Analyse acoustique de la voix émotionnelle de locuteurs lors d'une interaction humain-robot

Marie Tahon

Thèse présentée pour obtenir le grade de Docteur de l'Université Paris-Sud

Spécialité: Informatique

Analyse acoustique de la voix émotionnelle de locuteurs lors d'une interaction humain-robot

Marie Tahon

soutenance le 15 novembre 2012

Philippe Martin	Rapporteur	Université Paris-Diderot VII
Gaël Richard	Rapporteur	LTCI-CNRS, Telecom Paris Tech
Nick Campbell	Examinateur	Trinity College Dublin
Christophe d'Alessandro	Examinateur	LIMSI-CNRS, Université Paris-Sud XI
Laurence Devillers	Directeur de thèse	LIMSI-CNRS, Université Paris-Sorbonne IV
Claude Barras	Directeur de thèse	LIMSI-CNRS, Université Paris-Sud XI

Résumé

Mes travaux de thèse s'intéressent à la voix émotionnelle dans un contexte d'interaction humain-robot. Dans une interaction réaliste, nous définissons au moins quatre grands types de variabilités : l'environnement (salle, microphone) ; le locuteur, ses caractéristiques physiques (genre, âge, type de voix) et sa personnalité ; ses états émotionnels ; et enfin le type d'interaction (jeu, situation d'urgence ou de vie quotidienne). A partir de signaux audio collectés dans différentes conditions, nous avons cherché, grâce à des descripteurs acoustiques, à imbriquer la caractérisation d'un locuteur et de son état émotionnel en prenant en compte ces variabilités.

Déterminer quels descripteurs sont essentiels et quels sont ceux à éviter est un défi complexe puisqu'il nécessite de travailler sur un grand nombre de variabilités et donc d'avoir à sa disposition des corpus riches et variés. Les principaux résultats portent à la fois sur la collecte et l'annotation de corpus émotionnels réalistes avec des locuteurs variés (enfants, adultes, personnes âgées), dans plusieurs environnements, et sur la robustesse de descripteurs acoustiques suivant ces quatre variabilités. Deux résultats intéressants découlent de cette analyse acoustique : la caractérisation sonore d'un corpus et l'établissement d'une liste "noire" de descripteurs très variables. Les émotions ne sont qu'une partie des indices paralinguistiques supportés par le signal audio, la personnalité et le stress dans la voix ont également été étudiés. Nous avons également mis en oeuvre un module de reconnaissance automatique des émotions et de caractérisation du locuteur qui a été testé au cours d'interactions humain-robot réalistes. Une réflexion éthique a été menée sur ces travaux.

Abstract

This thesis deals with emotional aspect of voices during human-robot interaction. In natural interaction, we define at least four main kinds of variables : environment (room, microphone) ; speaker, his physic characteristics (gender, age, voice type) and personality ; his emotional states ; and finally the kind of interaction (game scenario, emergency, everyday life). Using acoustic features from audio signals collected in different conditions, we try to match speaker with his emotional state characterisation. It is a real challenge to find which features are essential and which are to be avoided, for it means to be able to deal with a high number of variables as rich and diversified as possible. The two main results of our work are first the collection and the annotation of natural emotional corpora that have been recorded from different kinds of speakers (children, adults, elderly people) in various environments, and second to find how reliable are acoustic features according to the four variables previously defined. This analysis led to two interesting aspects : the audio characterisation of a corpus and the drawing up of a "black list" of features too unsteady to be kept. Emotions are just a part of paralinguistic features supported by the audio channel. Other paralinguistic features have been studied such as personality and stress in the voice. We have also built an automatic emotion recognition and speaker characterisation module that has been tested during realistic human-robot interactions. Ethical aspects have not been neglected in our work.

Remerciements

Réaliser une thèse, c'est du même ordre que préparer un récital de clarinette... Défrichage, déchiffrage, temps trop long, trop court, références, expériences, essais, erreurs, etc... puis vient la réalisation finale : le manuscrit (ou le concert) et la soutenance. C'est du 150%, de la prise de tête, mais finalement et surtout c'est passionant ! surtout la fin quand tout s'imbrique (presque) parfaitement ! ! Pour en arriver là, je dois absolument remercier tout un chacun ayant contribué de près ou de loin à ce travail, ceux qui l'ont rendu possible, ceux qui l'ont soutenu, ceux qui l'ont rendu captivant.

Laurence Devillers et Claude Barras, pour leur recul sur mon travail et la recherche autour des interactions et des descripteurs, leurs expertises sur des domaines qui m'étaient a priori étrangers et surtout pour leur soutien et leurs exigences même dans les moments difficiles ;

Les émotion-philes, Agnès, Mariette, Christophe et Clément du thème "dimensions affectives et sociales des interactions parlées) du LIMSI pour les stress et les détentes, les pique-niques et les démos, mais également tous les annotateur-trices qui se sont arrachés les cheveux ou qui ont déprimé d'entendre des gens s'exprimer dans des situations difficiles et les stagiaires ayant participé de près ou de loin au projet ROMEO ;

Mes nouveaux collègues du CNAM pour le soutien salutaire pour les derniers moments de rush et pour m'avoir remise entre les mains des équations fondamentales de l'acoustique ;

Toutes ces personnes volontaires de l'Institut de la Vision, du LIMSI, les enfants du CESFO, mes cousines, des amis de passage, des étudiants en STAPS, qui ont accepté de se prendre au jeu à faire l'acteur, discuter avec le robot NAO, stresser en public au cours des enregistrements de corpus ;

Et... ne pas oublier Guillaume, Clément (et ?), la famille au sens plus large (entre autres les relecteurs Christophe et Anne-Hélène), ainsi que tous les copaings musiciens, chercheurs de poissons, de nano-structures, farlousiens ou militants ;

Je rajouterai aussi des remerciements à la musique, à la pratique pas toujours quotidienne de la clarinette, à l'organisation d'un récital en février 2012, autant de soupapes pour faire sortir la fumée.

Table des matières

.

Première partie

Introduction générale

Préambule

Par la voix, les êtres humains peuvent transmettre plusieurs messages, en particulier, du sens et de l'émotion. La perception d'une émotion dans la voix revient souvent à percevoir consciemment ou non, la "musique de la voix". Le neurologue Olivier Sacks parle très bien de la perception de la musique et de son caractère émotionnel dans son livre *Musicophilia* [Sacks 09] :

> Nous autres, êtres humains, sommes une espèce musicale non moins que linguistique. Si différentes que soient les formes prises par notre musicalité, nous sommes tous (à de très rares exceptions près) capable de percevoir la musique : percevant les sons, le timbre, les intervalles, les contours mélodiques, l'harmonie et le rythme (qui est peut-être la plus élémentaire de ces données), nous intégrons tous ces éléments en nous servant de parties distinctes de notre cerveau. A cette appréciation structurelle largement inconsciente de la musique s'ajoute une réaction émotionnelle aussi intense que profonde, le plus souvent.

Cependant aux données élémentaires citées par Olivier Sacks lors de la perception de la musique, il faut ajouter d'autres paramètres liés au contexte de l'écoute. La musique nous touche d'abord d'autant plus qu'elle s'exprime dans un langage connu. La musique contemporaine attire moins le public que la musique classique ou romantique par le fait qu'elle utilise des références et des codes très spécifiques (par exemple l'atonalité, l'arythmie). Même lorsque le langage musical nous est proche, on peut apprécier l'œuvre différemment selon l'interprète, l'instrument sur lequel il joue (ou sa voix), son interprétation musicale. Enfin le contexte dans lequel l'œuvre est écoutée peut influencer sa perception : on ne reçoit pas de manière similaire une musique qui passe à la radio, jouée dans une salle de concert, en plein air ou chez soi sur une chaine Hi-Fi. Ainsi une œuvre musicale nous touche différemment suivant le langage, l'interprète et le contexte.

La perception des émotions, est très liée à celle de la musique dans la voix. Les mêmes paramètres, souvent inconsciemment, entrent alors en jeu. La perception d'une émotion dépend (i) de la culture : langue, comportements sociaux, (ii) du locuteur qui s'exprime (mais aussi de celui qui perçoit) : de ses caractéristiques physiques, son expérience et son histoire, sa personnalité, ou encore d'une éventuelle pathologie (les personnalités alexithymiques par exemple, ne ressentent ni n'expriment d'émotions). L'émotion perçue dépend également très fortement (iii) du contexte, de l'environnement sonore ou du type d'interaction (cinéma, situation d'urgence, scènes de vie quotidienne, etc.). Ce sont ces sources de variabilités que nous allons développer tout au long de cette thèse.

Contexte de mes travaux de recherche

Les interactions humain-robot

Mes travaux de recherche sur l'analyse de la perception de la musicalité de la voix émotionnelle (les sons, le timbre, les intervalles, les contours mélodiques, le rythme comme le décrit Sacks) se situent dans un contexte assez spécifique : les interactions humain-robot. Ce domaine de recherche vise à développer les interactions entre des êtres humains et des machines telles que des ordinateurs, des agents virtuels ou des robots.

Ce domaine est extrêmement actif aujourd'hui grâce notamment au développement des capacités mécaniques et informatiques des machines. Les robots doivent pouvoir comprendre à la fois leur environnement mais aussi l'utilisateur dans le but de prendre ensuite une décision et d'agir en conséquence (action mécanique, message oral, synthèse de voix expressive, etc.).

Les robots et les émotions est un sujet d'étude pluridisciplinaire : un robot doit à la fois capter l'émotion du locuteur mais également pouvoir en exprimer. Il existe plusieurs robots capables d'exprimer des émotions soit par la parole, soit par des gestes. C'est le cas de ASIMO (Honda) ou de Leonardo et de Kismet robots conçus au MIT par l'équipe de Cynthia Breazeal. Plusieurs équipes travaillent sur l'utilisation d'entrées afin de construire un signal social lors d'une interaction homme-robot. Nous pouvons citer les travaux de Miwa sur WE-3RV [Miwa et al. 03] dont le robot traite des entrées audio et visuelles, les études sur iCat se focalisent sur les expressions faciales du locuteur [Castellano et al. 10], et enfin le robot Kismet est équipé d'un système de détection automatique de visages, de couleurs et de mouvements [Breazeal 99]. Cependant peu d'entre eux savent reconnaître les émotions dans la parole d'un humain lors d'une interaction.

Le projet ROMEO (2009-2011)

Cette thèse a été financée principalement par le projet ROMEO [1], un projet FUI Cap Digital (région Ile de France). L'objectif du projet est de développé un robot social (ROMEO) à la fois joueur et assistant qui serait le grand frère de NAO conçu par l'entrerpise Aldebaran. Le cadre du projet a permis de mettre en place un contexte précis pour mes recherches autour des interactions humain-robot. Un projet de cette ampleur regroupe un grand nombre de partenaires, industriels et laboratoires de recherche.

Nous nous plaçons dans un contexte d'interaction entre un ou plusieurs humains et un robot. Ce robot aura au moins deux rôles distincts : robot-joueur dans le cadre d'interactions avec un ou plusieurs enfants et robot-assistant dans le cadre d'une interaction avec une personne mal-voyante (en lien avec l'Institut de la Vision, partenaire du projet). Ainsi le public visé est très large : des enfants aux personnes âgées en passant par des personnes actives. Le contexte définit également le cadre de l'interaction : des scènes de la vie quotidienne, dans un environnement quelconque. On se limitera à un environnement intérieur (pas de prise de son en extérieur). Ce qui peut impliquer des environnements très variés : cuisine, chambre, salon, bureau avec des ambiances acoustiques très différentes les unes des autres. Une autre caractéristique de l'interaction est son déroulement temps réel : la prise de son se fait en continu, la reconnaissance du locuteur et de ses émotions doit se faire le plus rapidement possible. Cet aspect nous limitera notamment en temps de calcul.

Dans une interaction réaliste, nous pouvons définir au moins quatre grandes sources variabilités (que nous nommerons par la suite "variabilités") : (1) le contexte qui englobe l'environnement (salle, microphone) et la situation (type d'interaction, acté/spontané) ; (2) le locuteur, ses caractéristiques physiques (genre, âge, type de voix) et sa personnalité ; (3) le type d'émotions, quel neutre et quelles émotions peuvent être exprimées ; et enfin (4) le type de tâche, scénario de jeu, de situation d'urgence ou de vie quotidienne. Lorsque ces variabilités sont contraintes, la spécificité (ou "*typicality*") de l'interaction est plus forte.

1. www.projetromeo.org

D'après Marchi [Marchi et al. 12], la spécificité est corrélée également à la disponibilité des corpus correspondants. Plus la spécificité est importante, plus il est facile de collecter des données. Nous nous situons dans un contexte où la spécificité est très faible et les variabilités sont peu contraintes, il est donc difficile d'obtenir une grande quantité de données correspondantes à de telles interactions humain-robot.

Challenges

La reconnaissance d'indices paralinguistiques

Dans ce contexte, en plus de comprendre le message transmis par l'humain, la machine doit pouvoir interagir de manière fluide avec lui. Elle doit alors être capable non seulement de comprendre les informations linguistiques mais également paralinguistiques. Ces informations paralinguistiques sont d'au moins deux types : (i) la reconnaissance de traits caractéristiques physiques (âge, genre, éventuellement identité), psychologiques (personnalité), mais aussi liés à un état ponctuel (fatigue vocale, stress), ou à des pathologies particulières, (ii) la reconnaissance d'indices sociaux : émotions, attitudes et humeurs, marqueurs interactionnels (affinité, prise de pouvoir ou *dominance*). La détection et la reconnaissance de traits humains et d'indices sociaux est un courant de recherche émergent. Ce domaine de recherche a été mis en avant notamment lors de l'organisation de compétitions internationales (ou challenges) (2009 : détection d'émotions [Schuller et al. 09b], 2010 : détection de l'âge, du genre et de l'affection [Schuller et al. 10a], 2011 : détection de niveaux de fatigue vocale ou d'alcoolisme [Schuller et al. 11b], 2012 : détection de catégories de personnalité [Schuller et al. 12b]).

La reconnaissance des émotions dans la parole est un sujet d'étude assez récent puisqu'il n'a qu'une quinzaine d'années, cependant beaucoup de recherches sont actuellement faites dans ce domaine. La plupart d'entre elles se concentrent sur des émotions actées et prototypiques dans un contexte de laboratoire exprimées par quelques locuteurs. De plus la plupart de ces indices sont reconnus séparemment. Un des challenges lié au contexte est de reconnaitre les émotions sur un nombre important de locuteurs et qui plus est très diverses (enfants, adultes, personnes âgées, personnes souffrant de pathologies particulières, etc.). Un autre est de pouvoir reconnaitre un locuteur (son identité, son genre) alors qu'il exprime des émotions diverses et variées. Un premier défi soulevé par ce domaine de recherche est : ***l'imbrication entre reconnaissance du locuteur et de son état émotionnel.***

Les variabilités propres aux interactions humain-robot

Un autre challenge et non des moindres, de la reconnaissance des émotions, réside dans l'utilisation de plusieurs corpus émotionnels. La plupart des recherches actuelles se focalisent généralement sur un seul corpus qu'il soit acté, induit ou spontané. Quelques études néanmoins ont été réalisées en croisant des corpus afin de tester la robustesse des systèmes de reconnaissance automatique des émotions dans des situations diverses. Nous avons eu l'occasion de participer à la collecte de cinq corpus distincts les uns des autres par (i) le type de locuteurs enregistrés (enfants, personnes actives, personnes mal-voyantes,

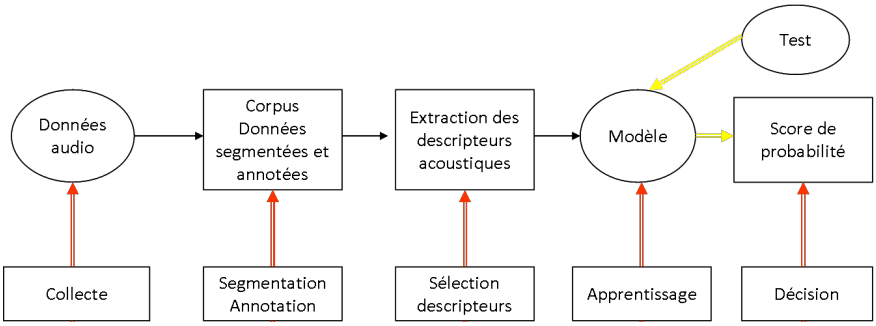


FIGURE 1 – La reconnaissance automatique d'indices paralinguistiques dans la voix

personnes âgées), (ii) la tâche (acté, induit, spontané), (iii) les conditions d'enregistrement (studio réverbérant, salon, laboratoire, etc.). La possibilité d'avoir tous ces corpus à disposition permet de relever un deuxième défi : ***la mise en évidence de la variabilité acoustique liée aux émotions, aux locuteurs, à la spécificité de la tâche ou de l'environnement sonore***.

Organisation du document

La reconnaissance automatique d'indices paralinguistiques dans la voix nécessite un certain nombre d'étapes complexes décrites sur le schéma 1.

La collecte des données audio nécessaires pour mettre en évidence les variabilités propres aux interactions sera abordée dans la partie 1 (la collecte au chapitre 1 et les annotations au chapitre 2). La deuxième partie du document vise à proposer un ensemble de plusieurs descripteurs acoustiques caractérisant le signal émotionnel (un état de l'art des descripteurs acoustiques utilisés en parole est donné au chapitre 3, une étude de ces descripteurs et de propositions originales suivant les différentes variabilités est abordée au chapitre 4). Dans l'optique d'une interaction entre un être humain et une machine, des systèmes de reconnaissance automatique ont été évalués suivant les quatre variabilités dans la partie 3. D'abord un système de reconnaissance automatique du locuteur et particulièrement la reconnaissance du genre (chapitre 5), puis un système de reconnaissance automatique d'indices paralinguistiques (chapitre 6). Les indices paralinguistiques que nous avons étudié sont composés des émotions (particulièrement de la valence), du stress et de la personalité.

Le document se conclut avec la présentation du module SysRELL (Système de Reconnaissance Emotion Locuteur du LIMSI) mis en place avec l'équipe du LIMSI et avec quelques réflexions éthique liées à l'utilisation de telles technologies (chapitre 7).

Deuxième partie

Émotions, locuteurs, corpus et annotations

Dans le cadre d'une interaction humain-robot, nous devons répondre à la question "quelle émotion peut exprimer l'humain ?" mais également à "quelle émotion la machine doit-elle percevoir et reconnaître ?". Le contexte est d'autant plus restreint que nous nous sommes limités à deux applications : celle du robot-joueur avec des enfants et celle du robot-assistant avec des personnes en perte d'autonomie (personnes âgées, personnes atteintes de déficience visuelle). Afin de pouvoir étudier les émotions exprimées dans ces contextes précis et d'élaborer des méthodes pour les reconnaître de manière automatique, nous avons besoin d'en collecter des échantillons. Nous cherchons donc à collecter des émotions les plus naturelles et spontanées possibles (données réelles), sur des locuteurs de type assez différents et dans des conditions acoustiques variables.

Pour cela, nous avons créé deux types de scénario qui se déroulent en interaction avec le robot NAO, l'un avec des personnes mal-voyantes dans lequel le robot joue le rôle d'un assistant, l'autre avec des enfants (moins de 13 ans), NAO jouant avec eux. Dans ces deux types d'interactions les émotions collectées ne seront évidemment pas identiques : des émotions assez spontanées, plutôt positives et marquées chez les enfants (les enfants seront plus facilement à l'aise surtout si ils sont deux à jouer), des émotions également spontanées, mais plus ténues (souvent annotées comme sans émotion) et plus complexes, les adultes ayant plus tendance à se contrôler.

Un schéma d'annotation des émotions perçues a été mis en place afin de vérifier que les émotions attendues ont bien été collectées, mais également pour définir plus précisément le contenu émotionnel de chaque type d'interaction. Nous souhaitons également faire varier les conditions acoustiques, nous avons donc collecté deux corpus pour chaque type d'interaction (robot-joueur ou robot-assistant) l'un dans une salle d'acoustique standard (salon, salle expérimentale), l'autre dans une pièce où l'acoustique est moins contrôlée (studio vide réverbérant, grande salle de jeu).

La variabilité liée à la tâche a ainsi été étudiée suivant deux types d'interactions avec le robot. Nous avons souhaité ajouter une tâche de jeu émotion (le participant doit exprimer une émotion de façon à ce qu'un système de détection la reconnaisse) sur un grand nombre de locuteurs, afin de pouvoir comparer avec des données induites ou spontanées avec des données plus prototypiques.

Chapitre 1

Théorie des émotions et acquisition de corpus

Il m'est souvent arrivé de ressentir des émotions lors de l'écoute d'une oeuvre en situation de concert (avec "le Sacre du printemps" de Stravinski, ou "Coro" de Berio). Le contexte fait que l'on ne doit pas faire entendre les émotions que l'on a ressentie, aucun son n'est alors émis. Par contre, l'augmentation du rythme cardiaque, l'accélération de ma respiration, une espèce d'euphorie envahissante. Le contrôle est très fort mais l'émotion parvient à s'exprimer.

Qu'est-ce que l'émotion ? D'abord l'émotion ne se définit pas seule, elle dépend de celui qui la ressent et l'exprime et éventuellement de celui qui la perçoit. Le fait d'exprimer une émotion implique un grand nombre de changements physiques et physiologiques (neuronal, activation de certaines zones du cerveau, augmentation du rythme cardiaque, etc.). Les émotions n'existent qu'en réaction à des événements qu'ils soient extérieurs ou intérieurs. Elles se situent donc dans un contexte. Une émotion peut ne pas être exprimée, entièrement intériorisée ou contrôlée. Lorsqu'elle s'exprime, ce peut être suivant des moyens très différents : vocal, gestuel, facial, physiologique, etc. Si l'on ne considère que l'émotion vocale, celle qui engendre des sons parlés ou non, l'expression d'une émotion passe par des modifications physiques du conduit vocal et de l'articulation, des changements au niveau de la respiration, de la salive ou encore par des mots ou des sons.

Une émotion n'est perçue que si il y a interaction entre la personne qui ressent l'émotion et une autre et la perception peut varier énormément en fonction du contexte, du lien entre celui qui s'exprime et celui qui perçoit, et des personnalités des deux personnes. Afin d'étudier plus en détail les émotions, il convient d'abord de les fixer, les replacer dans un contexte, analyser leur dépendance à celui qui s'exprime, et celui qui perçoit. C'est ce que nous présentons dans ce premier chapitre.

1.1 Théories des émotions

Les trois principales théories sont : la représentation à partir d'étiquettes verbales [Plutchik 84], les représentations à partir de dimensions abstraites [Mehrabian 96] et le

modèle d'évaluation cognitive [Scherer 99]. Une théorie perceptive plus récente développée dans la thèse Bänziger [Bänziger 04] propose une vision des émotions lors d'une interaction, prenant en compte les interactants de manière plus centrale. Un débat important dans le discours psychologique et théorique porte sur la modélisation des émotions ; doit-elle se faire avec des catégories ou bien des dimensions et combien de catégories ou dimensions sont nécessaires pour une modélisation optimum ? La plupart des communautés scientifiques tendent aujourd'hui à utiliser les deux représentations pour parler des émotions, par exemple au niveau du processus neuronal [Said et al. 10].

Nous ne considèrerons que les émotions d'un point de vue vocal et nous nous focaliserons sur les émotions perçues en interaction avec un robot dans des contextes de jeu, d'appels à l'aide ou de vie quotidienne. Une émotion étant un trouble lié à un événement extérieur ou intérieur d'une personne, l'expression et la perception des émotions sont très fortement liées aux individus eux-mêmes, leur personnalité, leur histoire, une éventuelle pathologie, mais également au contexte.

Les émotions font plus largement partie des indices paralinguistiques tels qu'ils ont été définis dans le challenge Interspeech 2010 [Schuller et al. 12a].

1.1.1 Définition d'une émotion : une question pluridisciplinaire

Les émotions correspondent à un phénomène humain de la vie quotidienne, trouver une définition simple est un véritable défi. Les émotions peuvent être étudiées de plusieurs points de vue : la perception des émotions, le ressenti et l'expression. Certains points de vue supposent l'existence d'une interaction entre au moins deux êtres humains : la perception et le ressenti.

Le phénomène émotionnel a été principalement étudié par les sciences humaines et les sciences de la vie. Ainsi la définition d'une émotion est un sujet de controverse sur lequel se sont penchés d'abord des psychologues, des philosophes ou des neurologues, et plus récemment des chercheurs en informatique. Beaucoup de travaux ont été réalisés dans le domaine de la neurobiologie, sur la reconnaissance d'états émotionnels chez les autres par le cerveau humain [Damasio 94, Ledoux 89]. Il existe aussi un nombre important de recherches et de théories psychologiques sur la perception et la production des émotions, en particulier, la théorie de l'évaluation [Scherer 99].

Une difficulté des travaux sur les émotions est la définition même de l'émotion et des états émotionnels ; émotion primordiale, émotion primaire, affect, humeur, attitude, etc. Toutefois, trois composantes sont généralement acceptées par tous comme constituants de la réaction émotionnelle : la réaction physiologique, l'expression émotionnelle et le sentiment subjectif ou "feeling". La majorité des travaux théoriques ne partage pas la même terminologie et la diversité des modèles d'expression des émotions montre à la fois la complexité et la richesse du sujet d'étude.

Nous pouvons raisonablement dire qu'une émotion est exprimée et perçue comme étant la même si les deux interacteurs communiquent avec des languages similaires, et que l'émotion est commune. Les émotions peuvent être classées en deux grandes catégories : les émotions primaires (innées selon Damasio [Damasio 94], et universelles) et les émotions dérivées acquises au cours de l'existence de chaque individu. Parmi les émotions primaires, nous pouvons citer la colère, la peur, la joie, etc. Parmi les émotions dérivées, nous avons la honte, le désespoir, le soulagement, l'amour, etc. Cowie utilise le terme d'états

émotionnels [Cowie et al. 01] et englobe également des états émotionnels dérivés comme les humeurs. Scherer propose une distinction entre des émotions "abouties" qui impliquent un grand nombre de fonctions humaines (cognitive, physiologique et expressif) d'autres états affectifs (comme les humeurs, les attitudes ou les dispositions affectives) en fonction de leur intensité, de leur durée et de leur focus [Scherer et al. 03]. Les émotions "abouties" sont alors une réaction à un événement bref et ponctuel.

Selon Scherer, "les émotions sont les interfaces de l'organisme avec le monde extérieur". Les aspects les plus importants du processus émotionnel sont l'évaluation de la situation par rapport aux besoins et projets de l'individu, la préparation physiologique et psychologique aux actions liées à la situation, la communication des états et intentions de l'individu à son environnement social.

1.1.2 Les principales théories émotionnelles

La théorie de l'évaluation Cette théorie propose une spécification détaillée de dimensions d'évaluation utilisées pour évaluer des événements qui provoqueraient des émotions (nouveauté, amabilité, pertinence par rapport au but recherché, etc.). Dans cette théorie, la nature de l'émotion est déterminée par une évaluation cognitive. Le "component process model" de Scherer [Scherer and Peper 01] offre une spécification détaillée des dimensions d'évaluation qui sont supposées être utilisées dans l'évaluation des évènements antécédents à l'émotion. C'est un modèle à plusieurs composantes En effet, un épisode émotionnel implique d'évaluer un évènement comme étant pertinent par rapport à l'atteinte de buts, crée un état de préparation à l'action chez l'individu et s'accompagne de l'expérience subjective d'états mentaux distincts de nos états habituels. La principale différence avec les représentations communes du phénomène émotionnel est de considérer que l'émotion est un phénomène dynamique et non un état statique défini par une catégorie statique. Pour Scherer, une émotion est une séquence de changements inter-liés et synchronisés de différents systèmes de l'organisme en réponse à l'évaluation d'un stimulus externe ou interne :

> L'émotion est un processus cumulatif, une unité temporelle pendant laquelle les différents composants de l'émotion se synchronisent et se désynchronisent.

Ces différents composants sont la composante d'évaluation, de tendances à l'action, d'activité physiologique, d'expression motrice et enfin la composante de sentiment subjectif ou "feeling" reflétant les changements survenus dans les autres composants. Scherer définit cinq principaux critères d'évaluation des stimuli ("Stimulus Evaluation Check", SEC) allant des plus universels aux plus spécifiques :

- nouveauté (soudaineté, familiarité, prévisibilité) : caractère inattendu ou non de l'évènement,
- agrément (intrinsèque ou global, désirabilité) : expérience plaisante ou déplaisante,
- rapports aux causes et buts (causalité interne, causalité externe, pertinence, degré de certitude dans la prédiction des conséquences, attentes, opportunité, urgence),
- potentiel de maîtrise (contrôle de l'évènement, contrôle des conséquences, puissance, ajustement) : possibilité de s'adapter,
- accord avec les standards (externes, internes) : accords aux normes sociales et concepts de soi.

Chercheurs	Émotions primaires
Plutchick	apathie/surprise, dégoût/confiance, joie/tristesse, peur/colère
Arnold	colère, aversion, courage, découragement, désir, désespoir, peur, haine, espoir, amour, tristesse
Ekman	colère, peur, joie, tristesse, dégoût, surprise
Frijda	désir, intérêt, bonheur, surprise,
Gray	rage, terreur, anxiété, joie
Izard	colère, mépris, dégoût, détresse, peur, culpabilité, intérêt, joie, honte, surprise
James	peur, douleur/chagrin, amour, rage
Darwin	colère, peur, joie, tristesse, dégoût
Mower	Souffrance, plaisir
Oatley	colère, dégoût, inquiétude, bonheur, tristesse

TABLE 1.1 – Les principales catégories d'émotions primaires, extrait de [Tato 99]

La théorie cognitive de Scherer est très intéressante pour comprendre le processus dynamique de génération des émotions. Par contre, elle est très difficile à utiliser en perception des émotions. Une annotation de clips vidéo en français et en anglais a été menée suivant les critères d'évaluation de la théorie cognitive de Scherer, [Devillers et al. 06] montre que cette annotation est fiable majoritairement sur les dimensions d'activation et de valence.

La théorie des émotions discrètes Elle est fondée sur l'idée d'un nombre limité d'émotions primaires. Plutchick postule pour huit émotions primaires [Plutchik 84] (joie, acceptation, peur, surprise, tristesse, dégoût, colère, et anticipation) a inscrit dans une roue (figure 1.1) ses émotions élémentaires qui, en se combinant, produisent des émotions secondaires : l'amour serait la résultante des émotions élémentaires de joie et d'acceptation ; la soumission résulterait de l'acceptation et de la peur. La distinction entre émotions primaires (ou élémentaires) et secondaires (ou sociales) est largement utilisée. Celles qui sont dites primaires seraient universelles et difficilement contrôlable par la volonté alors que les émotions secondaires résultent d'un processus de construction plus personnel, social et conscient.

Les étiquettes utilisées pour décrire les émotions primaires ne peuvent pas être consensuelles puisqu'elles sont très fortement liées au contexte. Une étiquette verbale ne suffit pas pour décrire une émotion (tableau 1.1). Pour simplifier la théorie des émotions discrètes, nous pouvons considérer les étiquettes des émotions primaires comme des "macro-classes" au sein desquelles peuvent se décliner plusieurs émotions secondaires en fonction du contexte et de la perception de l'émotion. Par exemple, dans la classe "colère", nous pouvons trouver de la colère forte, de la colère froide, de l'agacement, de l'énervement, etc.

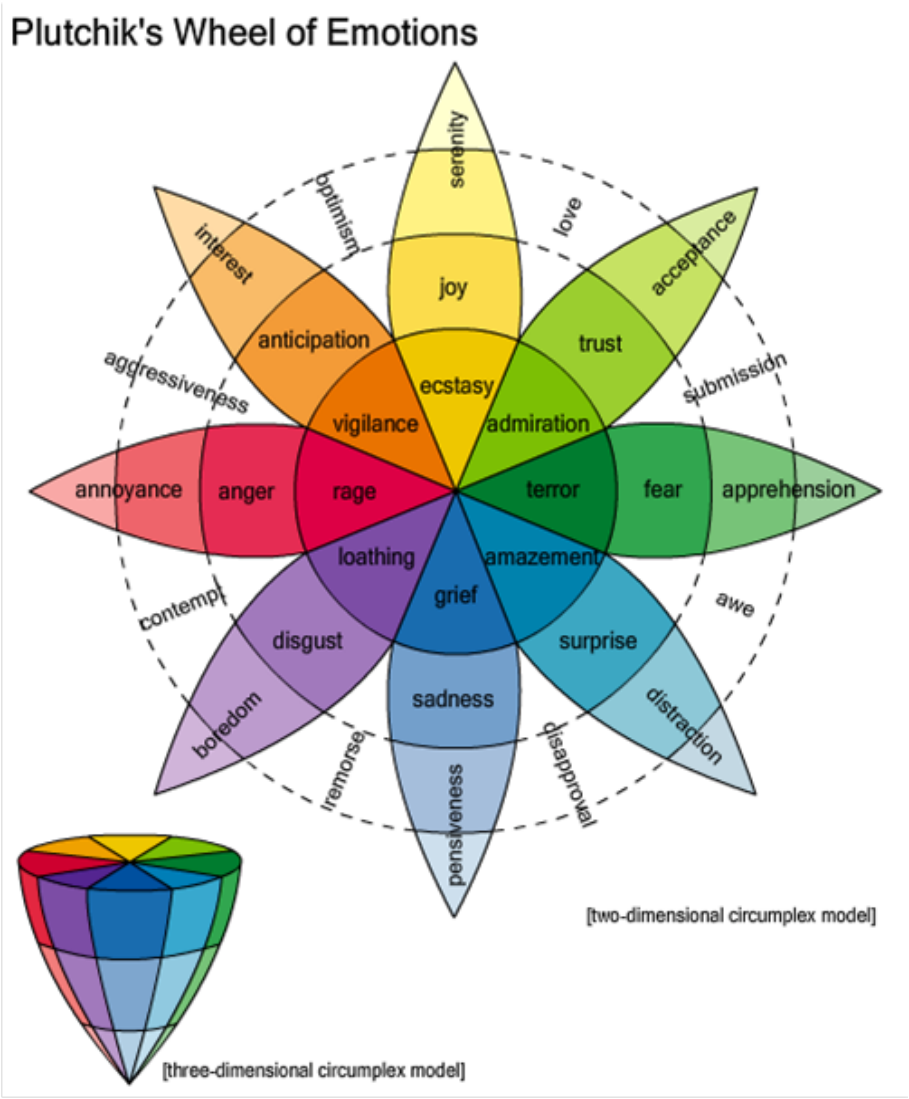


FIGURE 1.1 – Le cône des émotions de Plutchick

La théorie dimensionnelle Originalement, elle décrit l'état émotionnel suivant trois dimensions : envie-aversion (positif/négatif), excitation-apaisement (actif/passif), tension-soulagement [Wundt 13]. Ces dimensions se retrouvent de manière amplifiée dans l'écriture de la musique savante depuis l'époque classique : une succession de tensions, apaisements suivant des échelles plus ou moins importantes. Osgood [Osgood et al. 75] reprend ces trois dimensions : évaluation (ou valence, négatif à positif), activation (passif à actif) et pouvoir (faible à fort). Chez Russel, le pouvoir correspond également à l'ouverture (ou dominance) [Russel 97], et permet de distinguer les émotions donnant lieu à des réactions d'approche et de combat (comme la colère) de celles engendrant des comportements d'évitement, de fuite (comme la peur). Les autres dimensions subjectives sont par exemple le contrôle, l'intensité, etc. Ainsi on peut définir les émotions à partir de dimensions continues abstraites plutôt que de les nommer explicitement avec des catégories discrètes. La figure 1.2 montre le schéma le plus largement utilisé est fondé sur deux dimensions perceptives : l'évaluation (ou la valence) et l'activation. Cependant d'autres dimensions sont nécessaires, par exemple pour distinguer la Peur de la Colère. De nouvelles dimensions abstraites ont été suggérées par les théories de l'évaluation (citées plus haut), par exemple un axe par rapport au but favorable ou défavorable de l'évènement déclencheur d'émotions, i.e. le potentiel de chaque individu à gérer les conséquences de l'évènement émotionnel. Ces dimensions représentent des dispositifs descriptifs qui sont intéressants en eux-mêmes mais qui ne nous fournissent pas d'éclaircissements sur la nature de l'émotion.

1.1.3 Les "affect bursts"

Les "*affect bursts*" ont été introduit par Scherer [Scherer 94], ils sont définis comme "des marqueurs émotionnels non verbaux très brefs, discrets, présents dans la voix et le visage, déclenchés par des événements très clairement identifiables". Les rires, les raclements de gorge, les pleurs, les hésitations, les onomatopées, les répétitions, etc. sont autant d'exemples d'affect bursts. Ces marqueurs émotionnels sont très importants dans les interactions réalistes. Schröder [Schröder 03] a montré que les affect bursts ont une signification émotionnelle très forte.

1.1.4 Le locuteur et les théories émotionnelles

Les théories émotionnelles élaborées par les communautés de sciences humaines cherchent à être les plus générales possibles, mais rares sont celles qui prennent en compte les spécificités du locuteur. Le modèle de lentille permet d'adapter la caractérisation des émotions en fonction de celui qui s'exprime et ressent l'émotion et de celui qui la perçoit dans le cadre d'une interaction.

Le modèle de lentille Développé en premier lieu par Brunswick, le modèle de lentille (ou "*lens model*") fait le lien entre un phénomène émis et un phénomène perçu par l'intermédiaire d'un certain nombre d'indices. Ce modèle a été utilisé d'abord pour la vision (1956) puis adapté au phénomène émotionnel par Scherer [Scherer et al. 03], puis Bänziger [Bänziger 04] (à gauche l'encodage, à droite le décodage sur la figure 1.3). Les états internes d'un locuteur s'expriment par des modifications physiologiques (respiration,

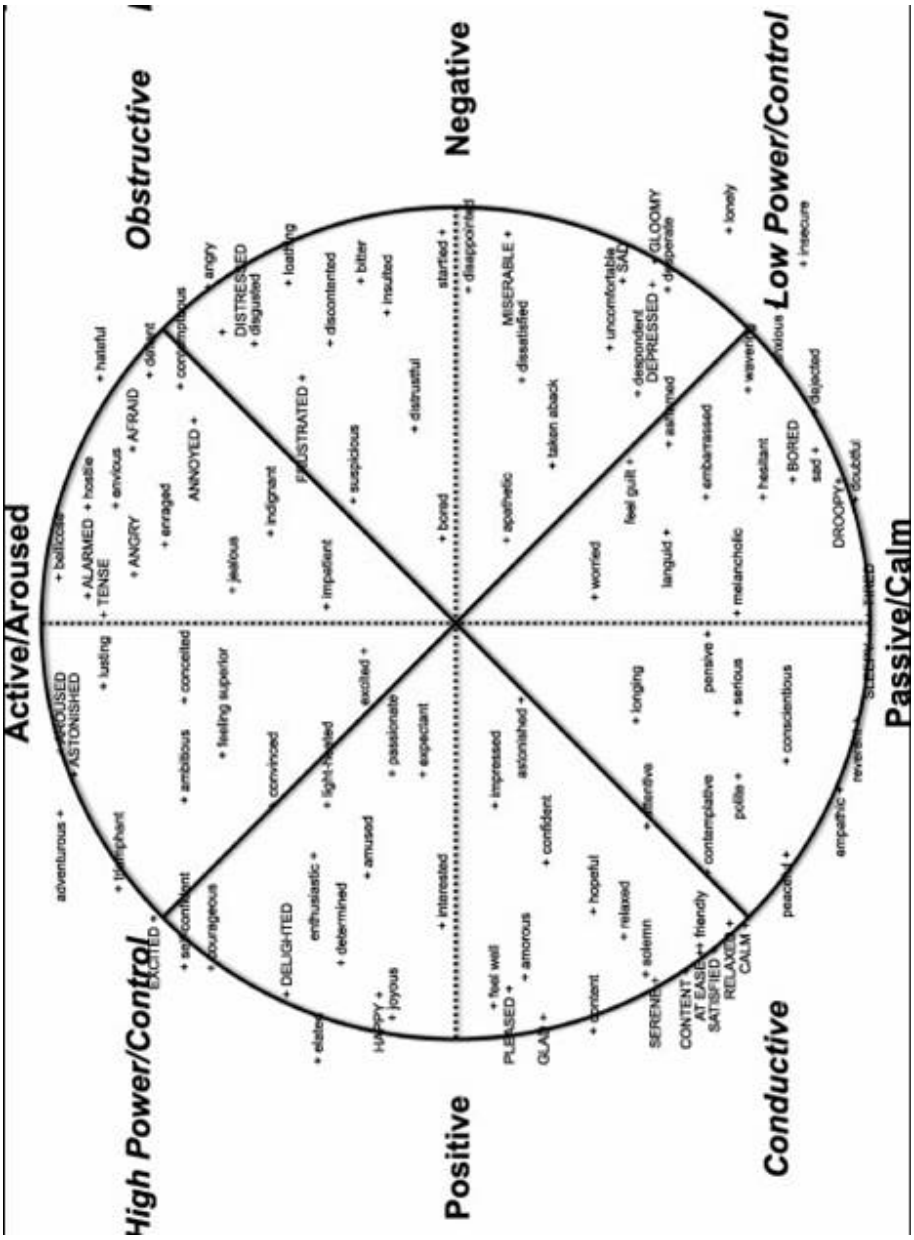


FIGURE 1.2 – Dimensions de Scherer (extrait des travaux de l'association HUMAINE)

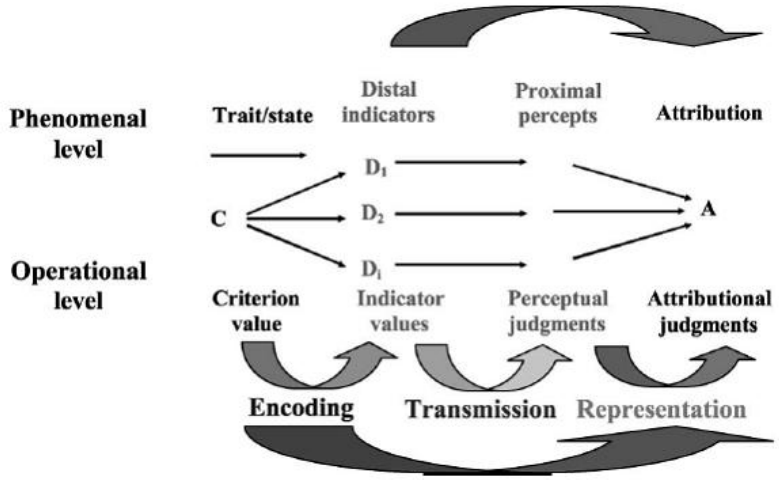


FIGURE 1.3 – Le modèle de lentille adapté aux émotions (adapté de [Scherer et al. 03])

phonation, articulation) et sont encodés par des indices mesurables par un observateur (indices acoustiques dans le cas de la voix) appelés indices distaux dans le modèle. Ces indices sont à la fois dus à des réactions involontaires ou "*push effects*" (effet des changements physiologiques caractérisant la réponse émotionnelle sur la voix : tremblement de la voix par exemple) et à une communication intentionnelle des états internes ou "*pull effects*" (régulation de la vocalisation pour des raisons stratégiques). Ils sont transmis jusqu'à l'oreille d'un observateur et perçus par le système perceptif auditif. L'observateur traite ces indices (nommés indices proximaux dans le modèle) et les représente par des percepts qu'il utilise pour inférer l'état du locuteur. Les indices proximaux sont censés être identiques aux indices distaux pour que l'émotion perçue soit également celle qui a été ressentie. Seulement ils peuvent être modifiés au cours de la transmission, soit par des phénomènes extérieurs à l'interaction (distance, bruit environant), soit par des spécificités de l'un ou l'autre des interacteurs (physiologiques, psychologiques, sociales ou culturelles). Ce modèle a l'avantage de prendre en compte l'ensemble des facteurs qui peuvent influer sur la perception d'une émotion.

1.1.5 Conclusion

Nous avons vu qu'il existait plusieurs modèles émotionnels se complétant les uns avec les autres. Le modèle d'évaluation est extrêmement complexe et difficile à utiliser en perception des émotions, ce n'est donc pas celui que nous retiendrons. Par contre les deux théories catégorielles et dimensionnelles permettent de définir les émotions suivant des étiquettes utilisable en perception. Etant donné la très grande variabilité de locuteurs et de situations dans lesquelles peuvent avoir lieu le phénomène émotionnel, nous n'utiliserons pas les émotions secondaires. Par contre, nous regrouperons sous forme de macro-classes des émotions primaires et des émotions secondaires (plus fines et spécifiques à chaque situation). Ces macro-classes pourront être combinées avec deux dimensions : la valence

et l'activation.

1.2 État de l'art : stratégie d'acquisition de bases de données locuteur et émotions

La plupart des théories émotionnelles en cours aujourd'hui, ont été grandement etayées par des études sur des corpus émotionnels. Dans le cas du modèle de la lentille, l'ensemble des critères ont été évalués sur des bases de données émotionnelles. Ces bases de données utilisent généralement les supports audio et vidéo, mais certaines sont également constituées d'images médicales ou de données physiologiques. Nous nous limiterons au support audio et à l'expression vocale des émotions et de l'identité d'un locuteur. Une première revue de l'état de l'art permet de faire un bilan des stratégies utilisées pour collecter des données émotives, et des données locuteurs. Ces deux aspects étant la plupart du temps traités séparemment.

1.2.1 Bases de données émotionnelles

Bases de données actées Les premières bases de données enregistrées pour la reconnaissance des émotions étaient généralement très actées avec relativement peu de locuteurs. Les émotions actées sont généralement hors contexte, on demande alors aux locuteurs d'exprimer des émotions particulières sur un support lexical figé. L'avantage de ce type de corpus est d'obtenir un grand nombre de données très prototypiques mais hors contexte. Parmi ces corpus, on peut citer :

- le Danish Emotional Speech Database : en danois, 4 sujets lisent 2 mots, 9 phrases, un extrait avec 5 émotions différentes (colère, joie, tristesse, surprise, neutre) [Engberg et al. 97].
- le Berlin Dataset : en allemand, 10 locuteurs lisent chacun 10 phrases suivant 7 émotions (Colère forte, ennui, dégoût, peur-panique, joie, tristesse, neutre) [Burkhardt et al. 05].

Ces corpus actés hors contexte sont très utiles pour évaluer des modèles, cependant ils sont inutilisables dans des contextes plus spontanés. En effet, plusieurs études montrent qu'il y a une grande différence entre le fait d'acter une émotion sans avoir de données contextuelles et le fait d'exprimer une émotion de manière naturelle. Cette différence se fait à la fois au niveau phonatoire comme le montre [Erickson et al. 06] sur la tristesse et au niveau des descripteurs acoustiques utilisés pour construire les modèles émotionnels notamment sur la colère [Tahon and Devillers 10].

D'autres types de bases de données actées sont collectées en contexte. Dans ce cas, le locuteur s'imagine dans une situation et est libre du support lexical. Les émotions collectées restent alors prototypiques mais peuvent être utilisées dans des contextes spontanés.

Bases de données réalistes Obtenir des émotions spontanées est une tâche très complexe. Plusieurs méthodes ont été expérimentées : regarder des clips vidéos, doubler des scènes de films (corpus CINEMO [Rollet et al. 09], Vera-Am-Mittag Corpus [Grimm et al. 08]), imaginer des situations dans lesquelles le sujet doit réagir spontanément. Mêmes si ces protocoles permettent d'obtenir des émotions plus réalistes que celles actées, nous ne pouvons pas être sûrs de leur spontanéité réelle. Notamment l'inter-

action est totalement absente. Les corpus réalistes collectés sont généralement basés sur une interaction humain-humain ou humain-machine.

Les corpus vocaux en interaction H-H les plus courants sont ceux collectés en centre d'appel. Cette méthode permet d'avoir un grand nombre de données ainsi que de nombreux locuteurs. Parmi les données collectées en centre d'appel, nous pouvons citer quelques références représentatives :

- corpus EmoVox, enregistrement en centre d'appel de service après vente [Vaudable et al. 10], humain-humain
- corpus CEMO, enregistrement en centre d'appel d'urgence [Devillers et al. 05b], humain-humain.

Par la suite, nous nous concentrerons sur les corpus collectés à partir d'interactions H-M. Ces corpus sont généralement enregistrés à partir de scénarios, les émotions obtenues alors ne sont pas génériques mais spécifiques au contexte dans lequel elles ont été exprimées. La machine peut être en fonctionnement autonome ou contrôlée par un opérateur humain. Lorsque la machine est controllée par un opérateur, on appelle ce dernier le "Magicien d'Oz" ; l'opérateur n'est pas visible par le participant, mais il agit sur la machine comme si elle fonctionnait de manière autonome en connaissant l'ensemble du contexte dans lequel à lieu l'interaction. Il est souvent situé derrière une vitre sans teint (sinon on peut retransmettre sur écran la scène), ce qui lui permet de voir le contexte, sans être vu. Ce type de protocole permet de mettre en place une situation où l'expérimentateur n'apporte pas de biais dans l'interaction. Ces interactions ont lieu dans des conditions très contrôlées en fonction de l'application que l'on vise par la suite. Les interactions homme-machine les plus typiques sont collectées en laboratoire. La tendance actuelle est au développement des corpus d'interaction entre un ou plusieurs humains et un robot ou agent virtuel. Ces corpus peuvent être collectées in situ, comme c'est le cas du corpus AIBO [Batliner et al. 04], enregistré au cours d'une interaction entre AIBO, le robot de Sony, et un enfant (entre 10 et 13 ans) dans deux environnements (une salle standard d'un premier collège et un gymnase d'un deuxième collège).

Ne sont cités ici que les plus représentatives des bases de données émotionnelles, on peut trouver des catalogues complets chez [Ververidis and Kotropoulos 03], [Cowie et al. 05], [Schuller et al. 09a], [Schuller et al. 11a]

Aujourd'hui, la plupart des bases de données exploitent plusieurs modes de production des émotions : l'audio est quasiment systématiquement enregistré, contrairement aux autres modes : vidéo (expression faciale, gestuelle, postures) ou signaux physiologiques (rythme cardiaque, sudation, images cérébrales). Il semblerait que la parole soit le support principal de l'expression des émotions alors que les autres canaux peuvent apporter des informations suplémentaires nuançant la décision, ou la contredisant. Par la suite, nous n'utiliserons que le canal audio même si d'autres supports ont été collectés.

1.2.2 Bases de données locuteur

Bases de données pour l'identification du locuteur Les corpus actuellement utilisés pour l'identification du locuteur sont généralement des corpus de grande taille avec plus d'une centaine de locuteurs. Les bases de données sur l'identification du locuteur sont globalement plus simples à collecter que celles contenant des émotions. C'est le cas des corpus utilisés avec succès dans le cadre de campagnes d'évaluation internationales

sur la reconnaissance du locuteur organisées aux Etats-Unis par le NIST depuis 1996, principalement sur des enregistrements collectés par le LDC. Les informations contenues dans ces corpus sont principalement le genre et l'identifiant anonyme correspondant au locuteur.

- conversations téléphoniques (en grande majorité),
- entretiens (dans un bureau avec différents microphones),
- séminaires (évaluation CLEAR).

Les technologies liées à la reconnaissance du genre et du locuteur (vérification ou identification) sont aujourd'hui relativement avancées. Les challenges soulevés par la communcauté scientifique portent sur la diversité des canaux (téléphone, type de microphone, distance au micro, type de salle, etc.) dont la variabilité acoustique est bien plus importante que celle liée aux locuteurs (enregistrements ambiances Switchboard). Mais également sur des segmentations aveugles en locuteurs ("*diarization*") ou groupes de locuteurs ou sur l'utilisation des technologies existantes sur de la parole émotionnelle. Ce dernier aspect étant encore très récent. De récentes campagnes NIST ont été réalisées pour étudier l'influence de l'effort vocal sur la reconnaissance du locuteur.

Bases de données paralinguistiques (âge et genre) L'identification du locuteur est une tâche relativement bien traitée aujourd'hui, alors que l'identification d'informations paralinguistiques telles que le genre ou l'âge est bien moins étudiée. Le challenge Interspeech 2010 sur les données paralinguistiques a permis entre autres de combler cette lacune [Schuller et al. 10b]. Les principaux corpus de données paralinguistiques sont :

- aGender [Burkhardt et al. 10] : un peu moins de 1000 participants ont été enregistrés sur 47h de parole téléphonique en allemand sans contrainte lexicale. Quatre groupes d'âge ont été consitués : des enfants (<14ans), des jeunes (entre 15 et 24ans), des adultes (entre 25 et 54 ans) et des seniors (>55ans).
- Vocal Aging Database (UF-VAD) [Harnsberger et al. 08] : 150 locuteurs ont été enregistrés lors de la lecture d'un texte en anglais américain. Trois groupes d'âge ont été consitués : jeunes (entre 18 et 29ans), adultes (40 à 55ans) et seniors (62 à 92ans).

1.2.3 Bases de données de personnalité

L'étude de la personalité d'un locuteur est un champ de recherche assez récent dans le domaine de l'informatique, il a surtout été étudié en sciences humaines ou sciences de la vie. Cependant, avec l'émergence des agents virtuels et robots sociaux, la reconnaissance automatique d'une personalité est aujourd'hui devenue essentielle pour tous les travaux touchant de près ou de loin aux interactions (entre humains ou avec des machines). Le Challenge Interspeech 2012 sur la personalité [Schuller et al. 12b] met en avant cet intérêt de la reconnaissance automatique de certains aspects de personnalité. Il faut souligner un point important, la reconnaissance automatique d'une personnalité est basée sur des catégories (dans le cadre du challenge, les catégories OCEAN (O : ouverture aux autres, C : caractère consciencieux, E : extraversion, A : agréement, N : nevrotisme [Wigging 96]) ont été annotées) et ses applications peuvent avoir des conséquences humaines et sociales non négligeables. Ces travaux doivent absolument être menés en parallèle d'une réflexion éthique.

- Speaker Personality Corpus (SPC) [Mohammadi et al. 10] : environ 300 participants ont été enregistrés sur des radios de langue française.

Ce domaine est aujourd'hui en pleine effervescence, et plusieurs autres facteurs sont étudiés en ce moment par la communauté : la fatigue vocale, le stress [Fernandez and Picard 03], certaines pathologies comme l'alcoolisme, etc.

1.3 Stratégie d'acquisition de nos corpus

Afin de pouvoir évaluer l'influence des variabilités que sont l'environnement, le locuteur, son émotion et la tâche, nous avons collecté plusieurs corpus dont les caractéristiques sont au plus près possible du contexte défini plus haut. C'est-à-dire qu'ils respectent les conditions suivantes :

- un nombre suffisant de locuteurs (au moins une vingtaine),
- enfants, personnes âgées, ou personnes en perte d'autonomie (par exemple malvoyants),
- interaction (simulée, induite ou réelle) avec un robot : robot-assistant ou robot-joueur,
- environnements acoustiques différents d'un corpus à l'autre,
- présence de parole neutre et émotionnelle,
- conditions d'enregistrement (matériel) similaire d'un corpus à l'autre.

Les corpus de parole émotionnelle spontanée sont relativement difficiles à collecter, à la fois pour des raisons éthiques (par exemple on ne peut pas faire peur à quelqu'un), et pour des raisons de contexte. Il est très compliqué pour une personne d'exprimer une émotion spontanément si elle n'est pas plongée dans un contexte plausible ou proche d'elle. Lorsque c'est le cas, le participant s'appuie sur sa propre expérience pour exprimer un ressenti. Si aucun contexte n'est proposé, alors les émotions exprimées sont plus prototypiques, il y a moins de variabilité entre tous les participants.

