilliard, A. (2008). Acted vs. spontaneous expressive speech: perception with inter-individual variability. *2nd International Workshop on Corpora for Research on Emotion and Affect*, Marrakech, Maroc, pp. 23-27.

Audibert, N., Aubergé, V. & Rilliard, A. (2008). How we are not all competent the same for discriminating acted from spontaneous expressive speech. *4th International Conference on Speech Prosody*, Campinas, Brésil, pp. 693-696.

Bõhm T., **Audibert N.**, Shattuck-Hufnagel S., Németh G. & Aubergé, V. (2008). Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles. *Proceedings of Acoustics'08*, Paris, France, actes CD-ROM.

Fék, M., **Audibert N.**, Szabo, J., Rilllard, A., Németh, G. & Aubergé, V. (2008). Multimodal Spontaneous Expressive Speech Corpus for Hungarian. *5th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Maroc.

Audibert, N., Aubergé, V. & Rilliard, A. (2007). When is the Emotional Information? A gating experiment for gradient and contours cues. *16th International Congress of Phonetic Sciences (ICPhS 2007)*, Saarbrücken, Allemagne, pp. 2137-2140.

Laukka, P., **Audibert, N.** & Aubergé, V. (2007). Graded structure in vocal expression of emotion: What is meant by "prototypical expressions"? *1st International Workshop on Paralinguistic Speech*, Saarbrücken, Allemagne, pp. 1-4.

Aubergé, V., **Audibert, N.** & Rilliard, A. (2006). Auto-annotation: an alternative method to label expressive corpora. *1st International Workshop on Corpora for Research on Emotion and Affect*, Gênes, Italie, pp. 45-46.

Audibert, N., Vincent, D., Aubergé, V. & Rosec, O. (2006). Evaluation of expressive speech resynthesis. *1st International Workshop on Corpora for Research on Emotion and Affect*, Genoa, Italie, pp. 37-40.

Audibert, N., Vincent, D., Aubergé, V. & Rosec, O. (2006). Expressive Speech Synthesis: Evaluation of a Voice Quality Centered Coder on the Different Acoustic Dimensions. *3rd International Conference on Speech Prosody*, Dresden, Allemagne (actes CD-ROM).

Burkhardt, F., **Audibert, N.**, Malatesta, L., Türk, O., Arslan, L. & Aubergé, V. (2006). Emotional Prosody - Does Culture Makes A Difference? *3rd International Conference on Speech Prosody*, Dresde, Allemagne (actes CD-ROM).

Audibert, N., Aubergé, V. & Rilliard, A. (2005). The relative weights of prosodic parameters for the expression of emotion in speech: a resynthesis study. *1st International Conference on Affective Computing & Intelligent Interaction*, Beijing, Chine, pp. 527-534.

Audibert, N., Aubergé, V. & Rilliard, A. (2005). The prosodic dimensions of emotion in speech: the relative weights of parameters. *9th European Conference on Speech Communication and Technology (INTERSPEECH 2005)*, Lisbonne, Portugal, pp. 525-528.

Aubergé, V., **Audibert, N.** & Rilliard, A. (2004). E-Wiz: A trapper protocol for hunting the expressive speech corpora in lab. *4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbonne, Portugal, pp. 179-182.

Aubergé, V., **Audibert, N.** & Rilliard, A. (2004). Acoustic morphology of expressive speech: What about contours? *2nd International Conference on Speech Prosody*, Nara, Japon, pp. 91-95.

Rilliard, A., Aubergé, V. & **Audibert, N.** (2004). Evaluating an authentic audio-visual expressive speech corpus. *4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbonne, Portugal, pp. 175-178.

Rossato, S., **Audibert, N.** & Aubergé, V. (2004). Emotional voice measurement: a comparison of articulatory-EGG and acoustic-amplitude parameters. *2nd International Conference on Speech Prosody*, Nara, Japon, pp. 749-752.

Aubergé, V., **Audibert, N.** & Rilliard, A. (2003). Why and how to control the authentic emotional speech corpora? *8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, Genève, Suisse, pp. 185-188.

Conférences et workshops francophones avec actes publiés

Audibert, N., Aubergé, V. & Rilliard, A. (2008). Emotions actées vs. spontanées : variabilité des compétences perceptives. *17èmes Journées d'Etude sur la Parole (JEP 2008)*, Avignon, France, pp. 257-260.

Audibert, N., Aubergé, V. & Rilliard, A. (2006). Synthèse vocale des émotions – Donner la parole émue à C-Clone. *2ème Workshop francophone sur les Agents Conversationnels Animés (WACA 2006)*, Toulouse, France, pp. 27-35.

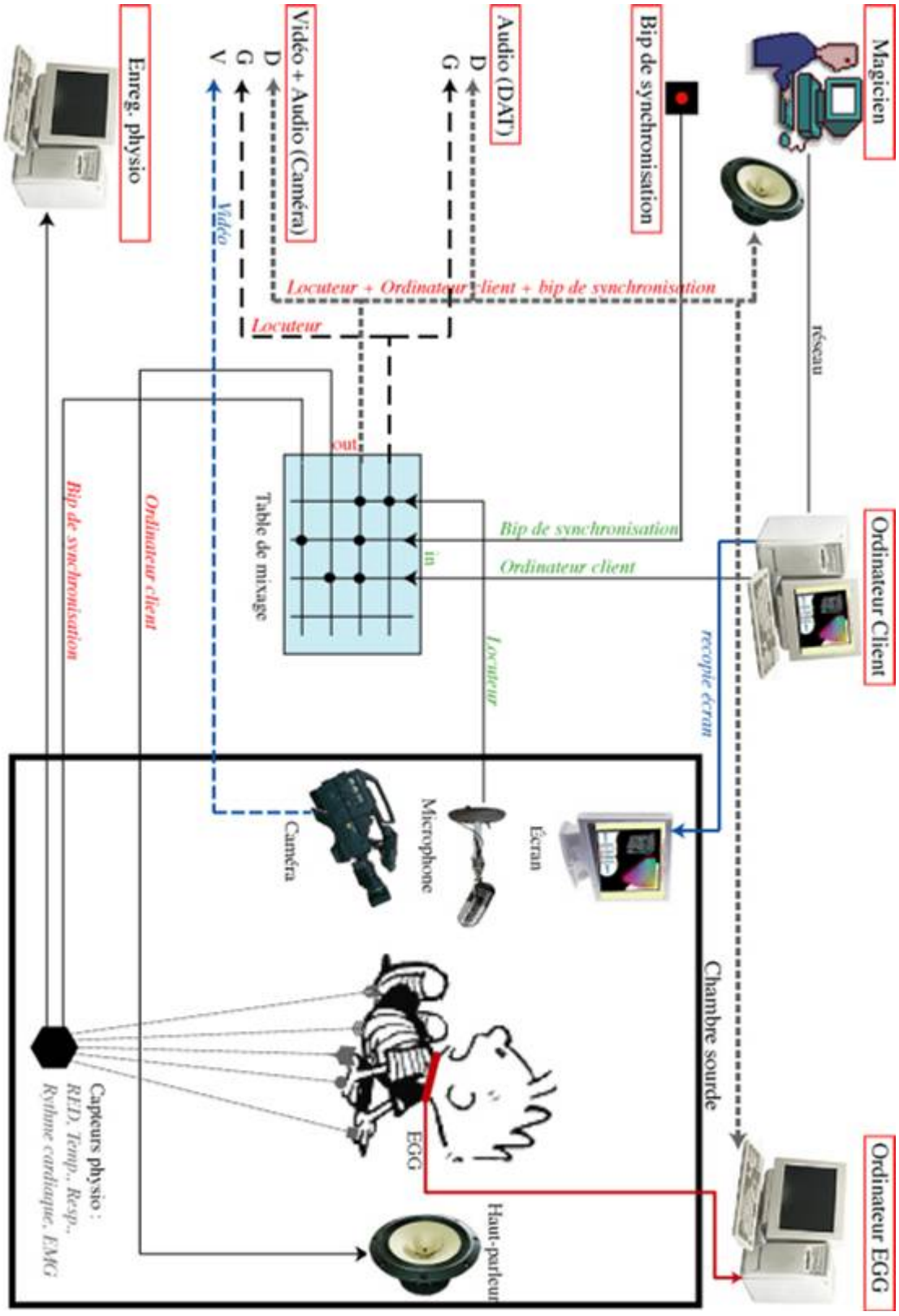
Audibert, N., Vincent, D., Aubergé, V., Rilliard A. & Rosec, O. (2006). Dimensions acoustiques de la parole expressive : poids relatifs des paramètres resynthétisés par Praat vs. LF-ARX. *16èmes Journées d'Etude sur la Parole (JEP 2006)*. Dinard, France, pp. 341-344.

Aubergé, V., Rilliard, A. & **Audibert, N.** (2005). De E-Wiz à E-Clone : méthodologie expérimentale pour la modélisation des émotions et affects authentiques. *1er Workshop francophone sur les Agents Conversationnels Animés (WACA'01)*, Grenoble, France, pp. 125-134.

Audibert, N., Rossato, S. & Aubergé, V. (2004). Paramétrisation de la qualité de voix : EGG vs. filtrage inverse. *15èmes Journées d'Etude sur la Parole (JEP 2004)*, Fès, Maroc, pp. 53-56.

Audibert, N., Aubergé, V. & Rilliard, A. (2004). EWiz : contrôle d'émotions authentiques. *15èmes Journées d'Etude sur la Parole*, Fès, Maroc, pp. 49-52.

Annexe I : Schéma du montage utilisé (Arnal, 2003)



Annexe II : Auto-annotation des locuteurs sélectionnés

Les informations reproduites dans cette section ont été recueillies auprès des acteurs par Aude Noiray dans le cadre de son mémoire de DEA (Noiray, 2003). Les grilles d'auto-annotation correspondent au protocole que nous avons défini pour l'ensemble des sujets. Les informations concernant la locutrice F2 n'ont pu être collectées.

Fiche d'informations de la locutrice F1

Informations générales

Age : 19 ans

Lieu de résidence : Grenoble

Langue maternelle : Français

Depuis combien d'années faites-vous du théâtre ?