La collecte des corpus est une tâche très importante, en coût et en temps. Afin d'optimiser les collectes, les scénarios et protocoles mis en place ont servis à plusieurs études. Elle a été réalisée en équipe avec notamment les membres du thème "Dimensions affectives et sociales des interactions parlées", à savoir Agnès Delaborde (étude du comportement du robot lors d'une interaction), Mariette Soury (étude des émotions chez des personnes atteints de pathologies particulières) et Clément Chastagnol (étude des émotions lors d'interaction avec des agents virutels et autres machines). Les annotations ont été réalisées par des experts en parole et entraînés sur les corpus collectés (Julietta Lencina, Caroline Benoît, Virgine Mouilleron, Nicolas Rollet).

1.3.1 Les corpus ROMEO

Les corpus ROMEO sont ceux que nous avons collectés dans le cadre du projet ROMEO pour répondre à l'application finale et qui nous ont permis d'étudier les facteurs de variabilité. Ces corpus sont enregistrés dans des conditions très similaires de celles définies dans le contexte, ils nous seront donc très utile pour évaluer nos systèmes d'identification du locuteur et de reconnaissance des émotions. Nous en avons enregistrés cinq correspondant à des types de locuteurs différents, des scénarios différents (assistance à la

Corpus	Lieu d'enregistrement	Age (#Genre)	Données	Réverbération
NAO-HR1	I-room, salle expérimentale du LIMSI (Orsay)	8-13 ans (6G, 6F)	spontané / acté	faible
IDV-HH	Appartement de la résidence St Louis (Paris)	23-79 ans (11H, 17F)	spontané / acté	très élevée
NAO-HR2	Centre aéré CESFO (Orsay)	6-10 ans (6G, 6F)	spontané / acté	élevée
IDV-HR	Appartement témoin, Insitut de la Vision (Paris)	28-80 ans (11H, 11F)	spontané / acté	faible
JEMO	Bureau, LIMSI	24-50 ans (27H, 35F)	acté	faible

TABLE 1.2 – Données caractéristiques des corpus ROMEO (H : homme, F : femme ou fille, G : garçon)

personne, jeux avec des enfants) et dans des environnements acoustiques différents (bureau, chambre, pièce à vivre). Tous les corpus ont été enregistrés avec un micro-cravate de bonne qualité (AKG PT40 Pro Flexx). Les caractéristiques de ce micro montrent que sa directivité est de type cardioïde, ce qui implique que le champ diffus peut être enregistré au même titre que le champ direct provenant de la voix. Les signaux audio sont échantillonnés (ou rééchantillonnés) à 16 kHz ce qui limite l'étude en hautes fréquences.

IDV-HH : corpus pour la reconnaissance du locuteur Le corpus IDV-HH [Tahon et al. 10] a été enregistré dans un appartement disponible de la résidence St Louis (11 rue Moreau, 75012 Paris) du 9 au 27 octobre 2009. Ce corpus contient 1 h 11 min 48 s de données actées et spontanées sur 28 locuteurs (11 hommes et 17 femmes) de 23 à 79 ans (une fois segmenté). Pendant une séance d'enregistrement, le participant doit d'abord répéter une série de mots afin de constituer un corpus de mots pour la reconnaissance vocale dans le cadre du projet ROMEO. Il doit ensuite aider l'équipe du LIMSI à améliorer son système de reconnaissance des émotions. Pour cela, il doit s'imaginer dans une situation qui lui est proposée (réveil du matin en forme, en mauvaise santé, urgence, déprime, bruits suspects). Ces situations se rapprochent du scénario final du projet ROMEO. Elles visent à induire chez le participant des émotions comme la tristesse, la peur, la douleur, la joie, le contentement ou l'agacement. Le système de reconnaissance des émotions est en fait piloté par un expérimentateur (ou Magicien d'Oz, *Wizard of Oz,* WoZ). Face au participant, le système piloté en WoZ, détecte l'émotion exprimée. De manière à inciter le participant à acter un peu plus une émotion, le WoZ peut se tromper, ou l'expérimentateur demander de recommencer. Les données récoltées seront donc actées (dans le cadre d'un scénario) ou spontanée (dans le cadre d'une interaction avec l'expérimentateur).

FIGURE 1.4 – deux enfants en interaction avec NAO lors de la collecte du corpus NAO-HR1

NAO-HR1 : corpus pour la détection des émotions sur les voix d'enfants Le corpus NAO-HR1 [Delaborde et al. 10] a été enregistré dans l'I-room (salle d'expérience du LIMSI – Orsay), entre le 23 septembre et le 7 octobre 2009. Ce corpus contient 31 min 7 s de données actées, spontanées et chantées sur 12 locuteurs (8 garçons, 6 filles) entre 8 et 13 ans, (une fois segmenté). Pendant une séance d'enregistrement, deux enfants (amis, frères et sœurs) jouent avec le robot Nao. Un maître du jeu est présent afin d'inciter les enfants à exprimer des émotions et permettre le bon déroulement de l'enregistrement. Un WoZ dirige Nao depuis une salle cachée. La séance se déroule en trois temps : jeu de question-réponses, jeu des chansons et jeu des émotions. Dans le premier jeu, un des joueurs (enfant1, enfant2 ou NAO) pose une question et les deux autres doivent trouver la bonne réponse, ce scénario permet d'induire des émotions spontanées. Dans le second jeu, les enfants doivent fredonner une chansonnette connue. Enfin dans le troisième jeu, le robot NAO demande à chacun des joueurs d'acter une émotion en contexte, cela permet d'obtenir des émotions induites. Les enfants actent une émotion proposée par Nao, et ce dernier doit la reconnaître.

IDV-HR : corpus pour la détection des émotions sur les voix de personnes en perte d'autonomie Le corpus IDV-HR [Tahon et al. 11] a été enregistré dans l'appartement témoin de l'Institut de la Vision (11 rue Moreau, 75012 Paris) du 11 au 16 octobre 2010. 22 locuteurs (11H, 11F) de 28 à 80 ans ont participé à cet enregistrement. La durée totale après segmentation est de 4 h 7 min et 43 s. Lors d'une séance d'enregistrement le participant est assis face à NAO (figure 1.5). Un expérimentateur est dans la salle, il l'accueille et propose un questionnaire. Un Magicien d'Oz commande le robot depuis une salle cachée. Le robot propose au participant une série de scénarios proche du réveil du matin avec différents états de santé (en forme, en mauvaise santé, urgence, déprime, joie). Chaque série est jouée plusieurs fois, NAO ayant des comportements à chaque fois différents (directif, dubitatif, encourageant, aimable, neutre, empathique). Le participant se place dans le contexte du scénario et essaie d'exprimer ses émotions de façon à se faire comprendre par le robot. Une séance complète permet de recueillir des émotions actées (scénarios) et spontanées (questionnaire).

NAO-HR2 : corpus de jeu sur les histoires interactionnelles avec des enfants et adultes NAO-HR2 [Tahon et al. 12b] est un corpus de voix de deux joueurs en interaction avec le robot NAO. Ce corpus dure 21 min 16 s après segmentation. 12 enfants de 6 à 11 ans ont été enregistrés. 4 adultes ont été enregistrés sur le même protocole afin de pouvoir étudier les différences liées à l'âge. Les enfants jouent par paire à ce qu'on appelle les histoires interactives 1.6. Une session de jeu consiste en 3 sections : d'abord le robot explique les règles en proposant des exemples, dans un second temps a lieu le jeu lui-même et finalement l'expérimentateur propose un questionnaire à chacun des joueurs. Dès que le robot commence à parler, l'expérimentateur n'intervient plus dans le déroulement de l'interaction entre NAO et les deux enfants. Un panneau de correspondance entre des mots et des émotions est situé derrière le robot. Dans la deuxième section, Nao raconte une histoire (ici *les trois petits cochons*). Au cours de l'histoire, si Nao prononce un des mots présent sur la tableau, il s'arrête de parler et un des joueurs doit acter l'émotion correspondante au mot. Si le robot détecte la bonne émotion, le joueur gagne un point. Pour l'enregistrement de ce corpus, il n'y a pas de détection automatique des émotions mais une entrée en WoZ par expérimentateur.

1.3.2 Les autres corpus utilisés pour nos études

JEMO : corpus de test et de démonstration Ce corpus [Brendel et al. 10] a été enregistré en laboratoire pour obtenir des émotions "réalistes" en contexte de jeu dans le cadre du projet ANR Affective Avatar. Le jeu consistait à faire reconnaître à la machine une émotion (colère, joie, tristesse, peur ou un état neutre) sans qu'aucun contexte ne soit indiqué. Les émotions collectées sont alors prototypiques. Le support lexical est totalement libre. Il a été enregistré en décembre 2010 au LIMSI. Sa durée totale est de 29 min. 62 locuteurs ont participé à l'enregistrement (27H et 35F).

Autres corpus Comme nous avons dit en 1.2.1, la collection de données réalistes étant une tâche relativement difficile et très dépendante du contexte et du scénario, les corpus peuvent être importants en durée mais pauvres en nombre de locuteurs. Ces corpus atteignent rarement la cinquantaine de locuteurs. Il peut alors être intéressant d'utiliser

FIGURE 1.5 – Dissposition du robot NAO et du matériel pour la collecte du corpus IDV-HR (haut) et participant en interaction avec le robot (bas)

FIGURE 1.6 – Interaction entre deux enfants et NAO lors de la collecte du corpus NAO-HR2

d'autres corpus collectés par d'autres membres de la communauté. Cette opération peut avoir plusieurs objectifs : étudier la robustesse des modèles créés sur un corpus ROMEO en les testant sur un nouveau corpus ; étudier l'influence de nouvelles caractéristiques (locuteurs, environment, émotions) sur les descripteurs acoustiques ou les performances de la détection ; ou encore les agglomérer pour en faire des modèles robustes à des conditions assez différentes.

Les autres corpus que nous utiliserons sont :

- CEMO [Devillers et al. 05b, Devillers and Vidrascu 06], corpus call-center enregistré dans un centre d'appel d'urgence en français (colère, peur, urgence, soulagement, neutre),
- CINEMO [Rollet et al. 09], corpus semi-acté enregistré à partir de séquences de films en français (colère, joie, peur, tristesse, neutre),
- AIBO [Steidl et al. 09], corpus de voix d'enfant jouant avec le robot Aibo de Sony en allemand (colère, empathie, neutre),
- SPC [Mohammadi et al. 10], challenge personnalité Interspeech 2012 (personnalité OCEAN),
- Comparse (en cours de traitement, voir section 2.2.4.6), corpus de stress dans la voix lors d'une prise de parole en public (projet ANR Comparse), avec différents types de stress (voir section 2.2.4.6).

1.4 Conclusion

Définir une émotion est une tâche extrêmement complexe et pluridisciplinaire. Plusieurs théories émotionnelles ont vu le jour, les plus connues étant la définition des émotions suivant des catégories, ou plutôt suivant des dimensions. Aujourd'hui, la plupart des applications utilisent un mélange de ces deux théories. La théorie de l'évaluation développée par Scherer définit le phénomène émotionnel comme une succession temporelle d'événements distincts. L'avantage de cette théorie est de prendre en compte un certain nombre de paramètres notamment la temporalité. Une théorie assez récente, le modèle de lentille, adaptée et développée par Bänziger permet de mettre en relation le locuteur avec le percepteur lors de la définition d'une émotion. Cette théorie nous semble très intéressante puisqu'elle met place le phénomène émotionnel au coeur de la communication

entre deux entités.

Pour être étudiées plus en profondeur, les émotions sont souvent collectées en bases de données. Nous avons choisi de ne présenter que les bases de données audio. Il faut d'abord rappeler que la plupart des corpus émotionnels aujourd'hui sont construits à partir de données prototypiques sur peu de locuteurs (moins de 10). La tendance des dix dernières années est d'enregistrer des émotions plus réalistes sur un relativement grand nombre de locuteurs (approcher les 50 locuteurs). Qui dit plus réalistes, dit également plus spécifiques et donc dépendantes d'un scénario. Nous avons également présenté des corpus de locuteurs, avec des enregistrements spécifiques pour l'âge ou le genre, et des corpus de personalité.

Nous avons collecté quatre corpus émotionnels réalistes dans le cadre du projet ROMEO. Ces corpus s'inscrivent dans la continuité de la tendance actuelle sur les bases de données émotionnelles. Cependant, ils ont l'avantage de s'appuyer tous sur des scénarios en interaction dans des conditions acoustiques différentes avec des locuteurs de différents types. Un cinquième corpus d'émotions prototypiques a été ajouté dans le contexte de ROMEO afin de pouvoir comparer des données réalistes avec des données plus stéréotypées. L'ensemble de cinq corpus forme un tout original et intéressant pour l'étude acoustique des émotions dans un contexte d'interaction humain-machine (voir le récapitulatif en Annexe A, A.1). Un des atouts de notre travail est d'avoir participé aux enregistrements, ce qui permet d'avoir une connaissance complète du contenu émotionnel, linguistique et paralinguistique des corpus utilisés par la suite.

La plupart de ces corpus (ROMEO et Comparse) seront également étudiés suivant un nouvel angle de recherche dans la thèse à venir d'Agnès Delaborde.

Chapitre 2

Annotation des émotions, contenu des corpus

Nous avons vu dans le chapitre précédent différentes théories émotionnelles. Nous avons vu que cette définition dépendait très fortement du contexte. Revenons à notre fil directeur musical, soit la réalisation d'une oeuvre musicale : comment la décrire le plus objectivement possible ? La première étape consiste à définir un certain nombre d'étiquettes possibles dans le contexte de cette réalisation (virtuosité, timbre, interprétation, etc..). La seconde étape serait de demander à un grand nombre de juges de décrire l'extrait musical avec les étiquettes disponibles et chacun suivant sa propre perception. On peut penser que plus le nombre de juges est grand, plus la subjectivité de chacun disparaît au profit d'une moyenne.

Nous allons procéder de même pour annoter les données émotionelles brutes. Le choix des étiquettes est orienté à la fois par les théories émotionnelles et par les utilisations technologiques qui seront faites de ces annotations. Cette étape d'annotation est primordiale dans l'analyse des émotions, puisque c'est elle qui définit l'ensemble des résultats futurs. Les contextes socioculturel et psychologique du récepteur (l'auditeur ou l'annotateur) entraînent une sensibilité différente dans la perception des émotions. La tâche d'annotation des émotions est également rendue difficile par la complexité du message oral communiqué par l'émetteur (le locuteur). Selon Scherer [Scherer et al. 80], la parole émotionnelle est conditionnée par deux effets pouvant donner lieu à des manifestations contradictoires : une excitation physiologique accrue "pousse" les vocalisations dans une certaine direction (effet push), alors que les tentatives conscientes de contrôle les "tirent" dans une autre direction et consistent en l'adoption de styles de langage culturellement acceptés (effets pull). Les différentes théories de la communication [Chung 00] ainsi que le modèle de lentille détaillé au chapitre 1, témoignent également de cette complexité à l'émission et la réception.

2.1 État de l'art : descripteurs émotionnels

La description des émotions soulève quatre principales difficultés : l'aspect dynamique, la possibilité d'émotions complexes (mélangeant plusieurs catégories émotionnelles), la grande dépendance au contexte, et à la manière dont chacun exprime ses propres émotions. La description des émotions se base principalement sur les théories émotionnelles mentionnées en chapitre 1.

Un schéma d'annotation a été défini dans le cadre d'HUMAINE : "Multi-level Emotion and Context Annotation Scheme" (MECAS) [Devillers et al. 05b], [Devillers and Vidrascu 06], [Vidrascu and Devillers 05]. Il permet de représenter des émotions réalistes et complexes dans l'audio et la vidéo suivant une hiérarchie des annotations en fonction de la précision des informations et sera détaillé en section 2.1.4.2.

Les travaux menés sur l'annotation et le choix des descripteurs émotionnels ont été majoritairement réalisés sur le signal audio, mais beaucoup d'entre eux proposent également d'annoter les informations linguistiques [Craggs and Woods 04, Devillers and Vidrascu 06], gestuelles [Kessous et al. 10], faciales [Audibert et al. 08, Devillers et al. 06, Caridakis et al. 10], physiologiques, séparemment ou simultanément. Nous traiterons en détail uniquement les travaux qui ont pour principal support le canal audio.

2.1.1 Annotation perceptive

L'annotation peut se faire de plusieurs manières : la plus courante et la moins coûteuse est de faire annoter les données émotionnelles par un petit nombre d'annotateurs experts. Experts signifie qu'ils sont formés pour l'annotation, qu'ils sont habitués avec cette tâche. On peut aussi utiliser un nombre beaucoup plus important d'annotateurs non-experts en proposant une annotation rémunérée via internet (ou *crowd sourcing*). L'inconvénient de cette méthode est la fiabilité des annotateurs (langue maternelle, attention) et l'impossibilité de souligner une perception biaisée de l'un d'entre eux. Un mélange de ces deux méthodes peut être envisagé ; valider l'annotation des experts par un test perceptif sur un échantillon représentatif du corpus.

On peut également demander à l'acteur ou au locuteur qui s'exprime d'annoter lui-même l'émotion qu'il a exprimée. On se situe alors plutôt sur du ressenti et moins sur de la perception. Il est relativement simple de demander à un acteur de s'annoter après chaque phrase qu'il prononce, cependant, dans le cas de la collecte de données réalistes, on ne peut demander au participant de s'annoter à chaque phrase. L'auto-annotation s'effectue alors a posteriori et demande un effort important de mémoire de la part du locuteur.

Ces deux types d'annotation sont forcément différentes, dans le cas de l'auto-annotation, on n'avoue pas exactement ce que l'on ressent consciemment ou inconsciemment. Le sujet va avoir tendance soit à masquer, soit à exagérer les émotions ressenties.

2.1.1.1 Mesures d'accord inter-juges

Si les données émotionnelles sont annotées par plus d'une personne, c'est pour essayer d'atteindre un consensus sur l'annotation des émotions perçues. Afin d'estimer le niveau d'accord entre les différents annotateurs ou juges, la mesure de kappa [Cohen 60] est souvent utilisée (eq. 2.1). P_o correspond à la proportion d'accord effectivement trouvée et

P_a à la proportion d'accord pour une annotation aléatoire.

$$\kappa = \frac{P_o - P_a}{1 - P_a} \tag{2.1}$$

Le kappa est égal à 1 lorsque les juges annotent avec les mêmes étiquettes. Dans [Callejas and Lopez-Cozar 08], plusieurs études ont été menées afin d'évaluer la robustesse de la mesure d'accord à la similarité entre les étiquettes, aux annotateurs, et à l'ajout d'informations de contexte. L'ajout d'information de contexte permet d'améliorer les scores d'agréement inter-juges.

2.1.1.2 Biais de l'annotation

L'annotation consensuelle est prise pour norme dans la construction des modèle, cependant, il ne faut pas oublier que cette norme ne correspond qu'à une moyenne. Selon Batliner [Batliner et al. 04], 10 annotateurs experts serait suffisant pour atteindre un consensus satisfaisant. L'annotation n'est peut-être pas identique sur tous les segments d'un même corpus pour un annotateur donné, son attention se relâche, il se familiarise avec les émotions exprimées. De plus il y a toujours cette différence entre émotion ressentie et émotion perçue. Les annotations à plusieurs juges ne permettent pas d'étudier les émotions ressenties. Une vérité absolue de l'émotion ressentie pourrait être obtenue par une fusion de signaux physiologiques : activité cérébrale, rythme cardiaque, sudation, etc. Cependant ces signaux ne peuvent être collectés sans biaiser la spontanéité du phénomène émotionnel.

2.1.2 Annotation du contexte et des informations locuteur

2.1.2.1 Description du contexte

Il a été montré que la variabilité dûe au contexte (support d'enregistrement, environement acoustique) était bien plus importante que celle liée au locuteur et donc à ses états émotionnels. Il est donc toujours très intéressant d'annoter le contexte. Il peut se définir comme tout ce qui a lieu en dehors du participant et des émotions qu'il exprime. Le contexte inclut donc le type de lieu, le matériel utilisé pour l'enregistrement, la distance entre le participant et le micro, la présence de bruits extérieurs (et lesquels), disposition du/des participants. On peut apporter des éléments de contexte interactionnel souvent très pertinents sur la parole : à qui s'adresse le participant (expérimentateur, machine, autre participant, lui-même), dirige-t-il sa voix vers le microphone (parole audible, bruits micro), semble-t-il à l'aise. La plupart du temps la définition du contexte se borne à celui dans lequel est collecté le corpus. En effet, dans le cas de corpus acté, le contexte est très similaire d'un participant à l'autre. Par contre, dans le cas de corpus réalistes collectés avec des scénarios, le contexte peut être très variable d'une session à une autre.

A titre d'exemple, dans le cas du schéma d'annotation MECAS appliqué au corpus collecté en centre médical d'appel d'urgence [Devillers et al. 05a], le contexte est décrit suivant des descripteurs lexicaux, dimensionnels et d'évaluation au niveau d'un tour de parole :

- l'origine de l'appel (patient, centre),
- la raison de l'appel (urgence médical, information médicale),

– la qualité acoustique de l'enregistrement (niveau de bruit, intérieur/ extérieur, téléphone mobile, fixe, radio).

2.1.2.2 Informations locuteur

Segmentation en tour de parole Les données brutes peuvent être enregistrées dans des conditions très variées, par exemple au cours d'un questionnaire, d'une interaction avec un expérimentateur, ou une machine. Il est alors nécessaire de segmenter en tour de parole afin de récupérer des signaux de parole où un seul locuteur parle. Lors de cette étape de segmentation, peuvent être supprimé également les superpositions de parole (ou overlaps), les bruits extérieurs (claquement de porte), ou même les toux, si le locuteur touche le micro, les saturations micro, etc.

L'avantage d'une segmentation propre est d'obtenir des données fiables permettant une analyse acoustique fine de la parole émotive. Cependant elle peut avoir l'inconvénient de ne pas prendre en compte la réalité de la prise de son. Si le système de reconnaissance des émotions n'a pas de détection d'activité vocale fiable ou bien si elle ne permet pas de repérer les changements de locuteur, peut-être vaut-il mieux avoir de ces événements dans les modèles.

Descripteurs locuteur L'expression d'une émotion étant très dépendante du locuteur, il peut être intéressant de caractériser ce locuteur pour comprendre comment varie l'expression d'une même émotion suivant plusieurs personnes. En fonction des corpus, les types de locuteurs sont plus ou moins variés. Les corpus de données actées n'ont généralement pas de variabilité forte des locuteurs puisqu'un petit nombre d'acteurs (2 à 10) sont enregistrés. La distinction du genre reste toutefois primordiale pour toute analyse acoustique de l'expression des émotions : les hommes et les femmes sont naturellement différents physiologiquement, ce qui implique des différences acoustiques (principalement au niveau de la fréquence fondamentale).

D'autres informations sont importantes pour comprendre les différences acoustiques de l'expression des émotions. Nous pouvons distinguer les annotations extralinguistiques des annotations paralinguistiques (état mental, état de santé).

Annotations extralinguistiques

– annotation agent/client [Devillers et al. 04, Vaudable et al. 10] dans un corpus de call-center : les enregistrements ne se font pas forcément par le même média, ce n'est pas non plus la même manière de s'exprimer (emploi de mot-clé, d'un ton standardisé, chez l'agent).
– annotation journaliste/interviewé [Vinciarelli et al. 10] sur corpus radiophoniques : les journalistes apprennent à s'exprimer en public suivant un certain ton, une manière très particulière.
– annotation de l'âge [Li et al. 10, Mohammadi et al. 10, Schuller et al. 10a] : la différence enfant/adulte relève presque plus de la question du genre que de celle de l'âge pour une simple question de morphologie. Cependant la catégorisation des locuteurs adultes suivant des groupes d'âge (jeunes, adultes, seniors) est de plus en plus utilisée en interaction homme-machine.

Annotations paralinguistiques

- annotation de l'état mental ou état de santé : certaines pathologies impliquent des comportements différents face à l'expression des émotions et à la communication en général (comme dans le cas des autistes [Ringeval et al. 08]). Cette information est très pertinente dans le cas des corpus d'assistance aux personnes.
- annotations paralinguistiques [Douglas-Cowie et al. 03, Douglas-Cowie et al. 07] dans les bases de données HUMAINE.

2.1.3 Unités temporelles pour l'annotation des émotions

Les émotions sont souvent exprimées ponctuellement dans le temps, elles peuvent également évoluer au cours d'un même tour de parole. On recense deux grandes tendances pour annoter la dynamique des émotions : (i) une annotation continue (ii) une annotation segmentale.

2.1.3.1 L'annotation continue

Ce type d'annotation permet de suivre l'évolution des émotions suivant un pas temporel défini. C'est le cas par exemple de l'outil d'annotation FeelTrace développé par Cowie [Cowie et al. 00] qui permet d'annoter les émotions en temps réel. A chaque instant, un certain nombre de catégories sont renseignées. L'inconvénient de cette méthode est son coût en temps : il faut faire un compromis entre la finesse de l'émotion annotée et le temps passé pour annoter [Vaudable 12, section 4.3].

2.1.3.2 L'annotation segmentale

L'annotation segmentale implique de définir un segment au cours duquel les étiquettes émotionnelles seront statiques. Un segment émotionnel se définit comme un segment émotionnellement homogène. Évidemment cette notion est subjective. La précision temporelle de l'émotion peut alors être plus fine qu'un simple tour de parole mais elle reste relativement grossière par rapport à une annotation continue. L'avantage de ce type d'annotation est évidemment le gain de temps, et donc on peut avoir plus d'annotateurs et/ou plus de finesse dans l'annotation de l'émotion.

Plusieurs études [Batliner et al. 11, Batliner et al. 10, Lacheret-Dujour and Victorri 02, Schuller et al. 11a, Clavel and Richard 10] ont été menées afin de définir l'unité adéquate pour l'annotation émotionnelle sur signal audio. On peut distinguer plusieurs niveaux :

1. le tour de parole : un seul locuteur parle,
2. le groupe de souffle
3. le segment émotionnel, notion intuitive définie par la personne qui segmente : l'émotion est fixe tout au long de la durée du segment. Ce type de segment est généralement également lexicalement cohérent,
4. la phrase, le groupe de mots, le mot : cette segmentation implique d'avoir une transcription fiable,
5. la syllabe, le phonème : cette segmentation implique également d'avoir une transcription fiable ou tout au moins un détecteur de syllabe ou phonème,

6. une durée fixe : de 10 ms à 5 s.

La question de la segmentation est très complexe et fondamentale pour la reconnaissance automatique des émotions. Elle implique des connaissances sur l'ancrage temporel des émotions, sur la durée nécessaire au cerveau humain afin de percevoir une émotion et sur la quantité de données nécessaire à la machine pour reconnaître l'émotion.

Quelle durée nécessaire pour la reconnaissance automatique des émotions ? Une étude intéressante menée lors d'une collaboration entre le chercheur Schuller et le LIMSI sur les corpus JEMO et CINEMO [Schuller et al. 10b], montre qu'une durée d'1 s serait nécessaire pour une reconnaissance satisfaisante des émotions sur ces corpus.

Une problématique intéressante sur la localisation temporelle du phénomène émotionnelle consiste à annoter l'ancrage de l'émotion [Grichkovtsova et al. 08]. C'est-à-dire précisément les instants temporels entre lesquels l'émotion peut être identifiée. L'ancrage ne correspond pas forcément à une accentuation ou proéminence dans la parole. Selon [Grichkovtsova et al. 09], l'émotion peut être identifiée très rapidement principalement grâce à la qualité vocale. Cependant cette notion d'ancrage est très dépendante de la notion de contexte. Une émotion est plus facilement et plus rapidement perçue lorsque des éléments de contexte sont apportés. Dans la situation, où l'on considère un segment émotionnel hors contexte, il est plus compliqué de définir précisement un point d'ancrage.

2.1.4 Annotation des informations paralinguistiques émotionnelles : des émotions fines aux macro-classes

L'annotation d'un corpus émotionnel a principalement deux objectifs : le premier est une approche sciences humaines de description des émotions présentes dans ce corpus afin de pouvoir les analyser, le second est une approche plutôt ingénieur de définition des classes qui permettront d'entraîner des modèles afin de construire un système de reconnaissance d'émotions. L'annotation fine permet de satisfaire le premier objectif de description tandis que l'annotation en macro-classe permettra aux systèmes de reconnaissance d'être plus performants. Ces macro-classes ne correspondent pas à des catégories émotionnelles réelles comme les émotions primaires mais regroupent un ensemble de catégories émotionnelles primaires et secondaires. Même si la performance des systèmes de reconnaissance guide la plupart des chercheurs en reconnaissance des émotions, elle ne doit pas nous faire oublier la réalité complexe décrite par l'ensemble des annotations.

2.1.4.1 Décrire et analyser les émotions

Suivant le contexte l'analyse des émotions demande une description plus ou moins fine. A la fois les catégories et les dimensions sont utilisées pour la description des émotions.

Descripteurs catégoriels L'approche catégorielle consiste en la dénomination des émotions par des labels lexicaux adaptés et prédéfinis. C'est la manière la plus intuitive pour décrire des émotions spécifiques, en utilisant des catégories issues du langage courant.

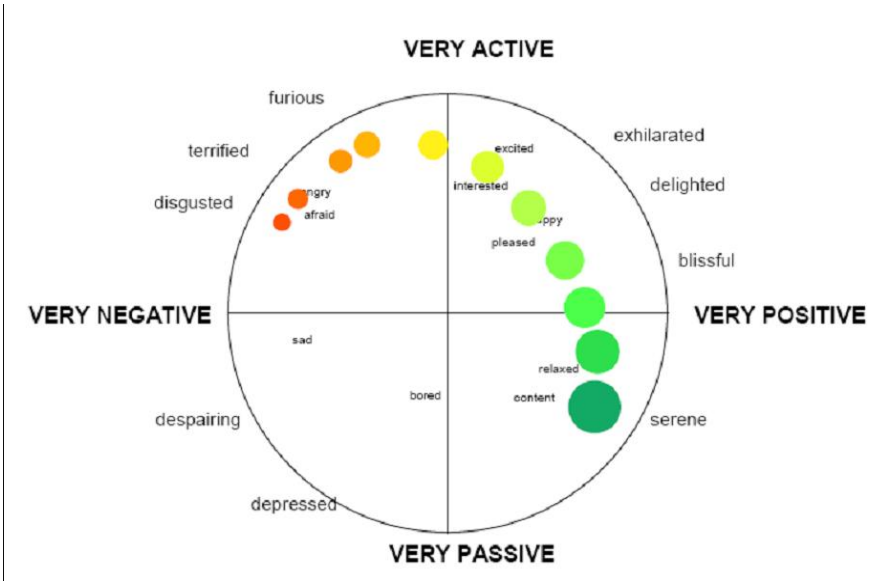


FIGURE 2.1 – Exemple d'une annotation continue avec l'outil FeelTrace, extrait de [Cowie et al. 00]

La définition de catégories émotionnelles consiste à tracer des frontières absolues dans l'espace perceptif. En fonction du corpus, un même label émotionnel représentera globalement la même émotion perçue, par contre il est fort possible que la manifestation acoustique de cette émotion soit très différente d'un corpus à l'autre. Par exemple, dans un corpus acté, la colère correspondra souvent à de la colère forte, explosive, alors que dans des corpus spontannés où le locuteur est en interaction avec un autre humain ou une machine, cette colère sera plus contrôlée et moins explosive.

Descripteurs dimensionnels Les dimensions les plus couramment utilisée sont valence et activation, mais le contrôle, l'intensité, la dominance ont été utilisée également. Dans une approche système, les dimensions sont discretisées afin de pouvoir utiliser une classification en classes. Les échelles de dimensions sont variables suivant les auteurs : de 1 à 7, de 1 à 5, de -2 à 2 ou même faible, moyenne, forte.

L'outil FeelTrace [Cowie et al. 00] permet d'annoter continuement (pas de discrétisation) suivant la valence et l'activation (figure 2.1).

2.1.4.2 Définition de macro-classes

L'approche ingénieur de la modélisation des émotions est surtout basée sur des critères de performances. En cela, il est nécessaire de réduire au maximum la complexité du phénomène émotionnel et des annotations. La question qui se pose est alors moins celle de la modélisation parfaite des émotions dans un contexte défini que la robustesse de la détection et classification des émotions.

Une approche pragmatique est de définir des macro-classes qui englobent un sous-ensemble d'émotions fines. Une hiérarchie peut être établie dans ces macro-classes.

1. émotionnel, état neutre,
2. positif, négatif, état neutre,
3. emotions primaires, ex : colère, joie, tristesse, peur, surprise, état neutre,
4. émotions secondaires (dont certaines émotions sont complexes, c'est-à-dire appartiennent à plusieurs macro-classes, ex : colère + joie donne de l'ironie),
5. émotions graduées, ex : colère forte, faible, moyenne,
6. autres dimensions : valence, activation, intensité, contrôle (elles peuvent être graduée : fort, faible, moyen, ou mesurées avec une échelle plus large).

En fonction du contexte dans lequel est collecté le corpus, les macro-classes seront différentes. Et la définition d'une catégorie pour un corpus est rarement reportable sur un autre corpus. Un des objectifs du réseau d'excellence HUMAINE est d'adopter des étiquettes consensuelles. Une liste de 20 catégories émotionnelles est disponible sur le site de l'association [1]. Tous les corpus utilisés pour l'analyse des émotions (chapitre 1) ont été annotés avec un certain nombre de catégories et/ou dimensions émotionnelles, le tableau 2.1 résume ces catégories. Le protocole d'annotation présente généralement un nombre élevé de catégories émotionnelles permettant de décrire au mieux un corpus.

Dans le cas du schéma d'annotation MECAS, un segment émotionnel sera représenté par une émotion Majeure (la première impression), une Mineure (permet de nuance l'émotion Majeure). Afin de minimiser le temps d'annotation, certaines dimensions abstraites sont déduites des émotions fines (par exemple la valence ou l'activation) certaines émotions fines restant ambigües (la surprise peut être positive ou négative). L'ensemble des étiquettes émotions sont organisées depuis une précision très fine, à un niveau plus macro. On peut également préciser si l'émotion perçue est ambigüe, évidente, conflictuelle (entre Majeure et Mineure). Grâce à des schémas d'annotation de ce type on peut analyser les émotions complexes [Mower et al. 09, Rollet et al. 09] comme l'ironie, les mélanges de positif/négatif, l'importance de certaines dimensions comme le contrôle ou l'activation potrès finesur certaines tâches.

2.1.5 Autres annotations paralinguistiques

Sans avoir recours à une transcription du signal de parole, certaines informations paralinguistiques peuvent être tout à fait pertinente à la fois pour préciser le contexte ou pour apporter des informations sur le locuteur. On peut citer les bruits de bouches, défauts de prononciation, accent régionaux, sociaux ou étrangers, qualité vocale (schéma d'annotation MECAS), mais bien d'autres étiquettes peuvent être utilisées pour l'annotation de la parole. Parmi elles, ont peut noter les affect bursts qui ne peuvent pas être annotés au même niveau qu'une émotion, mais qui vient apporter des informations complémentaires sur celle-ci.

Un catalogue synthétisant l'ensemble des annotations paralinguistiques peut être trouvé chez [Beller 09], on trouve aussi des annotations chez Douglas Cowie [Douglas-Cowie et al. 07].

1. http ://emotion-research.net/projects/humaine/ws/summerschool1/emotion%20words

Corpus	Langues	Année	Catégories émotionnelles	Acté/ Spontané
DES	Danois	2007	Colère, joie, tristesse, surprise, neutre	Acté
EMO-DB	Allemand	2005	Colère, ennui, tristesse, dégout, anxiété, neutre	Acté
SAL	Irlandais	2007	Valence, activation	Acté
eINTERFACE	Anglais	2006	Colère, dégoût, peur, joie, tristesse, surprise, neutre	Spontané
SUSAS	Anglais US	1997	Neutre, peur, stress	Mixte
SmartKom	Allemand / Anglais	2002	Colère, joie, surprise, impuissance, pensif, surprise	Spontané
VAM	Allemand	2008	Colère, joie, tristesse, dégoût, peur, surprise, neutre	Spontané
IrcamCorpus Expressivity	Français	2010	Neutre, colère, joie, peur, tristesse, ennui, dégoût, indignation, surprise, neutre	Acté
AIBO	Allemand	2008	Colère, empathie, motherese, neutre	Spontané

TABLE 2.1 – Principales catégories émotionnelles utilisés lors de l'annotation des corpus pour la reconnaissance des émotions

Sons non verbaux, respirations et voisement de la phonation	Effets de pitch et gutturaux	Restructurations	Affect bursts
inspiration	effet de pitch	long	rires
expiration	effet guttural	césure	pleurs
respiration nasale (reniflement)	autre effet que guttural ou de pitch	répétition	hésitation
bruit indéfini	non transcriptible		
bruits de bouche			
chuchotement,			
non voisé			

TABLE 2.2 – Annotations paralinguistiques, adapté de [Beller 09]

Le challenge Interspeech 2010 a permis de mettre au point un certain nombre d'annotations paralinguistiques consensuelles dans la communauté [Schuller et al. 12a]. Le tableau 2.2 synthétise les étiquettes utilisées pour l'annotation d'informations paralinguistiques relatives directement à la parole. Parmi les indices paralinguistiques peuvent également se retrouver des indices d'intention, de conversation, d'interaction, etc. [Schuller et al. 12a].

2.1.5.1 Annotations de personalités

La personalité du locuteur peut être un facteur intéressant pour les interactions homme-machine. L'annotation d'informations à caractère psychologique a commencé dans les années 2000 avec les corpus du Trinity College, Dublin [Douglas-Cowie et al. 03]. Ce domaine est aujourd'hui très en vogue avec par exemple le challenge 2012 Interspeech organisé par Schuller [Schuller et al. 12b]. Annoter la personalité suppose d'avoir des étiquettes définies et donc de définir des catégories correspondant à certains aspects de personalité. Les tests de personalités permettant de "mesurer" ces aspects, sont nombreux (Big Five Inventory [Plaisant et al. 10, John and Srivastava], mesures d'alexithymie TAS-20 [Bagby et al. 94a, Bagby et al. 94b], mesures de narcissisme NPI [Raskin and Hall 79], etc...) et peuvent être remplis soit par le candidat (ou le patient), soit par l'expérimentateur soit par un professionnel du soin. Ils permettent d'obtenir des scores suivant un certain nombre de dimensions. Les dimensions les plus classiques étant celles établies par Wigging [Wigging 96] : ouverture aux expériences, caractère consciencieux, extraversion, caractère agréable et névrotisme. L'utilisation de catégories ne doit pas faire oublier l'aspect extrêmement complexe de la personalité humaine.

2.1.5.2 Annotations de signaux sociaux

Breazeal [Breazeal and Aryananda 02] a travaillé sur l'annotation de dispositions affectives suivant le terme de Scherer : attention, approbation, neutral, interdiction, rassurant. Vinciarelli [Vinciarelli et al. 08] propose des annotations du signal social. L'objectif est ensuite de déterminer des marqueurs interactionnels.

2.1.6 Annotations linguistiques

Un point important pour ce chapitre concerne la transcription de ce qui a été dit. Cette transcription peut être automatique (en utilisant les systèmes de transcription automatique ou ASR) ou manuelle. Elle permet de collecter des informations linguistiques suivant des schémas d'annotation lexicaux [Craggs and Woods 04]. Les informations linguistiques sont très intéressantes pour reconnaître les émotions. D'après [Arunachalam et al. 01], les informations émotionnelles dans le texte se traduisent par l'utilisation de mots appartenant à un champ lexical particulier et d'altération grammaticale et syntaxiques (ex : "j'en ai marre"). Il a été démontré que le choix de l'annotation (automatique ou manuelle) influence énormément le système de reconnaissance des émotions lorsque celui-ci intègre des descripteurs linguistiques comme le nombre de mots [Clavel 07] ou l'utilisation de sac de mots [Vaudable et al. 10].

2.1.7 Les outils d'annotations

L'outil Transcriber [Barras et al. 00] permet d'annoter un dialogue dans son ensemble. Après une première phase de segmentation où l'annotateur définit les frontières des tours de paroles, overlaps, bruits extérieurs et segments émotionnels, la phase d'annotation se fait segment par segment. L'outil permet de définir des étiquettes à compléter pour chaque segment émotionnel. L'annotation se fait alors en contexte puisque l'annotateur a connaissance de ce qui a été dit avant et après le segment en cours.

L'outil ANVIL [2] [Kipp 01] est un outil conçu pour annoter le signal audiovisuel avant tout. Il propose une annotation multi-couche dont les catégories sont définies par l'utilisateur.

2.2 Corpus collectés : stratégie d'annotation des émotions et du contexte

La stratégie d'annotation que nous avons choisi pour annoter nos corpus est globalement identique pour tous les corpus. Elle dépend principalement des applications dans lesquelles les modèles liés au corpus seront utilisées. La segmentation et l'annotation ont été principalement réalisées avec l'outil Transcriber.

2.2.1 Unité pour l'annotation : le segment

Les annotations sont associées à un segment émotionnel, la segmentation est préliminaire à tout travail sur le corpus. Elle est réalisée par un seul expert pour des raisons de

2. http ://www.anvil-software.de/#

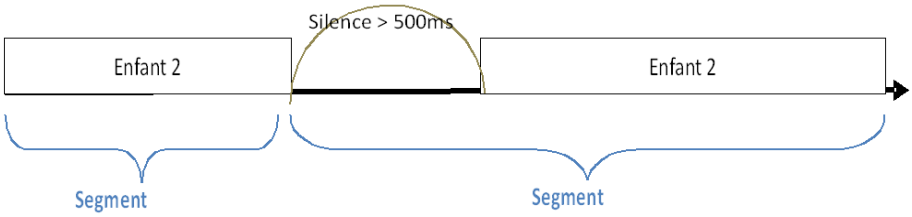


FIGURE 2.2 – Segmentation avec pause de plus de 500ms

gain de temps. On appelera session, l'enregistrement complet d'un même locuteur.

Les différentes sections de l'enregistrement sont segmentées et annotées en premier lieu (introduction, questionnaires, scénario). Puis les tours de parole sont segmentés (prise de parole d'un locuteur). Et enfin les segments émotionnels. L'étude mentionnée plus haut [Schuller and Devillers 10], montre qu'une durée de 1 s est suffisante pour la reconnaissance automatique des émotions. Il est donc préférable d'avoir des segments émotionnels de l'ordre de la seconde. Un certain nombre de règles ont été établies afin de définir précisement le segment émotionnel et d'éviter des cas atypiques :

- le segment contient la voix d'un seul locuteur (pas de superposition de voix),
- le segment ne contient pas de bruits trop importants (saturation, bruits de porte, bruits extérieurs, toux),
- le segment émotionnel ne doit pas excéder 5 s (afin d'éviter les segments trop longs), si il est véritablement plus long, le segmenter suivant des respirations, fin de phrase ou hésitations,
- le segment émotionnel ne doit pas être trop court (si possible durée supérieure à 0,5 s),
- le segment émotionnel ne doit comporter qu'une seule émotion,
- le segment ne doit pas comporter de pause de plus de 500 ms, ni de respiration, si il en existe une, couper le segment avant la pause (figure 2.2).

Ces règles permettent d'obtenir des segments émotionnellement homogènes, ne contenant qu'un seul groupe de souffle sur de la parole relativement propre.

Cette segmentation est très précise et demande beaucoup d'attention de la part de ceux qui réalisent la segmentation. Il arrive régulièrement que certaines des règles ne soient pas respectées, entre autres celles concernant les durées des segments. La figure 2.3 montre les différents niveau de signal (corpus, session, tour de parole et segment émotionnel) et les informations relatives à chaque niveau que l'on peut annoter.

2.2.2 Contexte et informations locuteur

Les corpus étant par essence relativement homogènes, on peut caractériser le contexte pour l'ensemble des sessions. Au niveau acoustique :

- type de salle dans laquelle se font les enregistrements,
- matériel utilisé,
- fréquence d'échantillonage,
- distance au microphone, etc.

Au niveau de la tâche :

neutre	colère	tristesse	joie/positif	peur	autres
neutre	colère	tristesse	joie	peur	ironie
intérêt	irritation	déception	positif	anxiété	surprise
	mépris		amusement	stress	provocation
	agacement		satisfaction	gène	excitation
			soulagement	inquiètude	maternage
			empathie		

TABLE 2.3 – Tableau d'équivalence entre macro-classes et émotions fines

- acté, spontané, mixte,
- support lexical fixé, libre,
- interaction humain-humain, humain-machine, pas d'interaction,
- contexte de jeu, d'assistance, de prise de parole en public, etc...

Par session, un questionnaire (oral ou écrit) permet de récupérer des informations importantes sur le locuteur. Ces informations serviront par la suite à interpréter les résultats émotionnels obtenus, à relativiser certaines émotions, mais aussi peuvent être utiles pour cerner la personalité de chaque participant (avec un questionnaire de personalité).

L'âge et le genre sont les principales données. Dans le cas des enregistrements des corpus IDV-HH et IDV-HR avec des personnes déficientes visuelles, une précision sur leur déficience leur est demandée. La situation personnelle permet également d'avoir des renseignements sur le locuteur : leur profession, leur situation maritale (célibataire, divorcé, vivant seul, en couple, avec/sans enfants). Le lieu d'habitation est également intéressant : résidence, maison de retraite, habitation particulière. On peut également demander les émotions les plus souvent exprimées par le participant afin de relativiser la proportion de certaines étiquettes. En fin de session, un questionnaire de retour d'expérience est posé : est-ce que le participant a bien compris les consignes, a-t-il compris l'intérêt de l'expérience, dans le cas d'une interaction homme-machine, est-ce que la machine a réagi de manière satisfaisante ?

Pour chaque segment, les informations de contexte sont annotées. Elles sont très importantes puisqu'elles permettent de relativiser ponctuellement le type d'émotions exprimées.

2.2.3 Annotation des émotions

2.2.3.1 Émotions fines et macro-classes

Lors que la phase d'annotation est terminée, il se peut que certaines étiquettes émotionelles ne soient que très peu représentées. On essaie alors de les regrouper avec des classes proches ou dans des macro-classes. De même le fait de n'avoir que deux annotateurs implique qu'il n'y a que très peu de consensus sur certaines classes particulières. Lorsque les deux annotations appartiennent à la même macro-classe, le segment est alors labellisé suivant cette macro-classe. Dans le tableau 2.3, les émotions fines classées dans "autres" correspondent à des émotions pour lesquelles la valence peut être soit positive, soit négative (cas de la surprise, l'excitation), ou des émotions complexes comme la provocation et l'ironie.

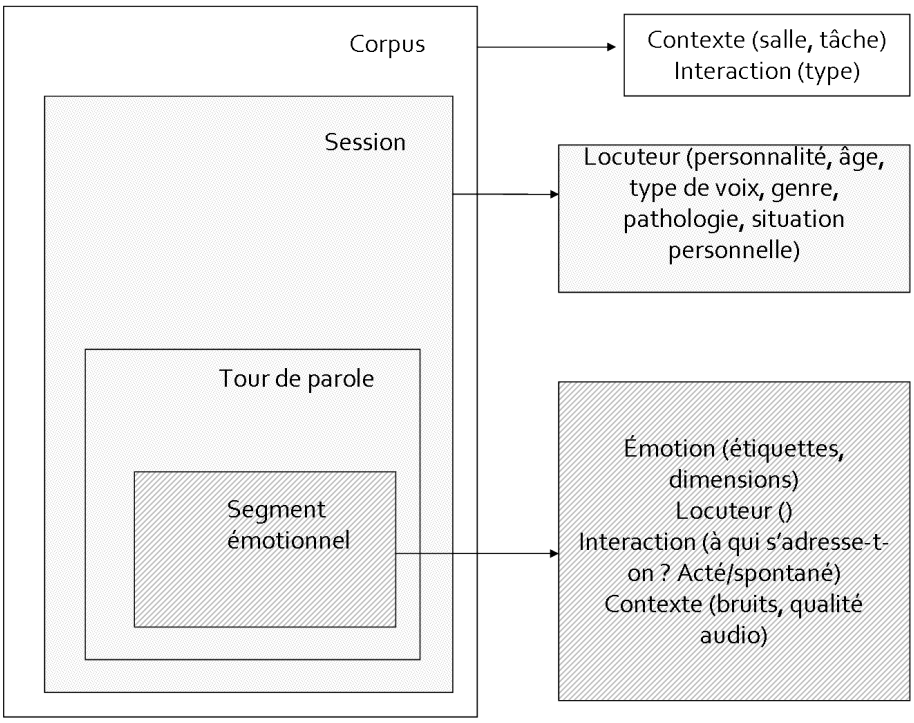


FIGURE 2.3 – Annotations et signal

2.2.3.2 Valence et activation

Les dimensions définies par la valence et l'activation sont relativement complexes à annoter. On choisira une annotation sur une échelle de -2 à 2. Pour certaines émotions, la valence est relativement claire. Le niveau d'activation est plus complexe : il ne correspond pas à l'intensité de l'émotion (qui est très fortement corrélé avec l'énergie contenue dans le signal) mais à la quantité de moyens phonatoires et articulatoires mis en oeuvre pour exprimer l'émotion, c'est-à-dire l'engagement du locuteur dans l'émotion exprimée. Une tristesse très forte correspondra à une activation forte alors que son intensité sera plutôt faible.

L'intérêt de faire annoter ces deux dimensions entre -2 et 2 est de pouvoir normaliser par rapport à l'annotateur. En effet, même si les annotateurs écoutent un échantillon servant de référence "neutre", leurs échelles peuvent varier. Pour recaler les échelles, il suffit de calculer la moyenne pour chaque des annotateurs et la mettre à 0. Le label négatif (resp. neutre, positif) sont alors attribués aux segments qui ont une valence entre -2 et -0.5 (-0.5 et 0.5 ; 0.5 et 2) en moyenne sur les deux annotateurs. Nous avons appliqué également ce principe sur l'activation.

2.2.3.3 Etiquettes émotionnelles utilisés

Afin de pouvoir être utilisable d'un point de vue de la classification, les annotations sont simplifiées au niveau du segment.

- macro-classe {neutre, colère, joie, tristesse, peur ou autre},
- valence {négatif, neutre ou positif},
- activation {actif, neutre ou passif},
- locuteur {nombre},
- genre {F ou M},
- éventuellement âge {nombre},
- éventuellement la qualité d'acteur {spontané, acté}.

Dans notre thèse, nous n'utiliserons aucune autre information au niveau d'un segment émotionnel. Par contre toutes les autres annotations peuvent servir à replacer certains segments dans leur contexte.

2.2.4 Contenu des corpus utilisés, données caractéristiques

Nous allons présenter dans cette partie plus précisement les corpus IDV-HR et NAO-HR1 que nous avons collecté. Les corpus IDV-HH et NAO-HR2 ont été traités de manière similaire.

2.2.4.1 Corpus IDV-HR

Le corpus IDV-HR [Tahon et al. 11] consiste en 22 sessions (une par locuteur) enregistrées dans l'appartement témoin de l'Institut de la Vision à Paris. Nous avons choisi de nous focaliser entre autres sur ce corpus parce qu'il regroupe des locuteurs ayant des qualités de voix très différentes dues principalement à leur âge. La segmentation et l'annotation a été réalisée par deux annotateurs experts.

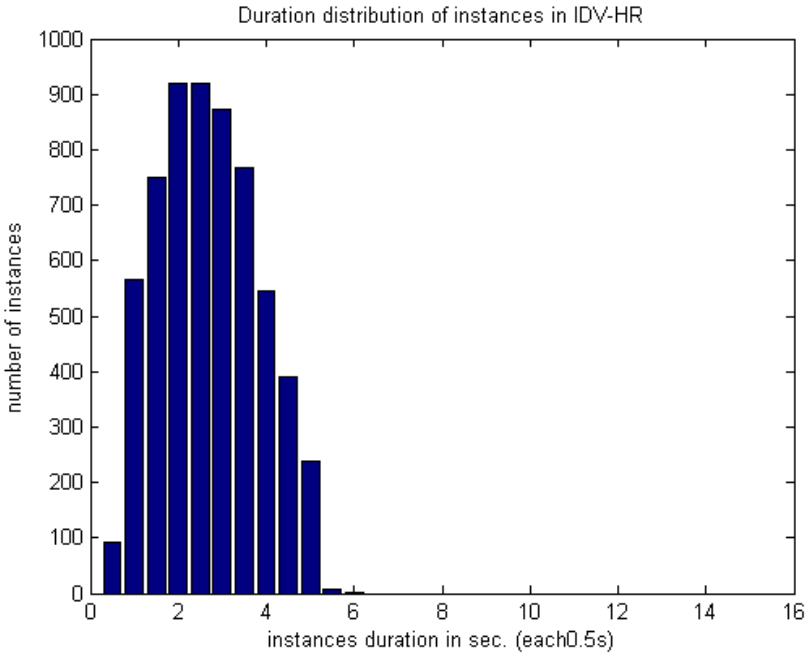


FIGURE 2.4 – Répartition des durées des segments sur IDV-HR

Organisation du corpus La collecte du corpus IDV-HR est basée sur l'alternance d'interactions entre le robot NAO et le participant et de questionnaires. Un premier questionnaire oral est proposé au participant afin de collecter des informations personnelles. Ensuite l'alternance de trois comportements du robot NAO positif (encouragement, empathie, amabilité), négatif (hésitant, neutre, directif) puis à nouveau positif suivi à chaque fois de questionnaires relatifs aux comportements du robot. La session se clôture par un questionnaire plus global sur l'aspect physique et comportemental du robot. Pour chaque comportement, 5 scénarios basés sur le lever du matin étaient proposés (en pleine forme, malade, déprimé, urgence, joyeux). Lors des phases d'interaction, la parole avait tendance à être plutôt induite (le sujet s'imagine dans une situation particulière), alors que lors des questionnaires, le sujet était bien plus spontané.

Après segmentation, le corpus représente 4 h 7 min 43 s. La durée moyenne des segments est de 2,45 s (min : 0,24 s, max : 5,94 s). Le nombre de segments est 6071. La figure 2.4 montre la répartition du nombre de segments.

Annotations des émotions et répartition des segments consensuels L'annotation du corpus a été réalisée en 2 étapes : les locuteurs 1 à 8 ont été annotées très précisement afin de définir les étiquettes et les dimensions qui semblent les plus pertientes à la fois pour obtenir un maximum de consensus entre les annotateurs et pour obtenir une bonne répartition entre les locuteurs. Pour les locuteurs 1 à 8, le schéma d'annotation est le suivant :

#labels *macro	Locuteurs 1 à 8 (labels)	Locuteurs 1 à 8 (macro)
Majeure #28 *7	0,22	*0,33
Mineure #28 *7	0,43	*0,44
Valence dim #5 *3	0,31	*0,39
Valence lbl #5	0,63	
Activation dim #5 *3	0,51	*0,55
Discours #2	0,97	
Direction #2	0,92	

TABLE 2.4 – Scores d'agréement entre annotateurs (2 annotateurs) locuteurs 1 à 8 sur IDV-HR (# nombre de labels, * nombre de macro-classes)

- émotion majeure et mineure : suivant le tableau 2.3 (ou ajoute une étiquette "poubelle" pour les instances que l'on ne peut pas annoter),
- valence et activation entre -2 et 2,
- valence : positive, négative, ambigüe, positif/négatif, neutre,
- discours : acté ou spontané,
- direction : au robot ou à l'expérimentateur.