Depuis sept ans.

Faites vous partie d'une troupe / école de théâtre ?

Je fais partie d'une ligue d'improvisation de Grenoble.

De manière générale, avez-vous une méthode pour exprimer des émotions ? Si oui laquelle ?

Je fais surtout appel à l'improvisation pour renforcer le naturel. J'ai fait aussi appel à l'imagination pour la 'tristesse'. J'ai essayé d'exprimer les émotions de manière instinctive, sans vraiment réfléchir.

Induction d'émotions via la plateforme « Sound Teacher »

Qu'avez-vous pensé de l'expérience du « Sound Teacher» ? Avez-vous des suggestions, critiques à nous faire ?

Non, juste que c'était vraiment long, à la fin, j'étais vraiment énervée !! Mais j'ai trouvé ça drôle de participer à ce genre d'expérience.

Pensez-vous que cette expérience soit un bon moyen de « piéger » les individus afin d'obtenir des émotions?

J'ai été piégé par ce logiciel. La mise en scène est bien faite et on y croît vraiment. Par contre, j'ai compris que mes résultats ne collaient pas avec mes vraies performances car souvent, je ne comprenais rien. Quand j'ai vu que je continuais à avoir des bons résultats, je me suis douté qu'ils étaient factices.

Avez-vous essayé de contrôler vos émotions durant cet exercice ?

Non mais j'ai essayé de rester concentrée. Je voulais bien faire !

Simulation d'émotions

Avez-vous des remarques, critiques sur cette expérience ?

J'ai trouvé cette partie assez dure car les phrases données n'aident pas du tout à exprimer les émotions. Elles n'ont aucun rapport. Je n'ai pas été gênée par la chambre sourde mais plutôt par le fait d'être assis. J'ai besoin du corps pour exprimer des émotions ou des sentiments. Cela paralyse un peu d'être assis.

Quelles sont les émotions qui ont été les plus difficiles à exprimer ?

J'ai trouvé difficile de nuancer la joie et l'amusement. J'ai eu aussi du mal à exprimer un dégoût sincère. Donc ces émotions sont à mon avis stéréotypées.

Avez-vous ressenti un quelconque stress durant cet exercice (du aux capteurs physiologiques par exemple) ?

Non pas du tout. Au début j'ai trouvé ça drôle et après, je n'y ai plus pensé.

Des suggestions ?

Non.

Grille d'auto-annotation de la locutrice F1

Temps	*Phase*	*Emotion*
En minutes	Introduction :	
	Phase 1 : les voyelles du français	
2,88	- prononciation	Surprise - perplexe
7,43	- perception ouverture	concentrée
9,78	- commentaires	amusement
10,41	- perception position Av /Ar	concentrée
11,68	- commentaires	Concentré
12,28	- perception arrondissement	Perplexe, concentrée
13,13	- commentaires	étonnée
13,68	- perception centralisation	concentrée
14,4	- commentaires	idem
14,7	- résultats généraux	perplexe
	Phase 2 : les voyelles proches de celles du français	
15,95	- perception ouverture	Surprise, agacement
17,11	- commentaires	blasée
17,78	- perception position Av /Ar	concentration
18,7	- commentaires	idem
19,2	- perception arrondissement	Agacement, concentrée
20,13	- commentaires	amusement
20,45	- perception centralisation	Agacement, concentrée

21,5	- commentaires	Amusement
22,1	- prononciation	Amusement, concentrée
24,05	- résultats généraux	Concentrée, largement agacée
	Phase 3 : généralisation aux langues du monde	
25,16	- perception ouverture	agacée
26,7	- commentaires	idem
26,95	- perception position Av /Ar	idem
27,81	- commentaires	idem
28,18	- perception arrondissement	idem
29	- commentaires	idem
29,46	- perception centralisation	idem
30,53	- commentaires	idem
30,9	- prononciation	idem
31,61	- Warning	Surprise des résultats mais je m'en fout
	Phase 4 : Vérification perceptive	
33,31	- perception ouverture	Concentrée au début puis agacée
34,7	- commentaires	agacée
35,1	- perception position Av /Ar	Agacée mais une pointe de concentration, je veux bien faire
35,8	- commentaires	idem
36,26	- perception arrondissement	idem

36,88	- commentaires	idem
37,41	- perception centralisation	idem
37,75	- commentaires	idem
38	- Warning perceptif	idem
38,81	- prononciation	Très agacée
40,03	- Warning production	Blasée des résultats
40,51	- Fin	

Commentaires supplémentaires :

Fiche d'informations de la locutrice F3

Informations générales

Age : 19 ans

Lieu de résidence : Grenoble

Langue maternelle : Français

Depuis combien d'années faites-vous du théâtre ?

Depuis six ans.

Faites vous partie d'une troupe / école de théâtre ?

Je fais partie du Conservatoire de théâtre de Grenoble.

De manière générale, avez-vous une méthode pour exprimer des émotions ? Si oui laquelle ?

Je fais appel à l'improvisation, à l'imagination. Parfois, j'utilise la mémoire affective.

Induction d'émotions via la plateforme « Sound Teacher »

Qu'avez-vous pensé de l'expérience du Sound Teacher» ? Avez-vous des suggestions, critiques à nous faire ?

J'ai trouvé que l'expérience était bien. Parfois, je me sentais complètement perdue. Je voulais bien faire mais je ne comprenais pas bien ce qui se passait.

Pensez-vous que cette expérience soit un bon moyen de « piéger » les individus afin d'obtenir des émotions?

Oui, je me suis laissé complètement piégé. La mise en scène avec votre complice était très bien faite ! J'avais vraiment l'impression de participer à un test d'un logiciel. En plus, avec tout le matériel, le contexte du laboratoire, on ne s'attend pas du tout à une farce !

Avez-vous essayé de contrôler vos émotions durant cet exercice ?

Non mais j'ai essayé de rester concentrée. Au début, je me suis sentie un peu stressée par tout ce matériel, les gens etc. mais après, ça allait mieux.

Simulation d'émotions

Avez-vous des remarques, critiques sur cette expérience ?

Non, je me suis sentie à l'aise.

Quelles sont les émotions qui ont été les plus difficiles à exprimer ?

Il m'a paru parfois difficile de nuancer certaines émotions : joie/amusement par exemple. Le fait que l'on exprime des émotions sur des phrases isolées, relativement courtes ou même des mots et non sur un texte est parfois gênant pour exprimer des émotions de manière naturelle.

Avez-vous ressenti un quelconque stress durant cet exercice (du aux capteurs physiologiques par exemple) ?

Non, j'étais plus impressionnée par l'environnement : le matériel, le contexte du laboratoire…

Des suggestions ?

Non.

Grille d'auto-annotation de la locutrice F3

Temps	*Phase*	*Emotion*
En minutes	Introduction :	
	Phase 1 : les voyelles du français	Stress, intimidée - soucieuse de faire bien
0,73	- prononciation	Surprise - étonnement
5,38	- perception ouverture	Ennui (longueurs), agacée - énervée
7,86	- commentaires	Très surprise, flattée, amusée Amusée compte tenu du résultat
8,81	- perception position Av /Ar	déconcertée
9,61	- commentaires	idem
10,58	- perception arrondissement	idem
11,11	- commentaires	idem
11,6	- perception centralisation	rien
12,3	- commentaires	Surprise du résulta et un peu saoulée, long
12,81	- résultats généraux	idem
	Phase 2 : les voyelles proches de celles du français	
13,61	- perception ouverture	Agacée - fatiguée
14,38	- commentaires	Concentrée - attentive
14,98	- perception position Av /Ar	Agacée, long, j'en ai marre !
15,45	- commentaires	idem

15,86	- perception arrondissement	idem
16,5	- commentaires	idem
16,98	- perception centralisation	Agacée - fatiguée,
17,48	- commentaires	Perturbée, perdue
17,96	- prononciation	surprise
18,16	- résultats généraux	Très flattée et contente
	Phase 3 : généralisation aux langues du monde	
20,3	- perception ouverture	Déconcertée - perdue
21,03	- commentaires	Stressée, déçue
21,43	- perception position Av /Ar	saoulée
22,06	- commentaires	idem
22,28	- perception arrondissement	idem
22,71	- commentaires	idem
23,33	- perception centralisation	idem
24,63	- commentaires	idem
25,2	- prononciation	idem
25,6	- Warning	Mal à l'aise
	Phase 4 : Vérification perceptive	
26,73	- perception ouverture	agacée
27,5	- commentaires	idem
28,06	- perception position Av /Ar	Fatiguée mais toujours concentrée
28,8	- commentaires	idem

29,35	- perception arrondissement	déçu
29,8	- commentaires	ennui
30,26	- perception centralisation	Idem, mal à l'aise
30,51	- commentaires	Idem, incertitude
30,9	- Warning perceptif	idem
31,23	- prononciation	idem
32,2	- Warning production	idem
32,71	- Fin	idem

Commentaires supplémentaires :

Fiche d'informations du locuteur M1

Informations générales

Age : 21 ans

Lieu de résidence : Grenoble

Langue maternelle : Français

Depuis combien d'années faites-vous du théâtre ?

Faites vous partie d'une troupe / école de théâtre ?

Fait partie de la Ligue d'Improvisation de Grenoble.

De manière générale, avez-vous une méthode pour exprimer des émotions ? Si oui laquelle ?

Je fais appel à l'imagination et la mémoire émotive surtout pour me replonger dans un état émotionnel précis.

Induction d'émotions via la plateforme « Sound Teacher »

Qu'avez-vous pensé de l'expérience du « Sound Teacher» ? Avez-vous des suggestions, critiques à nous faire ?

Non, mais j'ai trouvé ça très long et je commençais à perdre patience.

Pensez-vous que cette expérience soit un bon moyen de « piéger » les individus afin d'obtenir des émotions?

Oui, je pense que c'est une bonne technique. On est dans un laboratoire, l'expérience est présentée de manière sérieuse. Moi, je n'ai pas senti que c'était une blague !

Avez-vous essayé de contrôler vos émotions durant cet exercice ?

Non pas du tout, j'étais vraiment dans le logiciel. Je suis resté concentré même si des fois, j'en avais vraiment marre !

Simulation d'émotions

Avez-vous des remarques, critiques sur cette expérience ?

Non

Quelles sont les émotions qui ont été les plus difficiles à exprimer ?

Je crois que j'ai fais un amalgame entre stress et peur, c'était vraiment difficile de distinguer les deux quand on n'a pas de contexte autour qui nous aide. J'ai essayé de rester naturel mais je pense que certaines émotions sont stéréotypées : la peur par exemple.

Avez-vous ressenti un quelconque stress durant cet exercice (du aux capteurs physiologiques par exemple) ?

Non mais au début j'étais un peu impressionné par le contexte. Ensuite, j'ai trouvé mes marques et ça allait mieux.

Des suggestions ?

Non

Grille d'auto-annotation du locuteur M1

Temps	*Phase*	*Emotion*
En minutes	Introduction :	
1,1	**Phase 1 : les voyelles du français**	Anxiété avant de commencer, soucieux de bien faire
3,15	- prononciation	concentration
5,21	- perception ouverture	idem
7,38	- commentaires	anxiété
7,9	- perception position Av /Ar	Ennui - concentration
9,13	- commentaires	idem
9,63	- perception arrondissement	idem
10,6	- commentaires	ennui
11,01	- perception centralisation	Concentration, agacement
11,6	- commentaires	Ennui, concentration
12,11	- résultats généraux	Contente, surprise
	Phase 2 : les voyelles proches de celles du français	
13,05	- perception ouverture	concentrée
14,41	- commentaires	Ennui - concentré
14,8	- perception position Av /Ar	ennui
15,58	- commentaires	idem
15,81	- perception arrondissement	idem
16,88	- commentaires	idem

17,06	- perception centralisation	Concentration, ennui
18,03	- commentaires	idem
18,55	- prononciation	Ennui, incompréhension
20,03	- résultats généraux	surpris - content,
	Phase 3 : généralisation aux langues du monde	
22,01	- perception ouverture	Concentration, ennui
23,2	- commentaires	Surprise, nervosité
23,73	- perception position Av /Ar	Concentration puis ennui
24,51	- commentaires	Surpris, agacé
24,81	- perception arrondissement	concentré
25,61	- commentaires	Agacé, ennui, remise en cause du logiciel
26,08	- perception centralisation	Ennui, concentration
27,18	- commentaires	agacé
27,63	- prononciation	concentration
28,53	- Warning	Agacement puis amusement
	Phase 4 : Vérification perceptive	
29,55	- perception ouverture	Ennui, impatient de finir,
30,96	- commentaires	Idem, agacé
31,46	- perception position Av /Ar	idem
32,38	- commentaires	Agacé, impatient de finir
32,71	- perception arrondissement	ennui
33,48	- commentaires	Ennui, impatience de finir

33,65	- perception centralisation	Idem, agacé
34,25	- commentaires	Agacé, amusement
34,41	- Warning perceptif	impatient
35,08	- prononciation	Amusement, agacement
36,1	- Warning production	Amusement agacement
36,35	- Fin	Mon incompréhension est justifiée !

Commentaires supplémentaires :

Fiche d'informations du locuteur M2

Informations générales

Age :

Lieu de résidence : Grenoble

Langue maternelle : Français

Depuis combien d'années faites-vous du théâtre ?

Depuis 9 ans.

Faites vous partie d'une troupe / école de théâtre ?

Oui, je fais partie d'une compagnie de théâtre amateur.

De manière générale, avez-vous une méthode pour exprimer des émotions ? Si oui laquelle ?

Induction d'émotions via la plateforme « Sound Teacher»

Qu'avez-vous pensé de l'expérience du « Sound Teacher» ? Avez-vous des suggestions, critiques à nous faire ?

J'ai été très impressionné par la mise en scène, je n'ai pas vu venir le piège. Pour ce qui est des critiques, j'ai trouvé qu'il faisait chaud et c'était un peu long mais sinon, j'ai trouvé ça très bien.

Pensez-vous que cette expérience soit un bon moyen de « piéger » les individus afin d'obtenir des émotions?

Oui je pense, enfin pour moi, ça a bien marché, j'étais à fond dedans.

Avez-vous essayé de contrôler vos émotions durant cet exercice ?

Non

Simulation d'émotions

Avez-vous des remarques, critiques sur cette expérience ?

Non juste que ce n'était pas évident de produire toutes ces émotions sur des phrases aussi courtes.

Quelles sont les émotions qui ont été les plus difficiles à exprimer ?

La tristesse, ça n'a pas été évident. Je pense que je l'ai sur-joué. J'ai eu du mal à distinguer amusement et joie. Je crois que pour l'amusement, j'ai trop forcé sur le rire.

Avez-vous ressenti un quelconque stress durant cet exercice (du aux capteurs physiologiques par exemple) ?

Non

Des suggestions ?

Non.

Grille d'auto-annotation du locuteur M2

Temps	*Phase*	*Emotion*
En minutes	Introduction :	
	Phase 1 : les voyelles du français	Un peu perdu, essaie de comprendre, concentré
3,21	- prononciation	Pas sûr de ce qu'il faut faire, perdu - surpris du résultat
5,2	- perception ouverture	Un peu perdu, commence à comprendre, plus confiant
7,96	- commentaires	Encore perdu, pas sûr de moi
9,91	- perception position Av /Ar	Content de comprendre, d'y arriver
11,46	- commentaires	Content du résultat, amusement
12,43	- perception arrondissement	Pas sûr de comprendre,
14,21	- commentaires	Etonné du résultat - content
14,88	- perception centralisation	Pas sûr de comprendre
16,26	- commentaires	Etonné - content
17,13	- résultats généraux	Très étonné des résultats, pensait avoir moins
19,95	**Phase 2 : les voyelles proches de celles du français**	
20,68	- perception ouverture	confiant
22,35	- commentaires	Content, pense avoir compris
24,26	- perception position Av /Ar	Concentré, essaie

		d'améliorer mes résultats
24,81	- commentaires	Un peu étonné que mes résultats baissent
26,16	- perception arrondissement	Fait du hasard ou n'ai-je pas compris ?
27,31	- commentaires	
29,48	- perception centralisation	rien
30,51	- commentaires	
32,01	- prononciation	
32,93	- résultats généraux	Content, confiant, détendu mais reste étonné
	Phase 3 : généralisation aux langues du monde	
	- perception ouverture	Rien, tout se passe bien
	- commentaires	Grosse surprise du faible résultat
	- perception position Av /Ar	concentré
	- commentaires	Très surpris de nouveau, ne comprend pas- perdu
	- perception arrondissement	Las, Concentré, essaie de comprendre
	- commentaires	idem
	- perception centralisation	idem
	- commentaires	Etonné des résultats, pensait avoir compris
	- prononciation	concentré
	- Warning	Déçu - étonné des résultats, comprend pas, résigné, las

	Phase 4 : Vérification perceptive	Essaie de nouveau de comprendre et me concentrer
	- Perception ouverture	concentré
	- Commentaires	Incompréhension totale – perdu, agacé
	- perception position Av /Ar	Anxieux - un peu peur
	- Commentaires	Inquiet des résultats
	- Perception arrondissement	Essaie de nouveau de comprendre
	- Commentaires	Incompréhension
	- Perception centralisation	Concentré, un peu inquiet
	- Commentaires	Un peu perdu, ne comprend pas,
	- Warning perceptif	Idem, un peu peur mais toujours concentré
	- Prononciation	Concentré, essaie de trouver des raisonnements différents pour m'en sortir
	- Warning production	Déçu, comprend pas, fatalité ?
	- Fin	

Commentaires supplémentaires :

Fiche d'informations du locuteur M3

Informations générales

Age : 21 ans

Lieu de résidence : Grenoble

Langue maternelle : Français

Depuis combien d'années faites-vous du théâtre ?

Depuis 9 ans.

Faites vous partie d'une troupe / école de théâtre ?

Fait partie de la Ligue d'Improvisation de Grenoble.

De manière générale, avez-vous une méthode pour exprimer des émotions ? Si oui laquelle ?

Je fais appel à la mémoire pour exprimer certaines émotions comme la tristesse. Je joue aussi sur le mimétisme, l'imitation.

Induction d'émotions via la plateforme « Sound Teacher »

Qu'avez-vous pensé de l'expérience du « Sound Teacher» ? Avez-vous des suggestions, critiques à nous faire ?

J'ai trouvé que c'était une bonne expérience, un peu trop longue surtout qu'il faisait chaud. Pas d'autres commentaires

Pensez-vous que cette expérience soit un bon moyen de « piéger » les individus afin d'obtenir des émotions?

Oui, je me suis laissé complètement piégé. J'étais intéressé pour participer au test d'un futur produit technologique et en même temps impressionné par le matériel, le laboratoire… je n'ai pas du tout pensé à un piège.

Avez-vous essayé de contrôler vos émotions durant cet exercice ?

Non mais j'ai essayé de rester concentré tout le long.

Simulation d'émotions

Avez-vous des remarques, critiques sur cette expérience ?

Je me suis senti un peu stressé : le fait d'être assis sans pouvoir bouger dans une toute petite pièce, la chaleur. J'ai trouvé que les phrases étaient trop courtes, c'était difficile de lancer l'émotion.

Quelles sont les émotions qui ont été les plus difficiles à exprimer ?

La flatterie a été très difficile à exprimer ainsi que la 'tristesse' car les phrases données n'étaient pas porteuses de l'émotion. J'ai essayé de faire au maximum mais je crois avoir surjoué certaines émotions comme la 'tristesse'.

Avez-vous ressenti un quelconque stress durant cet exercice (dû aux capteurs physiologiques par exemple) ?

Non

Des suggestions ?

Au début, je me sentais stressé car j'étais confiné dans une petite pièce, le fait d'être filmé aussi, de ne pas pouvoir bouger. Ensuite, je me suis détendu.