Les scores de kappa (tableau 2.4) ont été calculés suivant l'équation 2.1. L'accord sur l'émotion majeure est relativement faible par rapport à celui sur l'émotion mineure (on peut remarquer que l'accord sur l'émotion mineure porte essentiellement sur l'état neutre). Les taux d'accord sur le type de discours et sur la direction sont très élevés. Ce résultat coïncide avec celui de [Callejas and Lopez-Cozar 08] qui montre la robustesse des coefficients kappa avec l'annotation du contexte. De plus, pour ces annotations, il n'y avait la possibilité que sur deux étiquettes ce qui réduit la marge d'erreur. La valence a été annotée suivant deux protocoles, l'utilisation d'étiquettes et l'utilsation d'une graduation continue entre -2 et 2. Au vu des valeurs de kappa du tableau 2.4, il semblerait que l'annotation catégorielle soit plus consensuelle. Les étiquettes les plus consensuelles sont celles correspondant aux valences positive, négative ou neutre. Les étiquettes de valence "ambigü" et "positif/négatif" semblent plus subjectives. On peut remarquer que les niveaux moyens des dimensions valence et activation ne sont pas les mêmes pour les deux annotateurs (valence (activation) moyenne : 0,06 (-0,28) pour l'annotateur 1 et 0,76 (-0,62) pour l'annotateur 2). Pour améliorer les scores d'accord inter-annotateurs plusieurs solutions peuvent être envisagées :

- regrouper les étiquettes en macro-classes (cf colonne macro du tableau 2.4),
- normaliser les dimensions pour chaque annotateur par sa valeur moyenne,
- utiliser les étiquettes plutôt que les dimensions dans la mesure du possible.

Le tableau 2.5 montre une baisse de kappa pour les locuteurs 9 à 22. Cela vient entre autre du fait que les données sont plus nombreuses. Le schéma d'annotation a été simplifié afin de diminuer la durée de la tâche d'annotation. Il est possible que la concentration des annotateurs ait été moins bonne sur cette seconde partie.

Etant donné l'effort fourni pour collecter des données émotionnelles, nous avons fait en sorte que la plupart des segments soient utilisables pour la reconnaissance des émotions. Pour cela, on peut définir des règles utilisant les annotations majeure et mineure pour

#Macros	Locuteurs 1 à 8	Locuteurs 9 à 22	Tous (moyenne)
Majeure #7	0,34	0,23	0,28
Valence #3	0,39	0,25	0,32
Activation #3	0,55	0,24	0,39

TABLE 2.5 – Scores d'agréement entre annotateurs (2 annotateurs) sur les macro-classes, tous locuteurs sur IDV-HR

macros	%moyen par locuteur	écart-type entre les locuteurs
neutre	60,18	17,54
colère	5,61	3,35
négatif	6,97	7,94
tristesse	5,06	3,24
pos-neg	5,05	4,45
peur	3,05	2,48
joie	14,07	7,58

TABLE 2.6 – Répartition des segments émotionnels du corpus IDV-HR suivant les locuteurs

déterminer une macro-classe qui soit consensuelle pour les deux annotateurs. Un exemple de règle sur les émotions majeures est : joie & (tristesse | colère | peur | négatif) = positif-négatif. Si l'une des émotions majeures est neutre, on peut regarder si il y a consensus sur les émotions mineures.

Malgré ces règles certaines annotations restent non consensuelles. Une triple annotation (voire une quadruple) sur la macro-classe et l'activation permet alors de réannoter ces segments afin de pouvoir conserver la plus grande partie des enregistrements. Seules les instances consensuelles sont conservées par la suite, ce qui réduit également la durée totale du corpus. La répartition des macro-classes sur le corpus IDV-HR est présentée sur la figure 2.6. Junk correspond aux instances qui ne pouvaient pas être annotées : présence de bruits, saturation, durée trop courte, pas de consensus trouvé, souffles, toux, etc. Plus de la majorité des instances consensuelles sont neutres. L'annotation est basée uniquement sur l'acoustique, on demande aux annotateurs de ne pas interprêter le support lexical. Les segments sont tous annotés en contexte avec l'outil transcriber. Une annotation hors contexte aurait pu être réalisée avec un autre outil que Transcriber permettant d'annoter des segments dans un ordre aléatoire.

Les émotions attendues dans le corpus IDV-HR étaient principalement des émotions relatives à la vie quotidienne comme l'énervement, le soulagement, la satisfaction ou l'ennui et quelques émotions relatives à l'urgence, à la douleur. Les émotions attendues n'étaient pas prototypiques, mais assez masquées du fait que les participants se contrôlent durant la session. Il a été assez difficile d'obtenir de la douleur et de l'urgence, sans doute parce que le contexte de l'enregistrement était trop éloigné d'un contexte réel de scénario d'urgence. Par contre, l'ensemble des autres émotions ont été globalement collectées.

2.2.4.2 Corpus NAO-HR1

Le corpus NAO-HR1 [Delaborde et al. 10] consiste en 12 locuteurs répartis par paires en 6 sessions enregistrées dans une salle expérimentale du LIMSI (i-room). Après segmentation, ce corpus contient 1275 segments pour une durée totale de 1 h 2 min 6 s. La durée moyenne des segments est de 1,46 s (entre 0,23 s et 8,38 s). Ce corpus est très intéressant puisqu'il est consistué de voix d'enfants en situation d'interaction homme-robot réelle. La segmentation et l'annotation ont été réalisées par deux annotateurs experts suivant les mêmes schémas que le corpus IDV-HR.

Organisation du corpus Chaque session collectée dans le cadre du corpus NAO-HR1, mettait en scéne une paire d'enfants qui se connaissaient (amis, famille), un maître du jeu (membre de l'équipe du LIMSI) et le robot NAO piloté en WoZ. Le maître du jeu prenait en charge le déroulement d'une session, il avait pour rôle de donner les consignes, lancer le robot à partir de mots clés repérés par le WoZ, et éventuellement relancer les enfants au cours de chaque partie. Une session se déroulait en trois parties. Au cours de la première, les enfants devaient jouer à un jeu de question-réponse, NAO posait les questions et accordait les points en fonction des réponses des enfants. Dans la seconde partie, chacun à leur tour, les enfants devaient fredonner un air connu que NAO devait reconnaître. Enfin, lors de la dernière partie, les enfants devaient jouer à un jeu émotion, acter une émotion afin d'être reconnus par NAO. NAO pouvait se montrer injuste, se tromper, favoriser un des enfants ou être honnête dans l'attribution des points et dans la reconnaissance de la réponse attendue. L'objectif de cette manipulation était d'induire des émotions fortes chez les enfants dans une situation de jeu compétitif.

Annotations des émotions L'annotation du corpus NAO-HR1 a été réalisée sur le même principe que le corpus IDV-HR :

- émotion principale : suivant le tableau 2.3 (ou ajoute une étiquette "poubelle" pour les instances que l'on ne peut pas annoter),
- émotions majeure et mineure : suivant le tableau 2.3 (ou ajoute une étiquette "poubelle" pour les instances que l'on ne peut pas annoter),
- intensité, activation et contrôle entre 0 et 5,
- valence : positive, négative, ambigüe, positif/négatif, neutre,
- discours : acté ou spontané.

La valeur de kappa très élevée pour l'émotion mineure est dûe au fait que l'étiquette "neutre" a été utilisée très souvent. Les étiquettes de macro-classe les plus consensuelles sont celles de la "joie" et du "neutre" (puis la "tristesse") plutôt que la colère et les étiquettes "autre" et "négatif". D'après le tableau 2.7, il est évident que lorsqu'on réduit le nombre de classes on obtient des annotations dont le kappa est plus satisfaisant.

L'annotation de ce corpus a mis en évidence des différences de traitement suivant chaque annotateur, par exemple l'étiquette "intérêt" a été beaucoup utilisée par un des annotateurs, alors que l'autre annotateur a préféré utilisé l'étiquette "neutre". Si on cherche les annotations consensuelles uniquement, on risque de réduire le corpus d'un tiers de ses segments.

Les émotions attendues dans ce contexte de jeu entre des enfants et un robot ont été globalement collectées. Nous attendions de l'ennui, de l'intérêt, de l'agacement (face aux

#labels *macro	Locuteurs 1 à 12	Locuteurs 1 à 12 (macro)
Principale #24 *7	0,39	0,48
Majeure #24 *7	0,35	0,41
Mineure #24 *7	0,93	0,93
Intensité dim #5 *3	0,44	0,46
Activation dim #5 *3	0,55	0,59
Contrôle dim #5 *3	0,43	0,45
Valence lbl (#5) *3	0,46	0,65
Discours#2	0,97	

TABLE 2.7 – Scores d'agréement entre annotateurs (2 annotateurs) sur NAO-HR1

modes de jeu du robot) mais également de la satisfaction (grâce au système de comptage des points au cours du jeu).

2.2.4.3 Corpus IDV-HH

Organisation du corpus Le corpus IDV-HH [Tahon et al. 10] est organisé en deux phases. La première consiste à répéter des mots correspondant à des ordres qu'on pourrait donner à un robot assistant tel que devrait l'être le futur ROMEO. Dans la seconde phase, le participant doit se placer dans le contexte de six scenarii différents correspondant à différents états émotionnels lors de scènes de la vie quotidienne.

Annotation émotionnelle et répartition des segments Après segmentation, le corpus IDV-HH est constitué de 1 h 23 min 45 s de parole émotionnelle, soit 2898 segments. 27 locuteurs de 20 à 89 ans) ont été enregistrés.

2.2.4.4 Corpus NAO-HR2

Organisation du corpus Après une courte introduction [Tahon et al. 12b], les enfants étaient placés par deux devant NAO qui expliquait lui-même les règles du jeu des histoires interactives. Une fois que le jeu était terminé, l'expérimentateur proposait à chacun des enfants un questionnaire de retour d'expérience.

Annotation émotionnelle et répartition des segments Après segmentation, le corpus est consitué de 603 segments émotionnels pour un total de 14 min 16 s (21 min 16 s si l'on ajoute les 4 adultes qui ont été enregistrés sur le même protocole). 12 enfants ont été enregistrés. Le temps de parole par enfant est très faible puisqu'il correspond à peu près aux réponses émotionnelles qu'ils devaient faire au cours de l'histoire interactive avec NAO.

Les réponses au questionnaire était excessivement brèves contrairement à ce qui était attendu (par rapport aux autres corpus collectés). Ce faible temps de parole est très certainement du au jeune âge des enfants et au fait qu'ils étaient assez peu préparés à l'expérience ;

	# AB (TT)	moyenne AB (TT) par locuteur
6-7 ans	30 (85)	6,0 (17,0)
8-11 ans	19 (80)	3,8 (16,0)

TABLE 2.8 – Nombre de segments affect bursts (AB) par rapport au nombre total de segment (TT) par locuteur, corpus NAO-HR2

Affect bursts Une étude préliminaire sur le corpus NAO-HR2 [Tahon et al. 12b] a permis de mettre en avant certains aspects liés aux affect bursts. Une grande majorité des segments émotionnels correspondent plus précisemment à des affect bursts (d'où également leur très courte durée). Les enfants n'avaient absolument aucun support lexical donné, ce qui tend à leur faire exprimer des affect bursts, comme des rires, des "grrr" pour l'expression de la colère.

Le tableau 2.8 résume le nombre d'affect bursts dans le corpus suivant deux classes d'âge de 5 enfants chacune : 6-7 ans et 8-11 ans.

2.2.4.5 Corpus JEMO

Le début de la collecte de ce corpus [Brendel et al. 10] date de 2008 dans le cadre du projet ANR Affective Avatar, il a été régulièrement complété au cours de mes travaux de thèse et de nouveaux locuteurs seront amenés à être enregistrés prochainement. Etant donné que nous l'utilisons largement dans nos travaux, nous le présentons ici au même titre que les corpus collectés dans le cadre du projet ROMEO.

Le corpus JEMO a été collecté dans le cadre d'un jeu émotion : les locuteurs devaient tester un système de reconnaissance automatique des émotions sur 4 catégories : colère, joie, tristesse et un état neutre ainsi que deux dimensions : valence (positif, négatif) et activation (actif, passif). Le support lexical et le temps de parole sont libres. Les émotions collectées sont prototypiques, ce qui permet d'avoir une annotation très consensuelle. Le corpus contient 1937 segments pour une durée totale de 1 h 4 min et 6 s. 64 locuteurs entre 23 et 54 ans ont été enregistrés. La durée moyenne d'un segment est de 1,98 s (entre 0,14 s et 16,03 s). Les enregistrements ont été réalisés dans des conditions de laboratoire avec un microphone en champ proche, dans un bureau du LIMSI.

2.2.4.6 Corpus COMPARSE

Dans le cadre du projet ANR Comparse, nous avons collecté un corpus de locuteurs exprimant du stress dans la voix lors d'une tâche de prise de parole en public. Dans une première phase de calibration du protocole 10 participants entraînés ont été enregistrés. Dans une seconde phase, 19 participants naïfs ont accepté de participer à l'expérience.

Lorsque le participant arrive dans la salle d'expérience, il est équipé d'un cardiomètre, d'un appareil mesurant la température cutanée, la sudation, le pouls, d'un micro-cravate. Deux caméras le filment en plan rapproché sur le visage et en plan large. Le participant est également placé au centre d'une plate-forme de force, permettant de mesurer les micro-déplacements de son centre de gravité. Deux personnes sont en face de lui. D'autres expérimentateurs sont dans la salle. Une fois qu'il est équipé, il doit lire un texte à haute voix puis la consigne lui est donnée. La tâche consiste à se présenter pour un entretien

d'embauche à un poste à définir soi-même. Il faudra également donner ses points forts et ses faiblesses. Il n'y a pas de temps de préparation. Une fois qu'il s'est présenté, les deux personnes en face de lui, en réalité des juges, lui posent des questions. Dans le cadre de l'expérience, il y a un juge positif, et un autre négatif.

Les enregistrements audio obtenus contiennent plusieurs phases distinctes (figure 2.5) : phase de lecture (environ 1 min), phase de présentation (environ 5 min), phase d'entretien avec les juges (durée variable). La phase de lecture est nécessaire afin d'obtenir une référence commune à tous les participants. Elle peut également servir de référence pour une éventuelle normalisation. Sur chacune de ces phases, une annotation du stress est en cours.

2.3 Conclusion

L'annotation d'une base de donnée audio est souvent fondée sur une appréciation perceptive. Plusieurs annotateurs (entre 2 et une dizaine) définissent un certain nombre de critères sur des unités temporelles choisies. La définition de ces critères est orientée par les différentes théories émotionnelles. Le choix des unités temporelles est une question complexe, elle dépend à la fois des applications souhaitées, du type d'émotions collectées, de la théorie choisie (dimensionnelle ou catégorielle) et du contenu de l'interaction. La définition de l'unité temporelle d'annotation n'est pas toujours étudiée avec précision, elle est pourtant d'une importance capitale si l'on souhaite que les applications en découlant soient satisfaisantes. Nous présentons un certain nombre des annotations existantes dans l'état de l'art : annotation du contexte, d'informations sur le locuteur, annotations linguistiques et paralinguistiques. Les annotations paralinguistiques sont celles qui nous intéressent le plus : émotions, personalité, interaction, signal social, etc... mais également informations paravarbales (bruit de bouche, respiration, type de voix, affect bursts).

Les bases de données collectées dans le contexte du projet Romeo ont été annotées perceptivement par deux annotateurs experts. Les indices émotionnels utilisés sont principalement des catégories d'émotions (émotions fines et macro-classes) mais également des dimensions (notamment valence et activation). L'annotation perceptive peut être appréciée grâce à une valeur de mesure d'agréement (ou kappa). Cet agréement peut être amélioré par une définition précise du segment émotionnel et par une définition des indices à utiliser également. Cette seconde définition passe par une phase "d'entraînement" des annotateurs.

La collecte et l'annotation des différents corpus "maison" nous ont permis d'étudier les différentes variabilités présentes lors d'une interaction homme-robot d'un point de vue acoustique (chapitre 3, paragraphe **??**). Ils ont également permis de mettre en place des mesures de comparaison entre différents corpus (chapitre 3, paragraphe **??**), comme la mesure de spontanéité. Mais également de pouvoir réaliser des expériences cross-corpus fortes afin d'étudier la généralisation des modèles issus de l'apprentissage (chapitre 5, paragraphe 6.2.2.2) et de construire une liste noire de descripteurs qui ne sont pas robustes à certaines variabilités (chapitre 3, paragraphe 4.4).

a)

b)

FIGURE 2.5 – Collecte de données de stress dans la voix, tâche de prise de parole en public (projet ANR Comparse) a) lecture, b) entretien avec les juges

Troisième partie

Analyse acoustique de la voix émotionnelle

La recherche de descripteurs acoustiques pertinents pour la reconnaissance des émotions dans la voix peut être une tâche d'analyse perceptive. En effet, une écoute experte des segments émotionnels permet de définir des paramètres intéressants pour caractériser les émotions et les voix qui les expriment. Etant moi-même musicienne, j'ai déjà développé une oreille sensible pour l'écoute de la musique (dans les domaines de l'interprétation musicale et de l'analyse musicale notamment). Cette écoute permet de mettre en relation des perceptions avec des éléments descriptifs comme le timbre, le rythme, la mélodie ou même la structure globale. Une telle capacité d'écoute est très intéressante pour rechercher des descripteurs dans l'audio quels qu'ils soient, en particulier dans la voix émotionnelle.

Cette partie sur l'analyse acoustique de la voix émotionnelle en contexte écologique porte essentiellement sur la description des indices acoustiques (en particulier le rythme et le timbre de la voix) et la proposition de nouveaux indices permettant de caractériser certains aspects émotionnels. Nous aborderons également l'influence des émotions, des locuteurs, de l'environnement et de la tâche sur certains descripteurs choisis. La mise en évidence de ces variabilités au niveau acoustique peut se faire à l'aide de classements des indices pour un corpus donné. Il sera assez aisé de faire la distinction entre les émotions et les locuteurs, par contre il est bien plus compliqué de montrer la différence acoustique au niveau de la tâche ou de l'environnement acoustique.

Nous abordons également l'intérêt de plusieurs nouveaux indices que nous proposons (indices de rythme et de timbre). Nous testons la robustesse d'un ensemble d'indices acoustiques sélectionnés pour la reconnaissance des émotions sur les différents corpus à notre disposition. Et enfin nous avons cherché à définir une liste "noire" d'indices qui ne sont absolument pas robustes à différents types de locuteurs, d'émotions, d'environnement ou de tâche.

Après un état de l'art sur l'ensemble des descripteurs acoustiques utilisés pour le traitement de la parole et particulièrement de la parole émotionnelle (chapitre 3), nous présentons plusieurs contributions importantes dans le chapitre 4 : l'ajout de descripteurs issus de domaines différents (transformation de voix, synthèse de voix ou analyse de signaux musicaux), puis l'apport de nouveaux descripteurs de rythme et de timbre. Ensuite l'utilisation de différentes méthodes pour sélectionner les descripteurs les moins robustes aux différentes variabilités liées à l'interaction.

Chapitre 3

Etat de l'art des descripteurs acoustiques pour la parole émotionnelle

Les descripteurs acoustiques sont un des éléments fondamentaux pour le traitement informatique de données audio. Ils sont utilisés dans plusieurs domaines du traitement du signal et majoritairement pour les signaux musicaux ou signaux de parole. Ils sont utiles à la fois pour la description des signaux (reconnaissance automatique, perception, etc.), pour la transformation de signal (par exemple la transformation de voix) ou pour leur synthèse (instruments, voix de synthèse). Ce chapitre regroupe un grand nombre de descripteurs utilisés par l'ensemble des chercheurs en traitement du signal, qui sont définis sur plusieurs niveaux temporels. Il n'a pas vocation cependant à être exhaustif, la quantité d'indices étant phénoménale.

3.1 Le signal de parole et ses modes de production

Dans cette première section, nous nous plaçons du point de vue de la parole et de sa production. Nous entrerons dans les détails des descripteurs acoustiques dans la section suivante.

3.1.1 Production de la parole

Le signal de parole est consitué d'une alternance entre des sons quasi-périodiques (voyelles, consonnes voisées), des sons apériodiques (fricatives) et des sons impulsionnels (plosives). Cette alternance relativement régulière s'effectue sur des durées allant de 5 à 500 ms. La durée moyenne d'un phone est de 70 ms pour un débit de parole normal.

L'appareil vocal humain est constitué de trois parties majeures (figure 3.1) : la structure sub-glottique (poumons, bronches et trachée), le larynx (cordes vocales et glotte) et le conduit vocal (pharynx, cavité bucale, nasale, joue et langue). Le souffle trachéal a pour principale fonction de mettre en vibration la glotte. Cette vibration produit du son

grâce à la mise en tension des cordes vocales, et crée un signal quasi-périodique synonyme de voisement (voyelles, consonnes voisées). La fréquence de vibration correspond à la fréquence fondamentale du son émis (F0). Mais elle peut aussi provoquer une impulsion générée par l'expiration phonatoire. L'ensemble du conduit vocal module le signal afin de créer des sons différents. Les lèvres et la langue sont également utilisées lors de l'articulation (plosives, fricatives). Etant donné que chaque être humain est différent physiquement, les voix de chacun sont également diverses et variées en fonction de l'âge, du genre, de la morphologie, etc. Le timbre d'une voix est principalement déterminé par les composantes morphologiques et l'utilisation des résonateurs. Le modèle de production le plus couramment utilisé est celui décomposant l'appareil vocal en un système source (structure sub-glottique et cordes vocales) - filtre (conduit vocal).

Suivant l'utilisation faite des cordes vocales, les modes de production de la parole peuvent être différentes :

- le voisement (les cordes vocales sont en vibration),
- l'absence de voisement : les cordes vocales sont en position écartée et ne vibrent pas,
- l'aspiration : courte période non voisée qui se produit pendant et immédiatement après un relâchement articulatoire dans les cavités supraglottiques,
- le murmure : les cordes vocales vibrent accolées,
- la laryngalisation : seule une petite partie des cordes vocales est en vibration,
- l'occlusion glottale : les cordes vocales sont maintenues l'une contre l'autre en position d'adduction,
- le chuchotement : les cordes vocales sont ouvertes pour laisser passer l'air. Il y a une constriction du conduit vocal.

Lorsque les cordes vocales sont en vibration, on peut distinguer plusieurs mécanismes de vibration (ou registres) correspondant chacun à un mode de vibration des cordes vocales. Il existe communément deux modes de vibration M1 (voix de poitrine, cordes vocales courtes et épaisses) et M2 (voix de tête, cordes vocales fines et longues) auxquels on peut ajouter le mode M0 (ou *fry*) et le mode M3 (ou sifflet, ressemblant à un cri). Ces mécanismes sont beaucoup travaillés par les chanteurs lyriques. Les chanteurs doivent être capables d'utiliser chacun de ces modes afin de pouvoir couvrir le plus grand ambitus possible, par contre dans un contexte de voix parlée, le mode M1 est le plus fréquemment employé. En fonction des styles de voix, des manières de s'exprimer et des états émotionnels, certains modes autres que M1 sont employés. Par exemple dans l'anglais américain, le mode M0 est souvent utilisé dans la voix parlée.

3.1.2 Aspects linguistiques

La langue parlée est porteuse de plusieurs informations dont le vecteur principal est linguistique. Ces informations sont complétées par l'intonation, la manière dont est organisé le discours dans le temps et bien sûr les émotions présentent dans la voix.

3.1.2.1 Les mots et la langue

L'aspect linguistique ne sera pas développé dans nos études, cependant il est important d'en souligner l'importance pour la reconnaissance des émotions. La plupart des théories

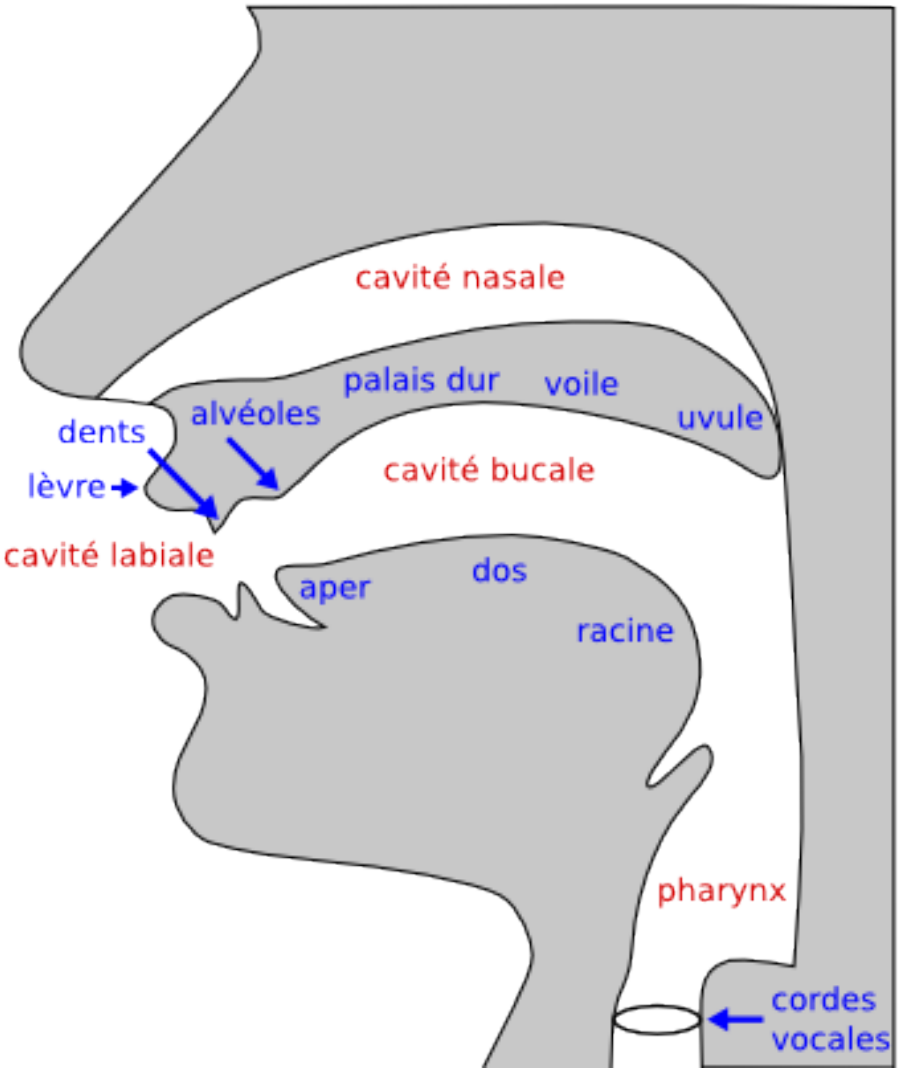


FIGURE 3.1 – Vue schématique d'une coupe de profil des organes phonatoires

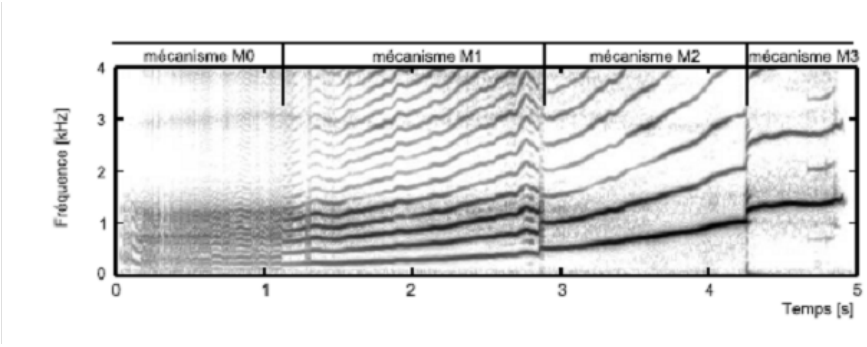


FIGURE 3.2 – Spectogramme d'un glissando réalisée par une femme dans les 4 mécanismes vibratoires (adapté de [Roubeau et al. 09])

se basent sur un schéma à deux niveaux, repris par Hirst [Hirst et al. 00] : physique (la phrase acoustique) et cognitif (représentation phonologique, syntaxique et sémantique).

Beaucoup d'études sur les émotions prennent en compte l'aspect sémantique de la parole : ces recherches utilisent une transcription automatique de la parole [Devillers and Vidrascu 06]. Des mots considérés comme à fort caractère émotionnel sont annotés. Certains corpus sont annotés au niveau du mot comme le corpus AIBO [Batliner et al. 10]. Craggs [Craggs and Woods 04] propose un schéma d'annotation du contenu linguistique d'un corpus de dialogues entre des patients atteints de cancer et des infirmières.

Les mots sont constitués de phones successifs énoncés suivant un rythme dont la stabilité dépend de son locuteur et de son état émotionnel. La segmentation automatique des syllabes n'est pas une tâche facile : elle repose principalement sur la détection des noyaux syllabiques [Xie and Niyogi 06, Lanchantin et al. 08].Une segmentation manuelle ou automatique des syllabes permet d'obtenir le débit syllabique [Audibert et al. 06, Beller 09, Obin et al. 08b, Clavel and Richard 10]. L'utilisation des syllabes comme fenêtre temporelle est également très courante dans les études sur les émotions [Batliner et al. 10, Dumouchel et al. 09, Li et al. 10, Ruiz et al. 10] mais aussi celles qui portent sur l'identification de locuteur [Leung et al. 08].

3.1.2.2 L'organisation temporelle de la parole

La parole est organisée sur deux structure qui interagissent et se complètent : le sens et la phonétique. Le sens s'appuie sur les mots, groupes de mots, phrases et au niveau du discours. Les aspects phonétiques se structurent sur plusieurs échelles temporelles (phones, syllabes, groupes de souffles, etc.).

Les études sur la métrique d'une langue sont très intéressantes pour approcher l'organisation temporelle du discours. Selon Di Cristo [DiCristo 03], la perception du rythme de la parole ou de la musique est liée à la régularité, l'accentuation, les proéminences, les battements, les groupements et le tout est hiérarchisé. Dans le cas de la parole, l'auditeur segmente le signal d'entrée en groupes, au sein desquels une régularité temporelle sous-jacente émerge sous la forme d'une structure de battement. Les auditeurs s'attendent à ce que les battements soient réguliers. La segmentation intiale permet de réaliser un

traitement hiérarchique entre chaque groupe, les groupes minimaux sont intégrés à des unités plus grandes. La structure des battements permet d'extraire les régularités et un relatif niveau de saillance (ou proéminence). Le battement peut se définir comme une variation soudaine d'un paramètre physique quelconque (fréquence fondamentale, intensité, timbre), elle engendre la perception d'une rupture et donc d'une unité perceptive de groupements. L'accent est lié à l'unité de base, la syllabe et correspond à un battement fort. Généralement les battements forts sont espacés de 3 à 4 battements faibles (environ 600 ms). Cette notion de battement introduit directement la recherche de proéminences dans le discours.

La perception du rythme dans la parole est à la fois basée sur les battements (temporalité fine) et sur les groupes de mots (temporalité globale). La perception d'un discours régulier, organisé, est fortement liée à la fluidité de la parole, les retours en arrière (redites), les hésitations, les bégaiements, etc. Il est difficile d'évaluer cet aspect au niveau acoustique uniquement sans avoir de transcription, il ne sera donc pas traité. D'autres approches de l'organisation temporelle de la parole ont été proposées : une analyse modulaire du discours et la segmentation prosodique des unités du discours chez Simon [Simon 02], ou par Martin [Martin 87].

3.1.3 Le signal de parole

Qui dit analyse acoustique, dit que le signal de parole est enregistré sur un support a priori numérique. La plupart du temps la fréquence d'échantillonage utilisée pour les signaux de parole est 16 kHz. L'information contenue dans les très hautes fréquences n'est alors pas du tout traitée. La fréquence d'échantillonage est adaptée en fonction des applications souhaitées. Les signaux enregistrés sur support radiophonique ou téléphonique sont à 8 kHz, on peut supposer qu'il manque alors une partie relativement importante du signal qui n'empêche cependant ni la reconnaissance du locuteur, ni la compréhension du discours.

Plusieurs outils d'analyse de la parole ont été développé comme WinPitch [Martin 05] ou Praat [Boersma and Weenink 09]. Pour pouvoir comparer les analyses avec la communauté de recherche sur les émotions, nous utiliserons l'outil Praat. La figure 3.3 montre un exemple des fonctions bas-niveau les plus couramment utilisées pour l'analyse de la parole :

- en haut : le signal acoustique (noir), les lignes bleues verticales correspondent aux pulses, c'est à dire aux instants de fermeture de glotte,
- au milieu : la fréquence fondamentale en bleue (en Hz) et l'intensité en dB en vert,
- en bas : le spectogramme (niveau de gris selon l'intensité), la fréquence fondamentale (bleue) et les formants 1 à 4 (en rouge).

On peut remarquer qu'il y a une rupture au niveau de la fréquence fondamentale (F0), elle correspond à un saut d'octave, c'est une des erreurs les plus fréquentes avec l'extraction de F0 avec Praat, avec une méthode d'autocorrélation puis un algorithme de filtrage des effets de sauts d'octave. La réduction des sauts d'octave est un compromis pour calculer au mieux la F0 dans le cas de la parole, dès que l'on commence à crier, l'extraction de la F0 ne correspond plus à la réalité. Les zones bleutés du signal correspondent aux parties voisées. Les formants sont calculés sur les zones d'intensité suffisantes mais pas uniquement sur les parties voisées.

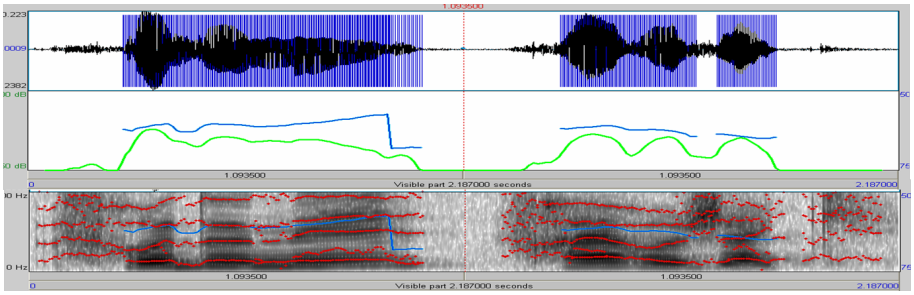


FIGURE 3.3 – Un exemple d'analyse acoustique de la parole avec Praat

A l'intensité, le spectre, la fréquence fondamentale, nous pouvons ajouter le cepstre et les durées temporelles de certains éléments comme descripteurs acoustiques de bas-niveau (*Low-Level Descriptors*, LLD). Les descripteurs de plus haut-niveau sont calculés majoritairement sur ces LLD, mais certains peuvent être déterminés directement à partir du signal temporel.

Le signal de parole est alternativement composé de phases d'activité vocale et de phases de silence. Il existe plusieurs manières de déterminer l'activité vocale. La plus simple repose sur l'utilisation du nombre de passage par 0 du signal (*Zero-Crossing Rate*, ZCR). Plus le ZCR est important, plus la quantité de bruit est forte, il y a donc de fortes chances que ce ne soit pas de la parole. Certaines consonnes comme les fricatives correspondent à un ZCR très élevé, un système de détection d'activité vocale uniquement basé sur le ZCR ne permet donc pas de détecter le signal de parole complet.

3.2 La prosodie

La prosodie peut se définir comme suit [1] :

> La prosodie est le domaine particulier de la phonétique qui s'occupe de décrire les sons du langage au niveau de l'énoncé (L'énoncé peut-être un mot, un groupe de mots ou une phrase.).
>
> La prosodie s'attarde plus précisément à l'impression musicale que fournit l'énoncé. On y observe des phénomènes prosodiques tels que : l'intonation, l'accentuation, le rythme, le débit et les pauses.
>
> Chacun de ces phénomènes prosodiques se manifeste par des variations au niveau de la fréquence, de la hauteur, de l'intensité et/ou de la durée.

Nous détaillerons plus en détail les aspects dynamiques de la parole dans la section 3.4 spécifique au rythme. La prosodie peut se décomposer en deux phénomènes : l'un sur une fenêtre temporelle correspondant au mot ou à la phrase (la macro-prosodie), l'autre sur une fenêtre temporelle bien plus réduite, de l'ordre de la centaine de milisecondes, du phone (la micro-prosodie).

1. http ://www.latl.unige.ch/safran/data/phono/mod9/1_def/index.htm

3.2.1 Fréquence fondamentale

3.2.1.1 Fonction bas-niveau

La fréquence fondamentale correspond à la fréquence de vibration des cordes vocales. On peut également définir la fréquence perçue par l'oreille humaine (généralement appellée "pitch") qui peut être différente de la fréquence fondamentale.

Pour un signal de parole, la fréquence fondamentale n'a de sens que sur des parties voisées. Deux outils d'extraction de la F0 sont couramment utilisées : méthode d'autocorrélation à court terme [Boersma 93] avec une réduction des erreurs de sauts d'octave ou l'algorithme YIN fondé sur l'autocorrélation puis la réduction des erreurs d'estimation de la F0 [de Cheveigné and Kawahara 02]. D'autres méthodes d'extraction de la F0 sont fondées sur des fonctions de différences moyennées (*Average Square Difference Functions*, ASDF), sur des méthodes d'estimation des maxima de vraisemblance [Doval and Rodet 93], sur l'algorithme de Viterbi ou encore sur l'estimation du cepstre.

La fonction F0 étant perçue par l'oreille humaine sur une échelle logarithmique, il peut être judicieux de la transposer sur une échelle linéaire : prendre une référence ($La_2 = 110Hz$, eq. 3.1) ou convertir en semi-tons avec une référence ($La_2 = 110Hz$, eq. 3.2).

$$f_{lin} = \frac{f_{Hz}}{110} \tag{3.1}$$

$$f_{semiton} = 12 \cdot log_2(\frac{f_{Hz}}{110}) \tag{3.2}$$

Les descripteurs de fréquence fondamentale sont généralement d'autant plus performants qu'ils sont proches de la perception humaine, et donc en utilisant une des deux conversions proposées. La référence est ici fixée à 110 Hz, mais on peut définir une référence dépendante du locuteur ou d'un genre. Une fréquence de référence de 1 Hz est également couramment utilisée.

3.2.1.2 L'intonation

L'intonation, c'est-à-dire la forme du contour de la fréquence fondamentale, est porteuse d'information concernant la structure du discours mais aussi de l'émotion du locuteur [Bänziger and Scherer 05].

Selon d'Alessandro [d'Alessandro and Mertens 95], la fréquence fondamentale calculée automatiquement ne correspond pas exactement à la fréquence perçue par l'auditeur. L'auditeur perçoit un ensemble de paramètres physiques : l'énergie perçue (*loudness*), la qualité de voix, la durée, les modulations de fréquences et d'amplitude. Un exemple simple serait le tremolo utilisé par le chanteur, l'auditeur n'entend pas précisement cette modulation de la fréquence fondamentale mais une fréquence moyenne.

Toutes les variations de fréquence fondamentale ne sont pas perçues. T'Hart [t Hart 81] estime à 3 semitons l'intervalle de fréquence significatif perçu en situation de communication. On peut définir le glissando comme étant le rapport entre l'intervalle de fréquence et la durée. Le seuil de perception des changements d'intonation a été estimé au glissando suivant : $G = \frac{0,16}{T^2}$ avec T la durée considérée.

Afin de pouvoir traiter automatiquement les contours intonatifs, il est important de pouvoir les modéliser. Il existe quatre modèles principaux de stylisation ou modélisation

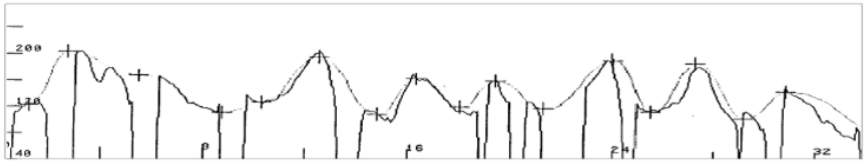


FIGURE 3.4 – Courbe de F0 et modélisation MoMel (extrait de [Hirst and Espesser 93])

FIGURE 3.5 – Codage de points cibles avec INTSINT (extrait de [Hirst et al. 00])

de la fréquence fondamentale (F0). Dans ces modèles, la F0 est la combinaison d'une composante macroprosodique (patron intonatif) et d'une composante microposodique (qui n'est souvent pas exprimée consciemment par le locuteur). Les modèles d'extraction de l'intonation cherchent à récupérer la composante macroprosodique.

- MOMEL [Hirst and Espesser 93] permet de réaliser une modélisation de la F0. Les parties voisées sont interpôlées, ce qui permet à la courbe d'intonation d'être continue (figure 3.4). La nature relativement neutre de cet algorithme a permis son utilisation large. Une stylisation phonologique a été mise au point à partir de cette modélisation. INTSINT [Hirst et al. 00] permet de coder des points cibles relativement au registre du locuteur (figure 3.5) : top (T), mid (M), bottom (B) ; ou codage relativement au point cible précédent : high (H), same (S), lower (L), up (U), down (D). Les points cibles étant les minima et maxima locaux de la courbe d'intonation modélisée (par exemple avec MoMel) ou réelle. L'inconvénient majeur de cette modélisation est le manque d'informations temporelles.
- ToBI [Pierrehumbert 80, Black and Hunt 96] est un codage symbolique de la F0 qui permet de générer les contours avec une regression linéaire. La description de l'intonation a pour but applicatif la synthèse de voix. La F0 est modélisée suivant une échelle discrète de tons (emprunt musical). Les modèles ToBI sont spécifiques à une langue donnée. Black a travaillé sur l'anglais. Ils peuvent s'appliquer à une courbe réelle de F0 ou une courbe stylisée par exemple avec l'algorithme MoMel.
- prosogramme [Mertens 04] est un algorithme fondé sur un modèle de perception tonale [d'Alessandro and Mertens 95] qui nécessite un alignement avec une transcription phonétique. La F0 est stylisée en un ou plusieurs segments de droite (figure 3.6) qui peuvent être plats ou avec une pente mélodique selon des seuils perceptifs de glissando réglables.

Les modèles présentés ont été exploités et adaptés dans des langues différentes de l'anglais : le coréen [Kim et al. 08] et au français [Campione et al. 98, Martin 06].

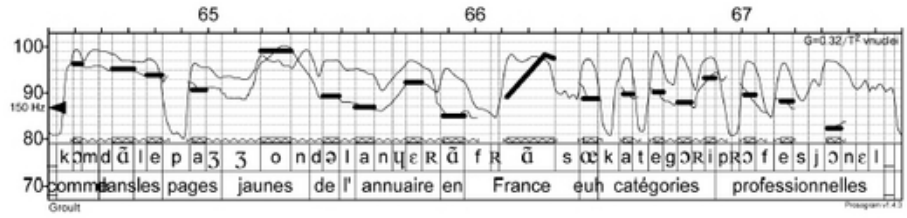


FIGURE 3.6 – Prosogramme (extrait de [Mertens 04])

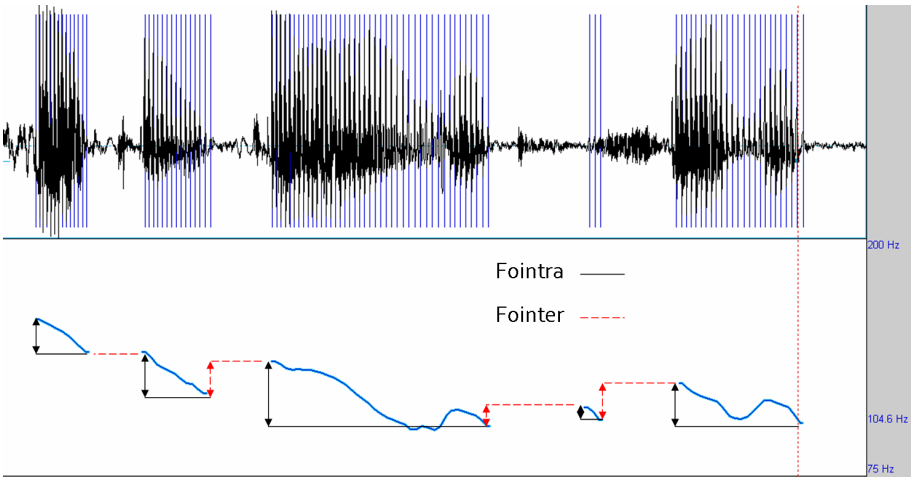


FIGURE 3.7 – Indices interF0 et intraF0 déterminés à partir de la courbe de F0 en Hz (bleue)

3.2.1.3 Indices haut-niveau

Sur la fonction F0, plusieurs descripteurs statistiques peuvent être définis, les plus courants étant la moyenne, le maximum, le minimum, l'intervalle (différence si échelle logarithmique, rapport si échelle linéaire). On peut aussi déterminer le nombre de pics de F0 sur un segment donné.

Dans une étude de Devillers [Devillers et al. 05b], deux indices supplémentaires sont introduits (fig. 3.7) : la variation de la F0 au sein d'une même partie voisée (intraF0) et celle entre deux segments voisés adjacents (interF0).

On peut définir également un glissando correspondant à la phase finale descendante de la courbe de F0 sur une partie voisée (eq. 3.3 et sur la courbe de F0 de la figure 3.8). Perceptivement, c'est la partie de la courbe qui est la plus entendue (puisque la dernière). Si la courbe d'intonation finit par une montée, le glissando sera défini par le minimum, si elle finit par une descente, on prendra en compte le maximum. Si le minimum ou le maximum correspond avec la valeur finale de F0, le glissando est mis à 0 par défaut (ce qui correspond à un grand nombre de cas).

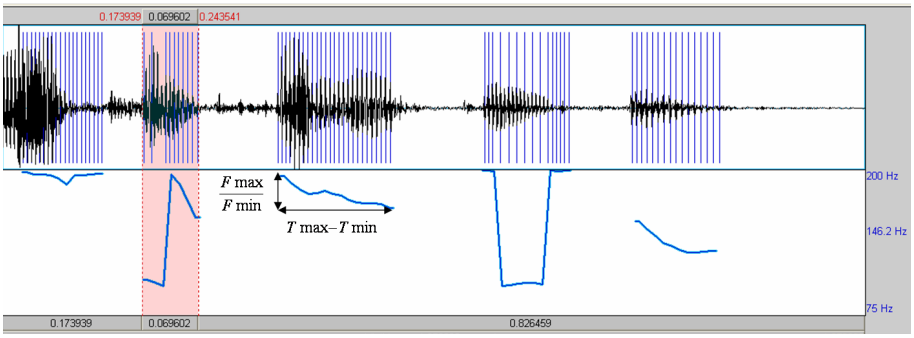


FIGURE 3.8 – Exemple d'erreur de sauts d'octave avec Praat

$$G = \frac{\frac{Fmax}{Fmin}}{Tmax - Tmin} \tag{3.3}$$

D'autres indices acoustiques de haut-niveau fondés sur la F0 sont proposés par l'équipe de Schuller (on peut notamment se référer au dernier set d'indices acoustiques proposés au challenge 2012 sur la reconnaissance automatique de personalité [Schuller et al. 12b]).

3.2.2 Energie

Il existe plusieurs fonctions représentant l'énergie d'un signal de parole. La plus courante est l'énergie RMS, elle correspond au niveau d'énergie moyen sur une fenêtre temporelle donnée, elle est homogène au carré de l'amplitude du signal. L'intensité (dB) est définie sur une échelle logarithmique. L'énergie perçue (*loudness*) correspond à l'énergie du signal à laquelle on applique un filtre correspondant à l'oreille humaine. Il existe donc autant de définition de loudness que de filtres utilisés. Pour la suite des travaux, nous utiliserons un filtre basé sur les bandes de Bark.

3.2.3 Proéminences et accentuation dans la parole

La proéminence est un aspect important du signal de parole. Les proéminences prosodiques se définissent principalement par une accentuation au niveau de la F0 et de l'énergie autour du noyau syllabique.

Tamburini [Tamburini 06] propose de détecter automatiquement les proéminences à partir de la F0, l'énergie globale autour d'une syllabe, la durée du noyau syllabique ainsi qu'un rapport de fréquence permettant de caractériser le timbre. Les proéminences sont spécifiques aux langues, des études ont été réalisées en allemand [Tamburini and Wagner 07], en anglais [Tamburini 06], en français [Obin et al. 08a] et en portugais brésilien [Barbosa 07]. Goldman a proposé un outil de détection automatique [Goldman 11] des proéminences pour différentes langues à partir des travaux sur le prosogramme de Mertens [Mertens 04].

Timbre	sombre / clair détimbré / timbré sourd / brillant dureté nasillard présence de souffle sur la voix
Oscillations de la F0	vibrato tremolo / tremor
Modes de production	voix grinçante voix criée voix chuchotée
Articulation	hyperarticulation

TABLE 3.1 – Différenciateurs sémantiques pour le timbre et la qualité de voix dans le chant lyrique et la voix parlée (adapté de [Garnier et al. 04] et [Abrilian 07])

De même que pour la fréquence fondamentale et l'énergie, plusieurs indices de haut-niveau sont définis sur la proéminence. Schuller [Schuller et al. 12b] propose notamment des indices fondés sur la détection de maxima locaux.

3.3 Timbre et qualité vocale

Le timbre d'un instrument se définit acoustiquement parlant par son spectre sur l'ensemble des notes qu'il joue. Le timbre de la voix est définit de la même manière, il est alors intrinsèque au locuteur sur l'ensemble des phones qu'il prononce, contrairement à la qualité vocale, qui peut varier d'un moment à l'autre de la journée en fonction de l'humeur, de l'âge, etc. Les deux aspects sont de toute manière complètement liés l'un avec l'autre.

3.3.1 Descripteurs sémantiques

3.3.1.1 Descripteurs sémantiques pour la voix chantée

On peut trouver des descriptions intéressantes chez les chanteurs lyriques, Garnier [Garnier et al. 04] propose une liste de descripteurs sémantiques bipolaires pour décrire la qualité de la voix chantée 3.1. La voix lyrique est une voix travaillée, très contrôlée puisqu'elle doit correspondre aux standards esthétiques de l'époque.

La brillance se définit par la position du centre de gravité des composante du spectre. Pour le chant lyrique, la brillance se traduit par l'émergence du formant du chanteur par rapport aux autres formants. L'aspect détimbré correspond à une atténuation spectrale entre 2 kHz et 4 kHz, ou au-delà de 4 kHz et un renforcement de l'énergie autour du fondamental. L'aspect clair/sombre ne se définit que sur les voyelles, il correspond à l'enrichissement ou non du spectre autour de la zone du formant du chanteur.

Evaluation de la Nouveauté	
Nouveau	Ancien
Interruption de la phonation Inhalation soudaine	pas de changements
Evaluation du plaisir intrinsèque	
Agréable	Désagréable
Relaxation du conduit vocal Energie en basses fréquences	Tension du conduit vocale Energie en hautes fréquences
Voix étendue	*Voix étroite*
Evaluation de la puissance face à l'événément	
Contrôle	Pas de contrôle
Tension de l'appareil vocal F0 élevée et beaucoup d'amplitude	Hypotension de la musculature F0 basse, peu d'amplitude
Voix tendue	*Voix relâchée*
Puissance	Pas de puissance
Registre de poitrine F0 basse et beaucoup d'énergie	Registre de tête F0 élevée et peu d'énergie
Voix pleine	*Voix fine*

TABLE 3.2 – Prédictions des changements vocaux en fonction des SECs (extrait de [Scherer 86])

3.3.1.2 Descripteurs sémantiques pour la voix parlée

Dans ses travaux de thèse, Abrilian [Abrilian 07] propose plusieurs descripteurs de qualité vocale en vue de la construction d'un schéma d'annotation multimodal de parole expressive. La liste de ces descripteurs sémantiques rejoint en partie celle présentée pour le chant lyrique (tableau 3.1).

3.3.1.3 Qualité vocale et théorie de l'évaluation

D'autres aspects sont directement issus des travaux de Scherer [Scherer 86]. Il prédit les changements vocaux en fonction des différentes situations d'évaluation (SEC) (tableau 3.2).

3.3.2 Descripteurs acoustiques

3.3.2.1 Descripteurs spectraux

Les descripteurs spectraux cherchent à représenter le timbre d'un signal, c'est-à-dire la répartition de l'énergie en fonction des fréquences. Ces descripteurs sont calculés trames à trames (généralement avec un pas d'environ 10 ms). Plusieurs représentations sont utilisées pour la reconnaissance des émotions :

- Le spectre à court terme (*Fast Fourrier Transform*, FFT).
- Les ondelettes [Mallat 00], elles permettent de mettre en évidence des irrégularités dans les échelles élevées et des tempi dans les échelles basses. Les onde-

lettes sont efficaces pour les signaux non-stationnaires comme la parole. Les ondelettes peuvent servir à coder le signal [Dellandrea et al. 03, Dellandrea et al. 02, Dellandréa et al. 03] pour ensuite utiliser la loi de Zipf afin de déterminer les irrégularités du signal. Selon He [He et al. 10] les ondelettes pourraient être utiles pour reconnaître les émotions et le stress au niveau des voyelles. D'après [Fernandez and Picard 03] les coefficients d'ondelettes seraient plus performants que les coefficients cepstraux pour la reconnaissance du stress dans la parole.

- Les coefficients LPC (*Linear Prediction Coding*) permettent d'obtenir les formants ainsi qu'une enveloppe spectrale contenant des informations relatives au conduit vocal [Ruiz et al. 08, Ruiz et al. 10].
- L'énergie par bandes de Bark ou par bandes de Mel. Ces deux échelles sont perceptives, elles sont utilisées aussi bien pour les signaux musicaux que pour les signaux de parole.
- L'énergie par bandes harmoniques [Xiao 08], l'échelle est définie par la fréquence fondamentale.

Plusieurs indices permettent de caractériser une enveloppe spectrale. Ces descripteurs sont très utilisés en musique : les fréquences roll off (5, 25, 50, 75, 95%), le barycentre spectral et la pente spectrale. La fréquence roll off x% correspond à la fréquence pour laquelle l'énergie spectrale est égale à x% de l'énergie totale. On peut également ajouter les notions de flatness, skweness et kurtosis qui décrivent la forme de l'enveloppe spectrale [Peeters 04].

3.3.2.2 Descripteurs cepstraux

Les descripteurs cepstraux sont dérivés du spectre du signal. Ils sont calculés trames à trames sur l'ensemble du segment émotionnel.

- Les coefficients MFCCs [Bogert et al. 63] sont connus pour leur robustesse au bruit de fond ainsi que pour leur capacité à mettre en évidence les changements ou les périodicités dans le spectre. Ils sont fondés sur le filtre perceptif de Mel. Ils sont dépendent très fortement du contenu lexical puisqu'ils capturent les fréquences de résonnance du conduit vocal.
- Les coefficients LPCC (Linear Prediction Cepstral Coefficients) sont des coefficients dérivés des LPC.
- Les coefficients PLP (Perceptual Linear Prediction) [Hermansky 90] correspondent à une amélioration des LPC puisqu'ils sont fondés sur le filtre perceptif de Bark.
- Les coefficients RASTA-PLP [Truong and van Leeuwen 07b], RASTA (Relative Spectral Transform) permettent un lissage des coefficients et sont conçus pour être plus robustes au bruit.

3.3.2.3 Descripteurs de qualité vocale

La qualité vocale est une problématique complexe qui ne fait pas consensus. Il existe cependant différents descripteurs de qualité vocale utilisés en reconnaissance des émotions, mais également en transformation de voix ou en synthèse vocale. Les descripteurs de qualité vocale sont globalement calculés au niveau d'une voyelle ou d'une partie voisée. Ils peuvent également être calculés sur d'autres types de signaux comme les fricatives.

Par contre, on ne peut généralement pas comparer les valeurs obtenues sur l'ensemble des différentes parties du signal de parole.

- Les jitter et shimmer mesurent les micro-variations de la fréquence fondamentale et l'énergie suivant les périodes de fermeture de glotte. Ils ont été largement utilisés en transformation de voix par [Beller and Marty 04]. Le lecteur trouvera la formule du jitter local utilisé dans Praat à l'équation 4.1. La même formule peut être appliquée au shimmer en remplaçant la période par l'énergie.
- NAQ (*Normalized Amplitude Quotient*) [Alku et al. 02] dépend de la fréquence fondamentale et de la configuration glottale au travers du Quotient d'Amplitude. Il est décrit dans [Campbell 04a] comme une mesure de souffle dans la voix. Ce coefficient a été utilisé pour la reconnaissance des émotions [Devillers et al. 05b, Audibert et al. 06, Clavel and Richard 10].
- Le rapport harmonique sur bruit (HNR) permet d'avoir des informations sur le voisement du signal, cependant ce paramètre dépend très fortement du bruit et donc des conditions acoustiques d'enregistrement du signal.
- Le coefficient de relaxation Rd est fondé sur un algorithme de détection des instants de fermeture et d'ouverture de glotte et par conséquent de la fréquence fondamentale et des paramètres du modèle LF [Fant et al. 85] (quotient d'ouverture, coefficient d'assymétrie, durée de la phase de retour et énergie). [Degottex et al. 08, Degottex 10] a développé une méthode d'estimation du paramètre Rd et des fonctions de distorsion de phase qui permettent de quantifier le relâchement de la voix (voir exemple sur la figure 3.9 pour de la joie dans le corpus IDV-HR). Plus la voix sera tendue, plus le Rd tendera vers 0,3, une voix normale se situera autour de 1,2 et une voix relâchée autour de 2,5.
- Dans [Farner et al. 09], les outils de transformation de voix sont principalement fondés sur l'enveloppe spectrale des voyelles uniquement. Celle-ci est extraite du signal grâce à une estimation de la *true-envelop* [Röbel and Rodet 05, Villavicencio et al. 09]. L'enveloppe spectrale (figure 3.10) est alors modifiée selon que l'on souhaite une voix chuchotée (pente positive en basses fréquences) ou une voix soufflée (pente spectrale élevée en hautes fréquences, aspiration du bruit autour de 2 kHz). On peut alors définir une fréquence (nommée VUF) qui correspond à la limite entre le signal quasi-périodique et le signal bruité. Ce VUF varie en fonction des différents modes de production et également de la qualité vocale.
- La modulation de fréquence (ou rugosité, *roughness*) [Shilker 09] est fondée sur les coefficients d'auto-corrélation pour déterminer des consonances/ dissonances. A partir des ratios entre fréquences, on obtient des harmoniques, sous-harmoniques et en fonction de la distance du ratio avec les ratios consonants ou dissonants on détermine la qualité vocale. La rugosité a été également estimée par Sun [Sun 02], à partir du rapport SHR (rapport entre les sous-harmoniques et les harmoniques). Ce rapport détermine aussi la F0 perçue. Si le SHR est important, la F0 est perçue un octave en-dessous de la F0 mesurée.
- Les coefficients issus directement de la modélisation de la source : Ee, Rd, Rk, Rg, Oq [Tao and Kang 05].
- Les ondelettes ont également été étudiées pour estimer la qualité vocale suivant la méthode LoMA (Line of Maximum Amplitude), initiée par Sturmel et d'Alessandro [Sturmel et al. 09]. Cette méthode permet de tracer des lignes plus ou moins

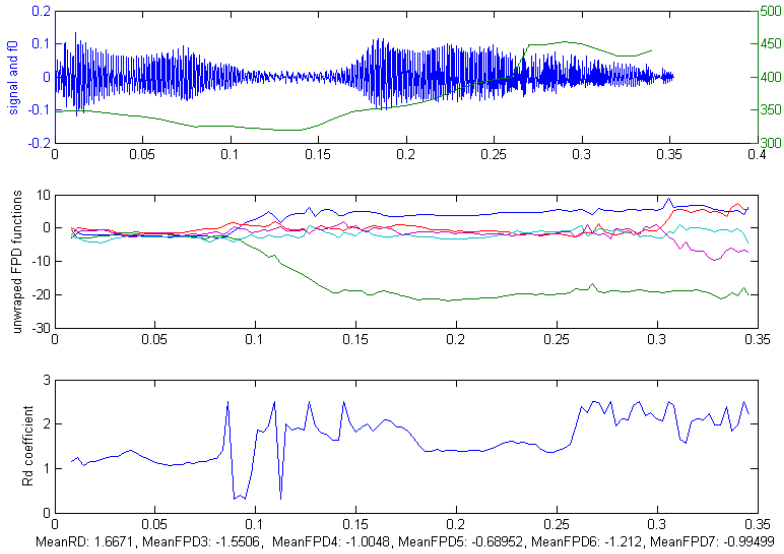


FIGURE 3.9 – Courbes de Rd et fonctions FPD (unwrap) sur un segment "joie" du corpus IDV-HR, locuteur 18.

parallèles à l'axe des échelles qui passent par les maximum d'amplitde de chaque fonction d'ondelette à échelle donnée. Ces lignes permettent de localiser assez précisement l'ouverture de glotte. Cette approche est très similaire à celle proposée pour l'estimation du coefficient de relaxation de Degottex [Sturmel 11, section 3.7].

- L'articulation peut se mesurer à l'aide du triangle vocalique [Beller et al. 08, Beller 09, Lee et al. 04] à condition de connaître a priori le type de syllabe prononcée (/a/, /u/ ou /i/).
- Le coefficient de Lyapunov est utilisé par Ruiz [Ruiz et al. 08, Ruiz et al. 10] pour quantifier le "chaos" d'une voyelle. Cette mesure étant très sensible au bruit, il est difficile d'obtenir des résultats concluants.

La plupart des études sur la qualité vocale ne prennent en compte que les parties voisées ou les voyelles. Kienast [Kienast and Sendlmeier 00] propose une mesure de qualité vocale pour les fricatives en calculant le barycentre spectral. Il montre que les émotions à forte activation (colère, peur, joie) ont une balance spectrale plus élevée qu'un état neutre.

3.3.2.4 Voix pathologiques

Les principaux descripteurs de qualité vocale sur des voix pathologiques (disphonie avec/sans lésion des cordes vocales, immobilité de la glotte, laryngite chronique, maladie de Parkinson, etc.) sont les jitter et shimmer, largement utilisés grâce à Praat. Cependant, si ces descripteurs présentent un avantage sur des voix pathologiques (jitter local

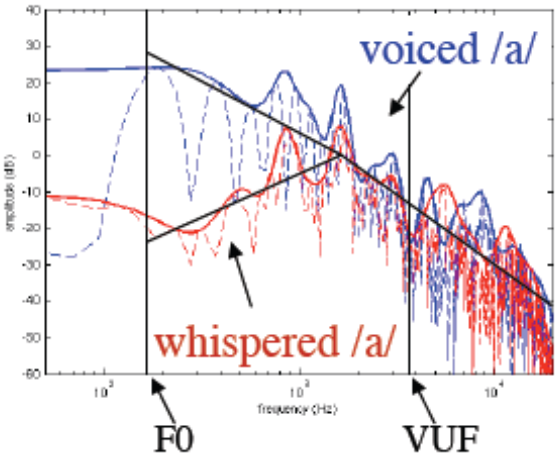


FIGURE 3.10 – Spectogrammes avec enveloppes spectrales pour une voix voisée et une voix chuchotée sur un /a/ (extrait de [Farner et al. 09])

> 1,04%, shimmer local > 3,81%), ils semblent être trop grossiers sur des voix "normales". De plus la perception d'une variation du jitter ou du shimmer semble n'être pas significative [Kreiman and Gerratt 03]. Par contre, le tremor qui consiste en une modulation irrégulière de la fréquence fondamentale (semblable au tremolo des chanteurs mais non contrôlé) semble être un indice tout à fait intéressant pour les voix pathologiques [Kreiman et al. 03]. Cette notion de trémor pathologique doit pouvoir être étendue aux locuteurs âgés et à certains états émotionnels (voir un exemple en situation d'urgence figure 3.11).

3.4 Le rythme de la parole

L'étude du rythme dans la parole commence par sa perception. Les recherches sur le rythme dans la parole sont encore assez marginales, celles sur le rythme dans la musique sont beaucoup plus nombreuses.

Selon Zellner-Keller [Zellner-Keller 04], pour comprendre l'organisation temporelle de la parole, il faut d'abord se demander sur quelle base la structure temporelle de la parole est définie. La plupart des recherches (comme celle de Di Cristo [DiCristo 03]) sont fondées sur des analyses prosodiques et la structure temporelle de la parole est alors réduite à l'accentuation (ou battements fort chez Di Cristo). Zellner-Keller montre que le rythme dans la parole est plus proche d'une structure chaotique que d'une structure régulière. Une autre étude [Jarina et al. 02], montre que la recherche d'une structure périodique à long terme (au-delà de 4 s) permet de discriminer des signaux de parole et des signaux musicaux. L'article montre que contrairement aux signaux musicaux, la parole n'a pas de structure périodique à long terme.

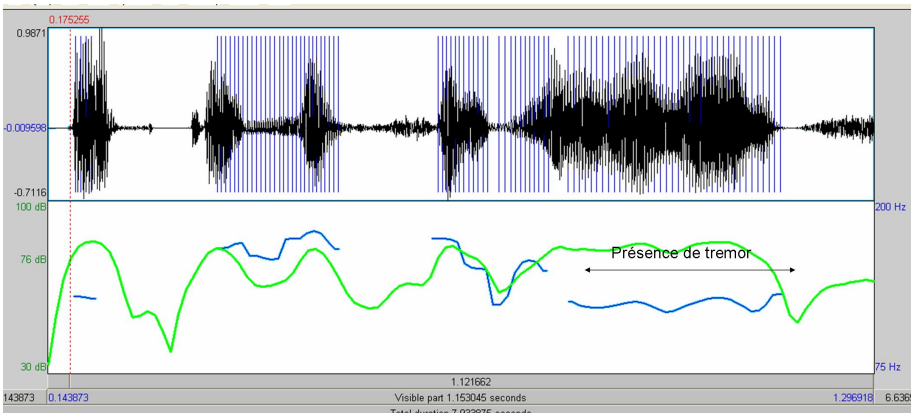


FIGURE 3.11 – Exemple de tremor, sur un locuteur en situation d'urgence (corpus CEMO), visualisation avec Praat.

3.4.1 Structure voisée

Les signaux musicaux se prêtent bien à la détection du rythme. La plupart des travaux existants se basent sur des algorithmes de détection de tempo [Yang 04], à partir des ondelettes [Smith and Honing 08] ou d'un filtrage en basse-fréquences. La recherche du rythme dans la parole ne sera pas fondée sur des méthodes aussi précises. On peut citer les travaux de Ramus [Ramus et al. 99] qui propose une mesure du rythme basée sur les parties vocaliques (%V proportion d'intervalles vocaliques, ΔV l'écart-type des intervalles vocaliques, ΔC l'écart-type des intervalles de consonnes par phrase). Différentes langues peuvent être discriminées dans les plans (%V, ΔV) et (%V, ΔC). Ces mesures ont été reprises par Galves [Galves et al. 02] auxquelles il ajoute la sonorité, c'est à dire si la parole est "claire" ou "sombre". Elles permettent de comparer plusieurs langues suivant leurs propriétés rythmiques. Des mesures relativement similaires de durée et de périodicité des voyelles ont été utilisées pour la classification des langues [Grabe and Low 02]. Ce type de mesure a également été utilisée par Patel [Patel et al. 06] pour comparer le rythme et la mélodie dans des signaux de musique et de parole française et britannique. Cela montre l'intérêt de ces mesures relativement simples et robustes pour l'analyse de rythme.

Le rythme peut être défini à partir des parties voisées :

- taux de fenêtres non-voisées sur un segment audio donné [Clavel and Richard 10], ce taux de voisement est également utilisé comme indice de qualité vocale par les mêmes auteurs [Clavel et al. 07].
- Vidrascu [Vidrascu 07] a utilisé pour ces travaux de thèse, le débit de parole défini comme l'inverse de la durée moyenne d'une partie voisée, ainsi que le nombre de silences définis comme une partie non-voisée entre 40 et 80 ms.

Mais il peut également être défini à partir des syllabes, ce qui implique une transcription ou une segmentation automatique en syllabe. Il semblerait que l'inverse de la durée moyenne d'une syllabe n'a de sens que dans un contexte de laboratoire [Beller 09, Obin et al. 08b], pour des données réalistes il faut pouvoir adapter le débit en fonction du temps. On peut également définir un débit fondé sur la fréquence d'apparition des

mots.

3.4.2 Loi de Zipf

D'autres études plus originales ont tenté de mesurer le rythme avec d'autres descripteurs bas-niveau. On peut citer entre autres les travaux de thèse de Xiao [Xiao 08]. La loi de Zipf est appliquée dans le domaine temporel et fréquentiel afin de récupérer des informations prosodiques. Suivant cette loi, la fréquence $f(p)$ d'un événement p est liée à son rang $r(p)$ suivant l'équation 3.4 où α et β sont des nombres réels.

$$f(p) = \alpha \cdot r(p)^{-\beta} \tag{3.4}$$

Zipf propose la loi de Zipf Inverse 3.5 pour les événements de faible fréquence d'apparition, avec f la fréquence d'apparition de l'événement, $I(f)$ le nombre d'événements distincts apparus avec cette fréquence. δ et γ sont deux nombres réels.

$$I(f) = \delta \cdot f^{\gamma} \tag{3.5}$$

Ces lois peuvent être linéarisées en appliquant la fonction logarithme. Elles sont appliquées à différents types de codage du signal : codage temporel, fréquenciel ou temps-échelle à l'aide des fonctions d'ondelettes [Dellandrea et al. 03, Dellandrea et al. 02, Dellandréa et al. 03]. Ces codages permettent de mettre en évidence par exemple les variations d'amplitude (up, flat, down). Dans sa thèse, Xiao propose plusieurs descripteurs issus de la loi de Zipf Inverse : l'entropie de la loi de Zipf Inverse, les coefficients polynômiaux issus d'une regression linéaire d'ordre 2 de la même loi (sous sa forme logarithmique).

3.5 Fonctionnelles et descripteurs de références

Dans cette section, nous décrivons les descripteurs acoustiques utilisés pour la reconnaissance des émotions. La plupart de ces descripteurs sont issus des travaux sur des signaux musicaux. Les descripteurs utilisés spécifiquement pour la parole sont ceux qui décrivent la prosodie (F0, énergie, rythme). Ils sont généralement cumulés avec des descripteurs statistiques (ou fonctionnelles) afin de décrire les variations d'un descripteur sur un segment émotionnel.

3.5.1 Fonctionnelles

Les systèmes de reconnaissance automatique des émotions utilisent une représentation vectorielle d'un segment émotionnel (sectionné ou entier). Le signal de parole doit alors être décrit par descripteurs prenant en compte la dimension temporelle. On cherche alors à définir un certain nombre de fonctionnelles (ou descripteurs statistiques) à appliquer aux descripteurs acoustiques de plus bas niveau afin de conserver des informations temporelles pertinentes.

Les fonctionnelles les plus courantes tentent de décrire statiquement ou dynamiquement une fonction acoustique bas-niveau (LLD : *Low Level Descriptor*) sur l'ensemble du segment émotionnel. On mentionnera donc les fonctions statistiques basiques : moyenne,

max, min, écart-type. L'utilisation d'une regression linéaire permet d'obtenir une information dynamique (pente). La recherche de pics notamment au niveau de l'énergie permet de dégager une périodicité.

3.5.2 Référence Challenge Personalité Interspeech 2012

L'organisation de compétitions (appellées *challenges*) autour de la reconnaissance automatique de traits de locuteurs, d'émotions, de personnalités a permis de proposer un ensemble de descripteurs servant de baseline. La méthode adoptée par les organisateurs est appellée brute-force puisqu'elle combine 6125 descripteurs bas-niveau, des fonctions appliquées à ces descripteurs pour obtenir des descripteurs haut-niveau et un grand nombre de fonctionnelles statistiques appliquées systématiquement à tous les descripteurs haut-niveau. Ces descripteurs sont calculés à partir d'un fichier audio grâce à l'outil OpenEar. On appellera OpenEarX la version correspondant au challenge de l'année X. L'ensemble des fonctionnelles et LLD utilisés pour la reconnaissance des émotions de la version OpenEar2012, sont décrits dans l'article [Schuller et al. 12b].

3.6 Conclusion de l'état de l'art

L'ensemble des indices acoustiques utilisés pour l'étude du signal de parole est très large. Certains d'entre eux nécessitent la connaissance du contenu linguistique : phones, syllabe, mots comme par exemple les descripteurs de proéminences et de battements. Ces notions de rythme sont très intéressantes pour l'étude des émotions. Nous aimerions définir ces notions plus spécifiquement aux émotions, dans ce cas, y a-t-il besoin de comprendre le contenu pour percevoir le rythme ? Après tout, on peut comprendre que quelqu'un est en colère même si on ne connaît pas sa langue dès que sa culture n'est pas trop éloignée de la notre. Il y a donc un modèle fondamentale regroupant un certain nombre d'indices extraits uniquement sur la musicalité de la voix. Nous n'avons pas utilisé d'alignement sur le contenu linguistique pour les modèles mis en place, par contre, certaines de nos analyses sur les descripteurs ont été menées sur des alignements phonétiques.

Un autre point important concerne l'analyse de l'ensemble des indices acoustiques cités sur de la parole réaliste. En effet, très peu d'études sur les indices ont été réalisées sur de grands corpus prenant en compte une relative variabilité de locuteur, d'émotions mais également de contenu linguistique et de tâches. La plupart des études sur les indices sont menées sur des corpus actés hors contexte et sur peu de voix différentes (entre 2 et une dizaine).

Chapitre 4

Analyse de descripteurs acoustiques appliquée à nos corpus

Les systèmes de reconnaissance automatique fondés sur l'audio ne "perçoivent" les informations contenues dans le signal que par les descripteurs acoustiques calculés sur ce signal. Les recherches présentées dans cette seconde partie se focalisent donc sur ce que peuvent mesurer les descripteurs acoustiques. Comment capturent-ils les informations émotionnelles, celles liées au locuteur, comment est-ce que l'environnement acoustique modifie les descripteurs et enfin comment le type de tâche (émotions actées, induites, spontanées) influence ces descripteurs. L'ensemble des analyses portées dans cette partie seront donc essentiellement des analyses sur les descripteurs acoustiques.

4.1 Les enjeux

Dans cette section, nous analysons les descripteurs acoustiques présentés dans l'état de l'art pour étudier leur intérêt dans la reconnaissance automatique des émotions mais également leur variation suivant les environnements acoustiques ou les locuteurs enregistrés. Les enjeux sur les descripteurs présentés dans cette section de ma thèse sont :

- de proposer de nouveaux indices, notamment de rythme (débit et précision) et d'articulation afin d'améliorer la classification,
- de sélectionner un sous-ensemble d'indices les plus fiables aux différentes variabilités pour de futures applications,
- de tester la robustesse des indices classiquement utilisés en classification automatique des émotions à différentes variabilités afin d'en extraire une liste des descripteurs les plus spécifiques à un contexte donné (*black-list*),
- et enfin de proposer des mesures de comparaison des corpus fondées sur la variabilité de certains indices.

La question du choix de la fenêtre temporelle pour le calcul des descripteurs sera abordée dans cette section. On ne fera pas de classification automatique des émotions, ce sera

l'objet du chapitre 5. L'influence de l'environnement acoustique et de la tâche sur les indices est analysée en section 3.2.3.1 [Tahon and Devillers 10]. La variabilité des locuteurs sur une voyelle fixée est étudiée en section 3.2.3.2. Les nouveaux descripteurs de rythme et d'articulation sont présentés en section 3.2.2.3 et 3.2.2.4. Et enfin les mesures de comparaison de corpus, ainsi que la sélection d'indices en fonction de leur robustesse aux différentes variabilités sont présentées en section 3.2.4.

Plusieurs méthodes seront utilisées pour évaluer la robustesse des descripteurs acoustiques aux variabilités liées aux interactions humain-robot. Deux mesures ont été développées dans le cadre de nos travaux (mesure de spontanéité et mesure de variabilité émotion/locuteur). Les deux autres sont des mesures standards (test d'analyse de la variance, ANOVA et utilisation d'algorithmes de sélection automatique de descripteurs).

4.2 Descripteurs acoustiques pour les émotions

Nous avons sélectionné un sous-ensemble des descripteurs détaillés dans l'état de l'art qui nous semble pertinents pour les applications écologiques que nous souhaitons faire. Nous appellons "écologique", une situation in situ à l'opposé d'un contexte de laboratoire. Cette sélection est fondée principalement sur l'écoute des corpus à notre disposition. Une étude perceptive experte ne permet pas de sélectionner des descripteurs trop complexes mais plutôt des descripteurs intuitifs. Les fichiers audio utilisés ont tous une fréquence d'échantillonage de 16 kHz (ou 8 kHz dans le cas du challenge sur la reconnaissance automatique de la personalité). Dans cette section nous décrivons l'ensemble de descripteurs de base que nous allons étudier pour la reconnaissance des émotions dans un contexte d'interaction homme-robot. Dans la communauté les ensembles de descripteurs de base sont définis par les versions de l'outil OpenEar lors des challenges Interspeech. Par exemple, lors du challenge Interspeech 2009, OpenEar comptait 384 descripteurs (soit 16 fonctions statistiques appliquées à 12 descripteurs bas-niveau, et leurs dérivées) [Schuller et al. 09b].

4.2.1 Descripteurs acoustiques usuels

Cet ensemble de descripteurs dits "usuels" correspond à la synthèse de l'ensemble des descripteurs acoustiques utlisés classiquement pour la reconnaissance des émotions par la communauté scientifique. Les descripteurs bas-niveau sont globalement identiques pour les chercheurs de la communauté, il peut y avoir des différences de calcul entre chacun. Nous avons choisi les descripteurs qui nous semblaient les mieux adaptés aux émotions, en privilégiant un point de vue perceptif. C'est-à-dire que nous prendrons la fréquence fondamentale en semitons et non en Hertz, l'énergie perçue basée sur les bandes de Bark (*loudness*) et pas l'énergie RMS.

4.2.1.1 Fonctions bas-niveau (LLD) :

Certaines de ces fonctions sont calculées sur l'ensemble du segment, d'autres uniquement sur les parties voisées. Nous avons choisi un pas d'échantillonage de 10 ms pour des fenêtres glissantes de 30 ms. Ce choix permet d'être synchrone avec les séries temporelles extraites avec Praat.

– F0 : extraction avec Praat (parties voisées),

- Formants : extraction avec Praat,
- MFCC : extraction à partir du signal audio,
- Spectre : extraction à partir du signal audio,
- Energy (*loudness*) : extraction à partir du signal audio,
- ZCR normalisé : extraction à partir du signal audio,
- Durée des parties voisées : extraction avec Praat.
- Jitter : extraction à partir de la F0.

4.2.1.2 Descripteurs haut-niveau

Toute la difficulté dans la définition et le calcul des descripteurs acoustiques réside plus sur le choix de la fenêtre temporelle que sur le choix du descripteur lui-même. En effet, les valeurs du ZCR calculées sur une voyelle (signal quasi-périodique, ZCR très faible) n'ont aucun rapport avec celles calculées sur une fricative (équivalent à un bruit blanc, ZCR très élevé). La distinction des fenêtres voisées/non-voisées est celle qui est la plus couramment utilisée. D'ailleurs un grand nombre de travaux de recherches ne définissent leur descripteurs que sur les parties voisées [Ruiz et al. 08]. Lorsqu'une transcription en syllabe est utilisable, la distinction entre syllabe, consonnes voisées et consonnes non-voisées est d'une grande utilité.

Nous proposons une première distinction entre fenêtres voisées et non-voisées.

4.2.1.3 Fréquence fondamentale

La fréquence fondamentale que nous utilisons est extraite avec Praat (nous avons décidé de prendre les paramètres d'extraction par défaut). La fonction obtenue comporte bien sûr un certain nombre d'erreurs dont les plus fréquentes correspondent à des sauts d'octave. Afin d'éviter un grand nombre de ces erreurs, nous ne traiterons pas les parties voisées de durée inférieure à 50 ms. Cependant il restera des erreurs d'estimation de pitch qui constituent un biais important dans l'analyse de la F0 (figure 3.8).

La fréquence est très souvent utilisée en reconnaissance des émotions sur une échelle linéaire (en Hertz). Or l'échelle des semitons (st) est plus pertinente au vu de la perception par l'oreille humaine. Cette question est importante car elle détermine l'utilisation des fonctionnelles. Par exemple la différence entre deux fréquences maximum et minimum sur une échelle linéaire n'a pas de sens pour l'oreille humaine contrairement au rapport entre les deux. Avec l'utilisation de l'échelle en semitons c'est l'inverse. De plus la normalisation la plus couramment utilisée consistant à soustraire à la valeur de F0 sa moyenne (sur le locuteur, le genre, ou l'ensemble d'un corpus) est bien plus pertinente pour une échelle logarithmique que linéaire. Nous choisirons donc de travailler dans la mesure du possible avec une fréquence fondamentale en semitons.

4.2.1.4 Jitter

Le jitter est un descripteur ambigü comme nous l'avons décrit dans l'état de l'art. Il y a plusieurs manières de déterminer le jitter, nous choisirons la formule correspondant au jitter local sous Praat. Le jitter est définit sur une partie voisée d'une durée $N \cdot 0,01s$ si la F0 est extraite avec un pas de 0,01s. Nous calculons alors différentes fonctions statistiques pour obtenir des valeurs sur le segment entier. Le jitter local est défini comme étant le

Descripteurs bas-niveau	Fonctionnelles (#indices)	voisé	non-voisé	tout
F0 (st) (6)	Médiane, écart-type, max, min (4)	X		
	Intra/interF0* (2)	X		
Energie (loudness) (12)	Médiane, écart-type, max, min (12)	X	X	X
Spectre (enveloppe LPC) (61)	Roll Off (5%*, 25%, 50%, 75%, 95%*), barycentre, pente (14)	X	X	
	Energie par bandes de Bark* (42)	X	X	
	Energie par bandes harmoniques* (5)	X		
Cepstre (78)	Moyenne MFCC et ΔMFCC 0-12 (78)	X	X	X
Formants (st) (34)	F1, F2, F3, bF1, bF2, bF3, F2-F1*, F3-F1* (médiane, écart-type, max, min) (32)	X		
	Articulation (moyenne, écart-type) (2)	X		
Rythme (13)	punvoicedPraat (ou ratio durée voisée sur durée non-voisée) (1)			X
	Débit (8)			X
	Précision (4)			X
Qualité vocale (5)	Jitter, jitterLocalPraat (2)	X		
	ShimmerLocalPraat, HNRPraat (2)			X
	Nombre d'harmoniques (1)	X		

TABLE 4.1 – 208 descripteurs usuels, autres descripteurs (*), nouveaux descripteurs (italique)

rapport de la différence entre deux périodes fondamentales T_0 consécutives et la moyenne des périodes sur le segment (eq. 4.1). Ce paramètre est sans dimension et peut s'exprimer en %.

$$JitterLocal = \frac{N}{N-1} \frac{\sum_1^{N-1}(T_0(k+1) - T_0(k))}{\sum_1^N T_0(k)} \tag{4.1}$$

4.2.2 Autres descripteurs acoustiques

4.2.2.1 Variations de F0 dans/entre parties voisées

Ces paramètres ont été définis à la section 3.2.1. IntraF0 correspond au rapport entre la F0 finale et la F0 intiale d'une même partie voisée, InterF0 correspond au rapport entre la F0 intiale d'une partie voisée et la F0 finale de la partie voisée précédente.

4.2.2.2 Coefficient de relaxation pour la qualité vocale sur la valence

D'après Gendrot [Gendrot 04], il semblerait que la qualité vocale est un aspect important dans la reconnaissance de la valence. En effet, cette dimension émotionnelle reste extrêment difficile à classifier avec des scores convenables.

Nous avons réalisé une étude afin de vérifier cette hypothèse [Tahon et al. 12a]. Testés sur le corpus IDV-HR, plusieurs paramètres de qualité vocale montrent un pouvoir de discrimination important lorsqu'ils sont seuls (par exemple sur un test d'analyse de la variance, ANOVA [Hogg and Ledolter 87]). Par contre, ajoutés à des descripteurs prosodiques (coefficients cepstraux, F0 ou énergie), leur intérêt n'est plus significatif. Quatre descripteurs de qualité vocale testés sont extraits avec Praat (jitter, shimmer, punvoiced et HNR), les autres sont issus des travaux de Degottex [Degottex et al. 10] sur la relaxation de la voix (coefficient de relaxation Rd, moyenne et écart-type sur l'ensemble des parties voisées d'un même segment ; fonctions de distortion de phase (FPD) moyennées sur les parties voisées également).

4.2.3 Nouveaux descripteurs de rythme

Comme nous l'avons vu dans l'état de l'art, la perception du rythme est un phénomène très complexe. Nous l'avons décomposé en deux composantes mesurables automatiquement sur le signal de parole : la précision et le débit. Les descripteurs de rythme, parce qu'ils sont définis sur une fenêtre temporelle relativement large, devraient être très robustes aux changements de salles. Travailler sur le rythme est fondamental pour améliorer la reconnaissance automatique in situ, dans des situations variées mais sa définition et mesure sont loin d'être triviales.

4.2.3.1 La précision

La précision est une idée originale qui correspond à la netteté d'articulation avec laquelle est prononcé un mot, une syllabe. Le noyau syllabique se repère assez simplement par un pic sur la courbe d'énergie (cf état de l'art). Notre hypothèse de départ est que ce pic est plus ou moins étalé selon la précision de l'énonciation. Une émotion comme la

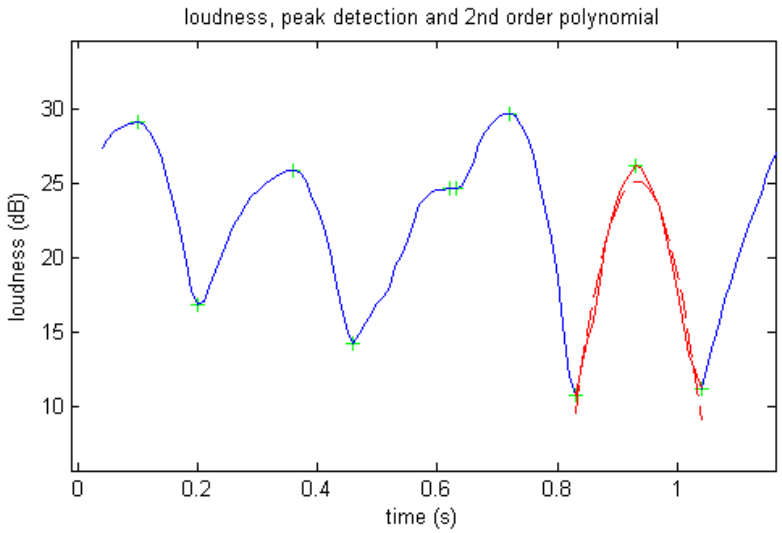


FIGURE 4.1 – Courbe de loudness et sa modélisation au niveau d'un noyau syllabique par un polynôme d'ordre 2.

tristesse pourrait correspondre à une précision assez faible, ainsi que la colère. Alors que pour un état neutre ou positif, la précision serait plus importante. Pour mesurer cette précision, nous proposons de modéliser le pic d'énergie (loudness définie sur les bandes de Bark) correspondant au noyau syllabique par un polynôme d'ordre deux (figure 4.1). Ce polynôme sera de la forme de l'équation 4.2 avec $p_1$et p_2 des nombres réels positifs. Le coefficient p_3 n'a pas d'intérêt puisqu'il dépend du niveau général de l'énergie perçue. Plus p_1 est grand, plus le pic est étalé et moins la précision est bonne.

$$L_{db} = -p_1 \cdot t^2 + p_2 \cdot t + p_3 \tag{4.2}$$

Cette approche nécéssite de sélectionner les sommets et les vallées de la courbe de loudness correspondant aux noyaux syllabiques. Pour cela, deux paramètres peuvent être utilisés : la différence minimum de loudness tolérée entre deux sommets consécutifs et un seuil minimal d'énergie permettant de ne pas repérer les pics qui ne seraient pas perçus par l'oreille humaine. La partie voisée comme fenêtre temporelle pourrait également être utilisée pour le calcul de la précision, mais la mesure sur le pic de loudness est plus précise.

4.2.3.2 Le débit

Nous avons déjà défini une mesure de débit relativement pertinente pour la parole : le ratio entre la durée des parties voisées et celle des parties non-voisées. Cette mesure permet d'avoir une première estimation du débit, cependant elle n'apporte aucune information sur la régularité. Nous proposons alors une mesure définie sur la distribution statistique de la durée des parties voisées, non-voisées, des périodes voisées, non-voisées. 8 descripteurs semblent être pertinents pour quantifier à la fois le débit et sa régularité :

– le rapport entre la moyenne et l'écart-type des durées des parties voisées (VD),

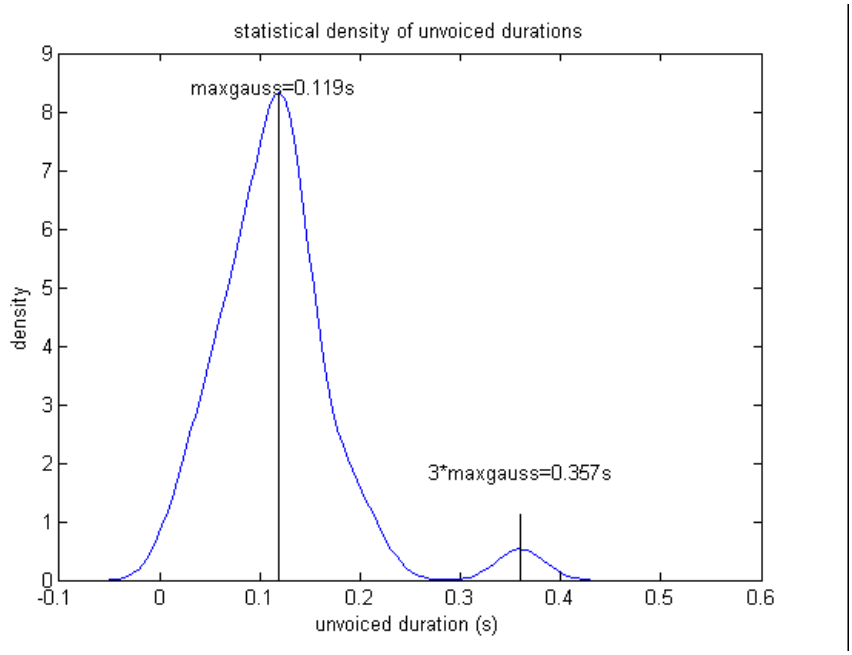


FIGURE 4.2 – Courbe de densité des durées des parties non-voisée, $UD_{max} = 0,1s$, $DUD_{max} = 8,2$, $3DUD = 0,5s$

- le rapport entre la moyenne et l'écart-type de la période entre deux parties voisées consécutives (VP),
- la durée non-voisée correspondant à une densité maximum (voir figure 4.2) (UDmax)
- la densité correspondante à UDmax (DUDmax)
- les quatre densités correspondant à une durée égale à n fois DUDmax (n allant de 2 à 5) (DUDn).

La valeur de UDmax correspond à peu près au débit de parole, la valeur de DUDmax donne une mesure de la régularité de ce débit. Plus DUDmax est proche de 1 plus la régularité est importante. Les valeurs de nDUD permettent de prendre en compte les doublements de période (comme si on avait à faire avec des blanches ou des rondes en musique plutôt que de ne traiter qu'avec des noires). La somme des DUD donne une mesure plus fiable de la régularité, plus elle s'approche de 1, plus on peut considérer que le débit est régulier, sachant que sa plus petite contenance est UDmax.

4.2.4 Nouveaux descripteurs d'articulation

La plupart des descripteurs de l'articulation sont fondés sur l'aire du triangle vocalique. Ce triangle est formé par les trois voyelles /a/, /i/ et /u/ dans l'espace formé par les deux premiers formants F1 et F2 [Lee et al. 04]. Plus l'aire du triangle est important plus l'articulation est forte. Dans l'hypothèse où nous n'avons pas accès à la transcription, nous ne pouvons pas construire ce triangle. Nous proposons donc une méthode permettant

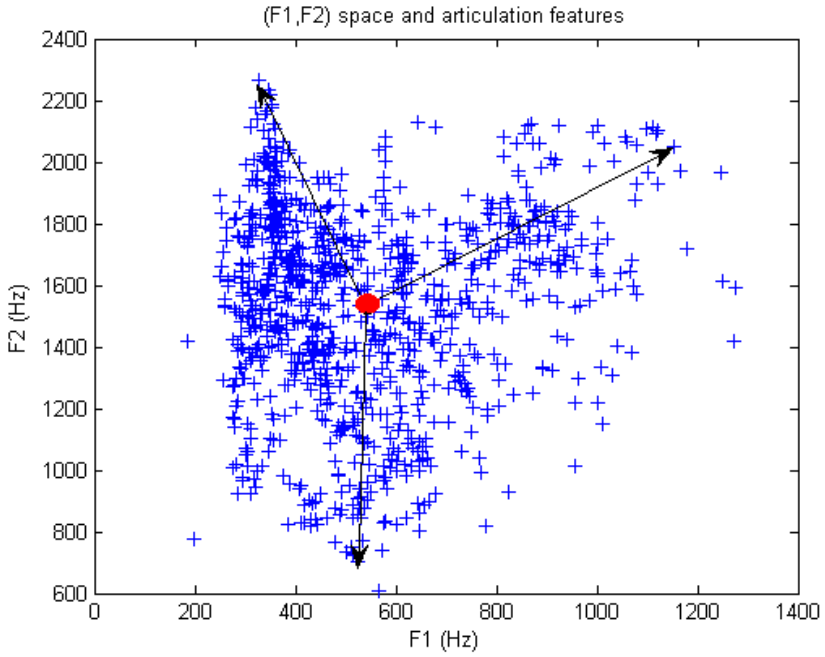


FIGURE 4.3 – Plan (F1,F2) pour une partie voisée, en rouge l'équibarycentre, en bleu les points (F1,F2), en noir la distance aux instants i.

d'estimer l'articulation, uniquement à partir des valeurs de F1 et F2 sur une partie voisée.

L'équibarycentre (bF1, bF2) (voir le point rouge sur la figure 4.3) est calculé sur une partie voisée à partir de l'ensemble des valeurs de F1 et F2. A chaque instant (ici le pas est pris à 10ms, ce qui correspond à l'échantillonage des formants fourni par Praat), la distance articulatoire est définie par l'equation 4.3.

$$d(i) = \sqrt{(F1(i) - bF1)^2 + (F2(i) - bF2)^2} \tag{4.3}$$

Ce descripteur est continu et défini uniquement sur les parties voisées, la moyenne et l'écart-type par partie voisée peuvent être utilisés pour décrire l'ensemble d'un segment, au même titre que la moyenne sur l'ensemble des points voisés d'un même segment. Cette distance est équivalente à une fréquence, elle peut donc être convertie en semitons. Comme la distance est calculée à partir de la définition d'un barycentre, elle n'est a priori que peu dépendante du locuteur.

4.2.5 Conclusion sur les nouveaux descripteurs de rythme et d'articulation

Les descripteurs que nous avons développé dans ce paragraphe ont l'avantage de ne pas être liés à une segmentation linguistique. Ils peuvent être utilisés sur les parties voisées,

mais seront plus précis lorsqu'ils sont calculés au niveau du noyau vocalique qui est plus stable.

Les descripteurs de débit proposés permettent d'estimer la répartition des durées voisées ou non-voisées ou sein d'une même phrase. Cet aspect est à relier avec la visualisation d'une partition de musique : par son simple graphisme, le lecteur peut estimer la quantité de valeurs courtes et de valeurs longues, savoir si les durées entre les notes sont régulières ou non. L'utilisation de la densité de probabilité est une idée originale qui permet de transposer l'aspect graphique de la partition à un signal de parole.

Plusieurs études ont été réalisées pour montrer l'intérêt de ces nouveaux descripteurs, notamment dans le cadre du challenge Interspeech 2012 sur la personalité (section 5.2.3.1, reconnaissance automatique de la personalité).

4.3 Robustesse des descripteurs pour la reconnaissance des émotions face à différentes variabilités

Nous souhaitons estimer la robustesse des descripteurs acoustiques pour la reconnaissance des émotions suivant trois variabilités : l'environnement acoustique, les locuteurs et le type de signal à partir duquel ils sont calculés. Pour cela nous avons utilisé un ensemble de 208 descripteurs qui sont déterminés soit sur les parties voisées, soit sur les parties non-voisées, soit sur l'ensemble du signal. Ils correspondent à l'ensemble des descripteurs usuels auxquel nous en avons ajouté d'autres (comme les intra/interF0, les bandes de Bark ou bandes harmoniques) et nos nouveaux descripteurs de rythme et d'articulation (tableau 4.1).

4.3.1 Influence du contexte sur les descripteurs

Dans cette section nous proposons deux expériences qui permettent d'analyser l'influence de l'environnement acoustique et du locuteur sur l'ensemble des descripteurs sélectionnés.

4.3.1.1 Expérience n°1 : Influence de l'environnement acoustique

Nous avons réalisé une première expérience afin d'étudier l'influence de l'environnement acoustique sur les descripteurs émotionnels. Pour cela nous avons utilisé un petit corpus enregistré dans deux salles différentes : une salle réverbérante, le studio de l'Institut de la Vision dans lequel nous avons enregistré le corpus IDV-HH, et une salle standard, l'i-room, salle d'expérimentation du LIMSI dans laquelle a été enregistré le corpus NAO-HR1. Pour chacune des salles le matériau sonore est identique : même microphone (micro-cravate AKG PT40 Pro Flexx), même locuteur, même support lexical : 5 phrases ont été répétées 5 fois, soit un total de 25 segments par salle. Nos 208 descripteurs sont évalués sur les 25 segments de chaque salle.

Si l'influence de l'acoustique de la pièce sur le calcul d'un descripteur est important, les valeurs obtenues sur l'ensemble des segments dans chacune des salles devraient être sensiblement différentes. Pour estimer cette différence, nous avons réalisé un test ANOVA pour chaque descripteur entre les deux salles. Pour un descripteur donné, la différence

entre les deux salles sera jugée significative, si la valeur du test est inférieure à 0,005. Les descripteurs qui semblent être les plus sensibles à l'acoustique de la salle sont listés dans le tableau 4.2.

D'après les résultats obtenus, on peut remarquer que la F0 n'est pas si robuste à l'environnement acoustique, de même que les formants. En effet, la réverbération d'une salle modifie l'ensemble du spectre en ajoutant notamment un délai, ce qui peut augmenter le nombre d'erreurs d'estimation. Nous pouvons également noter le fait que la réverbération affecte assez fortement les coefficients cepstraux. Nous retrouverons ce phénomène lors de l'identification du genre dans des salles acoustiquement différentes. Les valeurs de qualité vocale (ou micro-prosodie) extraites avec Praat, sont également à manier avec précaution car elles ne semblent pas très robustes aux changements d'environement acoustique.

Un résultat intéressant qu'apporte cette première expérience est que les bandes harmoniques semblent être plus robustes face à l'environnement acoustique que les bandes de Bark. Les descripteurs de rythme que nous avons mis en place semblent assez robustes face à l'acoustique de la salle sauf pour le coefficient de précision P2 (coefficient du 1er ordre de la modélisation polynomiale des pics de loudness, voir l'équation 4.2).

Lorsque les descripteurs sont normalisés, c'est-à-dire qu'on soustrait pour chaque segment et pour chaque descripteur la valeur moyenne de la salle correspondante au segment, la différence de valeur s'annule complétement entre chacune des salles. Mais dans le cadre d'une interaction homme-machine, cela suppose d'avoir une norme connue pour chaque pièce où a lieu l'interaction, ce qui est évidemment loin d'être anodin.

4.3.1.2 Expérience n°2 : Influence de la tâche, un exemple : acté/spontané

Parmi les données relatives au contexte, le type d'émotions collectées est susceptible d'influer très fortement la variation des descripteurs acoustiques. Plus la tâche sera définie, support lexical contraint, émotions actées, plus les résultats de la classification auront tendance à être performants. Scherer [Scherer et al. 03] a déjà montré que les scores de reconnaissance des émotions étaient bien meilleurs sur des données actées que sur des données spontanées. Dans notre étude [Tahon and Devillers 10], nous cherchons à caractériser le degré de spontanéité d'un corpus au travers des descripteurs acoustiques.

Pour cela, nous avons choisi trois corpus : JEMO (acté), CINEMO (induit) et CEMO (spontané). Nous proposons de comparer les valeurs des descripteurs acoustiques entre les signaux de colère et les signaux des autres émotions. En effet, la colère est l'une des émotions les plus marquées, elle est généralement présente dans tous les corpus. C'est la plupart du temps l'émotion la plus active d'un corpus. Pour chaque descripteur acoustique, nous avons défini une distance (équation 4.4) entre cette colère et l'ensemble des émotions d'un même corpus. Plus cette distance est importante, plus le descripteur aura une valeur éloignée entre la colère et les autres émotions. Pour un descripteur donné, nous pouvons alors dire que plus la distance est importante, plus l'expression de la colère est prototyique. Cette hypothèse se vérifie pour la plupart des descripteurs acoustiques (figure 4.4) utilisés dans cette étude (tableau 4.3).

$$Distance(d) = \frac{moyenne(Colère, d) - moyenne(All, d)}{moyenne(All, d)} \tag{4.4}$$

Les descripteurs pour lesquels la différence de distance entre le corpus acté JEMO et

Type	Voisé	Non-voisé	Tout
F0 (st)	max		
Loudness		écart-type	écart-type, min
Spectre	Bark1, 2 3 6, 16, 17	slope, Bark1	
Cepstre	moy MFCC10, 12		moy MFCC0, 2, 10, 12
	moy ΔMFCC8, 10		moy ΔMFCC5
Formants (st)	F3 (médiane, max), F1 (écart-type, max), bF1 (max), articulation (écart-type)		
Qualité vocale			jitterLocalPraat shimmerLocal-Praat, HNRPraat
Rythme			P2 (maxgauss, écart-type)

TABLE 4.2 – Descripteurs émotionnels peu robustes à l'acoustique de la salle

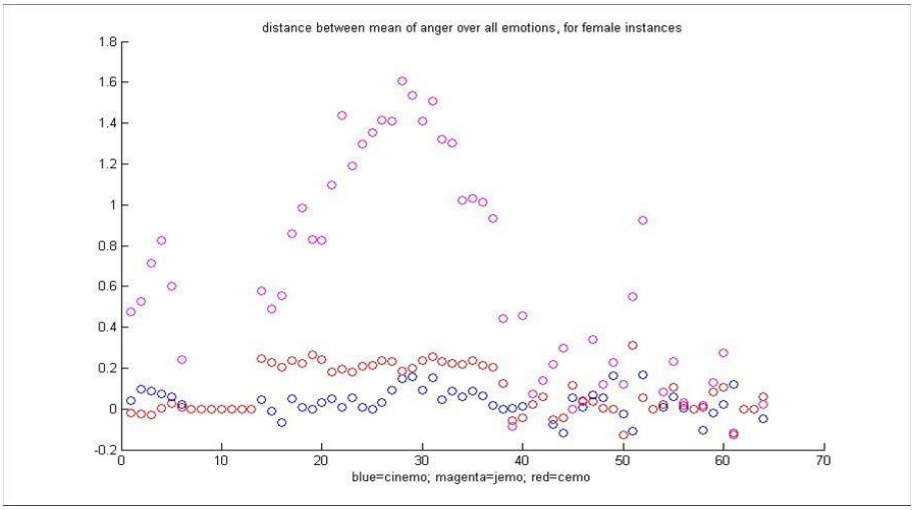


FIGURE 4.4 – Distance normalisée calculée pour les corpus JEMO (rouge), CINEMO(rose) et CEMO (bleu)

les autres est la plus élevée sont également ceux qui sont les plus sensibles à une différence de tâche. Les descripteurs qui varient le plus en fonction de la tâche sont des descripteurs spectraux (bandes de Bark, fréquences roll off, barycentre spectral), le ZCR et l'énergie (loudness).

4.3.2 Influence de la variabilité des locuteurs sur les descripteurs

Lors de la collecte de données émotionnelles du corpus IDV-HR, nous avons demandé à chaque participant de répéter trois fois la voyelle /a/ d'abord sans indication de force, puis en leur demandant de les répéter très fort. Cette demande était motivée d'abord pour des aspects de calibrage du microphone mais aussi dans l'espoir de pouvoir étudier plus précisemment l'influence de la variabilité des locuteurs sur les descripteurs émotionnels. C'est cette étude qui est présentée ci-dessous.

Nous avons donc 22 locuteurs différents homme et femme confondus. Chacun prononçant six fois la voyelle /a/ avec des intensités différentes (et donc des efforts vocaux différents). Les segments sont relativement courts, environ 0,15 s. Etant donné le matériau avec lequel nous avons réalisé l'expérience, les descripteurs non-voisés, de rythme, de variation entre parties voisées, n'ont pas de sens. De même les descripteurs calculés sur la totalité du signal sont équivalents à ceux calculés sur la partie voisée.

La plupart des 208 descripteurs étudiés varient énormément suivant les locuteurs, cependant un test ANOVA sur les 22 locuteurs, permet de montrer que certains semblent relativement invariants selon les locuteurs ($p>0{,}005$). Le tableau 4.4 présente uniquement les descripteurs les moins invariants suivant les locuteurs. Evidemment ces résultats sont fondés sur des /a/ prononcés de manière totalement artificielle. La difficulté de travailler avec des voyelles plus spontanées est de pouvoir gérer un signal extrêmement bref.

Sur ces signaux, la fréquence fondamentale ne varie quasimement pas pour l'ensemble

Type de descripteur	Nom (n° de référence)
Spectral	Rolloff 5, 25, 50, 75, 95% (1-5), barycentre (6), pente (7-8), bande de Bark (14-37), bande d'énergie (9-13)
Cepstral	MFCC 0-12 (39-50)
Energie	MeanLoudness (52), RMSenergy (53)
Domaine temporel	ZCR (51), Ratio voisé/non-voisé (61)
Qualité vocale, Formants	RatioF0 (54), varF0 (55), F2-F1 (56), F3-F2 (57), varF1 (58), varF2 (59), varF3 (60), Jitter (62), Shimmer (63), HNR (64)

TABLE 4.3 – Liste des descripteurs acoustiques pour l'étude de la distance entre colère et autres émotions ainsi que les n° correspondant à la figure 4.4.

Type	Voisé
F0 (st)	écart-type, intraF0
Spectre	rolloff (25, 50, 75, 95%), barycentre, Bark21
Cepstre	moy ΔMFCC5, 12
Formants (st)	F1 (médiane, max, min), F2 (médiane, max, min), F3 (max), articulation (mean)

TABLE 4.4 – Descripteurs émotionnels relativement invariants suivant les locuteurs pour des /a/ prononcés hors contexte.

des locuteurs, c'est pourquoi l'écart-type et intraF0 sont invariants. La voyelle étant la même suivant tous les locuteurs, les valeurs de formants et d'articulation sont sensiblement identiques pour l'ensemble des locuteurs. En parole spontanée, il se peut que cela varie plus suivant les locuteurs. Les valeurs des énergies dans des bandes hautes fréquences (que ce soit sur les Bark ou les MFCCs) n'ont pas beaucoup de sens, puisqu'elles ont tendance à capturer beaucoup de bruit.

Deux résultats intéressants de cette expérience doivent être notés. Les coefficients cepstraux n'apparaissent quasiment pas dans ce tableau, ce qui signifie que leur utilisation pour la détection du locuteur reste pertinente (voir chapitre 4). Plusieurs descripteurs spectraux (fréquences roll off et barycentre spectral) semble être assez invariants suivant les locuteurs contrairement à la F0. Est-ce que ces descripteurs pourraient être utilisés pour la reconnaissance des émotions sans être influencés par le locuteur ? Afin de pouvoir généraliser ces résultats à l'ensemble d'un discours spontané, il faudrait effectuer des tests similaires sur d'autres voyelles que des /a/ mais également en sélectionnant des voyelles spontanées et non pas artificielles comme nous l'avons fait. La difficulté de ce travail étant bien sûr la brêveté des voyelles spontanées.

4.3.3 Conclusion sur la robustesse des descripteurs

Nous proposons dans cette section une sélection de descripteurs qui semblent être robustes aux variations liées à l'environnement acoustique, au type de tâche (acté, spontané ou induit) et au locuteur. Cette sélection est issue des résultats des expériences présentées dans le paragraphe 3.2.3. avec les corpus que nous avons à disposition. Etant donné la complexité globale d'un signal de parole en interaction, une telle sélection est à prendre avec précaution.

Les bandes de Bark ne semblent être robustes ni au type de tâche, ni à l'environnement acoustique.

L'énergie (loudness) ne semble pas être robuste au type de tâche, ni à l'environnement acoustique lorsqu'elle est calculée sur l'ensemble du signal. Elle est plus intéressante lorsque la fenêtre de calcul est limitée aux seules parties voisées.

Les deux premiers formants F1 et F2 (en moyenne sur l'ensemble des parties voisées, pas d'écart-type, ni de maximum) semblent être robustes à l'environnement acoustique et à la tâche. Ainsi les descripteurs d'articulation proposés seront également robustes à ces variabilités.

Les premiers coefficients cepstraux (inférieurs à 8) semblent être robustes à l'environnement et à la tâche.

4.4 Proposition d'une mesure relative pour la variabilité des locuteurs et de l'environnement

Il est quasiment impossible de trouver les meilleurs indices décrivant les émotions tellement ceux-ci sont dépendants d'un très grand nombre de facteurs (type d'émotions, type de locuteurs, environnement, tâche). Nous proposons dans cette section une approche nouvelle : est-il possible de déterminer une liste des indices les moins fiables d'après un grand nombre d'expériences. Nous avons l'oppportunité d'avoir à notre disposition une quantité assez conséquente de corpus collectés dans des situations très diverses avec environ une vingtaine de locuteurs pour chacun. Nous pouvons donc faire varier suivant les corpus, le nombre et type de locuteurs, le contexte, l'environnement acoustique, le type de corpus (prototypique, spontané, induit). A partir de ces corpus, nous avons essayer de chercher les fonctionnelles, les indices dynamiques, statistiques, les moins fiables.