Grille d'auto-annotation du locuteur M3

Temps	*Phase*	*Emotion*
En minutes	Introduction :	
0 ,87	**Phase 1 : les voyelles du français**	Concentré - perplexe
3,01	- prononciation	
7,85	- perception ouverture	Perplexe - concentré, inquiet, perplexe pas sûr de lui
10,98	- commentaires	
11,75	- perception position Av /Ar	Concentré, las
13,05	- commentaires	
13,65	- perception arrondissement	
14,38	- commentaires	Surpris
14,7	- perception centralisation	Incertain
15,31	- commentaires	Inquiet
16,11	- résultats généraux	Surpris - flatté
	Phase 2 : les voyelles proches de celles du français	
16,9	- perception ouverture	Inquiet, dubitatif
18,98	- commentaires	Déçu
19,78	- perception position Av /Ar	
20,4	- commentaires	Etonné
20,63	- perception arrondissement	Exaspéré
21,7	- commentaires	Déçu

22,15	- perception centralisation	las - exaspéré
23,25	- commentaires	Las
23,81	- prononciation	Déconcerté
25,78	- résultats généraux	Surpris - flatté
	Phase 3 : généralisation aux langues du monde	
27,61	- perception ouverture	Hésitant
29,78	- commentaires	Lassitude
30,4	- perception position Av /Ar	Déception
31,4	- commentaires	Idem, surprise
31,95	- perception arrondissement	Sarcastique, déception
33	- commentaires	Idem
33,33	- perception centralisation	Exaspération - perdu
35,73	- commentaires	Exaspération
36,31	- prononciation	Grosse exaspération
37,2	- Warning	Exaspéré -inquiet
	Phase 4 : Vérification perceptive	
39,05	- perception ouverture	Exaspération, relent de concentration
40,65	- commentaires	Inquiet, soucieux
41,5	- perception position Av /Ar	Déconcentré
42,38	- commentaires	Exaspéré
42,51	- perception arrondissement	Perplexe, énervé
43,26	- commentaires	Exaspéré, inquiet

43,6	- perception centralisation	Exaspéré
44,21	- commentaires	Rassuré un peu, inquiet beaucoup
44,6	- Warning perceptif	Inquiet, amusé
45,45	- prononciation	Exaspéré
46,38	- Warning production	Rassuré
46,63	- Fin	Idem, perplexe

Commentaires supplémentaires :

Annexe III : Projection dimensionnelle des contours de F0 et d'intensité

Les données présentées dans cette annexe correspondent à l'évaluation perceptive des stimuli resynthétisés à l'aide de Praat.

Matrices de confusion brutes et après regroupement pour chaque condition de projection partielle

F0

	joie	satisfaction	joie/satisf.	anxiété	inquiétude	anx./inq	déception	résignation	tristesse	déc.-res.-tri.	dégoût	neutre
joie	**7.50%**	29.17%	**36.67%**	5.83%	14.17%	20.00%	1.67%	7.50%	2.50%	11.67%	1.67%	30.00%
satisfaction	71.67%	**21.67%**	**93.33%**	1.67%	1.67%	3.33%	0.83%	1.67%	0.83%	3.33%	0.00%	0.00%
anxiété	4.17%	8.33%	12.50%	**13.33%**	19.17%	**32.50%**	14.17%	18.33%	7.50%	40.00%	6.67%	8.33%
inquiétude	18.33%	56.67%	75.00%	2.50%	**7.50%**	**10.00%**	4.17%	1.67%	0.83%	6.67%	5.83%	2.50%
déception	0.00%	4.17%	4.17%	2.50%	10.00%	12.50%	**15.83%**	12.50%	15.00%	**43.33%**	5.00%	35.00%
résignation	0.83%	9.17%	10.00%	6.67%	16.67%	23.33%	2.50%	**9.17%**	13.33%	**25.00%**	2.50%	39.17%
tristesse	1.67%	5.00%	6.67%	14.17%	23.33%	37.50%	8.33%	5.00%	**29.17%**	**42.50%**	0.83%	12.50%
dégoût	1.67%	0.83%	2.50%	11.67%	12.50%	24.17%	15.00%	16.67%	31.67%	63.33%	**1.67%**	8.33%

Intensité

	joie	satisfaction	joie/satisf.	anxiété	inquiétude	anx./inq	déception	résignation	tristesse	déc.-res.-tri.	dégoût	neutre
joie	**0.00%**	5.83%	**5.83%**	2.50%	4.17%	6.67%	13.33%	16.67%	7.50%	37.50%	2.50%	47.50%
satisfaction	0.00%	**10.00%**	**10.00%**	5.83%	5.00%	10.83%	10.83%	25.83%	0.83%	37.50%	4.17%	37.50%
anxiété	0.83%	4.17%	5.00%	**3.33%**	8.33%	**11.67%**	20.00%	28.33%	10.00%	58.33%	8.33%	16.67%
inquiétude	0.83%	3.33%	4.17%	2.50%	**5.83%**	**8.33%**	21.67%	11.67%	14.17%	47.50%	4.17%	35.83%
déception	0.83%	1.67%	2.50%	0.00%	2.50%	2.50%	**14.17%**	9.17%	13.33%	**36.67%**	6.67%	51.67%
résignation	0.00%	5.83%	5.83%	1.67%	0.83%	2.50%	6.67%	**16.67%**	8.33%	**31.67%**	6.67%	53.33%
tristesse	0.00%	1.67%	1.67%	0.00%	0.83%	0.83%	19.17%	12.50%	**6.67%**	**38.33%**	5.00%	54.17%
dégoût	0.00%	4.17%	4.17%	2.50%	3.33%	5.83%	13.33%	15.00%	5.83%	34.17%	**6.67%**	49.17%

F0 et intensité

	joie	satisfaction	joie/satisf.	anxiété	inquiétude	anx./inq	déception	résignation	tristesse	déc.-res.-tri.	dégoût	neutre
joie	**10.00%**	32.50%	**42.50%**	5.83%	13.33%	19.17%	0.83%	6.67%	0.83%	8.33%	0.83%	29.17%
satisfaction	70.00%	**22.50%**	**92.50%**	1.67%	1.67%	3.33%	0.00%	2.50%	0.00%	2.50%	0.83%	0.83%
anxiété	1.67%	4.17%	5.83%	**18.33%**	17.50%	**35.83%**	17.50%	19.17%	5.00%	41.67%	4.17%	12.50%
inquiétude	25.00%	42.50%	67.50%	0.83%	**16.67%**	**17.50%**	0.83%	3.33%	0.83%	5.00%	5.00%	5.00%
déception	0.83%	3.33%	4.17%	5.00%	5.83%	10.83%	**6.67%**	15.00%	5.83%	**27.50%**	2.50%	55.00%
résignation	4.17%	15.83%	20.00%	10.00%	8.33%	18.33%	8.33%	**6.67%**	2.50%	**17.50%**	0.83%	43.33%
tristesse	9.17%	13.33%	22.50%	5.00%	13.33%	18.33%	7.50%	6.67%	**5.00%**	**19.17%**	0.83%	39.17%
dégoût	0.83%	5.00%	5.83%	5.00%	10.00%	15.00%	12.50%	19.17%	20.00%	51.67%	**3.33%**	24.17%

Qualité de voix et durée

	joie	satisfaction	joie/satisf.	anxiété	inquiétude	anx./inq	déception	résignation	tristesse	déc.-res.-tri.	dégoût	neutre
joie	**2.50%**	4.17%	**6.67%**	1.67%	4.17%	5.83%	11.67%	16.67%	7.50%	35.83%	11.67%	40.00%
satisfaction	30.00%	**12.50%**	**42.50%**	4.17%	9.17%	13.33%	3.33%	5.00%	0.83%	9.17%	3.33%	31.67%
anxiété	0.00%	0.83%	0.83%	**53.33%**	23.33%	**76.67%**	6.67%	10.83%	0.00%	17.50%	5.00%	0.00%
inquiétude	9.17%	8.33%	17.50%	15.83%	**19.17%**	**35.00%**	11.67%	10.00%	6.67%	28.33%	4.17%	15.00%
déception	0.00%	3.33%	3.33%	5.00%	5.83%	10.83%	**10.83%**	6.67%	32.50%	**50.00%**	6.67%	29.17%
résignation	2.50%	9.17%	11.67%	1.67%	2.50%	4.17%	5.00%	**19.17%**	1.67%	**25.83%**	8.33%	50.00%
tristesse	1.67%	1.67%	3.33%	6.67%	9.17%	15.83%	17.50%	5.00%	**43.33%**	**65.83%**	8.33%	6.67%
dégoût	0.00%	0.83%	0.83%	0.83%	0.00%	0.83%	10.83%	7.50%	45.83%	64.17%	**34.17%**	0.00%

Taux de reconnaissance par rapport au niveau du hasard

Tests t de Student bidirectionnels pour échantillons uniques, à 39 degrés de liberté. Chaque ligne évalue la significativité de la différence entre le taux moyen de reconnaissance après regroupement pour la condition de resynthèse et l'émotion correspondantes (colonne *Reconnaissance*), et le taux de reconnaissance au hasard fonction du nombre d'étiquettes dans le regroupement, indiqué dans la colonne *Hasard*. Les expressions neutres sur les énoncés [ʁuʒ] et [sabl], présentées en condition de contrôle seulement, sont indiquées respectivement comme *Neutre (R)* et *Neutre (S)*.

Colonne *Significativité* : les notations ***, **, * et n.s. correspondent respectivement aux niveaux de significativité p<.001, p<.01, p<.05 et non significatif.

Colonne *Direction* : les notations <, = et > correspondent respectivement à un taux de discrimination moyen significativement inférieur au hasard, non significativement différent du hasard et significativement supérieur au hasard.