4.4.1 Protocole

L'objectif est d'établir une liste "noire" de descripteurs qui ne peuvent pas être utilisés dans un grand nombre de cas. Pour cela nous avons calculé les 208 descripteurs définis dans le tableau 4.1 sur chacun des corpus spontanés étudiés : IDV-HH (26 locuteurs), IDV-HR (22 locuteurs), NAO-HR1 (10 locuteurs), AIBO-O (26 locuteurs) et AIBO-M (25 locuteurs). Nous distinguons les deux écoles dans lesquelles le corpus AIBO a été collecté, puisque ces deux sous-corpus correspondent à des conditions différentes. Nous avons testé avec ou sans normalisation au locuteur (z-norme définie et calculée sur les données du corpus lui-même). Pour déterminer les indices les moins fiables, nous avons sélectionné deux méthodes : la première consiste à estimer de manière chiffrée les variabilités liées aux locuteurs et aux émotions relativement. En effet, si un indice varie plus suivant le locuteur

Algorithme 4.1 Mesure de variabilité locuteur

$$\bar{d_L}(l) = \frac{\sum_{i=1}^{L(l)} d(i,l)}{L(l)}$$

$$\bar{d}_L = \frac{\sum_l \sum_{i=1}^{L(l)} d(i,l)}{\sum_l L(l)}$$

$$d_{L,norm}(l) = \frac{\bar{d}_L(l) - \bar{d}_L}{\bar{d}_L} \tag{4.5}$$

$$\sigma_L = \sqrt{E[d^2_{L,norm}] - E^2[d_{L,norm}]} \tag{4.6}$$

Algorithme 4.2 Mesure de variabilité émotion

$$\bar{d}_E(e) = \frac{\sum_{i=1}^{L(e)} d(i,e)}{L(e)}$$

$$\bar{d}_E = \frac{\sum_e \sum_{i=1}^{L(e)} d(i,e)}{\sum_e L(e)}$$

$$d_{E,norm}(l) = \frac{\bar{d_E}(e) - \bar{d_E}}{\bar{d}_E} \tag{4.7}$$

$$\sigma_E = \sqrt{E[d^2_{E,norm}] - E^2[d_{E,norm}]} \tag{4.8}$$

que suivant les émotions, il sera potentiellement moins efficace pour la reconnaissance des émotions. La seconde méthode consiste à faire une sélection automatique d'indices sur la valence. Nous avons choisi la valence, puisque les autres catégories émotionnelles ne se retrouvent pas dans les cinq corpus que nous souhaitons étudier.

4.4.2 Mesure de variabilité

La mesure de variabilité est ici envisagée pour un descripteur d donné pour un corpus choisi. Soit $d(i,l)$ la valeur du descripteur d de l'instance i pour le locuteur l avec $i \in [1 : L(l)]$ et $d(i,e)$ la valeur du descripteur d de l'instance i pour l'émotion e avec $i \in [1 : L(e)]$. $L(l)$ (respectivement $L(e)$) correspond au nombre d'instances du locuteur l pour une émotion neutre (respectivement au nombre d'instances de tous les locuteurs pour une émotion e qui n'est pas neutre).

Pour chaque locuteur l, la valeur moyenne du descripteur d est normalisé par la valeur moyenne de ce descripteur sur l'ensemble du corpus (eq. 4.5). La mesure de variabilité correspond alors à l'écart-type calculé sur l'ensemble des locuteurs (eq. 4.6). La varia-

bilité liée aux émotions est déterminée de manière similaire sur l'ensemble des émotions (algorithme 4.2). La variabilité liée au corpus peut être déterminée de manière similaire en prenant l'ensemble des corpus en compte. Le rapport entre les deux variabilités (eq. 4.9) permet d'estimer l'importance relative de la variabilité liée aux locuteurs par rapport à celle liée aux émotions.

$$R = \frac{\sigma_E}{\sigma_L} \tag{4.9}$$

4.4.2.1 Exemple

L'exemple suivant permet d'illuster la mesure de variabilité que nous proposons. Le corpus utilisé est NAO-HR1, il y a 10 locuteurs et 5 états émotionnels dont un état neutre.

Pour estimer le pouvoir de discrimination du descripteur suivant les émotions ou le locuteur, il faut comparer le rapport de variabilité à 1 :

- $R \approx 1$, le descripteur varie globalement autant suivant les émotions que suivant les locuteurs,
- $R < 1$, la variabilité liée aux locuteurs est importante relativement à celle liée aux émotions,
- $R > 1$, la variabilité liée aux émotions est importante relativement à celle liée aux locuteurs.

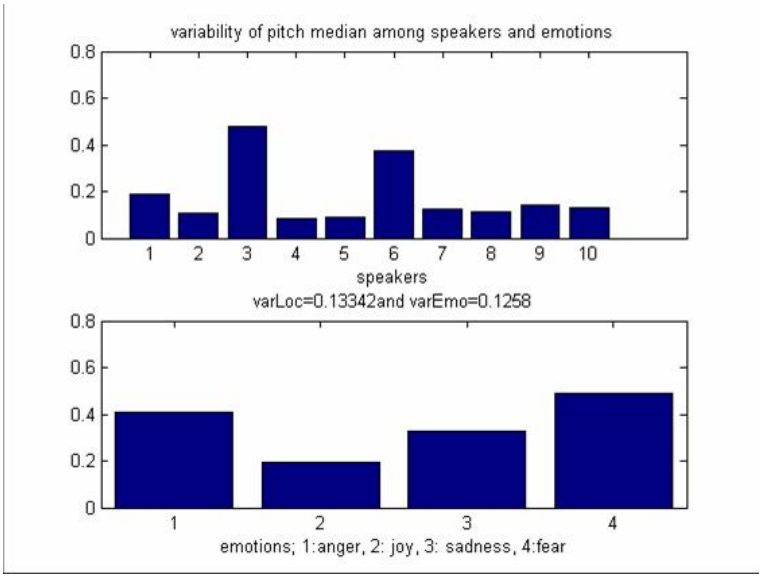
Dans notre exemple, le descripteur F0 médian est proche de 1, il n'est donc pas intéressant pour discriminer les émotions indépendamment du locuteur. Le descripteur MFCC12 est largement supérieur à 1, il est donc intéressant pour disciminer les émotions. La figure 4.5 détaille ces résultats : les valeurs de F0 médian normalisé sont du même ordre de grandeur ($\forall l, F0med_{L,norm}(l) < 0,5$ et $\forall e, F0med_{E,norm}(e) < 0,5$) suivant les émotions (bas) et les locuteurs (haut) alors que les valeurs de MFCC12 normalisé sont plus importantes suivant les émotions (particulièrement la colère, $MFCC12_{E,norm}(colère) \approx 1,5$) que suivant les locuteurs ($\forall l, F0med_{L,norm}(l) < 1$). Une fois normalisé au locuteur, le rapport de variabilité du descripteur F0 médian devient supérieur à 1 ($R = 1,15$). La normalisation au locuteur permet donc d'augmenter le pouvoir de discrimination des émotions du F0 médian. Ce n'est pas le cas du descripteur MFCC12.

En fonction de ce rapport, nous avons classé les 208 descripteurs groupés par familles suivant le type de fonction statistique, ou suivant le type de descripteur étudié. Ces familles regroupent plusieurs descripteurs de manière à ce que leur nombre soit relativement homogène entre les familles. Avant de faire ce classement, certains indices doivent être retirés de la liste car ils ne semblent pas pertinents pour une utilisation à grande échelle. C'est le cas de certains descripteurs de rythme (précision). A chaque famille est associé un rang moyen correspondant à la moyenne des rangs des descripteurs qu'elle contient.

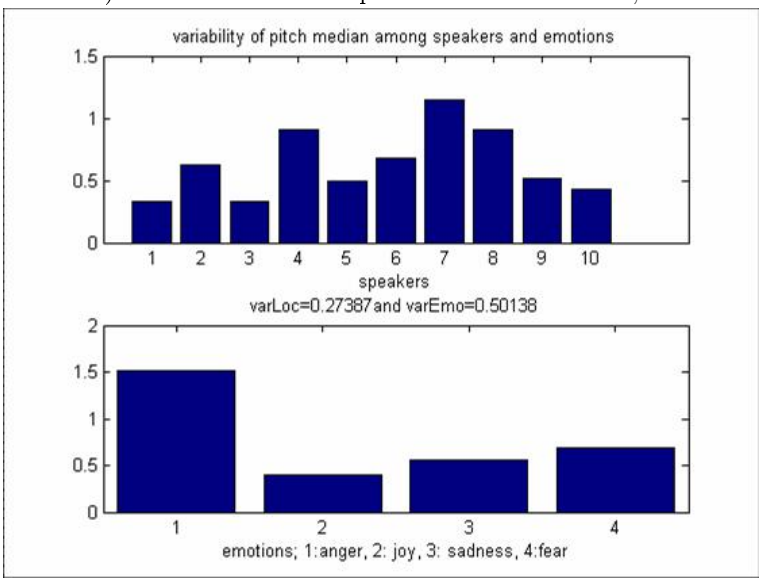
4.4.2.2 Résultats

Les tableaux en annexe B1 résume les moyennes sur l'ensemble des cinq corpus de ce rang moyen. La figure 4.6 permet de mettre en évidence les familles de descripteurs dont la variabilité liée au locuteur est plus forte que celle liée aux émotions. Cette variabilité pouvant être réduite avec une normalisation au locuteur.

Sans normalisation, les familles dont le rang est le plus élevé (supérieur à 110) sont :



a) Variabilité du descripteur F0 médian $R = 0,92$



b) Variabilité du descripteur MFCC12 $R = 1,83$

FIGURE 4.5 – Exemple de mesure de variabilité locuteur et émotion sur le corpus NAO-HR1 (10 locuteurs, 5 états émotionnels dont neutre) pour la F0 médiane (a) et le coefficient MFCC12 (b)

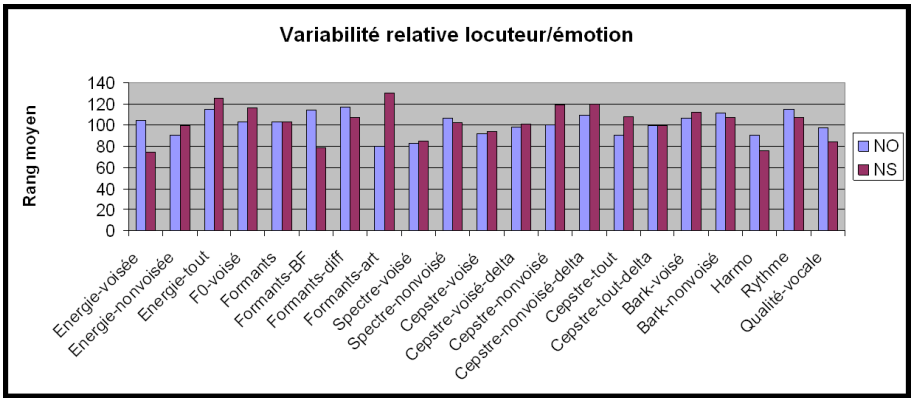


FIGURE 4.6 – Variabilité relative locuteur/émotion, influence de la normalisation locuteur et rang moyen par famille de descripteurs (NO : sans normalisation, NS : normalisation locuteur) sur l'ensemble des corpus Romeo (tableaux B1, B2).

- l'énergie sur tout le signal,
- les largeurs de bande des formants F1, F2 et F3,
- les différences de formants F2-F1 et F3-F2,
- les bandes de Bark sur le signal non-voisé,
- le rythme.

L'articulation semble être un paramètre relativement robuste au locuteur, il a été conçu pour être indépendant des locuteurs. Par contre une normalisation au locuteur lui est extrêmement préjudiciable. La normalisation au locuteur permet de rendre robuste un certain nombre de familles : l'énergie sur les parties voisées, les bandes de fréquence des formants, les bandes harmoniques et la qualité vocale. L'ensemble des descripteurs cepstraux voient leur rang augmenter avec la normalisation au locuteur. Globalement l'utilisation des portions non-voisées du signal pour le calcul des descripteurs (bandes de Bark, cepstre, descripteurs spectraux) donne un rang moins bon que l'utilisation des portions incluant les parties voisées (parties voisées uniquement ou tout le signal) sauf pour l'énergie. Les dérivées des coefficients cepstraux ont un rang moins bons que les coefficients cepstraux. Nous retrouvons ce résultat dans les expériences cross-corpus sur différents corpus (section 6.2.2.1).

La liste de descripteurs étudiée a été conçue de manière à être exhaustive, certains indices sont donc redondants. C'est le cas des bandes de Bark et des bandes Harmoniques. D'après le classement obtenu, il semblerait que les bandes Harmoniques soient plus intéressantes pour la discrimination des émotions que les bandes de Bark. La différence de classement sans normalisation entre les deux méthodes de calcul du jitter (Praat : 82,8 ; Matlab : 80,8) n'est pas significative sur l'ensemble des corpus. Nous pouvons donc considérer que le calcul proposé avec Matlab est pertinent (équation 4.1).

Le rang moyen calculé suivant des familles de fonctions statistiques (médiane, écart-type, maximum et minimum) est sensiblement identique quelles que soit les familles (en moyenne 111,2), ce rang diminue de 12 avec une normalisation au locuteur. Ainsi il semble difficile d'inclure certains fonctions statistiques dans la liste "noire".

4.4.2.3 Liste "noire" descripteurs

A partir de cette première expérience fondée sur les variabilités relatives liées aux locuteurs et aux émotions sur un ensemble de cinq corpus, nous pouvons établir une première ébauche de liste "noire" :

- bandes de Bark,
- descripteurs calculés sur les parties non-voisées,
- dérivées des coefficients cepstraux,
- différences de formants en semitons, et bandes de fréquences de ces formants.

Très peu d'études ont été réalisées sur de telles sélections de descripteurs. Nous pouvons néamoins citer à nouveau les travaux de Ruiz [Ruiz et al. 10] sur des signaux de parole enregistrés dans la cabine de pilotage des avions. Sa conclusion est qu'il reste extrêmement complexe de faire correspondre les variations de descripteurs acoustiques avec des phénomènes comme la fatigue ou l'engourdissement et donc de définir à partir d'analyses précises les descripteurs les plus robustes.

4.4.3 Sélection automatique des descripteurs

Nous avons également réalisé une sélection automatique des descripteurs sur le même principe que les travaux de [Polzehl et al. 10]. La performance individuelle de chaque descripteur étant donné sa classe émotionnelle est estimée suivant un algorithme de sélection automatique Information Gain Attribute Evaluation implémenté dans l'outil Weka. Cet algorithme permet d'évaluer chaque attribut (ou descripteur) individuellement grâce à une mesure d'information (calcul d'entropie) connaissant a priori la classe. Afin d'obtenir une estimation la plus générale possible, nous utilisons un système de cross-validation (10-folds). La sélection des descripteurs est d'abord réalisée corpus par corpus, puis les rangs de chaque descripteurs sont moyennés sur les cinq corpus étudiés. Les méthodes sont ainsi les mêmes que dans le paragraphe 4.4.2. Nous n'avons testé la sélection automatique des descripteurs uniquement sans normalisation au locuteur. Les résultats par familles de descripteurs sont détaillés en annexe B3.

4.4.3.1 Résultats

Les résultats obtenus avec une méthode de sélection automatique sont à interpréter avec précaution. En effet, ils peuvent être assez différents suivant le type d'algorithme de sélection utilisé. Nous ne présentons pas de résultats moyens sur plusieurs méthodes, étant donné que nous recherchons surtout à déterminer les descripteurs extrêmes, en particulier ceux qui sont les moins bien classés. Sans normalisation, les familles dont le rang moyen est le plus élevé (supérieur à 110) sont :

- les formants F1, F2 et F3,
- l'articulation,
- le cepstre dérivé déterminé sur les parties non-voisées du signal,
- le cepstre déterminé sur l'ensemble du signal,
- le cepstre dérivé déterminé sur l'ensemble du signal,
- les bandes de Bark déterminées sur les parties voisées,
- le rythme.

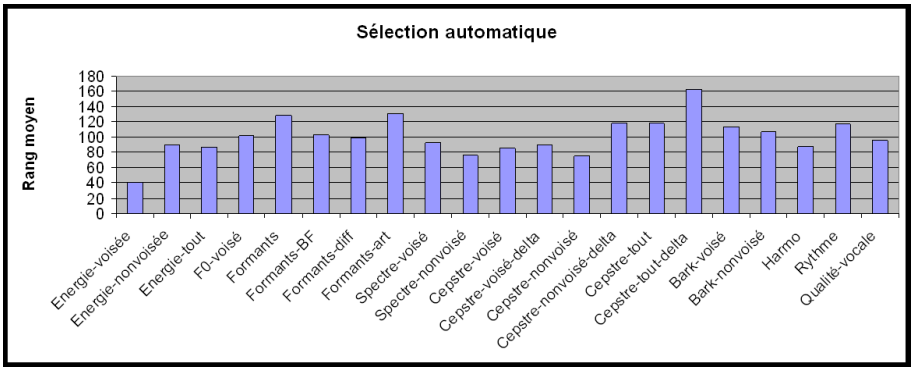


FIGURE 4.7 – Sélection automatique des descripteurs, sans normalisation, rang moyen par famille de descripteurs (tableau B3)

Nous retrouvons quelques résultats obtenus dans le paragraphe précédent, notamment en ce qui concerne les dérivés des coefficients cepstraux, les bandes de Bark. Par contre, l'utilisation des portions non-voisées du signal ne semble pas pénaliser le rang moyen de manière globale (uniquement sur l'énergie). Les descripteurs de rythme sont complexes à calculer automatiquement, leur calcul actuel ne permet sans doute pas de traiter des cas particuliers. De plus, il semble nécessaire d'optimiser certains paramètres nécessaires à leur détermination.

4.4.3.2 Liste "noire" de descripteurs

A partir de cette seconde expérience fondée sur un algorithme de sélection automatique sur un ensemble de cinq corpus, nous pouvons renforcer la première ébauche de liste "noire" :

- bandes de Bark,
- descripteurs calculés sur les parties non-voisées ou sur l'ensemble du signal,
- dérivées des coefficients cepstraux.

La liste de nos 208 descripteurs utilisée pour cette étude doit alors être réactualisée. C'est cette nouvelle liste qui sera utilisée pour les expériences de reconnaissance automatique présentées au chapitre 5.

4.5 Conclusion

Dans le chapitre 3, nous avons passé en revue un très large éventail de descripteurs acoustiques utilisés pour les études en parole que ce soit pour de la reconnaissance automatique, de la synthèse, de la transformation de voix, ou même de la linguistique. Certains descripteurs semblent très intéressants mais ne peuvent pas être utilisés tels quels dans nos travaux de thèse puisqu'ils nécessitent une connaissance a priori soit du contenu linguistique, soit d'une segmentation en unités linguistiques (comme les battements ou les syllabes).

Il existe de très nombreux descripteurs spectraux et cepstraux, les indices de rythme sont nettement moins nombreux. Cela est principalement lié à la structure quasi chaotique du rythme dans la parole. Cet aspect chaotique est probablement du à la perception d'une superposition de couches rythmiques intermêlées. Manière de parler, personnalité, humeur, émotion sont perçues différemment grâce à des structures rythmiques propres. Par exemple, le stress d'une personne pourrait se reconnaître suivant que son débit de parole est régulier ou non, alors que son état émotionnel se traduirait par l'allongement des voyelles (cas de la tristesse) ou vitesse du débit de parole. Ces couches rythmiques ont des structures différentes, mais portent également sur des durées différentes (très courtes dans le cas des émotions, longues pour la personnalité ou le stress). Il est donc a priori difficile de ne parler que d'un seul rythme et de définir à partir d'un niveau physique des indices efficaces pour décrire tel ou tel aspect affectif.

Nous proposons deux types de descripteurs de rythme pour la reconnaissance de la personnalité et du stress, l'un mesurant la précision c'est-à-dire un aspect très perceptif du rythme, l'autre mesurant la régularité du rythme dans la parole au travers de la répartition statistique des durées des parties voisées. Ces descripteurs sont conçus pour des durées temporelles relativement longues (plus de 10 s), il est donc nécessaire de les modifier pour qu'ils soient intéressants pour la reconnaissance d'états émotionnels.

Un autre travail important présenté dans ce chapitre est l'étude de la robustesse des indices face à plusieurs variabilités. Une première expérience cherche à déterminer la robustesse des descripteurs acoustiques à l'environnement acoustique avec des signaux d'un même locuteur sur un même contenu linguistique dans deux salles différentes. Une seconde expérience permet de valider une mesure de distance entre différents corpus fonction de la tâche (acté, spontané, induit). Plus cette distance est importante pour un descripteur donné, plus celui-ci est sensible aux variations de tâche. Enfin une dernière expérience basée sur des /a/ prononcés par différents locuteurs permet de mettre en évidence le caractère invariant de certains descripteurs face à plusieurs types de voix.

Le dernier enjeux consistait à obtenir une liste "noire" de descripteurs qui ne sont a priori pas robustes pour la classification des émotions (ici seule la valence a été étudiée) au travers d'une étude multi-corpus. Cette liste peut être établie par une sélection automatique de descripteurs qui peut varier énormément en fonction de l'algorithme choisi, ou bien par une mesure de variabilité que nous proposons. Cette mesure de variabilité utilise directement les valeurs des descripteurs, elle a donc l'avantage de ne pas dépendre d'un algorithme et de ses paramètres comme dans le cas de la sélection automatique de descripteurs. Au vu de l'ensemble de ces analyses, la liste "noire" de descripteurs contient les bandes de Bark, les indices d'énergie calculés sur l'ensemble du signal (voisé et non-voisé mélangés), les différences entre formants (F2-F1 et F3-F2), les largeurs de bande des formants et les dérivés des coefficients cepstraux. Les conclusions sur les autres descripteurs sont plus subtiles. Ces résultats sont relativement cohérents avec ceux de l'état de l'art : notamment Clavel [Clavel 07] a choisi de faire deux modèles séparés, l'un sur les parties voisées, l'autre sur les parties non-voisées, en aucun cas sur l'ensemble du signal. Le faible intérêt des coefficients cepstraux a été validé par une expérience de reconnaissance automatique [Tahon et al. 11]. Finalement dans cette approche, tous les descripteurs sont variables selon les corpus utilisés, cependant certains varient plus que d'autres. A la différence d'une approche focalisée sur quelques descripteurs pertinents pour des situations contrôlées, nous cherchons à définir les familles de descripteurs qui varient globalement

plus que d'autres afin de tirer des tendances utiles par la suite pour la reconnaissance des émotions et des locuteurs.

Mes travaux proposent plusieurs mesures pour évaluer la robustesse d'un descripteur donné à différentes variabilités. Contrairement à la plupart des analyses sur les descripteurs acoustiques, nous avons travaillé sur plusieurs corpus en parallèle, ce qui permet de valider ces mesures. Cependant il est nécessaire de vérifier leur validité sur plus de données encore (notamment des langues autres que le français). Les nouveaux descripteurs de rythme sont fondés sur un parallèle avec la musique, les propositions que nous avons faites peuvent encore être optimisées comme la régularité (notamment au niveau de la répartition statistique des durées). La liste "noire" de descripteurs proposée est intéressante puisqu'elle est basée sur une analyse des données indépendante d'un quelconque algorithme. Elle a été validée par une sélection automatique de descripteurs. Pour confirmer son intérêt dans la reconnaissance des émotions et du locuteur, elle doit encore être intégrée aux systèmes de reconnaissance automatique présentés aux chapitres 5 (caractérisation du locuteur) et 6 (reconnaissance des émotions).

Quatrième partie

Influence des variabilités présentes lors d'une interaction humain-robot sur la reconnaissance automatique d'indices paralinguistiques

Reprenons l'étude de nos variabilités au cours d'une interaction entre un robot et un humain dans un contexte réaliste. Le premier challenge consiste à étudier les émotions suivant les locuteurs et réciproquement. Nous cherchons à déterminer si il existe des descripteurs acoustiques invariants suivant les émotions puis suivant les locuteurs. Est-ce qu'un système de reconnaissance automatique peut faire la différence entre un nouveau locuteur et une nouvelle émotion ? Est-ce que les émotions et les locuteurs peuvent se caractériser par les mêmes descripteurs ou en existe-t-il des spécifiques aux locuteurs, spécifiques aux émotions ?

Dans le chapitre 5, après un bref état de l'art sur l'identification du locuteur, nous présentons l'ensemble de nos travaux sur la caractérisation de celui-ci à partir de parole émotionnelle. Et dans le chapitre 6, nous présentons nos résultats de reconnaissance automatique d'indices paralinguistiques indépendamment sur plusieurs corpus puis en croisant les corpus.

Chapitre 5

Caractérisation du locuteur dans un contexte émotionnel

L'identification du locuteur se fonde sur des indices acoustiques dont la mesure a l'avantage d'être simple et non intrusive, mais qui sont très loin de présenter la fiabilité de mesures biométriques comme les empreintes digitales ou la mesure de l'iris. En effet, la voix n'est pas une empreinte et ne traduit que de manière indirecte la configuration du conduit vocal au cours du geste phonatoire. Elle peut être déformée volontairement (imitateurs) ou non (maladie, stress, émotion,...) et varie lentement au fil du temps (veillissement des organes phonatoires) ; de plus les conditions d'enregistrements influent considérablement sur les caractéristiques acoustiques. Une identification parfaite par la voix est certainement impossible, mais dans un contexte applicatif particulier comme celui dans lequel nous sommes, son utilisation est parfaitement envisageable et intéressante.

Dans ce chapitre, nous verrons comment l'expression d'émotions peut influencer l'identification du genre et celle du locuteur. Nous évaluerons également l'influence de la durée de test, c'est-à-dire le temps que prend le système pour analyser un signal inconnu qui est une donnée très importante dans un contexte d'interaction humain-robot.

5.1 État de l'art

Les systèmes de reconnaissance du locuteur réalisent le plus souvent une analyse spectrale à court terme pour extraire le timbre, et caractérisent la voix du locuteur au moyen d'une distribution du timbre dans l'espace acoustique. Les paramètres spectraux sont généralement des paramètres MFCC (*Mel Frequency Cepstral Coefficients*, déjà présentés au chapitre 3) très souvent utilisés aussi en transcription automatique ou en identification de la langue car ils concentrent des informations sur le timbre (fréquences de résonances ou formants) dans un nombre réduit de paramètres (typiquement des vecteurs de dimension 10 à 15, échantillonnés à 100 Hz sur le signal audio). A partir de l'enregistrement de la voix d'une personne connue, il est possible de calculer une densité de probabilité au moyen d'un mélange de gaussiennes (*Gaussian Mixture Models*, GMM) sur l'ensemble des vecteurs spectraux calculés sur l'enregistrement. Disposant d'un modèle

associé à chacun des locuteurs connus, il est alors possible d'identifier le locuteur d'un nouvel enregistrement en comparant cet enregistrement à chacun des modèles et en sélectionnant le meilleur modèle. Cette approche par GMM, popularisée dans les années 1990 par Reynolds [Reynolds and Rose 95], reste dominante aujourd'hui avec un certain nombre de variantes et d'améliorations [Reynolds et al. 00] : on dispose de deux enregistrements (environ 2 minutes de conversation par exemple dans les campagnes NIST SRE) et on souhaite savoir si les deux enregistrements proviennent ou non d'un même locuteur (vérification du locuteur), sur des corpus comportant plusieurs centaines de personnes différentes. Malheureusement, comme toutes les approches par apprentissage statistique, les systèmes reconnaissent ce qui est similaire aux conditions d'apprentissage, et les performances se dégradent rapidement dans des configurations non observées.

5.1.1 Les différentes tâches associées à la reconnaissance du locuteur

Identification du locuteur On cherche à connaître l'identité d'un locuteur (son nom si il est connu, ou son genre)

Segmentation en locuteur ("*speaker diarization*") Sur un flux de parole, on cherche à identifier qui parle quand. Deux tâches se superposent : une segmentation en tour de parole, et ensuite une tâche d'identification d'un locuteur connu, ou de regroupement en locuteur identiques.

Vérification du locuteur Vérifier si le locuteur L est bien celui qui parle.

5.1.2 Paramètres acoustiques

Le spectre du signal reflète la structure du conduit vocal d'une personne qui est le facteur physiologique prépondérant pour distinguer une voix d'une autre. Pour reconnaître une voix, on pourrait donc utiliser la représentation spectrale LPC, cependant ces coefficients peuvent être très affectés par le bruit. Les coefficients MFCCs présentent l'avantage de modéliser également le conduit vocal tout en étant relativement robustes au bruit [Reynolds and Rose 95]. La plupart des travaux de recherche en identification du locuteur se basent donc sur l'extraction des coefficients MFCCs tous les 10 ms sur une fenêtre de 30 ms. Une fenêtre de hamming est souvent utilisée pour l'extraction des paramètres acoustiques. Le nombre des coefficients est variable suivant les chercheurs (entre 10 et 15). Certains travaux utilisent les coefficients cepstraux PLP [Barras et al. 07]. On peut également ajouter l'énergie (dont l'information est déjà présente dans le coefficient 0 des MFCCs) ainsi que les dérivées des paramètres usuels.

La reconnaissance du locuteur est bien sûr plus efficace si elle n'est réalisée que sur les voyelles porteuses de l'information sur le conduit vocal, mais on peut également extraire les paramètres acoustiques sur l'ensemble des signaux d'énergie suffisante ou sur les parties voisées. Ces signaux peuvent être détectés grâce à une détection d'activité vocale (VAD) ou une détection de voisement.

5.1.3 Modélisation par mélange de gaussiennes (GMM)

La modélisation par mélange de gaussiennes consiste à représenter une classe (ici un locuteur) par un ensemble de gaussiennes sur plusieurs observations (ici descripteurs acoustiques). Il s'agit alors de déterminer les paramètres de chaque gaussienne (moyenne, variance et amplitude) en fonction d'un critère de vraisemblance. Ces paramètres sont optimisés suivant l'algorithme EM (expectation, maximisation). Le calcul de la densité de probabilité de chacun des mélanges permet de comparer les résultats obtenus sur le signal de test par rapport aux différents modèles existants.

On suppose que chaque observation x suit une loi combinaison linéaire de K gaussiennes, ou loi normales (voir équation 5.1).

$$x \sim \sum_k \pi_k \cdot N(\mu_k, \sigma_k) \tag{5.1}$$

Les paramètres π_k, μ_k, σ_k (probabilité a priori, paramètres de la gaussienne k) sont inconnus.

5.1.3.1 Modélisation GMM

Mélange de gaussiennes Une variable x suit une loi normale si sa densité de probabilité se représente par la fonction suivante 5.2. La matrice de covariance est généralement considérée comme diagonale, c'est-à-dire que les variables sont considérés comme étant indépendantes.

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} = N(x; \mu, \sigma^2) \tag{5.2}$$

moyenne : $\mu = E[x] = \int_x x \cdot p(x) \cdot dx$
et dispersion : $\sigma^2 = E[(x-\mu)^2] = \int_x (x-\mu)^2 \cdot p(x) \cdot dx$
Un mélange de gaussiennes est une combinaison linéaire de K lois normales (eq. 5.3) tel que $\sum_{k=1}^{K} \pi_k = 1$

$$p(x; \mu_1, \sigma_1, ..., \mu_k, \sigma_k) = \sum_{k=1}^{K} \pi_k \cdot p(x; \mu_k, \sigma_k) \tag{5.3}$$

Chaque classe sera modélisée par un mélange de gaussiennes, la probabilité que x appartienne à cette classe s'esprime alors suivant l'équation 5.4

$$p(x/C_q) = \sum_{k=1}^{K} \pi_{k,q} \cdot p(x, \mu_{k,q}, \sigma_{k,q}) \tag{5.4}$$

L'apprentissage consiste à déterminer pour chaque classe q, les poids $\pi_{k,q}$ et les paramètres des gaussiennes (moyennes $\mu_{k,q}$ et dispersion $\sigma_{k,q}$).

Algorithme 5.1 Algorithme EM appliqué à un mélange de gaussiennes

step E : Estimation
calcul de la vraisemblance et estimation des probabilités à posteriori
step M : Maximisation
calcul des paramètres des gaussiennes (moyenne, dispersion) sachant les probabilités à posteriori

Apprentissage des paramètres des mélanges de gaussiennes L'initialisation des paramètres des mélanges de gaussiennes se fait grâce à l'algorithme des k-moyennes. Cette première étape permet d'obtenir une bonne approximation des paramètres. L'optimisation est réalisé suivant l'agorithme E.M. (Estimation, Maximisation) [Dempster et al. 77] dont le nombre d'itérations est généralement fixé à 10.

Décision La classification est réalisée à partir d'une règle de décision fondé sur le maximum à posteriori. L'ensemble des classes sont considérées comme équiprobales. Le score de vraisemblance est déterminé suivant la formule de Bayes, par l'équation 5.5.

$$p(C_q/x) = \frac{p(C_q) \cdot p(x/C_q)}{p(x)} = \frac{1}{Q} \cdot \frac{p(x/C_q)}{p(x)} \tag{5.5}$$

La classe la plus probable pour l'observation x, sera celle correspondante au maximum de probabilité suivant l'ensemble des classes C_q. On détermine alors la log vraisemblance de l'observation x confrontée à chaque classe q. La classe correspondant le plus probablement à l'observation x est celle pour laquelle la log vraisemblance est maximum.

$$S(x, q) = log(p(C_{q/x})) \tag{5.6}$$

5.1.3.2 Normalisation des descripteurs acoustiques

Plusieurs types de normalisation ont été expérimentés [Barras and Gauvain 03]. La plus simple consiste à soustraire le cepstre moyen (Cepstral Mean Substraction ou CMS) pour chaque matrice d'entrée (apprentissage et test). Cette normalisation permet théoriquement de s'affranchir des variations dues au canal (salle d'enregistrement, différences de bruit, etc.).

La normalisation type "warping" [Pelecanos and Sridharan 01] permet de faire une projection de la distribution cepstrale observée suivant une distribution normale. Cette normalisation permet de réduire efficacement le taux d'erreur (de 8% par rapport à la normalisation CMS selon [Barras and Gauvain 03]).

La "Z-norme" permet de normaliser une valeur V par rapport à la moyenne et à la variance suivant l'équation 5.7.

$$V_{Z-norme} = \frac{V - \mu_v}{\sigma_v} \tag{5.7}$$

La "T-norme" permet de normaliser une valeur V en log par rapport à la moyenne et à la variance suivant l'équation 5.8. Cette norme permet par exemple de centrer-réduire

un score de vraisemblance par rapport à des scores d'imposteurs en identification du locuteur.

$$V_{T-norme} = \frac{log(V) - \mu_v}{\sigma_v} \tag{5.8}$$

5.1.3.3 Adaptation

Il existe deux grands types d'adaptation : l'adaptation MAP et l'adaptation MLLR [Leggetter and Woodland 95]. L'adaptation MAP est la plus couramment utilisée, nous avons donc choisi ce type d'adaptation.

L'adaptation MAP (Maximum A Posteriori) a été appliquée en identification du locuteur aux HMM et GMM par [Gauvain and Lee 94]. Elle permet d'adapter un modèle générique (Universal Background Model) a un locuteur ou un type de locuteur particulier. L'avantage de cette méthode est qu'il n'est pas nécessaire d'avoir un grand nombre de données du locuteur auquel on souhaite adapter le modèle, contrairement à la création de modèles non adaptés.

Dans [Barras et al. 07], l'adaptation du modèle UBM se fait sur quelques dizaines de secondes. L'expérience a été réalisé dans le cadre d'une campagne NIST sur un corpus de séminaires (évaluation CLEAR). La reconnaissance se fait sur plusieurs dizaines de locuteurs. Les résultats ci-dessous permettent d'avoir une idée des taux d'erreur (ER : error rate).

- apprentissage 15s, test 1s, ER =51.7% ; test 10s, ER=6.6%,
- apprentissage 30s, test 1s, ER=38.8% ; test 10s, ER=2.1%.

5.1.3.4 Mesures de performances

Deux types de mesures sont généralement utilisées sur la détection du locuteur. Dans le cas d'une tâche d'identification du locuteur, on peut avoir plus de deux résultats possibles, on utilise alors un taux d'erreur. Lors d'une tâche de vérification du locuteur ou même de reconnaissance du genre, le problème est binaire, on utilise alors des mesures standards mises en place entre autre lors des campagnes NIST SRE. Ces mesures prennent en compte les fausses alarmes (lorsqu'on croit reconnaître un locuteur) et les détections manquées (le locuteur n'a pas été reconnu comme tel).

La fonction de coût (DCF : *Detection Cost Function*) est définie comme la somme pondérée des deux probabilités (de fausses alarmes et de détections manquées). Le coût normalisé se définit alors suivant l'équation 5.9 avec α, β choisis suivant que l'on cherche à privilégier les détections manquées ou les fausses alarmes.

$$C_{norm} = \alpha \cdot P_{manquées} + \beta \cdot P_{FaussesAlarmes} \tag{5.9}$$

La fonction de coût dépend du seuil choisi a posteriori pour la meilleure identification et ne considère a priori qu'un seul point de fonctionnement, il convient alors de la mettre en parallèle avec une autre fonction (EER : *Equal Error Rate*) qui permet de considérer l'ensemble des seuils possibles.

Les courbes de détection d'erreurs (DET : *Detection Error Tradeoff*) représentent les détections manquées en fonction du nombre de fausses alarmes paramétrées par le seuil de décision. Cette courbe est équivalente aux courbes (ROC : *Receiver Operating*

Characteristics) qui, étant linéaires, ne permettent pas de visualisation efficace. Lorsque les observations suivent réellement une loi normale, la courbe DET se rapproche d'une droite (d'où sa meilleure visibilité).

Les mesures de performances doivent être accompagnées d'une mesure de confiance qui dépend du nombre d'instances testées et du résultats. Nous avons choisi d'utiliser une mesure d'intervalle de confiance à 95% définie suivant l'équation 5.10 avec p la probabilité obtenue et N le nombre d'instances testées [Montacie and Chollet 87]. La probabilité p est alors stable dans l'intervalle $p \pm confiance\%$.

$$confiance = 1,96 \cdot \sqrt{\frac{p \cdot (1-p)}{N}} \cdot 100 \tag{5.10}$$

5.1.3.5 Normalisation des scores

Afin d'obtenir des systèmes les plus robustes possibles aux types de données, aux différents modes de transmission de la voix existants, une attention particulière doit être portée sur les scores de vraisemblance. Une première normalisation des scores est de pondéré la log-vraisemblance par la longueur du segment testé. En effet, plus le segment est court moins le score sera fiable et la log-vraisemblance élevée (eq. 5.11 et 5.12) avec S la fonction de vraisemblance et $L(x)$ la longueur de l'observation $x = \{x_1, x_2, ..., x_{L(x)}\}$.

$$S(x/q) = \prod_{i=1}^{L(x)} S(x_i/q) \tag{5.11}$$

$$S'(x/q) = S(x/q)^{\frac{1}{L(x)}} \tag{5.12}$$

Dans le cas d'une adaptation MAP avec un modèle UBM générique R, on peut également normaliser la vraisemblance par celle du modèle UBM (voir eq. 5.13)

$$S(x/q) = logf'(x/q) - logf'(x/R) \tag{5.13}$$

Enfin des expériences ont été menées pour étudier l'impact d'une normalisation T-norm ou Z-norm de la distribution des scores de vraisemblance [Barras and Gauvain 03]. Cette normalisation nécessite d'avoir suffisamment de données pour faire l'apprentissage et un ensemble de développement pour estimer la distribution des scores a priori. Cette normalisation permettra entre autre d'être plus indépendante de la distribution des données d'apprentissage et d'adapter la normalisation a un éventuel sous corpus plus proche de l'application finale.

5.1.4 Prise en compte du contexte émotionnel

Les corpus utilisés pour la reconnaissance du locuteur sont généralement de très grande taille. De nombreux corpus disponibles sont des corpus enregistrés lors de conversations téléphoniques (données call-center), ou lors de séminaires ou réunions. Ces corpus sont donc généralement constitués de parole neutre.

La plupart du temps les études en reconnaissance des émotions tentent de rendre leurs modèles indépendants au locuteur par l'utilisation de bases de données de relative grande taille, ou par l'utilisation de normalisations au locuteur. Par contre les travaux

en identification du locuteur ne prennent que très peu en compte l'influence de l'état émotionnel du locuteur lors de son identification.

Afin d'améliorer la reconnaissance des émotions en prenant en compte la variabilité des locuteurs, les travaux de Ding [Ding et al. 12], proposent une méthode de *bootstrapping* fondée sur l'apprentissage de modèles émotionnels par locuteur. A partir de quelques échantillons de parole d'un locuteur cible, le meilleur modèle émotionnel est déterminé, puis adapté à ce locuteur suivant l'adaptation MAP. On peut alors reconnaître l'émotion d'une instance inconnue du même locuteur cible suivant ce modèle adapté. Ces travaux ont été réalisés sur le corpus AIBO et LDC Emotionnal Prosody. La durée des échantillons de parole nécessaires à l'adaptation au locuteur n'est pas précisée dans cet article. Dans un contexte d'interaction humain-robot, il est fondamental que cette durée soit la plus courte possible puisque le robot doit être capable de passer d'un locuteur à l'autre rapidement.

L'identification du locuteur peut être extrêmement dégradée par de la parole affective. Li [Li et al. 05] propose une méthode de conversion des paramètres acoustiques (F0 moyenne, ambitus, énergie moyenne, durée des parties voisées). Une analyse LPC est faite sur de la parole neutre, les coefficients LPC et le signal résiduel servent à modifier la F0 et la durée des parties voisées en accord avec les résultats de l'analyse prosodique. Une synthèse LPC ajoutée à une modification d'amplitude suivant l'analyse prosodique permettent d'obtenir une conversion de la parole émotionnelle en parole neutre. Sur la parole convertie, on extrait les MFCCs, la reconnaissance du locuteur se fait par une modélisation GMM. Un des points les plus délicats de cette méthode est la définition d'un neutre comme état de référence. Suivant les tâches et les locuteurs, l'état neutre peut être très variable. Dans un contexte d'interaction humain-robot, faudrait-il définir un ensemble d'états neutre correspondant aux différentes tâches possibles (vie quotidienne, jeu, acté) ?

5.2 Identification du locuteur dans une interaction réaliste

5.2.1 Protocole de construction des modèles

La difficulté principale rencontrée pour l'identification du locuteur sur les données que nous avons collectées est liée à la faible quantité de données par locuteur et surtout au petit nombre de locuteurs. En effet, le temps de parole par locuteur est très faible par rapport à celui habituellement rencontré pour l'identification du locuteur. De plus dans les corpus habituellement utilisé pour l'identification du locuteur, il y a généralement plusieurs centaines de locuteurs, nous n'en avons qu'une petite dizaine par corpus (27 dans IDV-HH, 22 dans IDV-HR, 12 dans NAO-HR1, et 12 dans NAO-HR2, cf chapitre 1, section 1.3.1). Pour contourner ce problème, nous avons choisi de privilégier les expériences indépendantes du locuteur, c'est-à-dire de réaliser l'apprentissage des modèles sur tous les locuteurs sauf un, puis de tester le locuteur restant (leave-one-speaker-out). L'erreur finale correspond alors à la moyenne pour tous les locuteurs. Ce protocole est possible pour l'identification du genre ou de l'âge mais il ne pourra pas être appliqué pour l'identification de locuteurs.

Apprentissage Dans l'état de l'art l'apprentissage se fait sur plusieurs dizaines de secondes, dans nos corpus le temps de parole neutre moyen d'un locuteur est de l'ordre d'une ou deux minutes. Nous pourrons donc choisir également des durées d'apprentissage de quelques dizaines de secondes tout en conservant suffisamment de données pour les tests. Dans [Barras et al. 07], deux durées d'apprentissage sont testées : 15 s et 30 s, nous avons choisi de tester les durées suivantes : 10, 30, 60 s.

L'ensemble des instances déterminées pour l'apprentissage sont concaténées pour ne faire qu'un seul signal. Les coefficients cepstraux (entre 0 et 12) et leurs dérivés (soit 26 descripteurs) sont calculés sur l'ensemble de ce signal à une fréquence d'échantillonage de 100 Hz. Les modèles GMM sont alors déterminés sur un vecteur de dimensions $26 * (durée_{app} \cdot 100)$. Le nombre de gaussiennes a été optimisé, nous avons utilisés des modèles à 256 gaussiennes. Avec 132 gaussiennes, l'apprentissage est plus rapide mais les modèles sont moins performants. Cette valeur sera à optimiser en fonction des performances souhaitées pour l'application finale.

Deux cas de figure se présentent alors : on cherche à identifier soit le locuteur lui-même, soit le genre. Dans le premier cas, les modèles sont créés suivant le protocole défini plus haut. Dans le second cas, nous avons pris un nombre identique de locuteurs homme et femme, l'ensemble des signaux concaténés sur des voix de femme sont à nouveau concaténés, de même pour les voix d'homme. Si l'on souhaite une durée d'apprentissage de 10 s, et que 10 voix de femme sont à disposition, il faudra prendre une durée d'apprentissage de 1 s pour chaque locuteur femme, de même pour les hommes.

Test Pour chaque locuteur, les instances qui n'ont pas été utilisées pour l'apprentissage sont concaténées pour ne former qu'un seul signal. Il peut être composé de parole neutre, de parole émotionnelle, d'une émotion particulière uniquement mais toujours pour un seul locuteur. Ce signal est ensuite segmenté en fonction de la durée de test choisie. Cette segmentation est totalement arbitraire et ne correspond pas avec la segmentation manuelle fondée sur des groupes de mots. Etant donné que le temps de parole n'est pas identique suivant les locuteurs, si la durée d'apprentissage est fixée, le nombre d'instances de tests d'une durée fixe sera variable suivant les locuteurs. La durée de test sera également variable. Dans [Barras et al. 07], la durée de test est égale à 10s. Dans le contexte d'une interaction homme-robot, la machine devrait être capable de reconnaître un locuteur sur un "bonjour", soit environ 1 s. Nous avons testé des durées de 1 s à 20 s.

Pour les expériences d'identification du genre, les locuteurs sont testés un par un, c'est-à-dire que les modèles de genre seront construits sur l'ensemble des locuteurs sauf un. Ce dernier locuteur sera testé, cela permet d'utiliser l'ensemble des instances de ce locuteur pour le test. L'opération est réalisée autant de fois qu'il y a de locuteurs.

Décision Les taux d'erreur par locuteur testés sont pondérés selon le nombre d'instances testé pour ce locuteur et sont donnés en fonction des durées d'apprentissage et de test. La décision se fait sur les valeurs de vraisemblance sur les modèles homme et femme. On considérera que si l'équation 5.14 est vraie alors, le genre reconnu est masculin. La valeur de seuil est fixée à 0 par défaut, mais nous verrons plus tard l'impact qu'elle peut avoir sur les taux d'erreur.

$$S(x/M) - S(x/F) > Seuil \tag{5.14}$$

Cette valeur correspond à la moyenne sur l'ensemble des locuteurs du taux d'erreur du genre pondéré par le nombre d'instances qui ont été testées. Le fait de pondérer l'erreur en fonction du nombre d'instances testées permet de tenir compte de la variabilité des temps de parole de chaque locuteur.

5.2.2 Identification du genre sur de la parole neutre (IDV-HR)

Expérience n°1 : Influence des durées d'apprentissage et de test sur l'identification du genre (IDV-HR) Dans cette expérience, nous avons testé systématiquement plusieurs durées d'apprentissage (10, 30 et 60 s) et de test (1, 2, 5, 10, 20 s) pour la reconnaissance du genre sur les 22 locuteurs du corpus IDV-HR. Ces expériences ont été réalisées en croisant les locuteurs. L'apprentissage et le test se font uniquement sur de la parole neutre soit environ 60% du corpus.

La figure 5.1 montre des résultats conformes à l'état de l'art. Globalement, plus la durée de test est importante et plus la durée d'apprentissage est longue, plus l'erreur diminue. On peut également remarquer que les scores de confiance pour une durée d'apprentissage de 10s sont relativement élevés, cela est dû principalement au fait que le taux d'erreur est important. Pour une durée de test de 20 s, il y a globalement peu d'instances testées (entre 4 et 14, 10 en moyenne) et l'intervalle de confiance diminue.

Dans le contexte d'une interaction homme-robot, nous souhaitons que la décision sur le genre se fasse le plus rapidement possible. Nous estimons qu'une durée maximum de 1 à 2 s serait souhaitée pour l'application finale. D'après nos résultats, afin que l'erreur de détection soit la plus faible possible, il faut donc augmenter la durée d'apprentissage. On peut cependant noter que la différence entre une durée d'apprentissage de 30 s et de 60 s est très faible. Une durée de 30 s paraît donc suffisante. La figure 5.1 montre également un résultat intéressant : il semblerait qu'il existe un seuil d'erreur minimum. Plus la durée d'apprentissage est grande, plus la durée de test pour laquelle ce seuil est atteint est élevée, par exemple, pour une durée d'apprentissage de 60 s, le seuil semble être atteint dans nos expériences, avec une durée de test de 10 s. Nos résultats restent cependant nettement en deça des performances de l'état de l'art. Une des raisons principales est que la variabilité des types de voix et des âges des locuteurs dans IDV-HR est très importante par rapport à celle que l'on peut rencontrer dans les corpus des campagnes NIST.

A ce propos, il semble important de souligner les différences notables d'identification du genre en fonction des locuteurs. Certains locuteurs sont très mal reconnus en particulier chez les femmes le locuteur F8 (erreur 42,3% sur 10s de test, 30s d'apprentissage). Les hommes M5, M9 et M13 ne sont également pas très bien reconnus. F8 correspond à une femme de 80ans, une voix âgée mais tonique. M5 et M13 ont des voix assez claires, que l'on peut aisément confondre avec une voix de femme. M9 correspond à la voix d'un homme âgé de 70 ans.

Nous pouvons émettre l'hypothèse que l'identification du genre sur des voix âgées est une tâche plus compliquée, ce qui doit être vérifié sur des corpus de voix jeunes et âgées de plus grande taille. De plus certaines voix ne seront jamais bien reconnues parce que classées dans le genre opposé (femmes avec une voix grave, hommes avec une voix claire, du à une morphologie particulière, une conséquence de maladie, de la cigarette, de l'alcool, etc.).

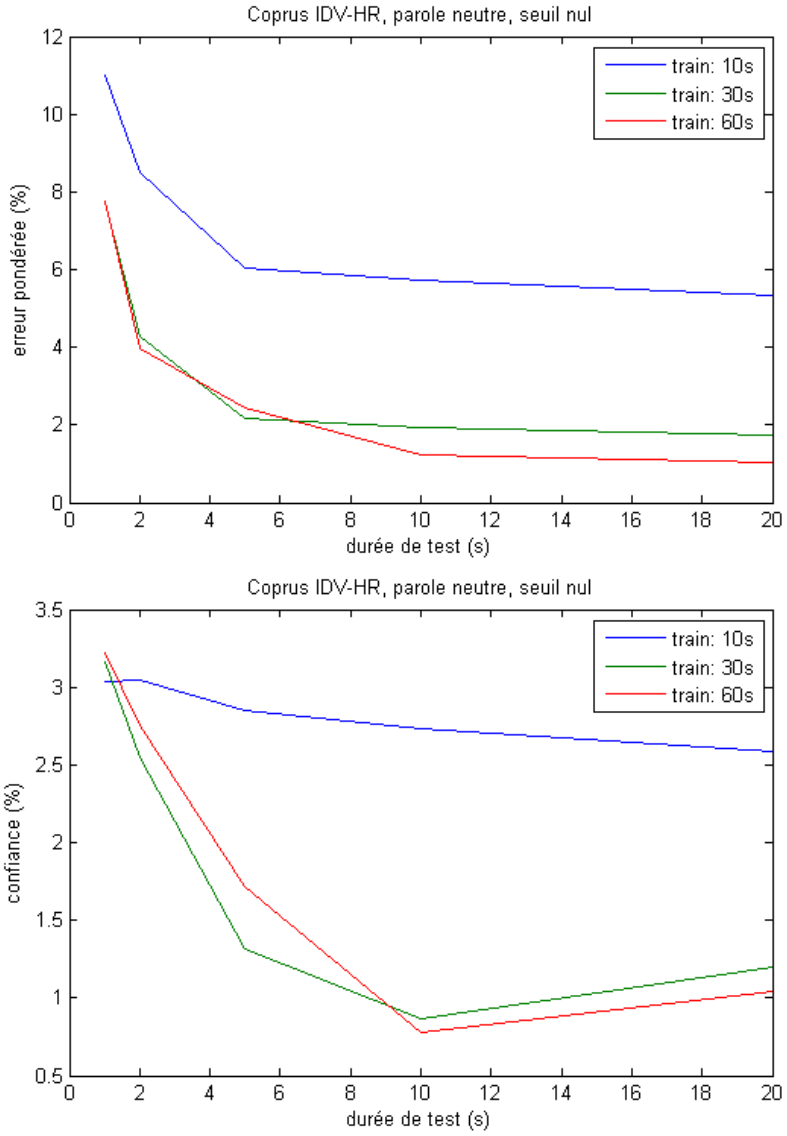


FIGURE 5.1 – Identification du genre sur le corpus IDV-HR, erreur pondérée (haut) et score de confiance (bas)

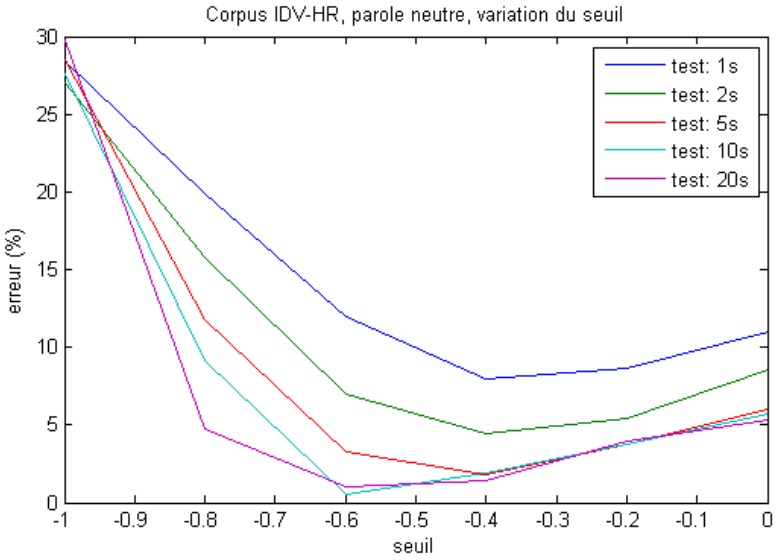


FIGURE 5.2 – Identification du genre, sur le corpus IDV-HR, erreur pondérée, variation de seuil de décision.

Expérience n°2 : Influence du seuil de décision Cette seconde expérience correspond à l'expérience n°1 pour laquelle la durée d'apprentissage est fixée à 30s et le seuil de décision est variable entre -1 et 0. Sur la figure 5.2, nous pouvons constater qu'un seuil fixé trop bas (inférieur à -0,8), l'équation 5.14 sera presque toujours vraie, et le genre masculin sera privilégié par rapport à l'autre. Si le seuil est fixé trop haut (supérieur à 0), l'autre genre féminin sera privilégié. Nous verrons dans les paragraphes suivants que ce seuil est dépendant également de l'environnement acoustique dans lequel a été enregistré le corpus.

5.2.3 Influence d'une parole émotionnelle sur l'identification du genre

5.2.3.1 En conditions normales (IDV-HR)

Peu d'études montrent l'influence de la parole émotionnelle sur la reconnaissance du locuteur, nous avons déjà cité celles de Ding [Ding et al. 12] et Li [Li et al. 05]. Nous allons traiter en particulier la reconnaissance du genre. Nous étudierons plus particulièrement les 4 macro-classes émotionnelles les plus représentées dans IDV-HR : état neutre, joie, tristesse et colère.