Emotion	Condition	Reconnaissance	Hasard	t	p	Significativité	Direction
Joie	Intensité	6%	22%	2.384	0.0220	*	<
	QV+durée	7%	22%	2.635	0.0120	*	<
	F0	37%	22%	8.889	0.0000	***	>
	F0+intensité	43%	22%	8.129	0.0000	***	>
	Contrôle	71%	22%	14.270	0.0000	***	>
Satisfaction	Intensité	10%	22%	3.290	0.0020	**	<
	QV+durée	43%	22%	8.859	0.0000	***	>
	F0	93%	22%	31.331	0.0000	***	>
	F0+intensité	93%	22%	33.005	0.0000	***	>
	Contrôle[34]	100%	22%				>
Anxiété	Intensité	12%	22%	3.747	0.0010	***	<
	QV+durée	77%	22%	14.617	0.0000	***	>
	F0	33%	22%	6.290	0.0000	***	>
	F0+intensité	36%	22%	6.306	0.0000	***	>
	Contrôle	78%	22%	15.099	0.0000	***	>
Inquiétude	Intensité	8%	22%	2.616	0.0130	*	<
	QV+durée	35%	22%	7.078	0.0000	***	>
	F0	10%	22%	3.053	0.0040	**	<
	F0+intensité	18%	22%	3.739	0.0010	***	<
	Contrôle	42%	22%	6.592	0.0000	***	>

[34] Du fait de l'écart-type nul la reconnaissance de l'expression de la satisfaction en condition de contrôle, reconnue comme joie ou satisfaction par 100% des auditeurs, ne peut être comparée au niveau du hasard par un test t. Toutefois il ne fait aucun doute que la probabilité d'erreur en affirmant que ce taux de reconnaissance est supérieur au hasard est nulle.

Déception	Intensité	37%	33%	6.841	0.0000	***	>
	QV+durée	50%	33%	8.872	0.0000	***	>
	F0	43%	33%	1.966	0.0570	n.s.	=
	F0+intensité	28%	33%	6.340	0.0000	***	<
	Contrôle	80%	33%	15.395	0.0000	***	>
Résignation	Intensité	32%	33%	6.790	0.0000	***	<
	QV+durée	26%	33%	5.613	0.0000	***	<
	F0	25%	33%	6.303	0.0000	***	<
	F0+intensité	18%	33%	5.089	0.0000	***	<
	Contrôle	43%	33%	6.678	0.0000	***	>
Tristesse	Intensité	38%	33%	8.347	0.0000	***	>
	QV+durée	66%	33%	13.516	0.0000	***	>
	F0	43%	33%	7.068	0.0000	***	>
	F0+intensité	19%	33%	4.577	0.0000	***	<
	Contrôle	44%	33%	7.289	0.0000	***	>
Dégoût	Intensité	7%	11%	2.206	0.0330	*	<
	QV+durée	34%	11%	5.888	0.0000	***	>
	F0	2%	11%	1.337	0.1890	n.s.	=
	F0+intensité	3%	11%	1.385	0.1740	n.s.	=
	Contrôle	62%	11%	9.040	0.0000	***	>
Neutre (R)	Contrôle	25%	11%	3.030	0.0043	**	>
Neutre (S)	Contrôle	38%	11%	4.603	0.0000	***	>

Annexe IV : Projection dimensionnelle à partir d'une modélisation ARX-LF

Matrices de confusion brutes et après regroupement pour chaque condition de projection partielle

F0 et intensité

	joie	anxiété	inquiétude	anx./inq.	déception	résignation	tristesse	déc./rés./tri.	dégoût	neutre
joie	**59%**	7%	7%	13%	0%	4.00%	1.33%	5%	0.00%	22.67%
anxiété	15%	**25%**	21%	**47%**	4%	8.00%	4.00%	16%	5.33%	17.33%
inquiétude	60%	9%	**25%**	**35%**	4%	0.00%	0.00%	4%	1.33%	0.00%
déception	5%	9%	11%	20%	**12%**	9.33%	4.00%	**25%**	2.67%	46.67%
résignation	32%	11%	17%	28%	8%	**1.33%**	2.67%	**12%**	1.33%	26.67%
tristesse	25%	9%	13%	23%	1%	4.00%	**38.67%**	**44%**	0.00%	8.00%
dégoût	4%	5%	4%	9%	19%	14.67%	22.67%	56%	**1.33%**	29.33%

Durée seule

	joie	anxiété	inquiétude	anx./inq.	déception	résignation	tristesse	rés./déc./tri.	dégoût	neutre
joie	**0.00%**	5.33%	4.00%	9%	33.33%	12.00%	9.33%	55%	8.00%	28.00%
anxiété	1.33%	**56.00%**	18.67%	**75%**	1.33%	5.33%	4.00%	11%	0.00%	13.33%
inquiétude	4.00%	5.33%	**12.00%**	**17%**	17.33%	22.67%	6.67%	47%	8.00%	24.00%
déception	0.00%	0.00%	0.00%	0%	**26.67%**	25.33%	5.33%	**57%**	38.67%	4.00%
résignation	5.33%	12.00%	13.33%	25%	18.67%	**8.00%**	13.33%	**40%**	6.67%	22.67%
tristesse	10.67%	4.00%	6.67%	11%	13.33%	6.67%	**53.33%**	**73%**	2.67%	2.67%
dégoût	0.00%	0.00%	0.00%	0%	10.67%	26.67%	13.33%	51%	**49.33%**	0.00%

Qualité de voix

	joie	anxiété	inquiétude	anx./inq.	déception	résignation	tristesse	rés./déc./tri.	dégoût	neutre
joie	**10.67%**	14.67%	6.67%	21%	16.00%	10.67%	5.33%	32%	2.67%	33.33%
anxiété	5.33%	**36.00%**	18.67%	**55%**	6.67%	10.67%	10.67%	28%	2.67%	9.33%
inquiétude	20.00%	12.00%	**25.33%**	**37%**	12.00%	18.67%	4.00%	35%	1.33%	6.67%
déception	0.00%	2.67%	2.67%	5%	**18.67%**	25.33%	5.33%	**49%**	5.33%	40.00%
résignation	1.33%	9.33%	6.67%	16%	17.33%	**16.00%**	0.00%	**33%**	6.67%	42.67%
tristesse	25.33%	10.67%	10.67%	21%	6.67%	8.00%	**36.00%**	**51%**	2.67%	0.00%
dégoût	1.33%	13.33%	2.67%	16%	8.00%	13.33%	2.67%	24%	**8.00%**	50.67%

Qualité de voix et durée

	joie	anxiété	inquiétude	anx./inq.	déception	résignation	tristesse	rés./déc./tri.	dégoût	neutre
joie	**30.67%**	2.67%	10.67%	13%	6.67%	12.00%	10.67%	29%	10.67%	16.00%
anxiété	0.00%	**56.00%**	26.67%	**83%**	1.33%	5.33%	8.00%	15%	2.67%	0.00%
inquiétude	13.33%	13.33%	**24.00%**	**37%**	12.00%	13.33%	9.33%	35%	6.67%	8.00%
déception	0.00%	0.00%	0.00%	0%	**25.33%**	53.33%	4.00%	**83%**	17.33%	0.00%
résignation	9.33%	30.67%	24.00%	55%	8.00%	**10.67%**	1.33%	**20%**	5.33%	10.67%
tristesse	20.00%	10.67%	12.00%	23%	9.33%	4.00%	**42.67%**	**56%**	1.33%	0.00%
dégoût	0.00%	0.00%	0.00%	0%	9.33%	29.33%	17.33%	56%	**42.67%**	1.33%

Source seule

	joie	anxiété	inquiétude	anx./inq.	déception	résignation	tristesse	rés./déc./tri.	dégoût	neutre
joie	**4.00%**	14.67%	14.67%	29%	10.67%	5.33%	20.00%	36%	2.67%	28.00%
anxiété	13.33%	**14.67%**	10.67%	**25%**	10.67%	13.33%	9.33%	33%	5.33%	22.67%
inquiétude	8.00%	14.67%	**8.00%**	**23%**	14.67%	8.00%	4.00%	27%	4.00%	38.67%
déception	0.00%	1.33%	2.67%	4%	**17.33%**	8.00%	2.67%	**28%**	2.67%	65.33%
résignation	1.33%	5.33%	6.67%	12%	13.33%	**16.00%**	24.00%	**53%**	1.33%	32.00%
tristesse	2.67%	8.00%	12.00%	20%	12.00%	12.00%	**20.00%**	**44%**	5.33%	28.00%
dégoût	1.33%	10.67%	10.67%	21%	18.67%	6.67%	24.00%	49%	**1.33%**	26.67%

Source et résidu

	joie	anxiété	inquiétude	anx./inq.	déception	résignation	tristesse	rés./déc./tri.	dégoût	neutre
joie	**4.00%**	8.00%	6.67%	15%	18.67%	9.33%	14.67%	43%	2.67%	36.00%
anxiété	4.00%	**16.00%**	21.33%	**37%**	8.00%	16.00%	6.67%	31%	4.00%	24.00%
inquiétude	9.33%	21.33%	**12.00%**	**33%**	12.00%	14.67%	6.67%	33%	2.67%	21.33%
déception	2.67%	4.00%	4.00%	8%	**22.67%**	12.00%	10.67%	**45%**	1.33%	42.67%
résignation	1.33%	4.00%	8.00%	12%	13.33%	**9.33%**	17.33%	**40%**	4.00%	42.67%
tristesse	6.67%	6.67%	13.33%	20%	4.00%	12.00%	**30.67%**	**47%**	4.00%	22.67%
dégoût	6.67%	9.33%	5.33%	15%	22.67%	6.67%	8.00%	37%	**5.33%**	36.00%

Confusions des expressions de tristesse présentées dans différentes conditions

	joie	anxiété	inquiétude	anx./inq.	déception	résignation	tristesse	rés./déc./tri.	dégoût	neutre
contrôle	22.67%	12.00%	9.33%	21%	10.67%	2.67%	**42.67%**	**56.00%**	0.00%	0.00%
durée	10.67%	4.00%	6.67%	11%	13.33%	6.67%	**53.33%**	**73%**	2.67%	2.67%
qv+durée	20.00%	10.67%	12.00%	23%	9.33%	4.00%	**42.67%**	**56%**	1.33%	0.00%
qv	25.33%	10.67%	10.67%	21%	6.67%	8.00%	**36.00%**	**51%**	2.67%	0.00%
src+res	6.67%	6.67%	13.33%	20%	4.00%	12.00%	**30.67%**	**47%**	4.00%	22.67%
src	2.67%	8.00%	12.00%	20%	12.00%	12.00%	**20.00%**	**44%**	5.33%	28.00%
f0+int	25.33%	9.33%	13.33%	23%	1.33%	4.00%	**38.67%**	**44%**	0.00%	8.00%

Taux de reconnaissance par rapport au niveau du hasard

Tests t de Student bidirectionnels pour échantillons uniques, à 24 degrés de liberté. Chaque ligne évalue la significativité de la différence entre le taux moyen de reconnaissance après regroupement pour la condition de resynthèse et l'émotion correspondantes (colonne *Reconnaissance*), et niveau théorique du hasard fonction du nombre d'étiquettes dans le regroupement, indiqué dans la colonne *Hasard*.

Les expressions neutres sur les énoncés [ʁuʒ] et [sabl], présentées en condition de contrôle seulement, sont indiquées respectivement comme *Neutre (R)* et *Neutre (S)*.

Colonne *Significativité* : les notations ***, **, * et n.s. correspondent respectivement aux niveaux de significativité p<0.001, p<0.01, p<0.05 et non significatif.

Colonne *Direction* : les notations <, = et > correspondent respectivement à un taux de discrimination moyen significativement inférieur au hasard, non significativement différent du hasard et significativement supérieur au hasard.

Emotion	Condition	Reconnaissance	Hasard	t	p	Significativité	Direction
Joie	Durée seule[35]	0%	13%				<
	Source seule	4%	13%	-3.844	0.0008	***	<
	Source+résidu	4%	13%	-2.9	0.0079	**	<
	Qualité de voix	11%	13%	-0.494	0.6259	n.s.	=
	QV et durée	31%	13%	2.53	0.0184	*	>
	F0+intensité	59%	13%	5.771	0.0000	***	>
	Contrôle	77%	13%	10.289	0.0000	***	>
Anxiété	Durée seule	75%	25%	8.049	0.0000	***	>
	Source seule	25%	25%	0.052	0.9593	n.s.	=
	Source+résidu	37%	25%	1.694	0.1033	n.s.	=
	Qualité de voix	55%	25%	3.646	0.0013	**	>
	QV et durée	83%	25%	9.419	0.0000	***	>
	F0+intensité	47%	25%	3.122	0.0046	**	>
	Contrôle	91%	25%	11.69	0.0000	***	>
Inquiétude	Durée seule	17%	25%	-1.61	0.1204	n.s.	=
	Source seule	23%	25%	-0.507	0.6168	n.s.	=
	Source+résidu	33%	25%	1.201	0.2415	n.s.	=
	Qualité de voix	37%	25%	1.586	0.1257	n.s.	=
	QV et durée	37%	25%	1.826	0.0803	n.s.	=
	F0+intensité	35%	25%	1.277	0.2139	n.s.	=
	Contrôle	44%	25%	2.118	0.0447	*	>
Déception	Durée seule	57%	38%	2.913	0.0076	**	>
	Source seule	28%	38%	-1.587	0.1257	n.s.	=
	Source+résidu	45%	38%	1.366	0.1846	n.s.	=

[35] Aucun auditeur n'a reconnu l'expression de joie resynthétisée en condition durée seule, d'où l'écart-type nul ne permettant pas une comparaison au niveau du hasard par un test t.

	Qualité de voix	49%	38%	1.581	0.1269	n.s.	=
	QV et durée	83%	38%	8.236	0.0000	***	>
	F0+intensité	25%	38%	-2.523	0.0187	*	<
	Contrôle	84%	38%	9.767	0.0000	***	>
Résignation	Durée seule	40%	38%	0.347	0.7315	n.s.	=
	Source seule	53%	38%	2.602	0.0156	*	>
	Source+résidu	40%	38%	0.411	0.6849	n.s.	=
	Qualité de voix	33%	38%	-0.653	0.5201	n.s.	=
	QV et durée	20%	38%	-2.742	0.0114	*	<
	F0+intensité	12%	38%	-5.998	0.0000	***	<
	Contrôle	28%	38%	-1.335	0.1945	n.s.	=
Tristesse	Durée seule	73%	38%	5.164	0.0000	***	>
	Source seule	44%	38%	1.032	0.3126	n.s.	=
	Source+résidu	47%	38%	1.436	0.1639	n.s.	=
	Qualité de voix	51%	38%	1.607	0.1211	n.s.	=
	QV et durée	56%	38%	2.351	0.0273	*	>
	F0+intensité	44%	38%	0.802	0.4302	n.s.	=
	Contrôle	56%	38%	2.351	0.0273	*	>
Dégoût	Durée seule	49%	13%	4.767	0.0001	***	>
	Source seule	1%	13%	-8.375	0.0000	***	<
	Source+résidu	5%	13%	-2.873	0.0084	**	<
	Qualité de voix	8%	13%	-1.291	0.2090	n.s.	=
	QV et durée	43%	13%	3.857	0.0008	***	>
	F0+intensité	1%	13%	-8.375	0.0000	***	<
	Contrôle	71%	13%	7.061	0.0000	***	>
Neutre (R)	Contrôle	44%	13%	4.310	0.0000	***	>
Neutre (S)	Contrôle	61%	13%	7.985	0.0000	***	>

Annexe V : Expérience de dévoilement progressif

Taux de reconnaissance par rapport au niveau du hasard

Tests t de Student bidirectionnels pour échantillons uniques, à 19 degrés de liberté. Chaque ligne évalue la significativité de la différence entre le taux moyen de reconnaissance après regroupement pour l'émotion, l'énoncé et la fenêtre de dévoilement correspondantes (colonne *Reconnaissance*) d'une part, et le niveau théorique du hasard fonction du nombre d'étiquettes dans le regroupement, indiqué dans la colonne *Hasard* d'autre part.

Colonne *Significativité* : les notations ***, **, * et n.s. correspondent respectivement aux niveaux de significativité $p<0.001$, $p<0.01$, $p<0.05$ et non significatif.

Colonne *Direction* : les notations <, = et > correspondent respectivement à un taux de discrimination moyen significativement inférieur au hasard, non significativement différent du hasard et significativement supérieur au hasard.

Emotion	Enoncé	Fenêtre	Reconnaissance	Hasard	t	p	Sig.	Dir.
joie	[ʒon]	1	70%	22%	4.5	0.0002	***	>
		2	55%	22%	2.9	0.0098	**	>
		3	65%	22%	3.9	0.0009	***	>
		4	60%	22%	3.4	0.0033	**	>
		5	75%	22%	5.3	0.0000	***	>
		6	70%	22%	4.5	0.0002	***	>
	[ʁuʒ]	1	30%	22%	0.7	0.4683	n.s.	=
		2	20%	22%	-0.2	0.8114	n.s.	=
		3	35%	22%	1.2	0.2573	n.s.	=
		4	55%	22%	2.9	0.0098	**	>
		5	65%	22%	3.9	0.0009	***	>
		6	75%	22%	5.3	0.0000	***	>
	[vɛʁ]	1	25%	22%	0.3	0.7826	n.s.	=
		2	25%	22%	0.3	0.7826	n.s.	=
		3	55%	22%	2.9	0.0098	**	>
		4	30%	22%	0.7	0.4683	n.s.	=
		5	40%	22%	1.6	0.1302	n.s.	=
		6	35%	22%	1.2	0.2573	n.s.	=

satisfaction	[ʒon]	1	90%	22%	9.8	0.0000	***	>
		2	95%	22%	14.6	0.0000	***	>
		3	90%	22%	9.8	0.0000	***	>
		4	95%	22%	14.6	0.0000	***	>
		5	100%[36]	22%				>
		6	95%	22%	14.6	0.0000	***	>
	[ʁuʒ]	1	75%	22%	5.3	0.0000	***	>
		2	75%	22%	5.3	0.0000	***	>
		3	95%	22%	14.6	0.0000	***	>
		4	100%	22%				>
		5	100%	22%				>
		6	100%	22%				>
	[vɛʁ]	1	40%	22%	1.6	0.1302	n.s.	=
		2	80%	22%	6.3	0.0000	***	>
		3	95%	22%	14.6	0.0000	***	>
		4	100%	22%				>
		5	100%	22%				>
		6	100%	22%				>
anxiété	[ʒon]	1	30%	22%	0.7	0.4683	n.s.	=
		2	20%	22%	-0.2	0.8114	n.s.	=
		3	45%	22%	2.0	0.0605	n.s.	=
		4	50%	22%	2.4	0.0256	*	>
		5	60%	22%	3.4	0.0033	**	>
		6	70%	22%	4.5	0.0002	***	>
	[ʁuʒ]	1	40%	22%	1.6	0.1302	n.s.	=
		2	20%	22%	-0.2	0.8114	n.s.	=
		3	20%	22%	-0.2	0.8114	n.s.	=
		4	35%	22%	1.2	0.2573	n.s.	=
		5	60%	22%	3.4	0.0033	**	>
		6	70%	22%	4.5	0.0002	***	>
	[vɛʁ]	1	20%	22%	-0.2	0.8114	n.s.	=
		2	40%	22%	1.6	0.1302	n.s.	=
		3	55%	22%	2.9	0.0098	**	>
		4	65%	22%	3.9	0.0009	***	>
		5	80%	22%	6.3	0.0000	***	>
		6	80%	22%	6.3	0.0000	***	>

[36] Les expressions de satisfaction reconnues par l'ensemble des auditeurs ont pour résultat un écart-type nul, ne permettant pas une comparaison au niveau du hasard au moyen d'un test t. Il ne fait cependant aucun doute que les taux de reconnaissance dans ces cas de figures sont très significativement supérieurs au hasard.