Expérience n°1 : Test en cross-validation sur les 22 locuteurs de IDV-HR sur de la parole émotionnelle. Nous testons des instances annotées émotionnellement ou neutre alternativement sur des modèles créés sur de la parole neutre et émotionnelle. D'après la figure 5.3, nous pouvons remarquer que l'utilisation de modèles fondées sur de

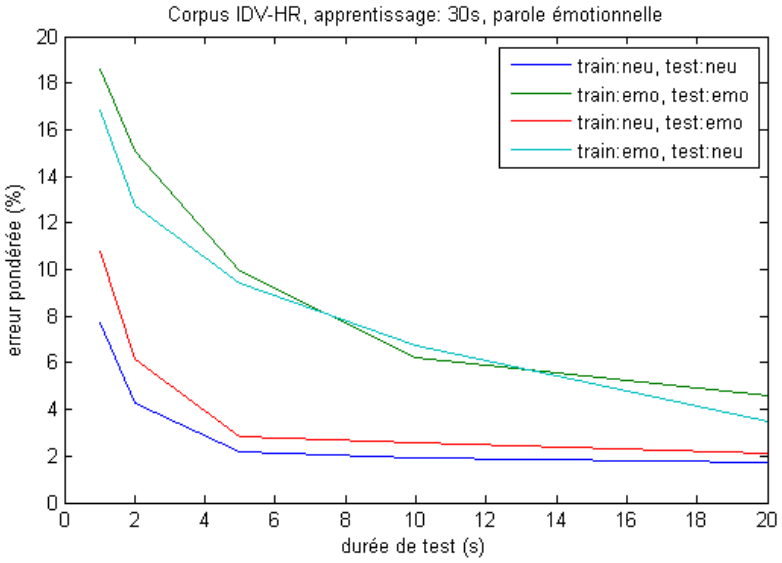


FIGURE 5.3 – Identification du genre sur de la parole émotionelle sur le corpus IDV-HR, erreur pondérée

la parole émotionnelle n'est pas très performant par rapport à ceux fondés sur de la parole neutre. Ces résultats sont valables pour des modèles neutre et émotionnel construits avec le même nombre de gaussiennes (256). Il est possible que les performances en apprentissage sur de la parole émotionnel soient légèrement différentes si le nombre de gaussiennes était optimisé sur ces modèles.

Résultat : Dans notre contexte, il semble qu'il vaut mieux utiliser des modèles de genre créés sur de la parole neutre, pour l'identification du genre avec ou sans parole émotionnelle.

Expérience n°2 : Influence des émotions sur l'identification du genre Toujours en cross-validation sur les 22 locuteurs du corpus IDV-HR, nous avons testé l'identification du genre en fonction de l'état émotionnel. Nous n'utiliserons que les instances émotionnelles, c'est-à-dire celles qui n'ont pas été annotées comme "neutre". Etant donné les résultats sur la parole neutre, nous utiliserons une durée d'apprentissage de 30 s sur de la parole émotionnelle. La durée de test reste variable (de 1 à 20 s).

- apprentissage : IDV-HR émotionnel (toutes émotions confondues), 1 min max sur neutre, 10 s sur colère, tristesse et joie, 5 s pour peur et 1 s pour négatif soit au maximum 46 s de parole émotionnelle par locuteur,
- test : IDV-HR émotionnel, uniquement colère, joie, tristesse et neutre sur une durée de 1 s.

Résultat : Lorsqu'on regarde précisement les émotions contenues dans les instances testées, (figure 5.4), la colère et la joie entraînent une hausse de l'erreur, particulièrement chez les hommes alors que la tristesse présente des résultats similaires à ceux obtenus sur

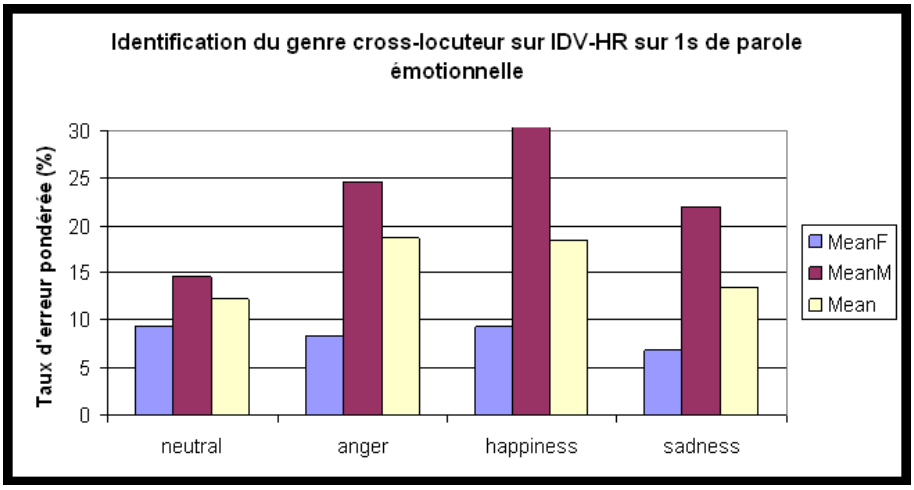


FIGURE 5.4 – Identification du genre (apprentissage et test sur de la parole émotionnelle, corpus IDV-HR) en fonction de l'émotion

un état neutre. La colère et la joie forte entraînent une modification du conduit vocal et généralement une hausse de la F0. L'effet se ressent plus chez les hommes qui sont alors confondus avec les femmes.

Un des moyens de s'affranchir de cet effet pourrait être de modifier le seuil de décision (il est ici fixé par défaut à 0) entre les deux classes hommes et femmes en fonction de l'émotion reconnue. Cela suppose de savoir détecter l'émotion sans connaître le genre a priori (pas de normalisation possible).

5.2.3.2 En conditions très réverbérantes (IDV-HH)

L'identification du genre sous des conditions acoustiques mauvaises semble être une tâche difficile. L'erreur monte à presque 50%, soit l'équivalent du hasard, lorsque les modèles sont construits et testés sur le corpus IDV-HH qui a été enregistré dans un petit studio avec une grande réverbération ($T_{60dB} > 100ms$). Dans ces conditions, un réglage fin du seuil de décision ou une durée de test plus élevée ne permettent pas d'améliorer les résultats.

Cela est principalement due à la répartition des valeurs de log-vraisemblance pour chacun des genres face à chacun des modèles. En effet, les deux densités correspondant aux hommes et aux femmes sont confondues sur le corpus IDV-HH (5.5, bas), alors qu'elles sont bien distinctes sur le corpus IDV-HR (5.5, haut). Il semble donc quasiment impossible d'identifier le genre sans compenser la réverbération dans des conditions réverbérantes.

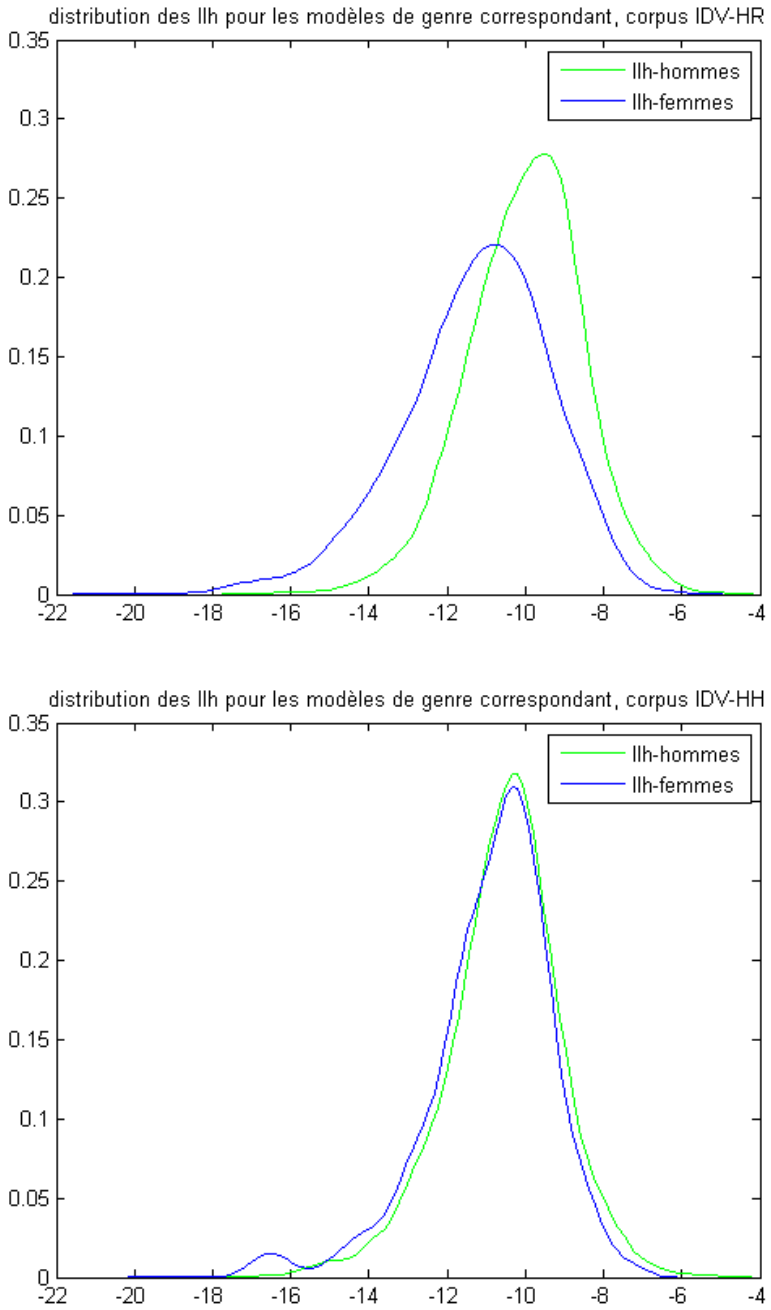


FIGURE 5.5 – Courbes de densité de répartition (ordonnée) des valeurs de log-vraisemblance (abscisse) pour les corpus IDV-HR (haut) et IDV-HH (bas)

5.2.4 Identification d'autres caractéristiques locuteur

5.2.4.1 Reconnaissance de l'âge

Dans le cadre du projet ROMEO, nous aurions aimé identifier si la personne qui parle au robot est un adulte ou un enfant. Pour cela, nous avons essayé de créer des modèles enfant à partir des corpus NAO-HR1 et AIBO (le corpus AIBO est en allemand mais cela ne pose pas de problème pour l'identification des locuteurs) et les modèles adultes à partir des corpus à notre disposition (IDV-HR, JEMO). Les corpus IDV-HH et NAO-HR2 ne seront pas utilisés étant donné les effets démontrés d'un environnement acoustique défavorable sur l'identification du locuteur. De plus, comme les corpus d'enfants et d'adultes ont été enregistrés dans des conditions acoustiques différentes, il est très difficile de savoir si la tranche d'âge sera bien reconnue parce que les conditions acoustiques sont reconnues ou réellement parce que la tranche d'âge est reconnue. Pour tenter de s'affranchir de cet effet, nous utiliserons des corpus différents pour l'apprentissage et le test. Cependant les différences entre les corpus peuvent être plus ou moins importantes et l'effet "corpus" ne pourra pas être totalement annulé. Les expériences seront faites uniquement sur des instances sans émotions, annotées "neutre".

Nous avons choisi de faire 3 modèles de "genre" : homme, femme et enfant. Une première expérience consiste à créer directement des modèles homme et femme à partir du corpus JEMO et le modèle enfant à partir du corpus AIBO. Dans une seconde expérience, nous avons utilisé un modèle générique UBM créé à partir de JEMO, que nous avons adapté ensuite aux hommes et aux femmes de JEMO, aux enfants de AIBO. Nous avons utilisé l'adaptation MAP. Aucun des deux protocoles n'a apporté de résultats satisfaisants, pour la simple et bonne raison, qu'au vu des valeurs de log vraisemblance, il semble quasiment impossible de distinguer les femmes et les enfants (figure 5.6).

La distinction adulte/enfant dans des conditions acoustiques très différentes est une tâche extrêmement complexe, elle semble même plus difficile que la reconnaissance du genre dans les mêmes conditions puisque les modèles de femme et d'enfant semblent se confondre.

L'identification de l'âge a été testé suivant le même protocole que celle du genre (modèles GMM et coefficients cepstraux uniquement). D'autres études sur l'identification de l'âge [Li et al. 10] ont présenté des résultats plus concluants en utilisant des descripteurs plus variés sur le modèle de ceux utilisés pour la reconnaissance des émotions.

5.2.4.2 Identification de locuteurs connus

Dans cette partie, nous cherchons à reconnaître un locuteur connu. La base de données IDV-HR est utilisée pour créer des modèles de locuteurs. Les modèles pour chacun des locuteurs sont adaptés d'un modèle UBM, fondé sur le corpus JEMO (62 locuteurs) : adaptation MAP avec un poids a priori de 20 sur 256 gaussiennes sur 5 s de parole neutre.

Nous avons choisi de n'utiliser que des données émotionnelles pour les tests alors que l'apprentissage est réalisé sur de la parole neutre. Pour avoir suffisamment de données pour chaque locuteur, nous devons limiter la durée de test à 1 s. Nous pouvons constater que l'erreur la plus forte est obtenue pour l'émotion colère, ce qui correspond au résultat obtenu au paragraphe 5.2.3.1 sur l'identification du genre sur le corpus IDV-HR. Par

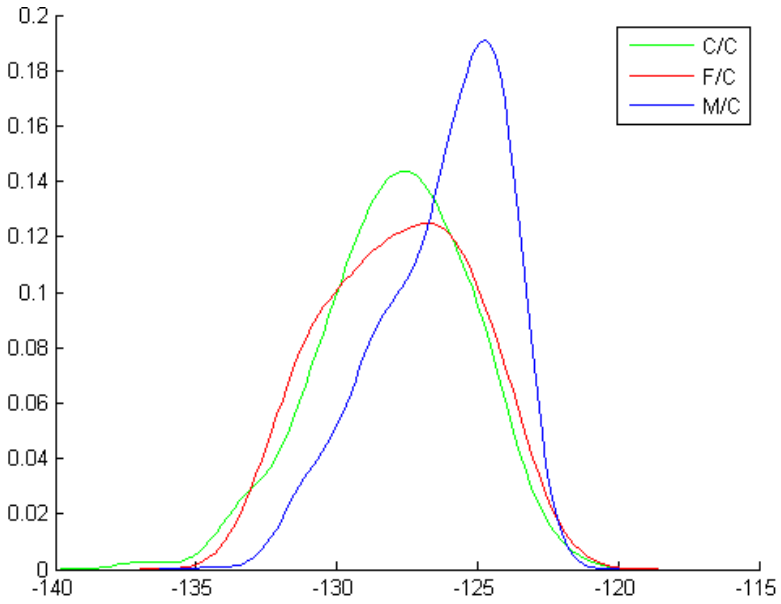


FIGURE 5.6 – Courbes de densité de répartition (ordonnée) des valeurs de log-vraisemblance (abscisse) lors du test des 3 genres (homme, femme, enfant) sur le modèle enfant.

test : 1 s	Confiance	Erreur
Colère	2,72	23,08
Joie	1,69	16,67
Tristesse	3,58	37,04

TABLE 5.1 – Identification d'un locuteur parmi 22 sur le corpus IDV-HR, parole émotionnelle

contre, la reconnaissance du locuteur sur de la joie semble être plus performante que sur de la tristesse contrairement aux résultats sur l'identification du genre. Lorsqu'un locuteur exprime de la joie, il semble plus facile de déterminer son identité à partir d'un modèle neutre que si il exprime de la colère ou de la tristesse.

5.3 Conclusion

Dans ce chapitre sur la caractérisation du locuteur lors d'une interaction humain-robot, nous avons principalement testé les méthodes classiques d'identification du genre et du locuteur sur les corpus à notre disposition (corpus ROMEO). Très peu d'études dans ce domaine utilisent de telles données avec une grande variabilité dans les types de voix et les contextes d'application puisque la grande majorité d'entre eux sont issus de centres d'appels téléphoniques. Nos corpus étant de petite taille (relativement à ceux utilisés par la communauté), nous avons mis en place un protocole de test croisant les locuteurs. Ce protocole permet de conserver l'ensemble de nos données, tout en restant indépendant des locuteurs.

Ainsi nous avons pu confirmer que l'identification du genre était une tâche plus complexe dans un contexte de parole émotionnelle par rapport à de la parole neutre. Certaines émotions fortes comme la joie ou la colère affectent de manière plus prononcée la dégradation des performances d'identification. De plus la reconnaissance du genre sur des voix âgées semble être plus difficile que sur des voix jeunes.

Un résultat important de nos travaux est que l'identification du genre est plus performante en utilisant des modèles de parole neutre, même si elle est testée sur de la parole émotionnelle. Ces résultats doivent néanmois être vérifiés en optimisant les paramètres pour chaque modèle (entre autres le nombre de gaussiennes).

Enfin, un point très important dans un contexte de robotique, la réverbération affecte très fortement la reconnaissance du genre (et sûrement des locuteurs). A tel point que les taux d'erreur se rapproche du hasard. Pour palier à ce problème, il faudrait mettre en place, soit des systèmes de filtrage amont de la réverbération, ce qui n'existe pas encore de manière satisfaisante, soit un système de détection de la réverbération qui permettrait de donner une valeur de confiance suplémentaire sur le résultat annoncé, quitte à ce que le robot dise "je ne sais pas".

Chapitre 6

Reconnaissance des émotions, du stress et de la personalité

Ayant à présent une connaissance approfondie des descripteurs acoustiques et des phénomènes liés à l'interaction réaliste qui peuvent les affecter, nous proposons d'étudier assez généralement la reconnaissance automatique d'indices paralinguistiques. Parmi ces indices, nous nous focaliserons sur des indices dans le signal social comme les marqueurs interactionnels, et de manière plus précise sur des indices émotionnels, mais également sur d'autres traits humains comme le stress ou la personnalité. Ce travail sur les indices paralinguistiques est totalement lié à une réflexion plus générale sur le déroulement d'une interaction humain-robot dans un contexte réel. La réalisation d'une reconnaissance automatique implique l'utilisation de descripteurs acoustiques qu'il faut choisir au mieux en fonction de l'application, et l'apprentissage de modèles à partir de corpus annotés, également proches de l'application.

Dans le contexte d'une interaction homme-robot réaliste, l'application n'est pas aussi bien définie qu'une application "laboratoire". C'est-à-dire que l'interaction peut avoir lieu dans différentes pièces de l'appartement, avec des enfants ou des adultes, le type de tâche peut également être très variable : du jeu, à la situation d'urgence, en passant par des scènes de vie quotidienne. Nous allons étudier comment ces variabilités peuvent influer sur les performances de la reconnaissance automatique en fonction de différents modèles et normalisations utilisés.

6.1 État de l'art

Les outils utilisés pour la reconnaissance automatique d'émotions sont ceux classiquement utilisés pour la reconnaissance des formes. Il existe deux grands types de classification : supervisé ou non-supervisé. Dans le cas d'une classification supervisé, les classes ainsi que les données sont fournies au programme d'apprentissage. Lors d'une classification non-supervisé, les classes sont déterminées automatiquement, en aveugle, en fonction de la structure même des données. La classification des émotions utilise essentiellement des apprentissages supervisés où les classes considérées sont des classes émotionnelles

déterminées à l'avance en fonction de l'application visée.

La classification automatique dans le domaine des émotions, mais également pour d'autres traits humains (personnalité, stress, pathologie, etc.) est un sujet de recherche en pleine ébullition. Il reste cependant très difficile de comparer les performances des systèmes mis en place par les différents laboratoires. Une première initiative de coopération, le CEICES [Batliner et al. 06] lancée dans le cadre de l'association HUMAINE en 2005 a permis de standardiser les différentes méthodes. L'initiative est principalement fondée sur l'utilisation d'un corpus d'apprentissage et de test commun sur lequel il est plus simple de comparer les performances. Cette initiative est depuis reprise régulièrement par l'équipe de Schuller en association avec d'autres chercheurs suivant les bases de données utilisées, lors des différents challenges Interspeech (premier challenge en 2009 [Schuller et al. 09b], puis 2010 [Schuller et al. 10a] et résultats publiés [Schuller et al. 11a, Schuller et al. 12a]).

6.1.1 La classification automatique d'indices paralinguistiques

6.1.1.1 Classifieurs

Il existe différents types d'approches pour l'apprentissage de modèles (supervisé / non-supervisé) et différentes modélisations (paramétrique / non-paramétrique). Nous n'entrerons pas dans les détails de ces algorithmes, mais nous décrirons surtout deux types d'approches utilisés dans nos travaux :

- une approche paramétrique type mélange de gaussiennes (GMM : *Gaussian Mixture Model*) ou modèles de Markov cachés (HMM : *Hidden Markov Model*),
- une approche par discrimination type réseau de neurones (NN : *Neural Network*) ou machines à vecteur support ou séparateurs à vastes marges (SVM : *Support Vector Machine*)

Les mélanges de gaussiennes (GMM) Les modèles GMM sont génératifs. Ils cherchent à modéliser une classe par un ensemble de gaussiennes (leur fonctionnement est détaillé au chapitre 4.1.3).

Les Séparateurs à Vastes Marges (SVM) Les modélisations SVM reposent sur une méthode discriminative. Elles sont très largement répandues dans la communauté de recherche sur les émotions. Le principe est assez simple : il s'agit de déterminer le meilleur hyperplan permettant de séparer deux classes en maximisant la distance entre les échantillons d'apprentissage et cet hyperplan. Parmi les modèles SVM on distingue le cas linéairement séparable et le cas non linéairement séparable. Pour surmonter les inconvénients du cas non linéairement séparable, l'idée des SVM est de transformer l'espace des données, afin de passer d'un problème de séparation non linéaire à un problème de séparation linéaire dans un espace de re-description de plus grande dimension 6.1. Cette transformation non linéaire est effectuée via une fonction, dite fonction noyau. Nous avons utilisé un noyau gaussien (noyau RBF : *Radial Basis Function*) définit par l'équation 6.1. Le paramètre γ détermine la capacité du modèle à englober les données. On définit généralement une constante de complexité c, plus elle est grande, plus la marge contenant les erreurs est petite. C'est ce type de modélisation que nous avons utilisé pour la reconnaissance automatique des émotions.

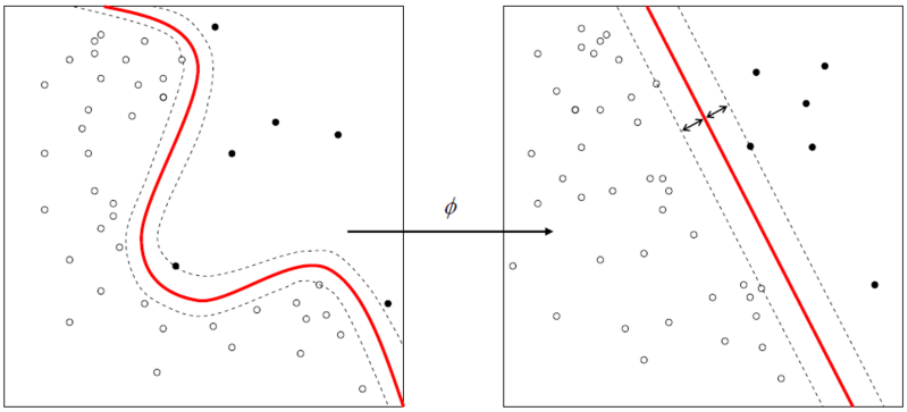


FIGURE 6.1 – Hyperplan de séparation dans un cas non linéaire (gauche) et dans un cas linéaire (droite)

$$K(x,y) = exp(-\gamma \cdot \|x-y\|^2) \tag{6.1}$$

Les modèles GMM sont moins utilisés que les SVM, mais apportent néanmoins des résultats satisfaisants pour la reconnaissance des émotions [Clavel et al. 07, Dumouchel et al. 09, He et al. 10]. Nous les avons principalement utilisés pour la caractérisation du locuteur (chapitre 4). Les modèles de Markov cachés (HMM) introduisent une dimension temporelle, le signal audio est alors analysé suivant une succession de plusieurs états. Ces modèles semblent être intéressants pour la segmentation en phonèmes [Lanchantin et al. 08, Xie and Niyogi 06], et sont également utilisés dans la classification des émotions [Lee et al. 04, Mower et al. 09, Yildirim et al. 04]. Parmi les autres classifieurs, nous pouvons citer les réseaux de neurones et les k plus proches voisins, qui sont nettement moins utilisés que les autres. Enfin, il existe des approches hybrides. Par exemple, Li [Li et al. 10] propose une modélisation SVM de vecteurs GMM.

6.1.1.2 Conditions d'apprentissage

Les résultats d'une classification ne dépendent pas que des classifieurs, il y a un grand nombre de points qui permettent de relativiser une mesure de performance donnée.

1. la base de données sur laquelle est réalisée l'apprentissage,
2. le nombre de classes émotionnelles, plus les classes sont nombreuses, plus la tâche est ambitieuse,
3. le type de classe émotionnelles : il est plus facile de classer les émotions actives et passives que les émotions positives et négatives par exemple, ou de classifier la colère et le neutre, plutôt que la tristesse et le neutre,
4. l'extraction des descripteurs et le type de descripteurs : acoustiques, linguistiques ou autre, et la normalisation de ces paramètres,
5. la constitution du corpus de test : cross-validation, cross-corpus, est-ce que les locuteurs testés sont également utilisés dans les modèles ?

Il n'est pas toujours évident d'obtenir ces informations afin de placer une mesure de classification dans son contexte.

6.1.1.3 Mesures de performances

Les mesures de performances ont été instaurées et stabilisées au cours des différents challenges. Ces mesures permettent de caractériser la matrice de confusion qui apporte une information complète et brute, mais peu utilisable en comparaison. La matrice de confusion contient généralement le nombre des instances pour chaque émotion testée.

émotion testée	émotion reconnue		
emo1	a_{11}	a_{12}	a_{13}
emo2	a_{21}	a_{22}	a_{23}
emo3	a_{31}	a_{32}	a_{33}

Pour chaque émotion, on peut déterminer le rappel (équation 6.2) et la précision (équation 6.3) :

$$Rappel_{emo_i} = \frac{a_{ii}}{\sum_j a_{ij}} \tag{6.2}$$

$$Precision_{emo_i} = \frac{a_{ii}}{\sum_j a_{ji}} \tag{6.3}$$

A partir de ces mesures, on peut déterminer le rappel moyen pondéré (WAR, équation 6.4) ou non (UAR, équation 6.5). La mesure pondérée WAR correspond aussi à la trace de la matrice de confusion (ou *accuracy*).

$$WAR = \frac{1}{\sum_i \sum_j a_{ij}} \cdot \sum_i (Rappel_{emo_i} \cdot \sum_j a_{ij}) = \frac{\sum_i a_{ii}}{\sum_i \sum_j a_{ij}} \tag{6.4}$$

$$UAR = \frac{1}{3} \cdot \sum_i Rappel_{emo_i} \tag{6.5}$$

Lorsque les différentes classes sont équilibrées, les deux mesures sont identiques. Un inconvénient majeur de ces mesures est qu'elles ne permettent pas de dire si la diagonale est majoritaire. En effet, si l'une des classes (le neutre) par exemple absorbe toutes les autres, il est possible d'obtenir des mesures UAR et WAR intéressantes, alors que la matrice de confusion montre un disfonctionnement. Nous avons proposé une autre mesure permettant de rendre compte de la diagonale de la matrice : la précision minimum.

Dans toutes nos études, les mesures de performances sont accompagnées d'une mesure de confiance, comme celle que nous avons utilisée au chapitre 4 (équation 5.10).

6.1.2 Émotions actées/ induites, types de classes, performances

Quelles performances peut-on attendre d'une expérience ? Pour répondre à cette question, l'équipe de Schuller a essayé de comparer les performances obtenues sur les mêmes émotions de plusieurs corpus [Schuller et al. 09a]. Pour cela, il a utilisé des corpus actés (DES, EMO-DB), induits (ABS, eNTERFACE) et réalistes (AVIC, SmartKom, SUSAS, VAM). Le contenu lexical est contraint ou libre, il peut y avoir interaction avec un autre

corpus	activation (%)	valence (%)
EMO-DB (acté)	96,8	87,0
VAM (spontané)	72,4	48,1

TABLE 6.1 – Scores UAR sur deux corpus l'un acté, l'autre spontané. Expériences indépendantes du locuteur, sur des modèles SVM, extrait de [Schuller et al. 09a].

interlocuteur humain, ou une machine. L'âge des locuteurs ainsi que les environnements acoustiques (acoustique de la salle, microphones, échantillonage) varient également selon les corpus. Les différentes émotions ont été regroupées suivant les deux dimensions les plus connues : l'activation et la valence. Deux classifieurs ont été évalués, une combinaison de HMM/GMM et les SVM.

Nous retrouvons le fait que les performances sont nettement plus élevées pour les corpus actés contenant des émotions prototypiques, alors que les corpus réalistes, plus compliqués à annoter également, apportent des résultats beaucoup moins élevés. Le tableau 6.1 montre clairement ces différences de performances sur deux corpus. Le corpus VAM [Grimm et al. 08] correspond à de la parole émotionnelle spontanée sur 47 locuteurs et le corpus EMO-DB [Burkhardt et al. 05] correspond à de la parole émotionelle exprimée par 10 acteurs professionnels. Globalement l'activation est mieux reconnue que la valence mais cela dépend encore des corpus. La valence sur les corpus spontanés semble être une tâche difficile. En effet, intuitivement, les émotions positives et négatives ne semblent pas homogènes acoustiquement parlant.

Les expériences cross-corpus Lorsque l'apprentissage des modèles est effectué sur un corpus (ou groupe de corpus) et le test sur un autre, le protocole est appelé cross-corpus. Ce type d'expérience permet de mettre en évidence les difficultés liées à l'utilisation des systèmes de reconnaissance automatique des émotions dans des conditions réelles. Le protocole permet également d'évaluer la robustesse d'un modèle face aux variabilités environnement, tâche, locuteurs et type d'émotions.

L'équipe de Schuller a beaucoup utilisé cette méthode pour estimer les performances lors d'un changement de tâche. Dans une étude menée par Eyben [Eyben et al. 10], quatre corpus spontanés sont testés SmartKom, Aibo, VAM et SAL. Ces corpus présentent des différences majeures entre eux : type d'interaction, adulte/enfant, unité d'annotation du mot au segment. L'apprentissage se fait sur trois des corpus et le test sur le quatrième. Les classes mises en jeu correspondent à la valence (positive, négative, neutre). Les performances moyennes de classification deux à deux (par exemple négative contre neutre et positive) sur les quatre corpus sont significativement meilleures que le hasard (50%) que dans le cas positive/négative (UAR=54,4%). Cela signifie que les expériences croisant différents corpus spontanés présentent un vrai défi à la communauté scientifique.

Dans [Wöllmer et al. 11], des expériences sur le même corpus Aibo ont été réalisées. Le corpus a été enregistré sur deux supports audio différents : micro-cravate (champ proche, signal bonne qualité) et caméra vidéo (champ plus diffus, signal bruité). Au travers de différents tests, Wöllmer a pu remarquer que si l'on souhaite tester des instances bruitées, il est préférable d'inclure ce type de bruit dans l'apprentissage. Marchi [Marchi et al. 12] propose également des expériences croisant différents corpus pour étudier l'intérêt des

descripteurs prosodiques pour la reconnaissance de la valence chez les enfants autistes.

Les expériences mélangeant les corpus restent un défi, pourtant elles consituent une étape nécessaire vers la mise en place d'applications réelles.

6.1.3 Extraction des descripteurs

Le choix des descripteurs est assez variable d'un laboratoire à un autre, il existe cependant quelques descripteurs communément utilisés comme les coefficients cepstraux (voir chapitre 3). Par contre le nombre de ces descripteurs est très variable puisqu'il va de quelques centaines à plusieurs milliers [Schuller et al. 12b].

6.1.3.1 Normalisation

Comme pour l'identification du locuteur, il existe différents types de normalisation pour la classification des émotions : t-norm, z-norm, warping, etc. (chapitre 4, section 5.1.3.2). Elles peuvent être appliquées aux locuteurs, au genre, à une classe d'âge, etc.

6.1.3.2 Sélection automatique des descripteurs

Généralement les systèmes utilisant un grand nombre de descripteurs, les sélectionnent automatiquement sur un sous-corpus de développement [Black et al. 11, Rong et al. 09, Wu et al. 10]. Cela permet de réduire le nombre de descripteurs à calculer (et donc le temps de calcul) mais également d'éviter les descripteurs redondants. Cette sélection ne se fait pas suivant une analyse acoustique, mais en utilisant des algorithmes puissants de sélection automatique. Il existe plusieurs algorithmes disponibles dans l'outil Weka. Le Sequential Fast Forward Selection (SFFS) algorithme [Pudil et al. 02] est un des plus largement répandu malgré son caractère "glouton" (ou sous-optimal).

Un des défis majeurs de ces algorithmes est lié à leur forte dépendance au corpus d'apprentissage et de développement. En effet, optimiser le nombre de descripteurs sur un sous échantillon de données ne permet pas d'affirmer l'optimisation des performances sur un échantillon inconnu. Cet aspect se retrouve dans les différences de scores obtenus dans Chastagnol [Chastagnol and Devillers 12] avec sélection de features sur les trois sous-corpus du Speaker Personality Corpus utilisé pour le challenge sur la personalité.

6.1.3.3 Fusion d'indices

La fusion d'indices est souvent opérée lorsqu'on se place dans un cadre multimodal. La fusion la plus courante est celle des indices acoustiques et linguistiques (et éventuellement vidéo). Clavel [Clavel 07] propose de fusionner les indices acoustiques extraits sur les parties voisées avec ceux extraits des parties non-voisées. Il existe deux grands types de fusion : précoce (au niveau des indices) ou tardive (au niveau de la décision).

6.2 Reconnaissance d'indices paralinguistiques en conditions d'interaction homme-robot

Les recherches actuelles sur les interactions homme-robot écologiques prennent en considération l'environnement acoustique, le type de locuteur et les émotions. L'ensemble des variabilités est infinie et on pourra croire la tâche irréalisable. Cependant, à force d'études et d'analyses, les chercheurs peuvent déterminer des niches de résistance face à toute cette instabilité, ou tout au moins proposer des pistes de solutions à partir de constats. Il est important de rappeler que la grande majorité de la recherche sur la parole neutre et émotionnelle aujourd'hui reste encore limitée à un petit nombre de données collectées dans des conditions de laboratoire très contrôlées. Cela permet de dégager un certain nombre de méthodes et de proposer des descripteurs nouveaux, mais leur utilisation en situation écologique est loin d'être anodine.

6.2.1 Protocole pour la reconnaissance automatique

6.2.1.1 Choix des descripteurs acoustiques

A partir de l'analyse acoustique que nous proposons au chapitre 3, nous avons défini un ensemble de descripteurs acoustiques qui paraissent être intéressants pour la discrimination des indices paralinguistiques en général. Il est évident que le choix de ces descripteurs doit être optimisé pour chaque type d'information que l'on cherche à discriminer (émotion, personnalité, stress). Ces descripteurs sont calculés sur les segments émotionnels qui ont été définis dans le chapitre 2 sur l'annotation des corpus.

Afin de pouvoir comparer nos résultats avec ceux de la communauté scientique, nous utiliserons également l'ensemble de 384 descripteurs acoustiques proposé lors du challenge Interspeech 2009 [Schuller et al. 09b].

6.2.1.2 Conditions d'apprentissage

Différentes conditions d'apprentissage ont été utilisées dans nos expériences. Les tests en cross-validation permettent de donner une bonne idée générale du contenu du corpus et de la difficulté de la tâche en cours. Les tests indépendants du locuteur permettent d'approcher les conditions d'une application réaliste en interaction. Enfin, les tests croisants plusieurs corpus (ou cross-corpus) sont absolument fondamentaux pour évaluer la robustesse des modèles à différentes conditions. Les tests en cross-corpus permettent d'évaluer la robustesse au type de tâche, à l'environnement acoustique.

6.2.1.3 Classification automatique

Nous avons choisi d'utiliser l'outil LibSVM développé par Chang [Chang and Lin 11] pour plusieurs raisons : une optimisation systématique de deux paramètres du modèle SVM (c est une constante de complexité, plus c est grand, plus la marge contenant les erreurs est petite, γ détermine la capacité du modèle à englober les données) est effectuée dans l'outil. De plus il est disponible et assez généralement utilisé par la communauté scientifique. Nous utilisons également l'outil Weka [Hall et al. 09] qui permet de visualiser

facilement les matrices de confusion. Les paramètres SVM optimisés par LibSVM seront utilisés pour Weka.

6.2.2 Reconnaissance automatique des émotions

La reconnaissance automatique des émotions est un domaine de recherche très large. Nous cherchons à mettre en place un système de reconnaissance le plus performant possible pour l'application visée. Nous devons donc prendre en compte les quatre variabilités émotion, locuteur, environnement, tâche présentes lors d'une interaction humain-robot.

6.2.2.1 Performances en cross-validation

Les expériences en cross-validation permettent d'avoir une idée générale des performances attendues pour un corpus donné et des dégradations possibles en fonction de la tâche, de l'environnement ou du type d'émotions à reconnaître. Les performances en cross-validation consituent en quelque sorte une signature du corpus. Nous proposons donc plusieurs résultats de performances sur les corpus ROMEO (sauf NAO-HR2 dont le nombre d'instances est trop faible) ainsi que sur le corpus AIBO à des fins de comparaison avec les résultats de la communauté. Nous distinguons les deux écoles dans lesquelles le corpus AIBO a été collecté, puisque ces deux sous-corpus correspondent à des conditions acoustiques différentes.

Protocole Pour chaque test, nous essayons, dans la mesure du possible, d'équilibrer les classes émotionnelles, ainsi les mesures UAR et WAR sont quasiment équivalentes (ce qui est faux lorsque les classes sont très déséquilibrées en taille). Nous ne donnons que les mesures UAR avec la valeur de confiance associée. L'intervalle de confiance est le même que celui utilisé en reconnaissance du locuteur (équation 6.6), avec p la probabilité et N le nombre d'instances total.

$$Confiance = 1,96 \cdot \sqrt{\frac{p \cdot (1-p)}{N}} \tag{6.6}$$

Nous proposons de donner les performances en cross-validation de la reconnaissance automatique de quatre émotions primaires (colère, joie, tristesse et un état dit "neutre"), ainsi que la valence et l'activation. Les descripteurs acoustiques utilisés pour cette étude sont ceux qui ont été développés au chapitre 3, ainsi que l'ensemble de descripteurs OpenEar [Schuller et al. 09b] avec (OE384) et sans les dérivées temporelles (OE192).

Le regroupement en macro-classes est une affaire extrêmement complexe. Plusieurs méthodes peuvent être employées. Tout d'abord une méthode fondée sur les théories catégorielles des émotions, qui permet d'intégrer des émotions secondaires dans des macro-classes d'émotions primaires. Il faut savoir que les émotions secondaires sont assez spécifiques à la tâche et que cette particularité peut se répercuter sur les macro-classes. Ces macro-classes émotionnelles risquent de ne pas être homogènes acoustiquement parlant. Une autre méthode consiste à étudier la matrice de confusion obtenue sur l'ensemble des émotions secondaires, et de regrouper les émotions en fonction du taux de confusion. Ainsi, les performances seront meilleures sur une tâche spécifique, mais cela ne permet pas de généraliser le regroupement à plusieurs corpus puisque certaines émotions secondaires

Corpus	OE384	OE192
JEMO	66,2 (2,5)	65,6 (2,6)
NAO-HR1	62,8 (6,8)	56,3 (7,0)
IDV-HR	51,3 (2,8)	47,5 (2,8)
IDV-HH	56,8 (4,5)	56,8 (4,5)

TABLE 6.2 – Reconnaissance automatique des émotions (colère, joie, tristesse et neutre)

peuvent être proches d'un point de vue système mais éloignées d'un point de vue sens. Enfin, on pourrait envisager une méthode plutôt fondée sur l'acoustique afin de regrouper les émotions fines, les plus proches d'un point de vue acoustique. Cela nécessite de pouvoir discriminer les émotions acoustiquement parlant de manière très fine. Nous ne sommes pas sûrs encore une fois, que cette méthode soit généralisable à un grand nombre de locuteurs et d'environnements sonores.

Nous avons choisi de faire un regroupement proche des théories émotionnelles. Cela permet de rester cohérent avec ce qu'exprime le locuteur et de pouvoir prendre une décision en rapport. Ainsi, en fonction des tâches, les macro-classes ne sont pas homogènes d'un point de vue système. Ce qui peut parfois dégrader les performances de façon significative. Pour palier cette dégradation, nous avons décidé de n'inclure dans les macro-classes que les émotions secondaires les moins ambigues en fonction de la tâche. Nous ne traiterons pas l'émotion peur qui peut inclure la gêne, le stress ou l'inquiétude parce que cette macro-classes contient un trop faible nombre d'instances dans l'ensemble des corpus.

Pour le corpus IDV-HR, la colère est représentée par les annotations colère et agacement, la joie par la joie, l'amusement, la satisfaction, le soulagement et les émotions positives, la tristesse par la tristesse, la déception et le neutre par le neutre et l'intérêt.

Pour le corpus IDV-HH, la colère est représentée par les annotations colère, agacement, ennui et autres émotions négatives, la joie par la joie, l'amusement, la satisfaction, le soulagement et les émotions positives, la tristesse par la tristesse, la déception et le neutre par le neutre et l'intérêt. Dans les corpus ROMEO, la valence a été annotée en tant que telle, c'est-à-dire qu'elle inclut toutes les émotions négatives (tristesse, colère, peur, etc.). Dans le corpus JEMO, la valence n'a pas été annotée, les émotions positives sont représentées par la classe joie, tandis que les émotions négatives par les classes colère et tristesse. Pour le corpus AIBO, nous n'avons pas accès à la valence telle que nous l'avons définie, nous montrons alors les résultats obtenus sur les émotions disponibles colère/positive/neutre de AIBO.

Nous avons choisi d'utiliser des modèles SVM (fonction SMO avec noyau RBF) de manière à utiliser les mêmes paramètres que ceux proposés dans les challenges Interspeech successifs. Les paramètres du modèle sont optimisés avec LibSVM [Chang and Lin 11], nous avons également utilisé l'outil Weka [Hall et al. 09] en utilisant les paramètres optimisés.

Résultats Au vu des performances UAR entre les ensembles OE384 et OE192, nous pouvons conclure que l'ajout systématique de dérivées temporelles dans les descripteurs acoustiques ne permet pas d'améliorer significativement les résultats. Cette remarque a déjà été faite dans l'article [Tahon et al. 11]. Par la suite, nous n'utilisons que le set

Corpus	OE384	OE192
JEMO	71,4 (2,3)	69,2 (2,4)
NAO-HR1	64,5 (3,6)	62,9 (3,6)
IDV-HR	62,3 (1,9)	61,2 (1,9)
IDV-HH	65,1 (2,3)	66,8 (2,3)
AIBO-O	37,2 (2,5)	37,5 (2,5)
AIBO-M	62,8 (2,9)	61,2 (2,9)

TABLE 6.3 – Reconnaissance automatique de la valence (positif, négatif et neutre)

Corpus	OE384	OE192
NAO-HR1	50,0 (3,4)	52,4 (3,4)
IDV-HR	54,8 (3,2)	55,8 (3,2)
IDV-HH	45,0 (2,5)	42,7 (2,5)

TABLE 6.4 – Reconnaissance automatique de l'activation (passif et actif)

OE192.

En accord avec l'état de l'art les performances de reconnaissance des émotions, de la valence et de l'activation sont plus élevées pour des émotions prototypiques (corpus JEMO) que pour des émotions plus spontanées. Les performances sont également plus élevées dans le cas du corpus de voix d'enfants NAO-HR1 alors que la valeur de kappa est la même que pour IDV-HH. Cela peut s'expliquer en grande partie par le fait que les enfants en situation de jeu, vont avoir tendance à s'exprimer de manière plus spontanée et moins contrôlée que des adultes et surtout que des personnes âgées. C'est également une des conclusions apportée par l'étude croisant des corpus de voix d'adultes et d'enfants [Tahon et al. 11].

Contrairement à nos attentes, les performances sont plus élevées sur le corpus IDV-HH que sur le corpus IDV-HR. Ces deux corpus ont été collectés sur des locuteurs similaires (entre 20 et 80 ans, mal-voyants) dans des conditions acoustiques très différentes. IDV-HH a été enregistré dans un studio extrêmement réverbérant alors que IDV-HR a été enregistré dans un salon d'appartement classique. Nous rappelons ici que les valeurs de kappa sont bien plus faibles pour le corpus IDV-HR (sur la valence 0,32 pour IDV-HR, 0,71 sur IDV-HH). Il semblerait alors que l'influence de l'annotation soit bien plus importante que la dégradation apportée par la réverbération.

6.2.2.2 Reconnaissance des émotions en cross-corpus

Nous proposons dans cette section des résultats de reconnaissance automatique sur plusieurs corpus. Ce type d'étude est fondamental pour évaluer la robustesse des systèmes face à des conditions très diverses.

Protocole Contrairement au protocole utilisé par Eyben [Eyben et al. 10], les modèles ne sont construits que sur un des corpus et non sur l'ensemble des corpus à disposition sauf celui qui est testé. Cela permet d'étudier la compatibilité entre les spécificités de

	NO	NC	NL
JEMO	44,3	58,9	58,7
NAO-HR1	24,4	38,1	37,3
IDV-HR	52,2	60,7	59,3
IDV-HH	45,5	48,2	45,3
AIBO-O	30,3	35,8	40,7
AIBO-M	41,6	43,6	42,8

TABLE 6.5 – Performances cross-corpus (moyenne sur les cinq corpus testés) avec différents types de normalisation (NO : aucune, NC : normalisation au corpus, NL : normalisation au locuteur). Score de confiance moyen 2,5. Valence : positif, négatif ou neutre.

chaque corpus. Pour pouvoir utiliser les six corpus à notre disposition, nous avons étudié la reconnaissance de la valence (positif, négatif ou neutre). Un des avantages de ce choix est que l'annotation de la valence est souvent plus consensuelle que celle des émotions et les valeurs de kappa sont relativement plus élevées que sur les autres tâches.

Etant donné que les résultats ne sont pas significativement meilleurs en cross-validation avec les dérivées temporelles, nous n'avons utilisé que le set OE192 pour comparaison. Les taux de reconnaissance sont obtenus en entrainant les modèles sur un des corpus avec Weka (fonction SMO, noyau RBF, paramètres optimum moyens obtenus à partir des expériences en cross-validation, $c = 2,0$ et $\gamma = 0,03125$) et en testant sur un des cinq corpus restants. Pour chaque corpus utilisé pour l'apprentissage, nous donnons le taux de reconnaissance moyen sur l'ensemble des cinq corpus testés (tableau 6.5). Le détail des taux de reconnaissance est donné en annexe B2.

Résultats Une simple normalisation au corpus (soustraction de la moyenne d'un descripteur donné sur l'ensemble du corpus) permet déjà d'améliorer les performances moyennes en cross-corpus de manière significative (de 2 points sur AIBO-M à 18,5 sur IDV-HR). Par contre, la normalisation au locuteur (soustraction de la moyenne d'un descripteur donné sur l'ensemble des échantillons d'un même locuteur) ne semble pas apporter d'amélioration significative. Il est possible qu'un autre type de norme (Z-norme ou T-norme) apporte un gain aux résultats obtenus.

La figure 6.2 propose une visualisation des taux de reconnaissance moyens en fonction du nombre d'instances dans chaque corpus. Globalement, plus le corpus contient d'instances, plus son pouvoir de généralisation de la valence est important et plus les taux de reconnaissance en cross-corpus sont élevés. Nous pouvons souligner quelques particularités de ce graphique. A nombre d'instances équivalent, le corpus JEMO obtient de bien meilleurs scores que les corpus AIBO-O, AIBO-M et IDV-HH, surtout après normalisation. Cela provient très certainement du fait que ce corpus contient des données prototypiques, qui ont de fortes chances de se retrouver de manières assez similaires dans les autres corpus. La variabilité liée au type de tâche (émotions actées ou spontanées) ne semble pas dégrader les résultats lors d'expériences cross-corpus. L'utilisation de modèles prototypiques permettrait de généraliser de manière plus efficace les émotions spécifiques de chaque corpus spontané.

Le corpus IDV-HH a été collecté dans un environnement sonore très réverbérant, on

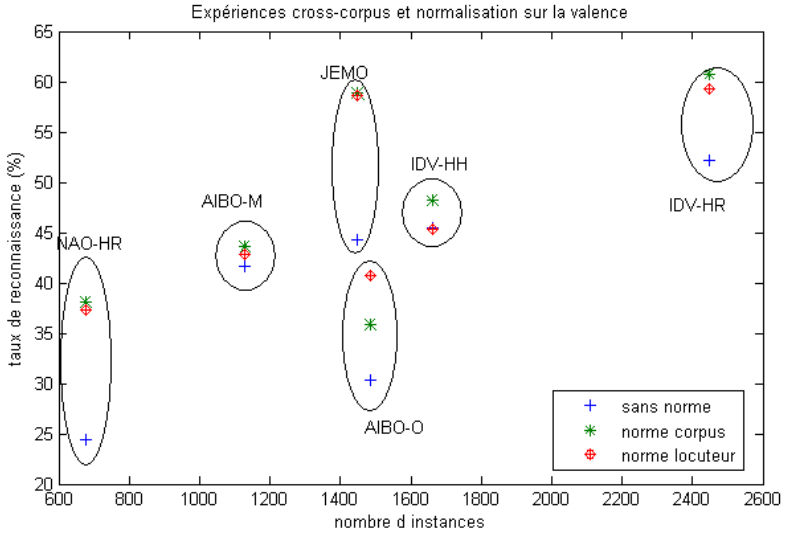


FIGURE 6.2 – Taux de reconnaissance (UAR) de la valence en cross-corpus en fonction du nombre d'instances et de la normalisation

pourrait penser que son utilisation en apprentissage apporterait de mauvais résultats, hors cela ne semble pas être le cas : seul le test du corpus IDV-HR sur le corpus IDV-HH donne un taux de reconnaissance inférieur au hasard (30,8%). Avec normalisation au corpus, ce taux grimpe à 46,5%. Cela nous porte à penser qu'une des différences majeures entre ces deux corpus concerne l'environnement sonore. Ainsi avec le set de descripteurs OE192 utilisé, une réverbération importante puisse être gommée grâce à une simple normalisation au corpus, contrairement à la normalisation CMS utilisée en identification du genre en conditions réverbérantes. Ainsi la variabilité "environnement" semble pouvoir se réduire avec une normalisation au corpus.

Les corpus AIBO-O et AIBO-M ne présentent pas des scores très intéressants, sauf lorsqu'ils sont testés et appris l'un avec l'autre. Même avec une normalisation, les résultats obtenus avec ces corpus qu'ils soient utilisés comme modèles ou comme tests sont de l'ordre du hasard (33%) en cross-corpus avec les corpus Romeo. Deux hypothèses peuvent expliquer cet aspect : la première vient du fait que les sous-corpus AIBO sont en allemand alors que les corpus Romeo sont en français. La deuxième viendrait du fait que la valence négative ne contient que l'étiquette colère dans AIBO, alors qu'elle est bien plus large dans les corpus Romeo. La variabilité "type d'émotion" nous semble alors très importante pour des expériences cross-corpus, et une simple normalisation ne permet pas de la réduire.

La variabilité "locuteur" ne semble pas perturber les scores de reconnaissance de la valence.

Ces deux corpus ont été collectés dans des conditions très différentes (notamment au niveau de la réverbération), normaliser par rapport au corpus permet d'augmenter le taux de reconnaissance de la valence de 53,4% à 68,1%.

Ces expériences cross-corpus sur le set de descripteurs OE192 avec plusieurs types de

normalisation permet de proposer une hiérarchie d'importance des variabilités entre les corpus et la normalisation qui permet de de réduire cette variabilité :

1. les émotions présentes dans le corpus (pas de normalisation classique possible),
2. l'environnement acoustique dans lequel le corpus a été enregistré (normalisation corpus),
3. le type de tâche, il vaut mieux mettre les données prototypiques en apprentissage,
4. le type de locuteur (normalisation non nécessaire).

6.2.2.3 Reconnaissance de la valence à partir de la qualité vocale

Selon Gendrot [Gendrot 04] la valence serait perçue essentiellement grâce à la qualité vocale. Nous avons testé cette hypothèse sur le corpus IDV-HR [Tahon et al. 12a]. Nous avons choisi un certain nombre de descripteurs de qualité vocale, certains assez couramment utilisés, d'autres beaucoup moins en reconnaissance automatique des émotions. Parmi ces descripteurs nous avons choisi les jitter, shimmer, taux de voisement et rapport harmonique sur bruit (HNR) qui sont calculés par le logiciel Praat. A ces descripteurs usuels, nous avons ajouter des descripteurs issus d'analyses en transformation de voix : le coefficient de relaxation (Rd) [Degottex et al. 08] et les fonctions de distortion de phase (FPD) [Degottex et al. 10] qui permettent de décrire le fonctionnement de la glotte. Ces descripteurs glottiques n'ont jamais été testé ni sur des signaux hors laboratoire (étant donné que leur application directe était la transformation de voix), ni sur des voix âgées. Leur utilisation sur des signaux enregistrés en contexte d'interaction homme-robot est donc totalement nouvelle.

Afin de limiter la complexité du problème nous n'avons utilisé que les instances annotées positives ou négatives avec une activation forte. Les descripteurs de qualité vocale sont calculés sur les parties voisées.

Un premier test ANOVA permet de montrer que les descripteurs de qualité vocale utilisés seuls ont un bon pouvoir de discrimination de la valence sauf les fonctions FPD, le jitter et le shimmer ($p>0.01$). Dans un second test, nous avons associé les descripteurs de qualité vocale à certaines familles de descripteurs issus du set OE192 (MFCC, F0, Energy). Chaque ensemble de descripteurs (par exemple OE-F0 + Rd) est utilisé comme entrée dans le système de classification automatique de la valence. Les résultats ne sont pas donné en mesures UAR ou WAR, mais en précision minimum. Cela permet de mettre en évidence la diagonale de la matrice de confusion. Si la diagonale n'est pas majoritaire, une des deux classes émotionnelles n'est pas reconnues à plus de 50%. Evidemment cette mesure est assez pessimiste.

Les résultats obtenus avec les MFCCs sont très mauvais, sans doute que les coefficients cepstraux ne permettent pas de reconnaître la valence. Les meilleurs résultats sont obtenus avec les sets OE-F0 + Shimmer et OE-F0+Energy+FPD (figure 6.3).

Un résultat important de cette étude concerne l'association de descripteurs plus ou moins redondants. En effet, lorsque l'on associe le Rd, le jitter, le punvoiced aux fonctionnelles OpenEar calculées sur la F0 alors que ces paramètres de qualité vocale sont dépendantes de la F0, les scores de classification diminuent. Si on les associe avec des fonctionnelles calculées sur l'énergie, les scores augmentent (figure 6.3). Sans doute faudrait-il restreindre la fenêtre de calcul des descripteurs de qualité vocale afin de les rendre plus robustes.