	[ʒon]	1	25%	22%	0.3	0.7826	n.s.	=
		2	10%	22%	-1.8	0.0918	n.s.	=
		3	35%	22%	1.2	0.2573	n.s.	=
		4	40%	22%	1.6	0.1302	n.s.	=
		5	50%	22%	2.4	0.0256	*	>
		6	45%	22%	2.0	0.0605	n.s.	=
inquiétude	[ʁuʒ]	1	40%	22%	1.6	0.1302	n.s.	=
		2	55%	22%	2.9	0.0098	**	>
		3	75%	22%	5.3	0.0000	***	>
		4	70%	22%	4.5	0.0002	***	>
		5	80%	22%	6.3	0.0000	***	>
		6	75%	22%	5.3	0.0000	***	>
	[vɛʁ]	1	55%	22%	2.9	0.0098	**	>
		2	45%	22%	2.0	0.0605	n.s.	=
		3	55%	22%	2.9	0.0098	**	>
		4	65%	22%	3.9	0.0009	***	>
		5	80%	22%	6.3	0.0000	***	>
		6	75%	22%	5.3	0.0000	***	>
déception	[ʒon]	1	40%	33%	0.6	0.5599	n.s.	=
		2	75%	33%	4.2	0.0005	***	>
		3	80%	33%	5.1	0.0001	***	>
		4	85%	33%	6.3	0.0000	***	>
		5	75%	33%	4.2	0.0005	***	>
		6	85%	33%	6.3	0.0000	***	>
	[ʁuʒ]	1	80%	33%	5.1	0.0001	***	>
		2	70%	33%	3.5	0.0025	**	>
		3	75%	33%	4.2	0.0005	***	>
		4	80%	33%	5.1	0.0001	***	>
		5	90%	33%	8.2	0.0000	***	>
		6	95%	33%	12.3	0.0000	***	>
	[vɛʁ]	1	75%	33%	4.2	0.0005	***	>
		2	45%	33%	1.0	0.3194	n.s.	=
		3	50%	33%	1.5	0.1625	n.s.	=
		4	85%	33%	6.3	0.0000	***	>
		5	80%	33%	5.1	0.0001	***	>
		6	85%	33%	6.3	0.0000	***	>
résignation	[ʒon]	1	30%	33%	-0.3	0.7549	n.s.	=
		2	20%	33%	-1.5	0.1626	n.s.	=
		3	25%	33%	-0.8	0.4122	n.s.	=
		4	25%	33%	-0.8	0.4122	n.s.	=
		5	25%	33%	-0.8	0.4122	n.s.	=

		6	25%	33%	-0.8	0.4122	n.s.	=
	[ʁuʒ]	1	10%	33%	-3.4	0.0031	**	<
		2	35%	33%	0.2	0.8803	n.s.	=
		3	15%	33%	-2.2	0.0374	*	<
		4	40%	33%	0.6	0.5599	n.s.	=
		5	45%	33%	1.0	0.3194	n.s.	=
		6	40%	33%	0.6	0.5599	n.s.	=
	[vɛʁ]	1	5%	33%	-5.7	0.0000	***	<
		2	35%	33%	0.2	0.8803	n.s.	=
		3	55%	33%	1.9	0.0729	n.s.	=
		4	50%	33%	1.5	0.1625	n.s.	=
		5	35%	33%	0.2	0.8803	n.s.	=
		6	40%	33%	0.6	0.5599	n.s.	=
tristesse	[ʒon]	1	5%	33%	-5.7	0.0000	***	<
		2	15%	33%	-2.2	0.0374	*	<
		3	20%	33%	-1.5	0.1626	n.s.	=
		4	10%	33%	-3.4	0.0031	**	<
		5	10%	33%	-3.4	0.0031	**	<
		6	20%	33%	-1.5	0.1626	n.s.	=
	[ʁuʒ]	1	10%	33%	-3.4	0.0031	**	<
		2	5%	33%	-5.7	0.0000	***	<
		3	10%	33%	-3.4	0.0031	**	<
		4	15%	33%	-2.2	0.0374	*	<
		5	20%	33%	-1.5	0.1626	n.s.	=
		6	15%	33%	-2.2	0.0374	*	<
	[vɛʁ]	1	25%	33%	-0.8	0.4122	n.s.	=
		2	15%	33%	-2.2	0.0374	*	<
		3	15%	33%	-2.2	0.0374	*	<
		4	15%	33%	-2.2	0.0374	*	<
		5	10%	33%	-3.4	0.0031	**	<
		6	10%	33%	-3.4	0.0031	**	<
dégoût	[ʒon]	1	15%	11%	0.5	0.6403	n.s.	=
		2	30%	11%	1.8	0.0883	n.s.	=
		3	60%	11%	4.4	0.0003	***	>
		4	25%	11%	1.4	0.1782	n.s.	=
		5	35%	11%	2.2	0.0418	*	>
		6	30%	11%	1.8	0.0883	n.s.	=
	[ʁuʒ]	1	40%	11%	2.6	0.0187	*	>
		2	40%	11%	2.6	0.0187	*	>
		3	30%	11%	1.8	0.0883	n.s.	=
		4	30%	11%	1.8	0.0883	n.s.	=

		5	25%	11%	1.4	0.1782	n.s.	=
		6	45%	11%	3.0	0.0079	**	>
	[vɛʁ]	1	50%	11%	3.4	0.0031	**	>
		2	50%	11%	3.4	0.0031	**	>
		3	85%	11%	9.0	0.0000	***	>
		4	70%	11%	5.6	0.0000	***	>
		5	75%	11%	6.4	0.0000	***	>
		6	80%	11%	7.5	0.0000	***	>
neutre	[ʒon]	1	15%	11%	0.5	0.6403	n.s.	=
		2	35%	11%	2.2	0.0418	*	>
		3	50%	11%	3.4	0.0031	**	>
		4	50%	11%	3.4	0.0031	**	>
		5	40%	11%	2.6	0.0187	*	>
		6	45%	11%	3.0	0.0079	**	>
	[ʁuʒ]	1	55%	11%	3.8	0.0011	**	>
		2	45%	11%	3.0	0.0079	**	>
		3	35%	11%	2.2	0.0418	*	>
		4	60%	11%	4.4	0.0003	***	>
		5	55%	11%	3.8	0.0011	**	>
		6	35%	11%	2.2	0.0418	*	>
	[vɛʁ]	1	25%	11%	1.4	0.1782	n.s.	=
		2	20%	11%	1.0	0.3448	n.s.	=
		3	35%	11%	2.2	0.0418	*	>
		4	45%	11%	3.0	0.0079	**	>
		5	20%	11%	1.0	0.3448	n.s.	=
		6	40%	11%	2.6	0.0187	*	>

Annexe VI : Sélection de stimuli pour l'étude de la typicalité des expressions

Instructions données aux auditeurs préalablement au pré-test

Dans cette expérience vous allez écouter des mots prononcés par différentes personnes. Vous devez déterminer quelles émotions les locuteurs expriment.

Si vous pensez que la personne exprime une émotion relative à la colère (par exemple de l'irritation, de l'agacement, …), donnez le niveau d'expression de cette émotion en réglant le curseur « colère » entre 1 et 10. Si vous pensez qu'elle exprime une émotion relative à la peur (par exemple du stress, de l'anxiété, …) vous devrez régler le curseur « peur » entre 1 et 10. Si vous pensez qu'elle exprime une émotion relative à la joie (par de l'amusement, de la satisfaction, …) vous devrez régler le curseur « joie » entre 1 et 10. Enfin si vous pensez que le locuteur exprime une émotion qui ne peut correspondre à aucune de ces 3 classes d'émotions, réglez le curseur « autre » entre 1 et 10.

Le niveau d'intensité pour chaque classe signifie une très faible intensité pour la graduation 1 jusqu'à une très forte intensité pour la graduation 10 (par exemple la terreur pour la classe « peur », l'exaltation pour la classe « joie », et la rage pour la classe « colère »). Notez bien que l'intensité émotionnelle n'est pas la même chose que la clarté avec laquelle l'émotion est perçue. Il est possible que pour une expression donnée vous ayez des difficultés à décider quelle catégorie d'émotions choisir, mais que vous puissiez facilement décider quelle intensité émotionnelle attribuer, c'est pourquoi nous vous demandons également de donner une valeur d'intensité émotionnelle lorsque vous décidez d'utiliser la classe « autre ».

La durée de ce test est d'environ 30 minutes. Si vous avez des questions quant au protocole n'hésitez pas à nous les poser.

Merci pour votre participation.

Distribution des 108 stimuli retenus

Ce tableau récapitule la distribution des 108 stimuli retenus suite au pré-test de catégorisation et d'évaluation de l'intensité émotionnelle perçue préalable à l'étude réalisée sur les déterminants de la typicalité des expressions vocales d'émotions (Laukka, Audibert, Aubergé, 2008) présentée en section X.

Les abréviations Acté n.p. et Acté p. correspondent respectivement aux expressions considérées comme actées non-prototypiques (ressenties au cours de l'induction et rejouées) et actées prototypiques (expressions des *Big Six*).

		Colère				Peur				Joie				Total
		Acté n,p,	Acté p.	Spontané	Total	Acté n,p,	Acté p.	Spontané	Total	Acté n,p,	Acté p.	Spontané	Total	
Intensité perçue	moyenne	5.03	6.05	4.25	5.11	5.4	4.23	3.47	4.37	6.19	4.49	3.74	4.81	4.76
	écart-type	2.12	1.71	1.30	1.85	1.37	1.61	1.01	1.54	1.61	1.52	1.55	1.84	1.76
Identification	moyenne	80.00%	90.56%	68.33%	79.63%	77.78%	78.33%	71.67%	75.93%	81.11%	79.44%	69.44%	76.67%	77.41%
	écart-type	21.08%	11.53%	18.88%	19.45%	18.93%	18.45%	20.13%	18.87%	26.03%	16.69%	22.10%	21.94%	20.01%
Sexe locuteur	# hommes	6	7	6	19	7	6	7	20	6	8	5	19	58
	# femmes	6	5	6	17	5	6	5	16	6	4	7	17	50
Enoncés	# monosyllabes	9	9	4	22	10	9	11	30	8	10	4	22	74
	# page suivante	3	3	8	14	2	3	1	6	4	2	8	14	34

Annexe VII : Discrimination d'expressions audiovisuelles actées vs. spontanées

Taux de discrimination pour chaque juge et chaque condition par rapport au hasard (50%)

Tests t de Student bidirectionnels pour échantillons uniques, à 47 degrés de liberté. Chaque ligne évalue la significativité de la différence entre le taux moyen de discrimination correcte pour le juge et la condition correspondants (colonne *Discrimination*) et la valeur de test 0.5 correspondant à la probabilité de discriminer correctement par hasard.

Colonne *Significativité* : les notations ***, **, * et n.s. correspondent respectivement aux niveaux de significativité $p<0.001$, $p<0.01$, $p<0.05$ et non significatif.

Colonne *Direction* : les notations <, = et > correspondent respectivement à un taux de discrimination moyen significativement inférieur au hasard, non significativement différent du hasard et significativement supérieur au hasard.

Sujet	Genre	Age	Condition	Discrimination	Écart-type	t	p	Significativité	Direction
S1	F	24	A	79%	41%	4.924	0.000	***	>
			AV	88%	33%	7.774	0.000	***	>
			V	60%	49%	1.460	0.151	n.s.	=
S2	F	42	A	35%	48%	-2.090	0.042	*	<
			AV	38%	49%	-1.770	0.083	n.s.	=
			V	25%	44%	-3.958	0.000	***	<
S3	F	19	A	52%	50%	0.286	0.776	n.s.	=
			AV	46%	50%	-0.573	0.569	n.s.	=
			V	46%	50%	-0.573	0.569	n.s.	=
S4	F	26	A	71%	46%	3.142	0.003	**	>
			AV	71%	46%	3.142	0.003	**	>
			V	63%	49%	1.770	0.083	n.s.	=
S5	F	24	A	71%	46%	3.142	0.003	**	>

			AV	88%	33%	7.774	0.000	***	>
			V	52%	50%	0.286	0.776	n.s.	=
S6	F	33	A	69%	47%	2.773	0.008	**	>
			AV	88%	33%	7.774	0.000	***	>
			V	73%	45%	3.535	0.001	**	>
S7	F	38	A	52%	50%	0.286	0.776	n.s.	=
			AV	73%	45%	3.535	0.001	**	>
			V	56%	50%	0.864	0.392	n.s.	=
S8	F	46	A	71%	46%	3.142	0.003	**	>
			AV	79%	41%	4.924	0.000	***	>
			V	75%	44%	3.958	0.000	***	>
S9	F	27	A	81%	39%	5.489	0.000	***	>
			AV	90%	31%	8.883	0.000	***	>
			V	71%	46%	3.142	0.003	**	>
S10	M	34	A	67%	48%	2.424	0.019	*	>
			AV	83%	38%	6.132	0.000	***	>
			V	56%	50%	0.864	0.392	n.s.	=
S11	M	27	A	65%	48%	2.090	0.042	*	>
			AV	73%	45%	3.535	0.001	**	>
			V	50%	51%	0.000	1.000	n.s.	=
S12	M	40	A	46%	50%	-0.573	0.569	n.s.	=
			AV	75%	44%	3.958	0.000	***	>
			V	58%	50%	1.159	0.252	n.s.	=
S13	M	25	A	50%	51%	0.000	1.000	n.s.	=
			AV	65%	48%	2.090	0.042	*	>
			V	56%	50%	0.864	0.392	n.s.	=
S14	F	26	A	60%	49%	1.460	0.151	n.s.	=
			AV	63%	49%	1.770	0.083	n.s.	=
			V	60%	49%	1.460	0.151	n.s.	=
S15	F	25	A	79%	41%	4.924	0.000	***	>
			AV	85%	36%	6.880	0.000	***	>

			V	77%	42%	4.418	0.000	***	>
S16	M	33	A	60%	49%	1.460	0.151	n.s.	=
			AV	73%	45%	3.535	0.001	**	>
			V	65%	48%	2.090	0.042	*	>
S17	M	19	A	48%	50%	-0.286	0.776	n.s.	=
			AV	71%	46%	3.142	0.003	**	>
			V	50%	51%	0.000	1.000	n.s.	=
S18	M	32	A	67%	48%	2.424	0.019	*	>
			AV	85%	36%	6.880	0.000	***	>
			V	75%	44%	3.958	0.000	**	>
S19	M	67	A	58%	50%	1.159	0.252	n.s.	=
			AV	63%	49%	1.770	0.083	n.s.	=
			V	58%	50%	1.159	0.252	n.s.	=
S20	M	50	A	42%	50%	-1.159	0.252	n.s.	=
			AV	75%	44%	3.958	0.000	***	>
			V	67%	48%	2.424	0.019	*	>
S21	F	22	A	63%	49%	1.770	0.083	n.s.	=
			AV	73%	45%	3.535	0.001	**	>
			V	63%	49%	1.770	0.083	n.s.	=
S22	M	44	A	48%	50%	-0.286	0.776	n.s.	=
			AV	60%	49%	1.460	0.151	n.s.	=
			V	54%	50%	0.573	0.569	n.s.	=
S23	M	47	A	65%	48%	2.090	0.042	*	>
			AV	44%	50%	-0.864	0.392	n.s.	=
			V	38%	49%	-1.770	0.083	n.s.	=
S24	M	31	A	48%	50%	-0.286	0.776	n.s.	=
			AV	44%	50%	-0.864	0.392	n.s.	=
			V	50%	51%	0.000	1.000	n.s.	=
S25	F	22	A	63%	49%	1.770	0.083	n.s.	=
			AV	67%	48%	2.424	0.019	*	>
			V	46%	50%	-0.573	0.569	n.s.	=

S26	F	50	A	73%	45%	3.535	0.001	**	>
			AV	67%	48%	2.424	0.019	*	>
			V	38%	49%	-1.770	0.083	n.s.	=
S27	M	28	A	88%	33%	7.774	0.000	***	>
			AV	85%	36%	6.880	0.000	***	>
			V	60%	49%	1.460	0.151	n.s.	=
S28	F	39	A	29%	46%	-3.142	0.003	**	<
			AV	46%	50%	-0.573	0.569	n.s.	=
			V	35%	48%	-2.090	0.042	*	<
S29	M	34	A	69%	47%	2.773	0.008	**	>
			AV	77%	42%	4.418	0.000	***	>
			V	67%	48%	2.424	0.019	*	>
S30	M	26	A	56%	50%	0.864	0.392	n.s.	=
			AV	35%	48%	-2.090	0.042	*	<
			V	27%	45%	-3.535	0.001	**	<
S31	F	25	A	54%	50%	0.573	0.569	n.s.	=
			AV	77%	42%	4.418	0.000	***	>
			V	48%	50%	-0.286	0.776	n.s.	=
S32	F	28	A	75%	44%	3.958	0.000	***	>
			AV	88%	33%	7.774	0.000	***	>
			V	75%	44%	3.958	0.000	***	>
S33	F	40	A	52%	50%	0.286	0.776	n.s.	=
			AV	77%	42%	4.418	0.000	***	>
			V	65%	48%	2.090	0.042	*	>

Résumé

Les récents progrès de la synthèse de la parole ont redéfini les enjeux majeurs de la communication personne-machine, qui s'articulent désormais autour du concept de pertinence du clone parlant à travers son expressivité. Nous défendons une hypothèse selon laquelle la dynamique des contours prosodiques serait porteuse d'information affective, en séparant les affects dont l'expression est contrôlée volontairement vs. involontairement. Pour cela a été développé E-Wiz, une plateforme dédiée à la capture de corpus multimodaux de parole expressive spontanée par des scénarios de Magicien d'Oz. 17 locuteurs francophones (dont 7 acteurs qui ont rejoué immédiatement après les affects ressentis) ont été capturés dans une tâche prétexte d'apprentissage des langues, le scénario Sound Teacher, visant l'induction d'états émotionnels de valence plutôt positive puis négative exprimés principalement sur des énoncés monosyllabiques isolés. Ils ont ensuite annoté eux-mêmes les affects exprimés. Nous avons identifié des contours de F0 distincts pour différentes expressions émotionnelles. Une évaluation perceptive en conditions audio et audiovisuelle des affects actés par un locuteur a montré que les monosyllabes étaient aussi bien reconnues que les énoncés plus longs. Deux expériences perceptives de projection dimensionnelle de stimuli expressifs sur des stimuli neutres ont montré qu'aucune dimension ne pouvait être liée directement à une classe particulière d'émotions. L'intensité et la typicalité d'expressions actées et spontanées produites par 6 locuteurs ont été mesurées, préalablement à une tâche de discrimination audio, visuelle et audiovisuelle de paires de stimuli actés vs. spontanés. Cette expérience a montré que des sujets naïfs étaient capables de discriminer les énoncés actés vs. spontanés sans différence notable entre classes d'émotion mais avec une importante variabilité interindividuelle, et sans que les différences entre intensités d'émotion perçues n'expliquent entièrement les scores de discrimination.

Abstract

Advances in speech synthesis have led to redefinition of the key issues of person-machine communication, the concept of relevance of the talking clone through its expressivity being nowadays central. We defend a hypothesis according to which the dynamics of prosodic contours would carry affective information, considering separately affects which expression is controlled voluntarily vs. involuntarily. For this purpose was developed E-Wiz, a platform dedicated to the collection of multimodal spontaneous expressive speech corpora with Wizard of Oz scenarios. 17 subjects, including 7 actors who were asked to replay the same affects immediately after the experiment, were tricked on a pretext language-learning task called Sound Teacher, aiming at the induction of both negative and positive emotional states, mainly on monosyllabic utterances. Speakers annotated themselves the affects they expressed. We could identify distinct F0 contours for different emotional expressions. A perceptual evaluation in audio and audiovisual conditions of affects acted by a speaker showed that monosyllables were as well identified as longer utterances. Two perceptual experiments based on dimensional projection of expressive stimuli on neutral utterances showed that no dimension could be directly linked to a particular emotional class. The intensity and typicality of acted and spontaneous expressions produced by 6 speakers were measured, prior to a task of discrimination in audio, visual and audiovisual conditions of pairs of acted vs. spontaneous stimuli. This experiment showed that naïve subjects were able to discriminate acted vs. spontaneous utterances without any significant difference between classes of emotion, but with a large inter-judge variability, while differences in perceived intensity of the emotion could not fully account for discrimination scores.