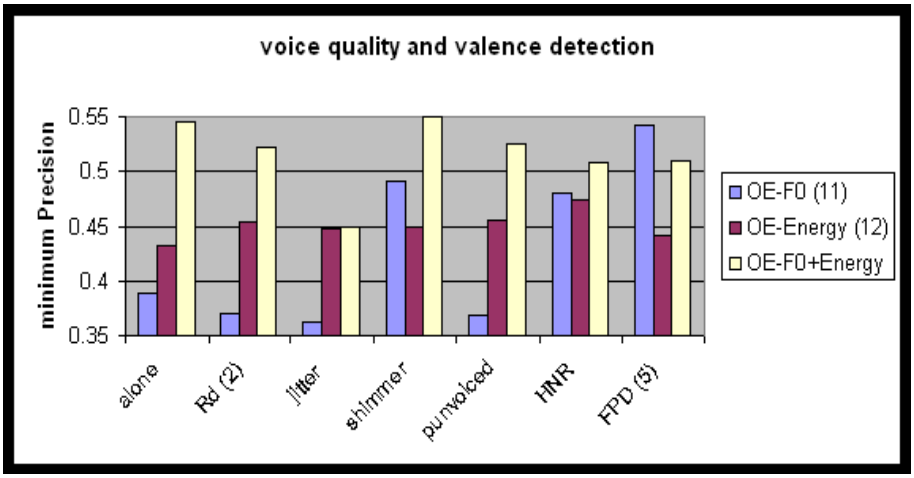


FIGURE 6.3 – Précision minimum obtenue après classification des classes négatives et positives sur IDV-HR avec plusieurs descripteurs de qualité vocale.

6.2.2.4 Reconnaissance d'affect bursts, exemple des rires

La reconnaissance automatique de rires a des applications très importantes dans un contexte d'interaction humain-robot. Les rires font partie des *affect bursts*, ils peuvent être à valence positive ou négative. Dans [Bachorowski et al. 01], Bachorowski propose une analyse très détaillée de rires clairs et sans ambiguité. Cette analyse propose plusieurs descripteurs acoustiques prosodiques (issus de la durée, la F0 et les formants) qui semblent être significatifs pour reconnaître un rire. Une étude menée par Devillers et Vidrascu [Devillers and Vidrascu 07] a permis de montrer après avoir fait un test perceptif sur une cinquantaine de rires que les rires à valence négative avaient tendance à être moins voisés, plus courts et moins forts que les rires à valence positive. Campbell [Campbell 04b] propose une première série d'indices acoustiques (à partir de la durée, la F0 et l'énergie du signal) fondés sur des études perceptives du rire, permettant une reconnaissance automatique.

Plusieurs méthodes de classification des rires ont été proposées par la communauté. Nous pouvons en retenir deux : un modèle GMM entrainé sur les coefficients PLP [Truong and van Leeuwen 07a] et un modèle SVM utilisant les coefficients cepstraux [Kennedy and Ellis 04].

Au-delà des rires clairs et sans ambiguité, il existe tout un ensemble de rires bien plus complexes dans la parole. Les plus difficile à cerner sont les rires superposés avec de la parole [Trouvain 01] qui sont largement majoritaires dans le discours (58% dans le corpus de rire en français de [Devillers and Vidrascu 07]). En français, les rires peuvent souligner l'ironie, ou d'autres émotions mixtes qui risquent d'être très difficiles à détecter automatiquement.

Protocole Les deux types de classifieurs GMM et SVM ont été expérimentés. Nous avons donc choisi de réutiliser les modèles GMM mis en place pour la reconnaissance des locuteurs, c'est-à-dire 26 coefficients MFCCs.

Un corpus de rire d'enfants a été créé à partir du corpus NAO-HR1, il est constitué de 125 segments de rires. Ces rires sont principalement de deux types, sonores et voisés ou de faible énergie et non-voisés. Deux modèles ont été entraînés sur l'ensemble 86 de ces rires, ainsi que sur 86 segments de parole émotionnelle et neutre sans rires. Le test est effectué sur 39 segments de rires et 45 segments sans rires.

Résultats Avec seulement 8 gaussiennes, on obtient un taux de reconnaissance de 85,7%. Ce résultat encourageant tend à montrer que la détection automatique d'un rire franc dans un flux de parole est tout à fait envisageable au cours d'une interaction humain-robot. Il n'en va pas de même des autres types de rires (superposition avec de la parole, rires liés à des émotions complexes comme la gêne masquant d'autres émotions, ...).

6.2.3 Reconnaissance automatique d'autres caractéristiques humaines

6.2.3.1 Reconnaissance de la personnalité

Nous avons testé nos nouveaux descripteurs (articulation et rythme) sur le corpus de personnalité SPC proposé [Mohammadi et al. 10]. Ce corpus a été annoté suivant cinq grands traits de personnalité : l'ouverture aux expériences (O : *Openness*), le fait d'être consciencieux (C : *Conscientiousness*), l'extraversion (E : *Extraversion*), l'agréabilité (A : *Agreableness*) et le névrotisme (N : *Neurotiscism*). L'idée de départ était de montrer que le rythme et l'articulation pouvait discriminer plusieurs types de personnalités.

Protocole Le corpus SPC est divisé en trois sous-corpus, l'un pour l'apprentissage, l'autre le développement et le dernier pour le test. Une première étape de notre étude consiste à évaluer le pouvoir discriminatoire des indices de rythme et d'articulation que nous avons développé au chapitre 3 (voir sections 4.2.3 et 4.2.4) grâce à des tests ANOVA. Dans une seconde étape, nous avons ajouté nos descripteurs aux 384 proposés comme baseline pour le challenge Interspeech 2009 [Schuller et al. 09b].

Résultats des tests de discrimination Les résultats du test ANOVA pour les descripteurs articulatoires sont présentés dans le tableau 6.6. Nous pouvons d'ores et déjà souligner le fait que suivant le sous-corpus choisi (ou la somme des deux), le pouvoir discriminant des indices d'articulation n'est pas le même. Cela montre le fait que les personnalités dans chaque sous-corpus ne sont pas les mêmes et ne peuvent pas forcément être généralisables d'un point de vue acoustique. Cette hypothèse rejoint également le fait que certains psychologues n'analysent pas la personnalité suivant des catégories mais avec des profils descriptifs uniques pour chaque personne. Cet aspect peut également compliquer grandement la tâche de sélection automatique d'indices. Nous pouvons néanmoins dire que l'articulation semble avoir un lien fort avec le fait d'être consciencieux ou pas.

En plus de l'articulation, le rythme nous semblait être un phénomène acoustique intéressant à utiliser pour la reconnaissance de la personnalité. Le test de discrimination n'est présenté cette fois que l'ensemble des sous-corpus apprentissage et développement mais les précautions à prendre pour l'analyse des résultats sont les mêmes que pour les paramètres d'articulation. Au vu des résultats des tableaux 6.7, le débit (descripteurs

sous-corpus	articulation	O	C	E	A	N
app	moyenne	X	X	X		
	écart-type		X		X	X
dev	moyenne	X	X			
	écart-type		X		X	X
app+dev	moyenne		X	X		
	écart-type		X			

TABLE 6.6 – Test de discrimination sur les indices d'articulation (une croix correspond à $p<0.001$)

Rythme	O	C	E	A	N
Débit					
durée des PV	X	X	X		
période des PV		X			X
durée des P non-V				X	X
densité maximum (P non-V)	X	X	X		
Précision					
P1.maxgauss	X	X			
P1.écart-type					
P2.maxgauss		X		X	
P2.écart-type					
P3.maxgauss					
P3.écart-type					

TABLE 6.7 – Test de discrimination sur les indices de rythme (une croix correspond à $p<0.05$)

de rythme fondé sur les parties voisées ou PV) paraît être un indice intéressant pour discriminer le fait d'être consciencieux et dans une moindre mesure l'ouverture aux autres, l'extraversion et le névrotisme. Par contre la précision ne semble pas être suffisamment discriminatoire, excepté pour le fait d'être consciencieux. Attention l'ensemble de ces résultats portent sur des aspects très précis de la personalité qui ont été expérimentés à partir d'un unique corpus. Ils ne permettent en aucun cas d'établir une généralité mais des pistes de travail.

Résultats de la classification automatique Les scores de confiance pour les résultats proposés dans le tableau 6.8 sont de l'ordre de 7%. Ainsi les gains réalisés sur la classification que ce soit entre les deux ensembles de descripteurs proposés par les organisateurs des challenges successifs (OE6125 et OE384), ou entre l'ensemble de base que nous avons choisi (OE384) et l'ajout de descripteurs d'articulation et de rythme), ne dépassent les 7% que dans peu de cas (névrotisme, agréabilité et le fait d'être consciencieux entre OE384 et OE6125).

L'ajout de nos descripteurs permet d'améliorer la classification des catégories O et

descripteurs (#indices)	O	C	E	A	N
Baseline (6125)	60,4	74,5	80,9	67,6	68,0
OE384 (384)	57,8	67,5	80,3	58,6	61,7
OE384 + articulation (386)	**58,4**	**68,2**	**80,3**	58,1	61,2
OE384 + ryt-precision (390)	**60,7**	**70,8**	77,1	57,6	60,6
OE384 + ryt-debit (392)	**58,3**	**68,4**	78,7	55,2	61,1

TABLE 6.8 – Classification automatique apprentissage sur le sous-corpus train, test sur le sous-corpus dev, scores UAR

C uniquement. L'amélioration de la classification de la classe C est en accord avec les résultats des tests de discrimination. Cependant l'articulation semblait être le paramètre le plus discriminant, alors que le score UAR avec la précision est le meilleur pour la classe C.

Conclusions Une des conclusions les plus importantes de cette étude est la différence d'approche entre tester le pouvoir discriminatoire d'un descripteur donné et l'ajouter à un ensemble de descripteurs existant pour en évaluer l'intérêt pour la classification automatique. L'approche discriminatoire ne prend pas en compte les relations de dépendance entre les différents descripteurs de l'ensemble choisi. Ainsi un paramètre peut se révéler peut discriminant, si il ne dépend d'aucun autre indice présent dans l'ensemble choisi, son ajout peut apporter une amélioration (c'est le cas de la précision pour la classification de la catégorie de personnalité C). Par contre si un paramètre n'est pas discriminant, son ajout ne permet pas d'améliorer les scores, et pourrait avoir tendance à les dégrader.

Au vu des différences obtenus sur les différents sous-corpus, la classification de la personnalité telle qu'elle a été abordée dans le contexte du challenge, est vraiment une tâche extrêmement complexe. La définition de la personnalité elle-même pose la question de la temporalité. La personnalité est une caractéristique humaine intrinsèque qui aura des conséquences plus ponctuelles et très liées au contexte, sur l'expression et la perception des émotions, sur l'élocution et sur le discours. Les enregistrements utilisés pour l'apprentissage et le test de la personnalité ne peuvent refléter l'ensemble des facettes d'une personnalité. De plus, il est difficile d'annoter la personnalité de quelqu'un sur un enregistrement de 10s correspondant à une situation particulière.

6.2.3.2 Reconnaissance du stress

Afin de continuer à tester nos paramètres de rythme et d'articulation, nous sommes intéressés à une autre tâche : la détection du stress dans la parole. Etant donné que l'annotation du corpus COMPARSE n'est pas terminée à l'heure de rendre ce manuscrit, nous n'avons pas eu l'opportunité de mener des tests de reconnaissance automatique du stress. Cependant nous proposons quelques pistes de recherches qui nous semblent pertinentes pour l'analyse du stress dans la voix.

Comme pour la personnalité, l'analyse du stress pose quelques problématiques au niveau de la temporalité. En effet, sur quel type de fenêtre temporelle, le stress dans la voix peut-il être annoté, analysé ? Quelle influence peut avoir le contrôle de soi sur l'expression du stress ? En effet, nous avons pu remarquer dans la majeure partie des

échantillons audio collectés, que le stress provoqué par la tâche de prise de parole en public est extrêmement ténu, il ne semble ce manifester que sur des indices très ponctuels et très précis. Par exemple, un léger tremblement de voix, une fréquence fondamentale plus haute que d'habitude ou encore une pause, une hésitation, sont des marqueurs de stress.

Protocole Nous avons donc choisi d'analyser certains descripteurs acoustiques sur quelques enregistrements de la collecte Comparse. Parmi ces descripteurs, nous avons sélectionné la fréquence fondamentale, l'énergie perçue (ou *loudness*), le rythme et l'articulation. Pour cette analyse, nous avons choisi les locuteurs 4 et 8 qui correspondent à des comportements ressentis comme extrêmes face au stress lors de la tâche de prise de parole en public. Le locuteur 4 est un homme relativement calme et posé alors que le locuteur 8 est une femme qui a été bien stressée par la tâche.

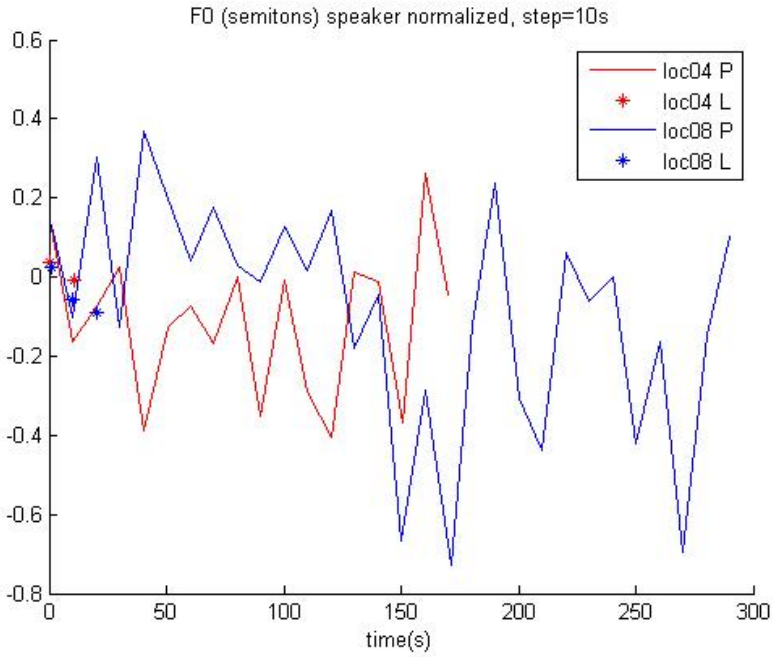
Les descripteurs d'articulation, de rythme et de fréquence fondamentale sont calculés par pas de 10s avec un overlap de 50%. Pour chacune des phases, nous représentons uniquement la moyenne temporelle. Le choix d'un pas de 10 s permet d'avoir suffisamment de signal pour déterminer de manière robuste les paramètres de rythme, cependant il semble que cette fenêtre temporelle soit inférieure à celle de la manifestation du stress.

Résultats La fréquence fondamentale est légèrement plus élevée au début de la phase de présentation que pour la phase de lecture pour les deux locuteurs étudiés. Au cours de la présentation, la F0 du locuteur 8 a tendance à diminuer ($-0,0015st/s$) alors que celle du locuteur 4 oscille autour de sa valeur moyenne. Pour le locuteur 8, il y a une baisse importante autour de 150 s de présentation, cela correspond avec un embrouillement au niveau du discours, des retours en arrières, des pauses, des hésitations. La baisse de la F0 sur un long temps de parole peut alors traduire la fatigue vocale, mais également une perte de confiance en soi.

Alors que le locuteur 4 a une F0 plutôt basse lors de ses réponses aux remarques négatives des juges, et plutôt haute sur les remarques positives, le locuteur 8 emploit la stratégie opposée. Il faudrait bien sûr étudier ces comportements sur l'ensemble des participants pour pouvoir généraliser ces tendances, mais nous pouvons d'ores et déjà dire qu'il existe des stratégies très différentes au niveau de la F0 en fonction des locuteurs.

L'articulation (figure 6.5) est en moyenne moins élevée lors de la présentation que lors de la phase de lecture, ce phénomène est relativement logique puisque lors de la lecture les mots sont déjà donnés. On peut néamoins remarquer que le locuteur 4 articule nettement moins que le locuteur 8 et son articulation a tendance à diminuer au cours de la phase de présentation ($-0,0002s^{-1}$). Il articule cependant plus sur la phase d'entretien et plus sur les questions négatives. Ce qui peut correspondre à une manifestation de stress. Le locuteur 8 va avoir une autre stratégie : comme au niveau de la F0, une rupture dans l'articulation se ressent également autour de 150 s. Lors des phases de questions négatives, le locuteur 8 aura tendance a moins articuler que pendant les questions négatives. Nous sommes donc en présence de deux comportements différents face au stress : l'un va hyper-articuler (stratégie de défense pour quelqu'un qui a confiance en soi ?) lors des remarques négatives alors que l'autre va au contraire sous-articuler (stratégie de fuite pour quelqu'un qui n'a peu confiance en soi ?).

a)



b)

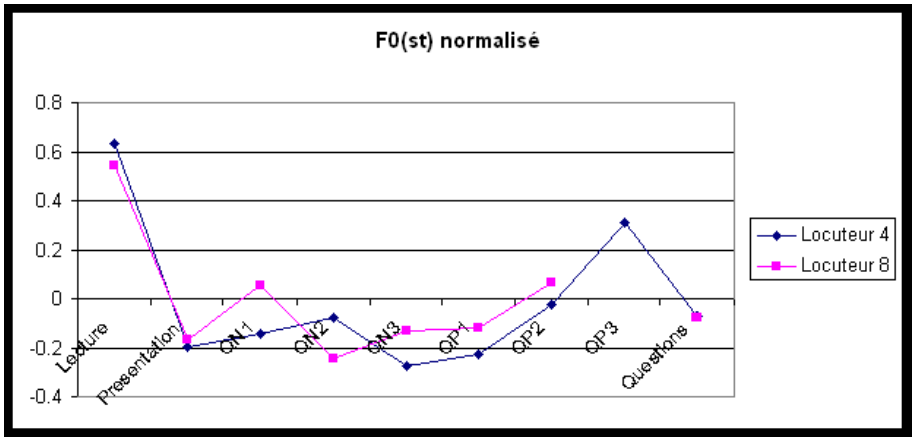
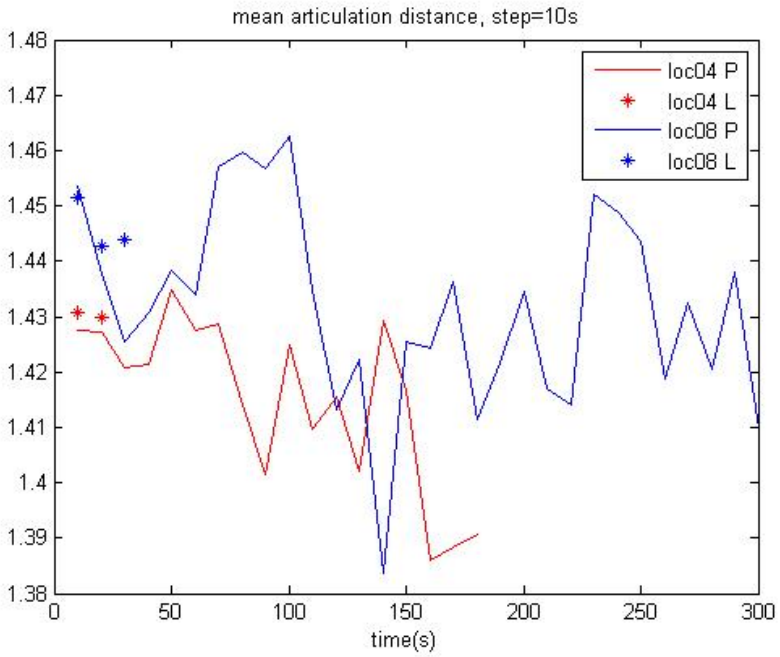


FIGURE 6.4 – Evolution de la fréquence fondamentale en semiton normalisée au locuteur pas à pas sur les phases de lecture (L) et de présentation (P) (a), en moyenne sur les différentes phases de lecture, présentation et questions (négatives QN et positives QP) (b)

a)



b)

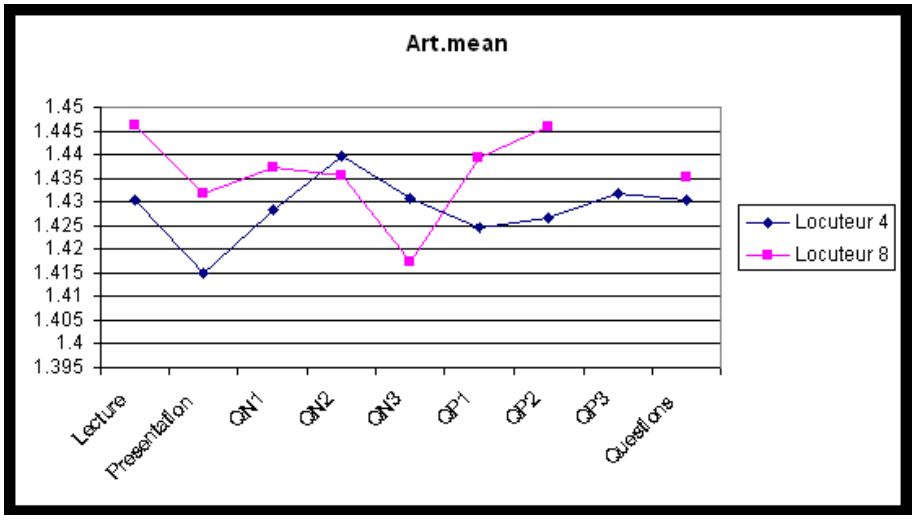


FIGURE 6.5 – Evolution de l'articulation pas à pas sur les phases de lecture (L) et de présentation (P) (a), en moyenne sur les différentes phases de lecture, présentation et questions (négatives QN et positives QP) (b)

Une étude similaire peut être conduite sur le débit de parole (figures 6.6). La mesure utilisée correspond à la période entre deux parties voisées consécutives. Plus la période est importante, plus le débit est faible et plus la personne parle lentement si elle parle régulièrement, plus elle fait de pause si elle parle de manière irrégulière. Nous pouvons remarquer que le locuteur 4 parle beaucoup plus lentement (et posément) que le locuteur 8. Cependant la période du locuteur 8 a tendance a augmenter $(+0,0001s/s)$ que ce soit sur la phase de lecture ou sur la phase de présentation. Pour ce même locuteur 8, la période va augmenter en moyenne sur les questions négatives, signe d'un débit plus lent ou d'une augmentation des pauses. Evidemment cette mesure de débit (ou de périodicité) doit s'accompagner d'une mesure de régularité. Ce que nous avons proposé également parmi nos descripteurs de rythme (densité de la période entre deux parties non-voisées consécutives). On peut remarquer que le locuteur 8 s'exprime de manière moins régulière que le locuteur 4.

Conclusions Les résultats présentés ci-dessus ne sont pas encore généralisables mais représentent des hypothèses de travail élaborées sur quelques locuteurs du corpus COMPARSE. Les descripteurs de fréquence fondamentale, de rythme et d'articulation que nous proposons semblent pertinents pour une analyse du stress dans la voix. Cependant ils n'ont pas été testés dans des systèmes de reconnaissance automatique du stress puisque le corpus n'était pas encore disponible. D'autres indices ponctuels doivent être ajoutés pour caractériser les tremblements de voix (jitter ou tremor). Cependant une des plus grandes difficultés de cette tâche reste dans le choix des fenêtres temporelles et des segments à analyser : pour la plupart des participants enregistrés, le stress ne se manifeste dans la voix que de manière extrêmement fine et ponctuelle.

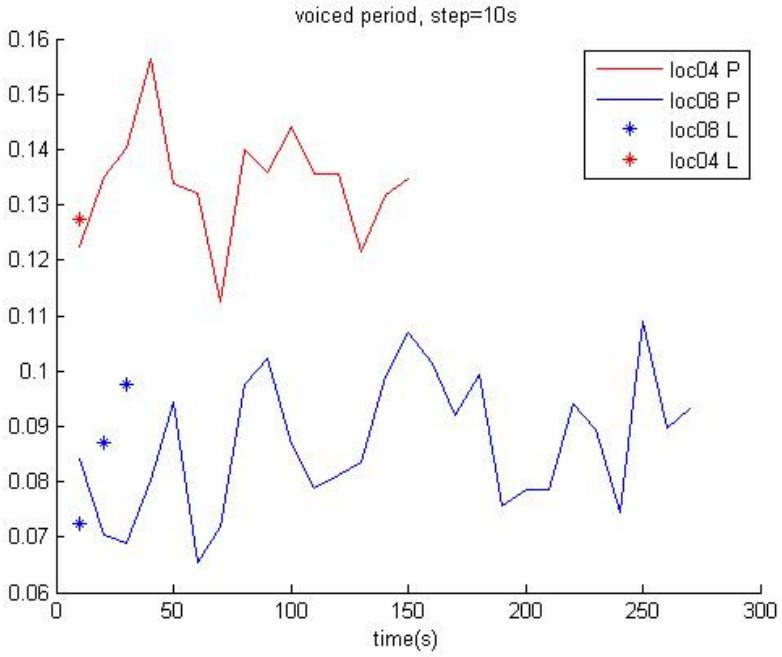
Nous n'avions pas à disposition les annotations de stress afin de pouvoir présenter des résultats sur les performances de reconnaissance. Une étude précédente sur la reconnaissance du stress dans la parole montre que les performances peuvent atteindre 51,2% sur quatre classes de stress [Fernandez and Picard 03] (modèle SVM, indépendant des locuteurs) en utilisant l'opérateur d'énergie Teager. Cet opérateur est également utilisé avec succès par Zhou [Zhou et al. 01] pour la reconnaissance de quatre classes de stress de la base de données SUSAS [Hansen and Bou-Ghazale 97]. Les performances de reconnaissance obtenues sont de 45-65% pour les types de stress non neutres. Une des perspectives pour cet axe de recherche serait de tester la classification automatique du stress en combinant nos descripteurs de rythme et d'articulation avec les des descripteurs comme les opérateurs Teager ou bien des filtres de Gabor [He et al. 10].

6.3 Conclusion

Ce chapitre met en avant l'influence des différentes variabilités présentes lors d'une interaction humain-robot sur la classification automatique d'indices paralinguistiques. Nous avons abordé plusieurs types d'indices présents dans le signal vocal : les indices émotionnels (émotions primaires, valence, activation et affect bursts), les indices de personalités et les indices de stress.

Les performances d'une classification automatique sont intrinsèquement liées aux annotations des corpus d'apprentissage des modèles. Dans le cas de la caractérisation du

a)



b)

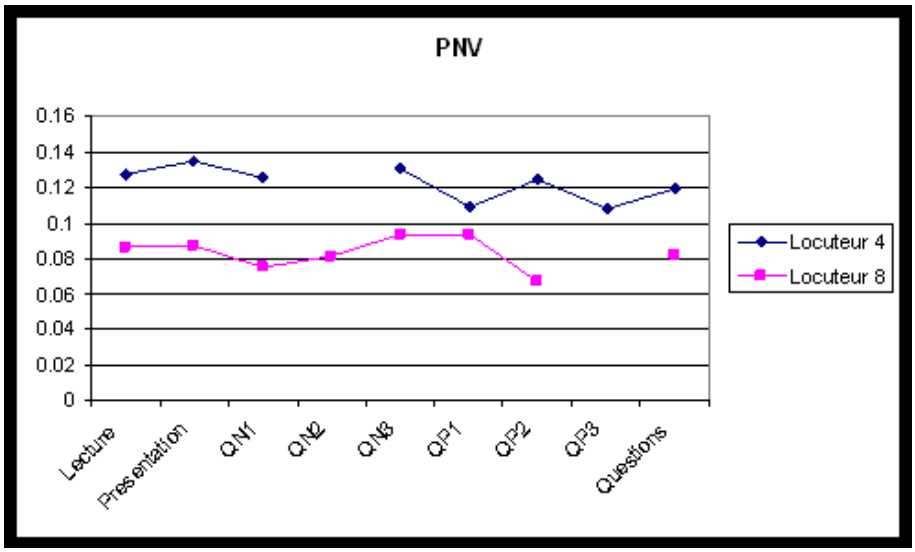


FIGURE 6.6 – Evolution de la période entre deux parties voisées consécutuves pas à pas sur les phases de lecture (L) et de présentation (P) (a), en moyenne sur les différentes phases de lecture, présentation et questions (négatives QN et positives QP) (b)

locuteur, l'environnement acoustique pouvait engendrer des dégradations très importantes des taux d'erreur, ce ne semble pas le cas avec les systèmes de reconnaissance automatique des émotions mis en oeuvre dans nos travaux qui utilisent des descripteurs acoustiques liés au segment et non pas à la trame courante.

Deux types d'expérimentation sur les indices émotionneles sont menées : en cross-validation sur un même corpus et en cross-corpus. Les expériences en cross-validation (section. 6.2.2.1) montrent que les dérivées des descripteurs acoustiques n'étaient pas d'importance pour la classification des émotions. Nous retrouvons également un résultat connu, à savoir que les émotions prototypiques se reconnaissent mieux que les émotions spontanées. Les expériences en cross-corpus 6.2.2.2) permettent de montrer l'intérêt d'une normalisation (au corpus plutôt qu'au locuteur). D'après nos résultats, il semble que le nombre d'échantillons contenu dans chaque corpus soit très fortement lié avec les taux de reconnaissance (UAR). Plus un corpus est important, plus son pouvoir de généralisation est fort et plus les performances obtenues sur des corpus de test différents seront bonnes. Cet aspect est à confirmer en utilisant des corpus d'apprentissage regroupant plusieurs corpus différents.

La reconnaissance des émotions a été testée dans deux langues différentes (français et allemand) sans que les performances cross-corpus en soient affectées. Cependant ces deux langues correspondent à deux cultures occidentales relativement proches. Des expériences mélangeant des cultures radicalement différentes (occidentale, asiatique ou africaine) permettraient de mettre en évidence la difficulté (voire l'impossibilité) d'obtenir des performances satisfaisantes en cross-corpus.

Nous avons également cherché à améliorer les scores de reconnaissance de la valence en utilisant des descripteurs de qualité vocale (section 5.2.2.3). Ces expériences permettent de conclure sur l'intérêt de la qualité vocale (paramètres glottiques, jitter, shimmer, taux de voisement, etc.) pour la valence lorsque les descripteurs sont isolés, mais il apparaît que lorsque ces descripteurs sont ajoutés à d'autres (F0, MFCC ou énergie), les résultats de classification ne sont pas meilleurs pour autant. La combinaison de descripteurs acoustiques est efficace lorsque ces descripteurs ne sont pas redondants.

Enfin nous avons étudié la reconnaissance d'indices de personalité et d'indices de stress. Les nouveaux descripteurs que nous avons proposé (chapitre 3, sections 4.2.3 et 4.2.4) apparaissent comme très pertinents pour ces deux types d'indices paralinguistiques. Des études sur des corpus plus variés de stress et de personalité sont nécessaires pour confirmer l'intérêt de nos descripteurs de rythme et d'articulation. Nous avons montré leur pertinence pour la reconnaissance du stress sur des échantillons du corpus COMPARSE, il faudrait les combiner avec d'autres types de descripteurs de timbre (opérateur Teager ou ondelettes) afin d'obtenir des performances de classification intéressantes.

Cinquième partie

Perspectives et conclusions

Chapitre 7

Vers une application réelle : intégration d'un module de reconnaissance automatique dans un robot et aspects éthiques

L'objectif de l'ensemble de ces travaux est essentiellement de mettre en place à plus ou moins court terme, un système de reconnaissance du locuteur et de son état émotionnel capable de fonctionner au cours d'une interaction humain-robot. Les études de la troisième partie permettent de souligner les avantages et les limites d'un tel système, ainsi que de proposer des pistes pour le rendre plus robuste à l'ensemble des variabilités présentes au cours de l'interaction. Malgré l'ensemble de ces inconnues, nous avons construit un système fonctionnel intégré au robot NAO. Ce système permet de montrer une certaine faisabilité au cours d'une application très cadrée. L'intégration d'un tel système dans le robot et son utilisation en temps réel, posent de nouvelles questions, telles que le choix d'un microphone ou la gestion de la distance entre le locuteur et le robot.

Rendre un système de détection automatique de traits de personnalités, d'états affectifs ou émotionnels, disponible au grand public pose également un certain nombre de question d'ordre éthique. Nous proposons une réflexion assez large sur la collecte de données nécessaire à la construction des modèles, mais également sur l'utilisation pratique qui peut être faite des systèmes automatiques d'identification du locuteur, de reconnaissance de la personnalité, des émotions. Enfin, l'apparition des robots dans le quotidien, comme robot-joueur ou robot-assistant soulève une quantité de questions d'ordre sociétal.

7.1 Le module SysRELL (Système de Reconnaissance Emotion-Locuteur du LIMSI)

Dans la partie précédente, nous avons vu l'effet de l'environnement, de la tâche, du type de locuteur et des émotions sur les descripteurs acoustiques eux-mêmes, mais aussi

sur les systèmes de reconnaissance automatique de reconnaissance du locuteur et de son état émotionnel. Nous avons mis en place un système automatique, temps réel, d'identification du genre et de reconnaissance des émotions pouvant s'intégrer dans le robot NAO, appelé SysRELL. Dans le cadre du projet ROMEO, le signal entrant dans le module de reconnaissance automatique des émotions est censé avoir été filtré en amont, segmenté suivant un niveau d'activité vocale et relativement peu sensible aux variations d'amplitude (liées à la distance entre le microphone et le locuteur, à la réverbération dans la salle, aux types de microphones). Nous n'avons donc pas étudié ces problématiques absolument fondamentales si l'on souhaite mettre en place un système autonome et fonctionnel. Nous avons fait en sorte que le signal entrant dans notre système soit relativement propre, ce qui impose un grand nombre de contraintes lors de tests en situation réelle (choix du microphone, de la salle). Dans cette section, nous avons réalisé des expériences hors ligne afin d'évaluer les performances théoriques de SysRELL (capture audio, segmentation et reconnaissance sur des signaux propres et connus). Par contre l'évaluation temps réel du module intégré au robot NAO (ou ROMEO) n'a pas pu être réalisé : nous n'avons pas eu accès aux modules de pré-traitement du signal capté (gestion de la distance entre le locuteur et le robot, filtrage amont des signaux pour supprimer les bruits du robot lui-même, articulation ou ventilation, réverbération, etc.)

Les performances réelles d'un tel système sont en fait un compromis entre une bonne reconnaissance des locuteurs et des émotions et la robustesse à différentes conditions. Un modèle sera d'autant plus performant qu'il sera proche de la situation de l'application. Cependant nous ne pourrons jamais avoir autant de modèles que de situations, la création d'un modèle passant par la collecte et l'annotation d'un corpus est une tâche très coûteuse. Une des pistes pour améliorer la robustesse de tels systèmes serait de mélanger l'adaptation des modèles à la situation présente, tout en ayant des modèles différents par locuteur, environnement, tâche, mais jusqu'à quel point ?

Même si la reconnaissance automatique des émotions est un domaine de recherche très actif aujourd'hui, relativement peu de laboratoires tentent l'intégration d'un système automatique de reconnaissance des émotions dans une machine aussi complexe que le robot.

7.1.1 Synopsis

7.1.1.1 Le projet Romeo

Le projet Romeo était un projet FUI financé par la région Ile de France par l'intermédiaire de Cap Digital. De nombreux partenaires industriels et laboratoires ont participé à la réalisation de ce projet piloté par Aldebaran [1], leader. L'objectif final était de construire un robot humanoïde doté d'un certain nombre de capacités motrices, cognitives et de communication. Nous nous sommes intéressés plus spécifiquement aux axes traitant de l'analyse audio du signal. Le robot ROMEO a été conçu pour évoluer dans un environnement domestique chez une personne en perte d'autonomie. Le public visé est très large : des enfants, adultes et personnes âgées (dont des mal-voyants). L'Institut de la Vision, localisé à Paris, était un des partenaires du projet et avait pour objectif de proposer à des personnes mal-voyantes de participer au projet. C'est entre autres, grâce

1. www.aldebaran-robotics.com

à ce partenariat que nous avons pu enregistrer les deux corpus de locuteurs mal-voyants IDV-HH et IDV-HR.

Les capacités de traitement de l'audio du robot ROMEO étaient très ambitieuses : le robot pouvait localiser une source parmi plusieurs, séparer différentes sources (le propriétaire qui parle, la télévision et la sonnette de la porte d'entrée par exemple) afin de pouvoir se diriger vers elle pour mieux analyser le signal. La capture du son a été étudiée par Telecom Paris Tech, la proposition finale était de doter le robot de 16 micros disposés en couronne autour de la tête. Le partenaire Telecom Paris Tech devait également fournir au robot un signal relativement propre et non-bruité, contenant une seule source (et donc un seul locuteur). Ce second volet n'a pas encore vu le jour, nous n'avons donc pas pu travaillé directement avec les signaux audio issus de la capture du robot. Pour des raisons pratiques (enregistrements de corpus, analyse fine, etc.), nous avons préféré travailler sur des signaux de bonne qualité enregistrés sur micro-cravate.

Le signal audio devait être traité par le LIMSI pour la reconnaissance des émotions et la caractérisation (ou l'identification) du locuteur. Voxler [2], partenaire industriel, pour l'ajout d'applications de jeux (reconnaissance de chansons), de transcription automatique mais surtout de segmentation du signal en locuteur et en pseudo-phrases. A un autre niveau, Spirops [3], également partenaire industriel, devait s'occuper du cerveau du robot, c'est-à-dire lui donner la capacité de prendre des décisions à partir d'entrées fournies. Spirops devait également créer des applications type agenda, appel téléphonique.

Dans le cadre du projet, nous avons mis en place un module de reconnaissance automatique des émotions et de caractérisation du locuteur. Etant donné le caractère ambitieux du projet Romeo et de l'ensemble des recherches qui devaient y être menées, il est évident qu'un certain nombre de contraintes se sont imposées assez rapidement. La plus importante concerne la qualité des signaux audio. Nous n'avons pas eu accès à des échantillons de signal capturés directement avec les microphones du futur robot, il a donc fallu travailler en déporté. Nous n'avions pas accès non plus aux filtres amont qui étaient censés rendre le signal plus propre. C'est pourquoi nous avons du faire des choix afin de pouvoir construire nos modules dans les meilleures conditions. Nous avons travaillé avec le robot NAO d'Aldebaran pour tout ce qui est intégration et communication avec les différents modules, mais la capture audio n'a pas pu être réalisée avec les microphones du petit robot.

7.1.1.2 Eléments de contexte

L'élaboration de SysRELL est un compromis entre les derniers résultats des recherches du groupe et une mise en oeuvre efficace. C'est-à-dire que nous souhaitons mettre en place un système intégré pour pouvoir étudier l'inflence des variabilités liées à l'interaction sur le système complet. Afin de limiter la complexité des variabilités, nous choisirons de fixer celles correspondant à la tâche, la capture du son est également fixée, par contre l'environnement acoustique reste variable.

1. Tâche fixe : nous avons choisi une tâche de jeu émotion où le locuteur doit acter une émotion qui doit être reconnue par le robot NAO, suivant le même protocole que la collecte du corpus JEMO [Brendel et al. 10]. Le public visé est alors un public

2. www.voxler.eu
3. www.spirops.com

FIGURE 7.1 – Interaction entre le robot NAO et une expérimentatrice

d'adultes actifs français (étudiants, chercheurs, industriels, etc.), nous nous plaçons alors hors de situations extrêmes (voix âgées, voix pathologiques). Les modèles de genre, de locuteurs et d'émotions seront construits sur la même base de données JEMO. Le choix d'un corpus acté permet de pouvoir s'adapter de manière plus efficace à l'application finale souhaitée (voir chapitre 5). L'application finale, interfacée avec le module de Spirops est conçue pour s'adapter à plusieurs types de tâche (jeu, scénario de la vie quotidienne par exemple le réveil du matin, etc.). D'après les résultats du chapitre 5, l'utilisation d'un corpus d'apprentissage de plus grande taille (ou une conbinaison de plusieurs corpus) devrait permettre au système d'être plus adaptatif. Nous n'avons testé qu'une seule tâche et donc réduit la quantité de données à traiter pour l'apprentissage en nous limitant au corpus JEMO.

2. Capture fixe : l'utilisation des microphones du robot lui-même apporte plusieurs problématiques nouvelles : la gestion de la distance entre le locuteur et le robot, le filtrage amont des signaux pour supprimer les bruits du robot lui-même (articulation, ventilation, etc.). Nous avons décidé de conserver le micro-cravate de bonne qualité (AKG PT40 Pro Flexx) utilisé pour la collecte de nos corpus. Ce choix permet de ne pas prendre en compte les problèmes liés à la distance entre le locuteur et le robot qui seront évidemment à étudier lors de l'utilisation des microphones du robot. Nos micro-cravate sont de type cardioïde, cela signifie qu'une partie du champ réverbéré est capturé avec la voix. Nous verrons que dans certains lieux, cela peut poser des problèmes importants.

Nous sommes donc dans une situation d'interaction entre un humain et un robot (figure 7.1), la capture du son ést réalisée avec un micro-cravate. Nous avons considéré qu'il n'y a pas de superposition de voix, c'est-à-dire que la situation ou deux personnes parleraient en même temps au robot n'est pas envisagée.

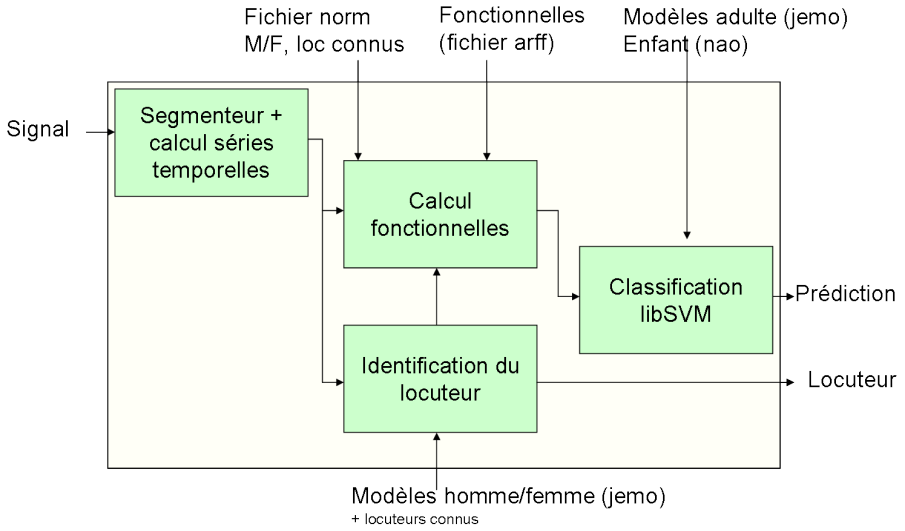


FIGURE 7.2 – Architecture globale de SysRELL

7.1.1.3 Architecture du système

Le module SysRELL a été implémenté en interne en C++ lors du stage de Guillaume Dulin (Master Pro), mais également par Mariette Soury, alors ingénieur informatique. C'est un executable autonome capable de tourner sur un PC. SysRELL (figure 7.2) permet d'analyser en temps réel un flux audio entrant :

1. de le segmenter,
2. d'identifier le genre et si le locuteur fait partie de la base de locuteurs connus ou non, si oui, le l'identité du locuteur,
3. de reconnaître l'émotion exprimée par le locuteur parmis 4 états émotionnels : neutre, colère, joie et tristesse, cette reconnaissance est plus performante lorsqu'une normalisation au genre lui est associée.

Les sorties de SysRELL sont alors : le genre du locuteur (homme ou femme), son nom (base de données de 8 locuteurs, inconnu sinon), l'état émotionnel (neutre, colère, joie ou tristesse), la valence (neutre, positive ou négative) et l'activation (actif ou passif).

Segmentation La segmentation du flux audio est réalisée par la PME Voxler en fonction de l'activité vocale (intensité du signal et énergie dans les hautes fréquences). Si un silence de plus de 800 ms est détecté ou si plus de 5 s de données sont collectées, le système traite le segment capté. La durée maximale d'un segment a été fixée à 5 s : c'est un compromis entre les capacités mémoire du système et la durée nécessaire pour une reconnaissance du locuteur satisfaisante. La librairie de Voxler permet également d'extraire plusieurs séries temporelles sur le segment (F0, Energie RMS, MFCCs et ΔMFCCs, ZCR, FFT).

Identification du genre et du locuteur Un module parallèle cherche à identifier le genre pour déterminer la norme à appliquer. Il permet également de reconnaître un locuteur si il appartient à la base de locuteurs connus.

Reconnaissance de l'état émotionnel Nous calculons 144 descripteurs à partir des séries temporelles extraites par la librairie Voxler. La sélection de ces descripteurs est un compromis entre un coup relativement bas en temps de calcul (et donc un relativement faible nombre de descripteurs) et la disponibilité des séries temporelles (par exemple, les formants ne font pas partie des séries temporelles extraites par la librairie Voxler, nous n'avons donc pas pu ajouter nos descripteurs d'articulation). Cet ensemble de 144 descripteurs peut encore être optimisé, certains fonctions doivent être ajoutées et d'autres supprimées. Les 144 descripteurs sont ensuite normalisés selon le genre du locuteur reconnu et utilisés pour identifier l'émotion, la valence et le niveau d'activation selon des modèles construits avec libSVM [Chang and Lin 11]. Afin de pouvoir utiliser un système de détection automatique des émotions en application réaliste, il faut qu'il puisse effectuer ses calculs en temps réel (on considère un temps inférieur à 3 s acceptable pour l'application visée, un jeu). Dans notre application, le calcul des 144 descripteurs acoustiques et la classification coûtent en moyenne 2,3 s (processeur intelcore2 Duo 1,6 gHz, 3,45 Go de mémoire), alors que la durée moyenne d'un segment dans le corpus JEMO est d'environ 1,9 s.

7.1.2 Evaluation de SysRELL en contexte de laboratoire

7.1.2.1 Identification du locuteur

La base de données utilisée pour l'apprentissage des locuteurs est JEMO. Le corpus contient 64 locuteurs ce qui est relativement peu en comparaison avec les centaines de locuteurs utilisées dans les campagnes NIST SRE. Cependant il semblerait que cela suffise à l'utilisation que nous souhaitons en faire.

L'identification du locuteur est fondé sur des modèles GMM et les coefficients MFCCs tels que nous l'avons étudié au chapitre 4. Nous avons construit un modèle par genre et un modèle par locuteur connu, soit sept modèles. La normalisation par soustraction du cepstre moyen (CMS) sur le segment analysé est systématique. Cependant, alors qu'elle est nécessaire lorsque les conditions acoustiques d'enregistrement sont différentes entre l'apprentissage et le test, elle ne l'est pas lorsque les conditions sont identiques. C'est pourquoi nous n'appliquerons pas la norme CMS lors des expériences en leave-one-speaker-out, alors qu'elle est implémentée dans SysRELL.

Pour des segments de test d'une durée de 2 s, tous les locuteurs connus du corpus JEMO sont parfaitement reconnus. Le taux d'erreur de reconnaissance du genre sur ces locuteurs connus est de 1,1%. Le seuil de log vraisemblance permettant d'identifier un locuteur connu ou inconnu a été optimisé sur les données JEMO à -97. Cette valeur est à optimiser en fonction de l'application.

7.1.2.2 Reconnaissance des émotions

Le corpus JEMO est également utilisé pour l'apprentissage des états émotionnels. Trois modèles sont créés sur la base des annotations faites sur le corpus : émotions, valence et

Type de descripteur	Fonctionnelles	voisé	non-voisé	tout
F0 (Hz) (10)	médiane, étendue, écart-type, ratio	X		
	différence inter/intra voisé médian, maximum, écart-type			
Energy (RMS) (3)	étendue, écart-type, fréquence des pics			X
Durée (Rythme) (5)	moyenne voisé, écart-type voisé, ratio voisé/non-voisé, moyenne silence, fréquence silence			X
FFT (68)	roll off 5, 95% médian, ratio Rolloff95/F0			X
	barycentre spectral médian			X
	bandes de Bark 1-22 médianes	X	X	
	bandes harmoniques 1-10	X	X	
Cepstral (52)	MFCC 0-12 mean	X	X	X
	ΔMFCC 0-12 mean			X
Qualité Vocale (6)	jitter médian, max, écart-type	X		
	shimmer médian, max, écart-type			X

TABLE 7.1 – Les 144 descripteurs acoustiques implémentés dans SysRELL

activation. L'outil LibSVM est utilisé pour la création des modèles ainsi que pour la classification d'une instance inconnue.

Parmi l'ensemble des descripteurs acoustiques qui ont été étudiés au chapitre 3, nous en avons choisi 144 qui nous paraissaient être intéressants (tableau 7.1). Des tests de sélection devront être menés sur ces 144 descripteurs pour supprimer ceux qui ne sont pas nécessaires pour l'application. L'objectif est en effet d'obtenir l'ensemble de descripteurs le plus réduit possible afin de consommer le moins de temps de calcul.

Les performances (WAR) hors ligne des modèles SVM implémentés dans le module SysRELL sont données dans le tableau 7.2. Ces performances correspondent aux résultats donnés en cross-validation par Weka (fonction SMO, noyau RBF avec les paramètres optimisés par LibSVM) avec ou sans normalisation par rapport au locuteur, avant et après avoir retiré les fonctionnelles spectrales sélectionnées par la classification automatique.

7.1.3 Conclusion

Cette section est entièrement dédiée à la mise en place d'un module de reconnaissance des émotions qui puisse s'intégrer dans une plateforme plus large de comportement du robot. Nos travaux montrent qu'il est effectivement possible pour une machine de reconnaître les émotions, seulement dans des conditions extrêmement calibrées. Le module SysRELL mis en place au LIMSI est aujourd'hui fonctionnel mais nécessite encore et toujours plus d'améliorations. En premier lieu, l'ensemble des 144 descripteurs doit être revu

Norme genre	Neutre/ Colère/ Joie/Tristesse	Valence (Positif/ Neutre/Négatif)	Activation (Passif/Actif)
Non	60,1	66,0	82,2
Oui	61,6	69,5	81,7

TABLE 7.2 – Performances (WAR) hors-ligne des modèles émotionnels de SysRELL

en fonction des conclusions des chapitres précédents et en ajoutant l'extraction de formants à la librairie de Voxler. Dans un second temps, un système d'adaptation à la tâche doit être mis en place afin de pouvoir rendre le système autonome sur plusieurs tâches. Cette adaptation peut être faite en utilisant une mesure de spontanéité comme proposée au chapitre 4 (section 4.3.1.2). Enfin la normalisation des signaux entrants dans le système doit être revue afin de corriger des éventuelles différences de gain entre l'apprentissage et l'application.

L'avantage d'avoir développé un module pouvant fonctionner en temps réel est d'avoir pu le tester dans des conditions variées avec des locuteurs différents. Des tests préliminaires (qui n'ont pas été présentés ici) ont permis de souligner l'importance de la réverbération et du choix de l'ensemble de la chaîne de mesure du signal de parole. Par exemple, une réverbération trop importante entraîne une mauvaise détection du genre, la normalisation au genre apporte donc des erreurs suplémentaires au niveau de la reconnaissance des émotions. De même lorsque le gain du microphone est trop fort, toutes les émotions risquent d'être reconnues comme étant de la colère. L'étape suivante serait bien sûr, de faire une évaluation du système intégré avec une centaine de locuteurs dans des conditions contrôlées, mais nous pouvons d'ores et déjà apporter quelques pistes de réflexion (tableau 7.3) sur les points à traiter avec précaution lors du passage d'un système élaboré en laboratoire à une application.

Un point n'a pas été abordé ici puisque c'est un sujet de recherche en lui même : comment prendre en compte la distance entre le locuteur et les microphones du robot ? Cette distance peut être estimée en utilisant le canal visuel, il faudrait alors normaliser le signal en conséquence. Mais lorsqu'une personne parle proche ou loin de son interlocuteur, son effort vocal diffère et transforme le timbre de sa voix. Ce point est important dans le cas de la construction de robots assistants présents dans le domicile d'une personne en perte d'autonomie.

.

7.2 Réflexion sur l'éthique

L'éthique correspond à une réflexion relativement récente autour de l'utilisation que l'on peut faire de certaines découvertes scientifiques. Cette discipline est apparue en premier lieu autour des biotechnologies : peut-on cloner un être humain, peut-on utiliser des cellules souches à des fins thérapeutiques ? Ces questions sont d'une extrême importance : elles permettent de poser des garde-fous autour des recherches scientifiques. On ne peut pas faire n'importe quoi avec la science.

Les réflexions autour de l'éthique sont aujourd'hui réalisées une fois les technologies

	Contexte labo	Contexte réel	Solutions proposées
Signaux de plusieurs locuteurs superposés	Supprimé (segmentation manuelle)	Possible	Piste de recherche (séparation de source)
Gain microphone	Contrôlé	Inconnu	Normalisation amplitude à 1, prendre un gain maximum sans saturation
Segmentation	Manuelle	Automatique (dépend de l'intensité et du gain)	Appliquer la même segmentation sur les données d'apprentissage
Locuteurs	Connus	Inconnus a priori	Identification du locuteur
Emotions	Modèles prototypiques	Dépend du contexte	Contraindre le contexte applicatif
Environnement acoustique	Connu	Inconnu a priori	Piste de recherche (prétraitement ou intégration du bruit ou de la réverbération dans les modèles)
Puissance de calcul	Quasi-illimitée	Limitée	Limitation du nombre de descripteurs et de la taille du segment analysé

TABLE 7.3 – Différences entre un contexte de laboratoire et une application réelle

trouvées, cependant elle devrait également se situer en amont, au niveau des décideurs de projets : on ne peut pas mener n'importe quel projet scientifique. Un des problèmes majeurs de l'éthique est qu'elle tend à s'imposer comme un consensus qui ne satisfait évidemment pas toutes les aspirations individuelles. Certains chercheurs voudraient aller plus loin dans des recherches sensibles alors que d'autres estiment que les recherches possibles vont déjà trop loin.

L'objectif de ce chapitre est d'apporter mes réflexions sur l'éthique des Technologies de l'Information et de la Communication (les TICs). La construction d'une éthique des TICs est très récente alors que l'utilisation de masse d'Internet a déjà plus d'une trentaine d'années. Un pas important a été franchi grâce à la mise en place au sein du CNRS et de l'INRIA de comités de réflexion sur l'éthique des TICs [Dowek et al. 09, Mariani et al. 09] ainsi que la CNIL.

Tels qu'ils sont conçus aujourd'hui, les systèmes de reconnaissance automatique sont fondés sur des modèles construits sur un grand nombre de données. La collecte elle-même de ces données peut poser des problèmes éthiques. L'utilisation que l'on fait ensuite de ces systèmes est évidemment très sensible. Nous verrons des exemples au travers des quatre traits humains sur lesquels j'ai travaillé : la voix d'un locuteur, sa personalité, son niveau de stress et ses émotions.

7.2.1 Collecte de données en vue de la construction de modèles

La collecte de données consiste dans notre cas, comme nous l'avons vu en chapitre 1 ou 2, à l'enregistrement audio principalement mais également vidéo de participants placés dans un contexte particulier. Deux aspects nécessitent une attention partciulière : (i) le participant doit donner son accord pour participer à l'expérience, (ii) il doit également être d'accord pour fournir des données individuelles (état civil, situation personnelle, profil de personalité, enregistrement audio, vidéo, rythme cardiaque, etc.).

7.2.1.1 Le consentement du participant

Le consentement est le résultat de réflexions sur l'expérimentation portées par les procès de Nuremberg en 1947. Le code de Nuremberg issu du procès, permet de poser un cadre légal autour des expériences scientifiques (et particulièrement médicales) sur la personne humaine. Ce code, à la base de la réflexion éthique moderne, a été pensé pour éviter de reproduire les expérimentations scientifiques inacceptables qui ont eu lieu par exemple sur les prisonniers des camps de concentration par les Nazis lors de la seconde guerre mondiale [Halioua 07] ou réalisées par l'Unité 731 (unité de recherche japonaise après l'invasion de la Mandchourie en 1932). A la base de l'éthique des expériences scientifiques sur des personnes humaines, nous trouvons alors :

- le consentement libre du participant,
- l'interdiction de porter atteinte à la dignité et à l'intégrité physique et mentale du participant,
- la responsabilité de l'expérimentateur,
- l'importance de motiver l'expérience par des raisons scientifiques et non de divertissement.

Pour qu'une personne accepte de participer à une expérience, il faut qu'il ait connaissance du déroulement de celle-ci, des risques et des dangers pour sa propre personne. Selon

l'article 5 de la Déclaration Universelle des Droits de l'Homme des Nations Unies (1948), "nul ne sera soumis à la torture, ni à des peines ou des traitements cruels, inhumains ou *dégradants*". Directement issue du code de Nuremberg, la loi de 2012, dit que toute personne doit avoir le choix de suivre ou non une expérience. Un sujet a également le droit d'arrêter l'expérience au moment où elle le souhaite. Une personne ayant participé à une expérience de recherche à le droit de demander les résultats de cette recherche. L'expérience ne doit pas être pratiquée au hasard et sans nécessité. Un essemble assez exhaustif de textes fondamentaux depuis le code de Nuremberg jusqu'aux plus récentes lois du droit français a été présenté par Emmanuel Hirsh, chercheur en éthique [Hirsh 12]. La description de l'expérience est une étape importante, il faut également s'assurer que le participant ait compris, ce qui n'est pas évident dans le cas de personnes dépendantes ou malades. A partir de ce constat apparaît alors une limitation dans les protocoles "éthiquement" corrects : une expérience ne peut pas être menée si un point qui peut mettre en danger le participant ne peut lui être communiqué sans fausser l'expérience. Evidemment la notion de "mettre en danger" un sujet est très subjective.

Deux exemples assez connus illustrent la nécessité de "cacher" certains aspects de l'expérience au participant. Lors de tests de médicaments à grande échelle, l'utilisation d'un placebo est très importante afin d'estimer le fait que le participant malade peut éventuellement se trouver mieux uniquement parce qu'il croit au médicament et non pas grâce aux effets de la molécule elle-même. Le participant malade ne doit pas savoir si la molécule qui lui a été donnée est un placebo ou non. De même lors d'expérimentations homme-machine (robot, agent virtuel), la machine est assez fréquemment pilotée en Magicien d'Oz, c'est-à-dire par un expérimentateur caché, la machine n'est alors pas autonome comme doit le croire le participant pour le bon déroulement de l'expérience. Dans ces deux cas, la description de l'expérience ne correspond pas exactement à la réalité. Cependant l' "intégrité physique et morale" de la personne semble respectée. Il est néanmoins important de souligner que dans le cas du Magicien d'Oz, l'image que se fait le participant de la machine est bien plus valorisante que la réalité. Dans les deux cas, le sujet scientifique n'est a priori pas mis en danger.

Une expérience scientifique est fondée sur la confiance que le participant a dans le scientifique. Dans cette relation, l'expérimentateur a évidemment un pouvoir sur la personne naïve. Cette obéissance de naïfs aux ordres de l'autorité scientifique a été étudié à partir des années 50 par Stanley Milgram [Milgram 63, Milgram 74].

7.2.1.2 Les données à caractère personnel

L'enregistrement de données relatives à une personne indivuelle ne peut se faire qu'avec son consentement.

Article 8 de la Charte des droits fondamentaux de l'union européenne, 18 décembre 2000 :

> 1. Toute personne a droit à la protection des données à caractère personnel la concernant.
>
> 2. Ces données doivent être traitées loyalement, à des fins déterminées et sur la base du consentement de la personne concernée ou en vertu d'un autre fondement légitime prévu par la loi. Toute personne a le droit d'accéder aux données collectées la concernant et d'en obtenir la rectification.

3. Le respect de ces règles est soumis au contrôle d'une autorité indépendante.

En France, c'est la CNIL qui régule les cadres légaux pour la collecte de données de ce type. Evidemment, l'enregistrement de données audio, vidéo, physiologiques implique également leur conservation sur un support numérique. Ces supports peuvent être sujets à plusieurs types d'intrusion : via Internet (les ordinateurs étant systématiquement connectés), via des utilisateurs non impliqués dans l'expérimentation. De plus le droit français garanti au participant la possibilité de modifier ou de supprimer les données à caractère personnel si il le souhaite. Cependant dans le cadre d'une recherche, les données peuvent être duppliquées sur plusieurs supports, sous plusieurs formats, elles peuvent être segmentées, transformées et même publiées. Une fois que le participant a donné son accord pour l'enregistrement, il est donc quasiment impossible de modifier ou supprimer ces données. Il faut donc qu'il soit conscient de cet aspect. La principale garantie que peut lui apporter le responsable des recherches est celle de l'anonymat. En effet, l'identité de la personne ne doit pas pouvoir être retrouvée à partir des données utilisées pour les recherches.

7.2.1.3 Cas de collecte de données dans le cadre de mes travaux de thèse

Afin de pouvoir collecter des corpus de parole émotionnelle spontanée avec des participants volontaires, nous avons donc besoin au préalable de leur autorisation et leur consentement. Cette autorisation comporte plusieurs paragraphes obligatoires : état civil, descriptif de l'expérience (objectif et durée), déclaration d'autorisation. L'autorisation est relativement exhaustive, elle insiste sur le fait que le participant comprend le sens de sa participation et garanti le laboratoire contre tout recours une fois l'autorisation signée.

extrait :

> Je déclare avoir été expressément informé(e) du projet du CNRS (Centre National de la Recherche Scientifique) (i) d'enregistrer ma voix, puis (ii) d'en faire un montage anonymisé aux fins de diffusions ultérieures. En conséquence de quoi et conformément aux dispositions relatives au droit à la voix et aux droits de la personnalité, j'autorise expressément le CNRS dans le cadre de cette recherche, à titre gratuit et non exclusif, à : (i) fixer et reproduire ma voix, à titre gracieux, par tous procédés, sur tous supports, en tous formats (tels que notamment photographies papiers ou numériques, films vidéos ou numériques) et ce en vue de leur communication au public (ci-après les "Enregistrements"). (ii) exploiter ou faire exploiter les Enregistrements, avec le logo et/ou la mention du CNRS et/ou du LIMSI, intégralement ou par extraits, à des fins de recherche ou de valorisation, sous toutes formes et tous supports par tous modes d'exploitation connus ou inconnus à ce jour, par tous réseaux de transmission ainsi que par tous réseaux de communications électroniques (par téléchargement et/ou en mode streaming) et en tous formats.
>
> Je garantis que je ne suis pas lié(e) par un contrat exclusif relatif à ma voix.
>
> Je comprends que ma participation n'est pas obligatoire et que je peux stopper ma participation à tout moment sans avoir à me justifier ni n'encourir aucune responsabilité. Mon consentement ne décharge pas les organisateurs de la recherche de leurs responsabilités et je conserve tous mes droits garantis par la loi.

> Je comprends que les informations recueillies sont strictement confidentielles et à usage exclusif des investigateurs concernés, pour des fins de recherche. J'ai été informé(e) que mon identité n'apparaîtra dans aucun rapport ou publication et que toute information me concernant sera traitée de façon confidentielle. J'accepte que les données enregistrées à l'occasion de cette étude puissent être conservées dans une base de données et faire l'objet d'un traitement informatisé non nominatif par l'UPR 3251.
>
> En conséquence de quoi, je garantis le CNRS contre tout recours et/ou action que pourraient former les personnes physiques ou morales estimant avoir des droits quelconques à faire valoir sur l'utilisation des Enregistrements. Je reconnais et accepte que le CNRS conserve toute liberté pour exploiter ou ne pas exploiter, intégralement ou par extraits, les Enregistrements.
>
> Je déclare que la présente autorisation est accordée (i) pour le monde entier et (ii) pour une durée de 5 an (s) à compter de la signature des présentes, renouvelable par tacite reconduction, sauf dénonciation de la part du signataire.

Dans le cadre de l'équipe Parole et Emotion dans lequel j'ai effectué mes recherches, les données ont été systématiquement anonymisées. C'est-à-dire que l'identité des participants n'apparait à aucun moment dans les données enregistrées. Seul un document renseignant le nom des participants, le numéro correspondant à leur enregistrement et d'autres données personnelles permettrait de retrouver son identité. Ce document n'est en aucun cas utilisé dans les recherches, il n'est conservé que pour ne pas enregistrer plusieurs fois la même personne, ou pour retrouver le numéro correspondant au cas où le participant souhaiterait accéder à ses données.

7.2.2 Systèmes de reconnaissance/détection automatique de traits humains

Les systèmes de reconnaissance automatique de traits humains se développent de manière importante. Par reconnaissance de traits humains, nous entendons l'identité d'une personne ([CNRS 08]), de son émotion, de sa personalité, d'une éventuelle pathologie (alcoolisme, fatigue, etc.) ou encore de son stress. Cette reconnaissance peut se faire à partir de support d'information diverses : empreintes digitales, forme de la main, rétine, voix pour la reconnaissance de l'identité ; voix pour celle des émotions ; voix et mesures physiologiques telles que rythme cardiaque, sudation, etc. pour la personalité. Il faut souligner que les méthodes utilisées pour effectuer cette reconnaissance ne sont pas infaillibles et qu'elles peuvent commettre des erreurs plus ou moins nombreuses selon la technologie employée. Il est évident que des réflexions éthiques doivent accompagner la recherche dans le domaine de la prise en compte de traits humains dans les Technologies de l'Information et de la Communication (TIC). Cette réflexion doit porter sur les effets sur l'homme et sur la société que peuvent avoir de tels systèmes de reconnaissance comme l'a fait le réseau d'excellence Humaine (Research on Emotions and Human-Machine Interaction) [4] dans le domaine des émotions.

4. http ://emotion-research.net/

7.2.2.1 Problème de la gestion des erreurs, leurs conséquences

Lorsque les machines deviennent capables de reconnaître un trait humain et de s'en servir, c'est là que l'éthique trouve toute sa place dans la recherche scientifique. Il faut se demander jusqu'où veut-on aller et quel contrôle l'homme pourra garder sur lui-même. Aujourd'hui, les machines ne sont pas encore dotées de raisonement éthique, leur autonomie face à la prise de décision fondée sur la reconnaissance automatique de traits humains est alors un sujet sensible. Pour le moment, la responsabilité des questions éthiques incombe principalement au concepteur de la machine et non à la machine elle-même. Il doit donc être conscient des risques liés à une mauvaise utilisation (intentionnelle ou non) de ces machines, ou à un dysfonctionnement lié à une modélisation statistique imparfaite du trait que l'on cherche à reconnaître.

Par exemple des systèmes de reconnaissance du stress dans la voix ont été développées afin d'être utilisées comme détecteur de mensonges. Ces systèmes sont utilisés en Grande-Bretagne pour détecter les fraudeurs aux indemnités chômage [Cox 08]. Les conséquences pour les usagers peuvent être très importantes.

On peut également citer, l'utilisation d'expertise en vérification du locuteur lors de procès en France à partir d'un signal vocal [Boë 00]. Les juges sont en mesure de demander à un expert (et qu'est-ce qu'un expert dans ce domaine? à ce sujet voir les travaux du COMETS sur l'éthique de l'expertise[5]) de donner son avis sur un signal de parole enregistré d'une personne suspecte. La vérification du locuteur sur des signaux de parole enregistrés dans des conditions et sur des supports différents est loin d'être sûre à 100%. Or il se peut que des "experts" sûrs de leurs outils, apportent une réponse définitive sans apporter de score de confiance relatif à la méthode utilisée.

7.2.2.2 Reconnaissance automatique sur des données de centre d'appel téléphonique

L'utilisation du support audio pour reconnaître un aspect humain a des applications nombreuses et variées dont certaines peuvent avoir des conséquences importantes sur les individus et la société. Par exemple, l'utilisation à des fins commerciales de conversations téléphoniques (âge, genre de la personne, pathologies), une surveillance téléphonique autonome qui pourrait renseigner un certain nombre de critères sur l'individu surveillé. La plupart de nos conversations téléphoniques vers des boîtes vocales sont enregistrées afin d' "améliorer le service". La conservation et le traitement de ces données à caractère personnel (appel vers Pôle Emploi, vers la Sécurité Sociale) a peu de chances de respecter les règles définies par la loi. En effet, comment demander la suppression d'une de nos conversations téléphoniques sans savoir consciemment qu'elle a été enregistrée. De toute manière, les appelants n'ont pas le choix si ils n'ont que ce numéro de téléphone, donc le droit à l'oubli ne peut pas être respecté. Si de telles quantités de données audio téléphoniques sont enregistrées, c'est aussi pour effectuer des recherches sur la reconnaissance automatique de satisfaction (données centre d'appel du projet VoxFactory par exemple[6]), des émotions, du genre, des identités, etc. dans le but d'améliorer les sevices rendus, mais aussi d'optimiser leur efficacité.

5. http ://www.cnrs.fr/fr/organisme/ethique/comets/docs/ethique_et_expertise.pdf
6. http ://www.capdigital.com/vox-factory/

Il est important qu'aujourd'hui la conception même des systèmes automatiques prennent en compte l'aspect éthique afin d'intégrer dès le début certaines possibilités techniques comme supprimer une information personnel, effacer une conversation, etc.

7.2.3 Ethique des robots

Donner au robots la capacité de reconnaître des émotions humaines ainsi que les signaux sociaux émis lors d'interactions, leur donner également la possibilité d'y apporter une réponse intelligente voire émotionnelle est un objectif de recherche très ambitieux qui peut être perçu comme intrusif et dérangeant.

7.2.3.1 Un robot acceptable

Une grande partie des questions éthiques sur les robots concerne plutôt l'aspect du robot. Les robots utilisés dans les domaines de l'industrie ne sont pas concernés par cette question puisqu'ils ne sont a priori pas doté de composante cognitive. Ce n'est pas le cas des robots à usage domestique, les robots d'assistance ou les robots joueurs. A côté des robots matérialisés anthropomorphes, se sont développés les agents virtuels, intelligents et conversationnels. Ces agents servent d'avatars dans le monde électronique, de compagnons artificiels ou de prothèses permettant d'augmenter les capacités d'un être humain (envoyer son agent à une conférence au Japon, alors que nous sommes physiquement au laboratoire). L'éthique des robots est également valable pour ces agents virtuels.

Les robots anthropomorphes sont conçus pour ressembler aux humains tant sur le plan physique que cognitif. Dans cette course à la ressemblance est apparu un phénomène paradoxal : au-delà d'un certain niveau de similitude avec l'être humain, le robot n'est plus considéré comme une machine merveilleuse bien qu'imparfaite mais comme un humain imparfait et monstrueux. Ce phénomène est appelé "vallée dérangeante" [Mori 70] et doit être pris en compte lors de la conception des robots.

Selon une étude suisse menée par Ray [Ray et al. 08], il semblerait que les robots soient vus négativement uniquement par 4% des répondants. Les personnes interviewés sont majoritairement des actifs (40-65ans) ayant une éducation de niveau second degré. L'article souligne le fait que les robots peuvent provoquer des craintes et des peurs (par exemple, peur de perdre son travail ou l'autonomie du robot), qui ne sont pas forcément considérées comme des aspects négatifs.

7.2.3.2 Les robots sociaux

Les robots sociaux sont de plus en plus présents dans nos sociétés occidentales. Ces machines sont généralement dotées d'intelligence sociale et affective qui leur permettent de s'adapter plus facilement aux situations les plus diverses. Les signaux sociaux, qu'ils soient émotionnels, interactionnels, informatifs ou communicatifs sont consitués d'expressions majoritairement inconscientes. Une machine qui peut reconnaître l'un ou l'autre de ces signaux, entre d'une certaine manière dans l'inconscient des êtres humains. C'est cet aspect qui peut être dérangeant et apporter une méfiance vis à vis des robots sociaux. Les machines dotées d'intelligence sociales ne sont pas uniquement des robots à proprement parlé, mais vise des applications très variées : téléphonie mobile, surveillance, jeux vidéo, assistance aux malades ou aux personnes âgées.

Parmis les premiers robots compagnons, nous pouvons citer le robot-chien de Sony Aibo ou le jouet Furby. Ces robots ne sont cependant pas suffisamment performants pour interagir affectivement et socialement avec leur propriétaire. Les robots d'assistance commencent à se développer et ouvrent des pistes de recherche très ambitieuses.

7.2.3.3 Les robots d'assistance

En France en particulier, les robots d'assistance se développent de manière importante. Parmi les différents projets, Romeo (Aldebaran), Armen[7] ou Robadom[8]. Ces projets visent à imaginer des robots ayant une bonne capacité motrice (se déplacer dans un environnement domestique connu, reconnaître et attraper un objet, servir de support pour aider une personne à se déplacer) mais également des capacités perceptives et cognitives (adapter son comportement gestuels et parole au comportement du propriétaire).

Une des principales questions soulevées par le projet Romeo qui vise à créer des robots d'assistance est, pourquoi a-t-on besoin de robots pour aider les personnes en perte d'autonomie dans leur quotidien ? Un élément de réponse nous vient du Japon, où la population est très largement vieillissante. Le nombre de personnes dépendantes augmente alors que le nombre relatif d'actifs diminue, les robots sont alors vus comme une solution pour palier à ce phénomène démographique.

Effectivement, il me semble qu'un robot peut apporter une aide technique sur un certain nombre de tâches de la vie quotidienne, rappeler un agenda, lire des étiquettes pour les mal-voyants, mettre et sortir un plat du four. Pour ces tâches très techniques, un robot social peut être un plus, si il sait s'adapter à l'humeur de son "maître", même si cette fonctionalité n'est pas nécessaire. Certaines personnes du corps soignant peuvent également témoigner de la pénibilité de certaines tâches, comme prendre la température, donner des médicaments, et voudrait pouvoir se concentrer sur des aspects plus humains.

C'est justement ces aspects humains qui nous intéressent ici. Dans le cas de personnes seules, il me semble qu'un robot même social ne peut suffir à une assistance de qualité. Il existe déjà des robots compagnons dont l'influence positive sur le moral de personnes âgées a été démontrée (comme le robot Paro), mais ces robots seuls ne suffisent pas. Certains objets techniques commes les téléphones portables (ou i-phones, ou ordinateurs portables, ou même télévision) sont déjà considérés plus ou moins comme des compagnons dans ce sens ou ils occupent leur propriétaire. La société peut confier aux robots la prise en charge de la transmission de données techniques (température, poids, message du "maître"), mais plus difficilement celle de prendre une décision concernant le diagnostic.

Prenons le cas d'un robot doté d'un système de reconnaissance des émotions, nous avons vu tout au long de cette thèse que ce type de système peut faire des erreurs de reconnaissance suivant le contexte. Peut-on lui confier la tâche de faire un appel d'urgence en fonction du nombre d'émotions négatives détectées par exemple ? Une des meilleures solutions seraient de laisser le maître choisir les fonctionnalités possibles du robot avec une adaptation très précise des différents modèles. Les constructeurs ont la responsabilité de créer ou non ces fonctionnalités qui peuvent être préjudiciable au propriétaire du robot mais également à ceux qui l'entourent.

Avec l'arrivée des robots dans le quotidien des personnes, se pose également la question

7. http ://projet_armen.byethost4.com/
8. http ://www.robadom.vermeil.org/

de qui ils vont remplacer. Cette problématique renvoie à l'organisation de la société de manière très globale. Il manque aujourd'hui de personnel soignant, ce n'est donc pas en introduisant des robots que l'on risque de les mettre au chômage.

Même si dans une interview au journal "le Point", le docteur Rumeau préfère vendre des services plutôt que des machines [9], je dirais qu'il faut considérer les robots comme des outils techniques et des supports, pas comme des compagnons afin de limiter leur pouvoir de décision et de diagnostic.

7.2.4 Conclusion

Les réflexions éthiques sont aujourd'hui de plus en plus appréciables tant par la communauté scientifique elle-même que par la société civile en général. L'éthique est apparue dans le domaine de la médecine (expérimentations sur des êtres humains, pharmacie, etc.) et reste aujourd'hui majoritairement portée par ce domaine. Cependant un nombre croissant de chercheurs publient leurs réflexions sur leurs sujets de recherche. C'est le cas dans les domaines de recherche touchant de près ou de loin la robotique. Il me semble fondamental aujourd'hui qu'un chercheur réfléchisse à l'impact de ses découvertes sur la société, c'est pourquoi j'apporte une grande importance à ce chapitre.

Les recherches sur les interactions humain-robot doivent, selon moi, se pencher sur un certain nombre de questions. Une des plus importantes concerne le consentement des volontaires lors de la collecte de données en situation d'interaction humain-robot. Les participants doivent être conscients des risques qu'ils peuvent encourir mais aussi des objectifs scientifiques de l'expérience. Les applications qui découlent de ses recherches peuvent être des systèmes de reconnaissance automatique de traits humains. Ces systèmes doivent être fournis avec des performances propres et des conditions d'utilisation. Il est important pour la société que ces applications ne mettent pas leurs utilisateurs dans des situations difficiles. Enfin, l'éthique des robots est aujourd'hui un sujet de préoccupation à part entière, notamment en ce qui concerne la responsabilité pénale du robot. Qui en est responsable ? le constructeur, l'utilisateur, le vendeur ? Un autre point nécessite de prendre le temps de réfléchir, est-ce que l'apparition de robots sociaux et particulièrement de robots d'assistance est bénéfique pour la société en générale ? La société japonaise utilise déjà des robots sociaux à plusieurs niveaux de la société.

Les travaux de thèse que j'ai réalisés s'inscrivent parmi ceux autour des interactions humain-robot. Alors qu'il faut encore me convaincre de l'utilisation de robots trop intelligents, je suis convaincue de l'utilité des robots en tant qu'outils techniques permettant d'aider une personne pour une tâche précise. Il me semble que les découvertes scientifiques publiées dans ce domaine sont riches et aident à mieux comprendre les relations entre les humains. L'utilisation d'une machine (par exemple un robot) dans ces recherches permet de mettre en place des protocoles stables pour l'étude des interactions.

9. http ://www.lepoint.fr/futurapolis/sante/docteur-pierre-rumeau-gcs-telesante-midi-pyrenees-les-robots-compagnons-de-sante-ne-sont-pas-des-machines-mais-un-service-02-11-2011-1391885_431.php

Chapitre 8

Conclusion générale

8.1 Conclusions

Dans le domaine de la reconnaissance des émotions, la tendance est plutôt à rechercher des données réalistes plutôt que actées dans un contexte de la laboratoire. Nos travaux se situent en plein dans cette problématique. L'utilisation de données réalistes est essentielle pour construire des modèles crédibles et utilisables par la suite dans des interactions réelles, cependant la complexité de ces données et l'infinie possibilité des situations dans lesquelles elles peuvent être collectées, rend la tâche extrêmement difficile.

Dans la première partie, nous avons vu que l'annotation de données réalistes étaient une tâche difficile : les scores d'agréement entre annotateurs sont souvent moins bons que lors de l'annotation de données actées prototypiques. Dans la seconde partie, nous proposons plusieurs méthodes permettant de mettre en évidence les différentes variabilités présentent lors d'une interaction parlée d'un point purement acoustique. Nous avons cherché à mettre en évidence des descripteurs les moins variables suivant l'environnement (acoustique de la salle principalement), les locuteurs (enfants, adultes, personnes âgées), et suivant la tâche (émotions actées prototypiques, induites ou spontanées). A partir des corpus collectés et d'un ensemble de descripteurs acoustiques sélectionnés, nous mettons en relief pour un descripteur donné, les variations liées au locuteur et aux émotions. Ce type de méthode permet de comparer les différents corpus et les émotions exprimées. Les résultats sur l'identification du genre et des locuteurs dans un contexte émotionnel montre une dégradation des performances sur les émotions fortes (colère, joie forte) uniquement. Un locuteur exprimant une émotion ne sera pas identifié comme un autre locuteur mais comme inconnu. L'identification du genre est également très sensible à l'environnement dans lequel elle est réalisée. Une forte réverbération dans la salle ne permet plus une identification correcte du genre. Les systèmes de reconnaissance automatique des émotions sont également très sensibles à une grande variabilité des voix des locuteurs, à l'environnement acoustique mais également à la tâche. Les performances obtenues sur des corpus actés prototypiques sont nettement meilleurs que sur des corpus réalistes. Ces performances doivent être relativisées par rapport à la difficulté de la tâche d'annotation qui conditionne les modèles.

8.2 Contributions

Mes contributions au domaine de recherche sur les interactions humain-machine sont principalement de trois types : sur les corpus eux-mêmes, sur les indices acoustiques et sur le pouvoir de généralisation des modèles obtenus par apprentissage par des expériences cross-corpus.

8.2.1 Les corpus

Nous avons déjà abordé la notion de spécificité d'un corpus, c'est une notion très intéressante et cependant peu étudiée aujourd'hui. L'idée générale que nous avons suivie pour l'étude des corpus est de leur donner une signature (acoustique, linguistique, ou autre). Cette signature permettrait de caractériser chaque corpus afin de pouvoir évaluer les complémentarités, les ressemblances et les divergences. L'utilisation de plusieurs corpus qui diffèrent suivant la tâche, l'environnement, le type de locuteur ou le type d'émotions est un outil essentiel à ce travail. Dans ce cadre, ***nous avons participé à la collection et l'annotation de quatre corpus de données naturelles en interaction humain-robot***. Ces données sont originales et essentielles pour l'étude des variabilités dans un contexte d'interaction par ***la diversité des locuteurs enregistrés*** (enfants, personnes âgées, mal-voyants), ***le choix des tâches*** (spontané, induit, scénarii de jeu, de situation d'urgence et de vie quotidienne) et ***l'utilisation de plusieurs environnements acoustiques*** (réverbération, studio, laboratoire). Nous avons également participé à l'enregistrement d'***un corpus de stress*** lors d'une prise de parole en public. Ce corpus est précurseur pour l'étude du stress par l'étendue des modes collectés (audio, vidéo, physiologique). L'ensemble de ces corpus a été collecté dans le cadre d'un travail d'équipe.

8.2.2 Les indices acoustiques

Les ensembles d'indices habituellement utilisés par la communauté sont relativement pauvres en descripteurs de timbre et de rythme, nous avons donc proposé d'intégrer des indices acoustiques venant d'autres disciplines (transformation de voix, synthèse, analyse de sigaux musicaux principalement). Mais nous proposons également des ***indices de timbre et de rythme nouveaux et intéressants*** principalement pour la reconnaissance des émotions, du stress ou de la personnalité. Nous proposons également de définir le rythme comme une superposition de différentes structures suivant le style de parole, l'humeur, la personnalité ou l'émotion. Cette définition complexe du rythme n'est pas sans rappeler celle de la musique, où tempi, phrases, mesures, motifs sont autant de structures intermêlées ayant chacune son sens propre. Les indices de rythme proposés doivent encore faire l'objet d'évaluation sur des données plus variés afin d'établir leur robustesse pour la reconnaissance d'indices paralinguistiques.

Une autre aspect important de mes travaux de recherche est l'analyse de la robustesse d'indices acoustiques face aux différentes variabilités. Déterminer quels indices sont essentiels et quels sont ceux à éviter est un défi complexe puisqu'il nécessite de travailler sur un grand nombre de variabilités et donc d'avoir à sa disposition des corpus riches et variés. A partir de nombreuses expériences réalisées sur plusieurs corpus de type très différents (actés, spontanés, en interaction avec un humain, avec un robot, dans des centre d'appels,

etc.) nous avons établi ***une liste "noire" d'indices peu robustes aux variabilités*** de ces corpus.

Les descripteurs acoustiques peuvent également servir à la signature de corpus. A partir des indices nous proposons ***une mesure de spontanéité à partir de la colère et une mesure de variabilité permettant de positionner un corpus dans un espace émotion/locuteur***.

8.2.3 Expériences cross-corpus : pouvoir de généralisation des modèles

Les modèles utilisés pour la caractérisation du locuteur sont habituellement appris et testés sur de la parole neutre. D'après nos résultats, ***le pouvoir de généralisation des modèles appris sur du neutre est plus important que celui des modèles appris sur de la parole émotionnelle***. Ce point est très intéressant puisque en pratique la parole neutre est bien plus abondante que la parole émotionnelle. Cependant certaines émotions fortes comme la colère ou la joie ne permettrons pas que le locuteur soit correctement caractérisé. La réverbération affectant fortement les indices acoustiques traditionnellement utilisé en identification du locuteur, l'environnement acoustique peut avoir des conséquences importantes sur l'identification des locuteurs.

Des expériences cross-corpus montrent qu'***un modèle appris sur des données prototypiques apporte de meilleurs résultats de reconnaissance automatique des émotions que lorsqu'il est appris sur des données spécifiques même en normalisant les descripteurs acoustiques***. La construction d'un ensemble de descripteurs acoustiques pertinents et non-redondants est fondamental pour une bonne reconnaissance d'indices paralinguistiques (émotion, personalité, stress).

Nous pouvons choisir des modèles différents selon les situations (modèles d'émotion par locuteur, modèles d'environnement acoustique, modèles correspondant à des tâches spécifiques) ce qui impliquerait une quantité effarante de modèles différents. Une autre piste serait d'avoir quelques modèles de référence qu'il faudrait adapter en fonction de la situation réelle. Enfin la dernière piste consiste à normaliser les données, les modèles, ou les descripteurs. Ces pistes sont évidemment complémentaires.

8.3 Perspectives

8.3.1 Perspectives à court-terme

Signatures de corpus Les mesures de spontanéité et de variabilité relative émotion/locuteur semblent intéressantes pour la classification et la caractérisation acoustique de corpus émotionnels. Un point nous semble très pertinent à approfondir, c'est la position dans l'espace des corpus sur les deux axes locuteurs et émotions. Nous pourrions également définir des axes correspondant à l'activation ou l'articulation sur lesquels pourraient se positionner l'ensemble des corpus de la communauté.

Les descripteurs acoustiques Concernant les descripteurs acoustiques, plusieurs améliorations peuvent être faites à partir de mes travaux. Les descripteurs de rythme, notamment ceux décrivant la répartition statistique des durées des parties voisées, peuvent être

optimisés (choix des seuils, choix des fenêtres temporelles). Il faudrait également définir de nouvelles structures rythmiques correspondant à différents états affectifs ou style de parole (manière de parler, humeur, personnalité ou émotion). Cette définition du rythme en plusieurs couches superposées pourrait permettre de comprendre plus en détail sa perception et permettre de structurer le quasi chaos.

Nous avons également proposé une liste "noire" de descripteurs très sensibles aux variabilités d'une interaction humain-robot. Cette liste doit être validée sur de plus nombreux corpus (par exemple le corpus Comparse), mais également par des systèmes de reconnaissance automatique des émotions et du locuteur intégrant les ensembles de descripteurs proposés.

8.3.2 Perspectives pour la suite du projet ROMEO : ROMEO2

Pré-traitement du signal entrant Par un système de filtrage adapté aux conditions réelles (bruit de fond, déréverbération, etc.) le calcul de l'ensemble des descripteurs acoustiques pourraient être amélioré. Une normalisation de l'amplitude du signal d'entrée est également un point à traiter. La segmentation du flux audio peut être encore plus performante avec des systèmes de détection d'activité vocale précis et couplés avec une détection du locuteur en temps réel (diarization). Cela permettrait notamment de pouvoir traiter les superpositions de parole ainsi que les changements rapides de locuteurs.

Développement de l'outil SysRELL L'outil SysRELL est aujourd'hui utilisable et permet de faire des tests dans des conditions réelles. Cependant un certain nombre d'améliorations peuvent lui être apportées. Notamment, la poursuite des tests de sélection des descripteurs pourrait permettre d'optimiser l'ensemble des 144 descripteurs existants aujourd'hui dans l'outil. Une évaluation générale de l'outil doit être mise au point, avec une centaine de locuteurs pour estimer la robustesse de l'outil aux locuteurs, aux conditions d'enregistrement (chaîne de mesure, réverbération, etc.) et pour permettre de choisir le type de normalisation souhaitable.

8.3.3 Perspectives à long-terme

Ces travaux de recherche décrivent de manière exhaustive la difficulté d'imbriquer locuteur et émotion dans des systèmes de reconnaissance automatique, ainsi que la très grande variabilité des descripteurs acoustiques lors d'une interaction humain-robot. Afin de rendre les systèmes de reconnaissance automatique plus robustes à l'ensemble de ces variabilités qui affectent le signal sonore, nous envisageons plusieurs pistes de recherches :

Adaptation des modèles L'utilisation de modèles différents en fonction des situations nécessite de détecter automatiquement ces différentes situations. On pourrait alors choisir le modèle en fonction de la situation. Cette méthode a le désavantage de nécessiter un nombre très important de modèles, et donc de collecter un grand nombre de bases de données représentatives des situations. Une autre méthode pourrait être d'adapter un modèle générique en fonction de la situation présente. Cette méthode a déjà fait ses preuves dans l'adaptation des modèles de locuteurs. Cependant, certaines situations, même avec adaptation, ne pourront pas permettre d'obtenir de bonnes performances, avec réverbération

par exemple. L'adaption devra pouvoir se faire au locuteur et à l'environnement acoustique. L'adaptation dynamique des modèles émotionnels à la tâche est plus complexe, mais reste une vraie piste de recherche.

Enrichissement des modèles et multimodalité Dans nos travaux, seul le canal audio a été pris en compte. De nombreux travaux de recherche portent sur d'autres modalités (linguistique, vidéo, physiologique) et sur des méthodes de fusion des indices. Il est évident que suivant les situations et en fonction des indices paralinguistiques que l'on cherche à reconnaître, l'audio n'est pas toujours suffisant.

Nous avons essentiellement travaillé avec des modèles GMM ou SVM, un de leurs inconvénients est qu'ils ne prennent pas en compte l'aspect dynamique de la parole émotionnelle. L'utilisation de HMM par exemple, ou bien la fusion d'indices à des niveaux temporels différents pourraient prendre en compte cette dimension. Pour l'instant l'utilisation des HMM seuls n'a pas apporté d'améliorations significatives.

8.4 Discussions

8.4.1 "Open-microphone" ou "effet loupe"

Les recherches "open-microphone" prennent en compte l'ensemble des signaux qui ont été captés par un microphone lors de l'enregistrement de données (bruit de fond, réverbération, bruit de bouche, etc.). A l'opposé, les recherches "effet loupe" sélectionnent des données extrêmement contrôlées. Ces deux approches, bien que très différentes, sont complémentaires. La première permet de s'approcher d'une application réelle et de prendre en considération des aspects inattendus qui peuvent perturber les résultats théoriques de manière significative. Cette approche fait ressortir des tendances alors que l'approche "loupe" permet d'obtenir des résultats fiables sur des phénomènes précis. Les travaux présentés ici reprennent beaucoup de résultats obtenus grâce à des analyses fines afin de les tester sur des conditions plus globales.

8.4.2 Des systèmes capables du meilleur en toutes circonstances ?

D'après les résultats exposés tout au long de ces travaux, il semble quasiment impossible de concevoir des systèmes de reconnaissance automatique d'indices paralinguistiques ayant des performances satisfaisantes dans tout type de conditions. Il est bien plus raisonable de penser que de tels systèmes doivent être utilisés dans des situations contraintes, les contraintes pouvant porter sur les types de locuteurs, le choix de la tâche, l'environnement acoustique ou les émotions susceptibles d'être exprimées. L'ensemble des futurs modèles pour la reconnaissance du locuteur et de son état émotionnel doit être pensé dans cette optique plus contrainte.

Sixième partie

Annexes

Annexe A

Caractéristiques des corpus ROMEO et autres corpus utilisés

Corpus	Lieu d'enregistrement	Age (#Genre)	Données	Durée totale	Réverbération
NAO-HR1	I-room, salle expérimentale du LIMSI (Orsay)	8-13 ans (6G, 6F)	spontané / acté	31 min	faible
IDV-HH	Appartement de la résidence St Louis (Paris)	23-79 ans (11H, 17F)	spontané / acté	1h 23 min	très élevée
NAO-HR2	Centre aéré CESFO (Orsay)	6-10 ans (6G, 6F)	spontané / acté	21 min	élevée
IDV-HR	Appartement témoin, Insitut de la Vision (Paris)	28-80 ans (11H, 11F)	spontané / acté	4h 07 min	faible
JEMO	Bureau, LIMSI	24-50 ans (27H, 35F)	acté	1h 04 min	faible

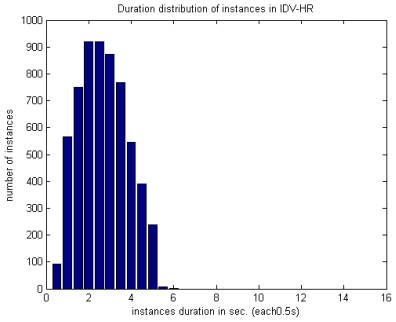
TABLE A.1 – Données caractéristiques des principaux corpus utilisés (H : homme, F : femme ou fille, G : garçon)

macros	%moyen par locuteur	écart-type entre les locuteurs
neutre	60,18	17,54
colère	5,61	3,35
négatif	6,97	7,94
tristesse	5,06	3,24
pos-neg	5,05	4,45
peur	3,05	2,48
joie	14,07	7,58

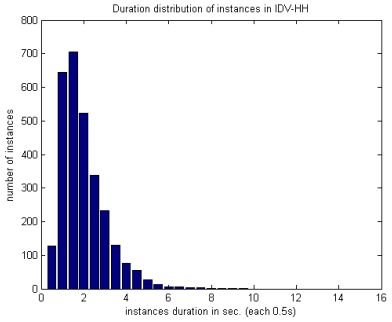
TABLE A.2 – Répartition des segments émotionnels du corpus IDV-HR suivant les locuteurs

macros	%moyen par locuteur	écart-type entre les locuteurs
neutre	45,87	15,78
colère	6,04	4,28
négatif	0,43	1,00
tristesse	2,74	3,03
pos-neg	6,95	5,07
peur	8,93	5,66
joie	13,53	8,01

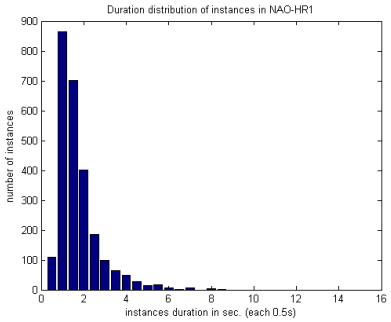
TABLE A.3 – Répartition des segments émotionnels du corpus IDV-HH suivant les locuteurs



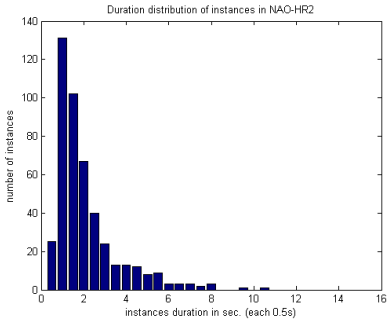
a) moy : 2,45s ; min : 0,24s ; max : 5,94s



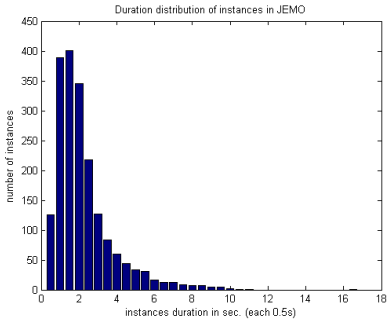
b) moy : 1,73s ; min : 0,18s ; max : 9,04s



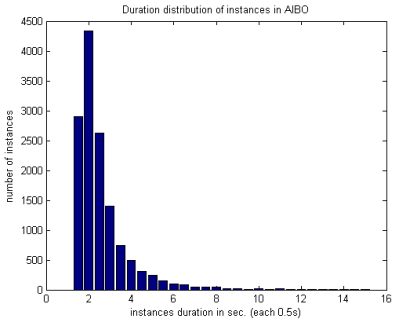
c) moy : 1,46s ; min : 0,23s ; max : 8,38s



d) moy : 1,83s ; min : 0,27s ; max : 10,47s



e) moy : 1,98s ; min : 0,14s ; max : 16,03s



f) moy : 2,34s ; min : 1,00s ; max : 38,12s

FIGURE A.1 – Répartition des durées dans IDV-HR (a), IDV-HH (b), NAO-HR1 (c), NAO-HR2 (d), JEMO (e), AIBO (f)

macros	%moyen par locuteur	écart-type entre les locuteurs
neutre	44,05	13,96
colère	3,02	2,18
négatif	0,09	0,28
tristesse	4,43	3,00
pos-neg	7,83	5,07
peur	3,37	3,54
joie	21,27	6,92

TABLE A.4 – Répartition des segments émotionnels du corpus NAO-HR1 suivant les locuteurs

macros	%moyen par locuteur	écart-type entre les locuteurs
neutre	45,87	15,78
colère	6,04	4,28
négatif	0,43	1,00
tristesse	2,74	3,03
pos-neg	6,95	5,07
peur	8,93	5,66
joie	13,53	8,01

TABLE A.5 – Répartition des segments émotionnels du corpus IDV-HH suivant les locuteurs

	Toutes classes émotionnelles	Macros classes	Valence	Activation
IDV-HR	0,22 (#28) locuteurs 1 à 8	0,28 (#7)	0,32 (#3)	0,39 (#3)
IDV-HH	0,39 (#28)	0,56 (#7)	0,71 (#3)	0,34 (#3)
NAO-HR1	0,39 (#24)	0,48 (#7)	0,65 (#3)	0,59 (#3)

TABLE A.6 – Scores d'agréement (kappa)

Annexe B

Classification des descripteurs et liste noire

	Moyenne	Ecart-type
Energie-voisé	104,2	27,1
Energie-non voisé	90,5	20,0
Energie-tout	*114,8*	*34,1*
F0-voisé	103,0	28,6
Formants	103,0	9,2
Formants-BF	*114,0*	*31,8*
Formants-diff	*116,9*	*12,6*
Formants-art	79,9	24,7
Spectre-voisé	83,1	34,6
Spectre-non voisé	106,2	9,0
Cepstre-voisé	91,6	12,5
Cepstre-voisé-delta	98,3	22,0
Cepstre-non voisé	100,5	16,8
Cepstre-non voisé-delta	109,5	18,5
Cepstre-tout	90,3	13,7
Cepstre-tout-delta	99,8	15,1
Bark-voisé	106,2	6,1
Bark-non voisé	*111,7*	*13,0*
Harmo	90,7	19,2
Rythme	*115,0*	*7,0*
Qualité vocale	97,4	42,6

TABLE B.1 – Rang moyen par famille de descripteurs sur cinq corpus spontanés suivant le rapport de variation locuteur/émotions sans normalisation au locuteur, *black list*

	Moyenne	Ecart-type
Energie-voisé	74,4	35,5
Energie-non voisé	99,9	35,0
Energie-tout	*125,2*	*20,0*
F0-voisé	*116,0*	*25,9*
Formants	102,9	15,2
Formants-BF	79,0	22,8
Formants-diff	107,3	10,1
Formants-art	*130,1*	*55,1*
Spectre-voisé	85,0	23,0
Spectre-non voisé	102,7	15,8
Cepstre-voisé	94,1	12,4
Cepstre-voisé-delta	100,8	13,2
Cepstre-non voisé	*119,3*	*15,2*
Cepstre-non voisé-delta	*119,9*	*16,0*
Cepstre-tout	107,8	18,9
Cepstre-tout-delta	99,3	25,2
Bark-voisé	*111,9*	*16,7*
Bark-non voisé	107,2	8,0
Harmo	76,1	9,4
Rythme	107,3	18,5
Qualité vocale	84,2	33,8

TABLE B.2 – Rang moyen par famille de descripteurs sur cinq corpus spontanés suivant le rapport de variation locuteur/émotions avec normalisation au locuteur, *black list*.

	Moyenne	Ecart-type
Energie-voisé	39,8	19,3
Energie-non voisé	89,3	49,6
Energie-tout	86,2	43,3
F0-voisé	101,5	36,1
Formants	*128,9*	*26,2*
Formants-BF	102,6	46,1
Formants-diff	99,2	60,2
Formants-art	*130,2*	*22,8*
Spectre-voisé	92,5	36,5
Spectre-non voisé	75,7	24,4
Cepstre-voisé	85,5	20,3
Cepstre-voisé-delta	89,1	45,5
Cepstre-non voisé	75,4	21,0
Cepstre-non voisé-delta	*118,4*	*40,0*
Cepstre-tout	*118,2*	*20,6*
Cepstre-tout-delta	*162,8*	*43,3*
Bark-voisé	*113,2*	*21,5*
Bark-non voisé	106,7	28,1
Harmo	87,7	32,3
Rythme	*117,0*	*18,0*
Qualité vocale	95,5	43,5

TABLE B.3 – Rang moyen par famille de descripteurs sur cinq corpus spontanés suivant la sélection automatique d'indices, sans normalisation au locuteur, *black list*.

Annexe C

Performances en reconnaissance automatique des émotions

C.1 Nombre d'instances utilisées pour chaque corpus

Nombre d'instances utilisés dans les expériences du chapitre 5 pour chacun des corpus étudiés.

	JEMO	NAO-HR1	IDV-HR	IDV-HH	AIBO-O	AIBO-M
Neutre	369	60	156	330		
Colère	297	33	102	306		
Joie	354	60	132	312		
Tristesse	297	42	66	276		
Positif	441	240	777	552	510	375
Neutre	495	225	858	528	501	399
Négatif	510	213	813	582	474	354
Actif		282	330	480		
Neutre		288	330	504		
Passif		264	240	522		

TABLE C.1 – Nombre d'instances utilisées pour la cross-validation et les expériences cross-corpus

C.2 Performances cross-corpus

Taux de reconnaissance de la valence (UAR) obtenus en croisant les corpus Romeo et AIBO. La classification est fondée sur le set de descripteurs OE192, en utilisant Weka (fonction SMO, noyau RBF, paramètres $c = 2,0$ et $\gamma = 0,03125$).

Sens de lecture : apprentissage JEMO, test NAO-HR1 $UAR = 46,6 \pm 3,7\%$.

	JEMO	NAO-HR1	IDV-HR	IDV-HH	AIBO-O	AIBO-M	Moyenne
JEMO	X	46,3 (3,7)	54,1 (2,0)	53,4 (2,3)	37,3 (2,5)	30,0 (2,7)	44,3
NAO-HR1	27,5 (2,3)	X	23,8 (1,7)	46,3 (2,4)	10,5 (1,6)	13,8 (2,0)	24,4
IDV-HR	51,6 (2,6)	55,1 (3,7)	X	59,5 (2,3)	50,3 (2,5)	44,6 (2,9)	52,2
IDV-HH	45,8 (2,6)	55,6 (3,7)	30,8 (1,8)	X	48,8 (2,5)	46,7 (2,9)	45,5
AIBO-O	30,7 (2,4)	12,8 (2,5)	27,8 (1,8)	21,2 (2,0)	X	58,9 (2,9)	30,3
AIBO-M	43,4 (2,6)	51,7 (3,8)	19,1 (1,6)	26,8 (2,1)	66,8 (2,4)	X	41,6

TABLE C.2 – Performances en cross-corpus sur la valence sans normalisation

	JEMO	NAO-HR1	IDV-HR	IDV-HH	AIBO-O	AIBO-M	Moyenne
JEMO	X	56,8 (3,7)	59,4 (1,9)	68,1 (2,2)	56,0 (2,5)	54,0 (2,9)	58,9
NAO-HR1	45,2 (2,6)	X	41,4 (1,9)	47,7 (2,4)	25,6 (2,2)	30,6 (2,7)	38,1
IDV-HR	62,0 (2,5)	56,0 (3,7)	X	66,5 (2,3)	58,8 (2,5)	60,4 (2,8)	60,7
IDV-HH	53,8 (2,6)	51,6 (3,8)	46,5 (2,0)	X	44,7 (2,5)	44,4 (2,9)	48,2
AIBO-O	33,5 (2,4)	23,6 (3,2)	28,3 (1,8)	34,3 (2,3)	X	59,4 (2,9)	35,8
AIBO-M	40,1 (2,5)	44,9 (3,7)	30,5 (1,8)	36,4 (2,3)	66,0 (2,4)	X	43,6

TABLE C.3 – Performances en cross-corpus sur la valence avec normalisation au corpus

	JEMO	NAO-HR1	IDV-HR	IDV-HH	AIBO-O	AIBO-M	Moyenne
JEMO	X	58,7 (3,7)	56,0 (2,0)	69,9 (2,2)	54,2 (2,5)	54,4 (2,9)	58,7
NAO-HR1	45,8 (2,6)	X	32,0 (1,8)	49,7 (2,4)	27,8 (2,2)	31,1 (2,7)	37,3
IDV-HR	60,7 (2,5)	49,9 (3,8)	X	63,5 (2,3)	62,8 (2,5)	59,7 (2,9)	59,3
IDV-HH	52,9 (2,6)	52,4 (3,8)	39,6 (1,9)	X	44,3 (2,5)	37,2 (2,8)	45,3
AIBO-O	38,2 (2,5)	37,3 (3,6)	29,8 (1,8)	36,3 (2,3)	X	61,9 (2,8)	40,7
AIBO-M	40,6 (2,5)	43,2 (3,7)	30,6 (1,8)	36,8 (2,3)	62,8 (2,5)	X	42,8

TABLE C.4 – Performances en cross-corpus sur la valence avec normalisation au locuteur

Abréviations utilisées pour la bibliographie

Eurospeech : European Conference on Speech Communication and Technology

LREC : International Conference on Language Resources and Evaluation

JEP : Journées d'Etudes sur la Parole

ICASSP : IEEE International Conference on Acoustics, Speech and Signal Processing

Interspeech : International Conference of the Speech Communication Association (ISCA)

ICSLP : International Conference on Speech and Language Processing (ISCA)

ACII : IEEE International Conference on Affective Computing and Intelligent Interactions

ICME : IEEE International Conference on Mulimedia

Bibliographie

[Abrilian 07] Abrilian, S. *Représentation de Comportements Emotionnels Multimodaux Spontanés : Perception, Annotation et Synthèse.* Thèse de doctorat, Université Paris-Sud 11, 2007.

[Alku et al. 02] Alku, P., Backstrom, T. and Vilkman, E. *Normalized amplitude quotient for parametrization of the glottal flow.* Journal of the Acoustical Society of America, vol. 112 (2), pages 701–710, 2002.

[Arunachalam et al. 01] Arunachalam, S., Gould, D., Anderson, E., Byrd, D. and Narayanan, S. *Politeness and frustration language in child-machine interactions.* In Eurospeech, Aalborg, Denmark, 2001.

[Audibert et al. 06] Audibert, N., Aubergé, V. and Rilliard, A. *Synthèse vocale des émotions, donner la parole émue à C-Clone.* In Workshop francophone sur les Agents Conversationels Animés (WACA06), numéro 2, pages 27–35, Toulouse, France, 2006.

[Audibert et al. 08] Audibert, N., Aubergé, V. and Rilliard, A. *Emotions actées vs. spontanées : variabilité des compétences perceptives.* In JEP, Avignon, France, 2008.

[Bachorowski et al. 01] Bachorowski, J.-A., Smoski, M. J. and Owren, M. J. *The acoustic features of human laughter.* Journal of the Acoustical Society of America, vol. 110 (3), pages 1581–1597, 2001.

[Bagby et al. 94a] Bagby, R., Parker, J. and Taylor, G. *The Twenty-Item Toronto Alexithymia Scale – I. Item selection and cross-validation of the factor structure.* Journal of Psychosomatic Research, vol. 38, pages 33–40, 1994.

[Bagby et al. 94b] Bagby, R., Taylor, G. and Parker, J. *The Twenty-Item Toronto Alexithymia Scale – II. Convergent, discriminant and concurrent validity.* Journal of Psychosomatic Research, vol. 38, pages 33–40, 1994.

[Barbosa 07] Barbosa, P. A. *From syntax to acoustic duration : A dynamical model of speech rhythm production.* Speech Communication, vol. 49, pages 725–742, 2007.

[Barras and Gauvain 03] Barras, C. and Gauvain, J.-L. *Feature and score normalization for speaker verification of cellular data.* In ICASSP, volume II, pages 49–52, Hong-Kong, China, 2003.

[Barras et al. 00] Barras, C., Geoffrois, E., Wu, Z. and Liberman, M. *Transcriber : development and use of a tool assisting speech corpora production.* Speech Communication, vol. 33(1), pages 5–22, 2000.

[Barras et al. 07] Barras, C., Zhu, X., Gauvain, J.-L. and Lamel, L. *The CLEAR'06 LIMSI acoustic speaker identification system for CHIL seminars.* Lecture notes in Computer Science, vol. 4122, pages 233–240, 2007.

[Batliner et al. 04] Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M. and Wong, M. *"You stupid tin box" - children interacting with the AIBO robot : A cross-linguistic emotional speech corpus.* In LREC, pages 171–174, Lisbon, Portugal, 2004.

[Batliner et al. 06] Batliner, A., Steidl, S., , Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., loic Kessous and Aharonson, V. *CEICES : Combining Efforts for Improving automatic Classification of Emotional user States : a « forced co-operation » initiative.* In Language and Technologies Conference, pages 240–245, Slovenia, 2006.

[Batliner et al. 10] Batliner, A., Seppi, D., Steidl, S. and Schuller, B. *Segmenting into adequate units for automatic recognition of emotion-related episodes : a speech-based approach.* Advances in Human-Computer Interaction, 2010.

[Batliner et al. 11] Batliner, A., Schuller, B., Seppi, D., Steidl, S., Devillers, L., Vidrascu, L., Vogt, T., Aharonson, V. and Amir, N. The automatic recognition of emotions in speech, chapitre Cognitive Technologies, pages 71–99. Springer, 2011.

[Beller and Marty 04] Beller, G. and Marty, A. *TALKAPILLAR : outil d'analyse de corpus oraux.* In Rencontres Jeunes Chercheurs de l'école doctorale 268, Paris, France, 2004.

[Beller et al. 08] Beller, G., Obin, N. and Rodet, X. *Articulation degree as a prosodic dimension of expressive speech.* In Speech Prosody, pages 681–684, Campinas, Brasil, 2008.

[Beller 09] Beller, G. *Synthèse vocale de l'expressivité.* Thèse de doctorat, Université Paris VI - Pierre et Marie Curie, 2009.

[Black and Hunt 96] Black, A. W. and Hunt, A. J. *Generating F0 contours with ToBI labels using linear regression.* In International Conference on Spoken Language Processing, Philadelphia, USA, 1996.

[Black et al. 11] Black, M. P., Katsamanis, A., Baucom, B. R., Lee, C.-C., Lammert, A. C., Christiensen, A., Georgiou, P. G. and Narayanan, S. S. *Toward automating a human behavioral coding system for married couples interaction using acoustic features.* Speech Communication, 2011.

[Bänziger and Scherer 05] Bänziger, T. and Scherer, K. R. *The role of intonation in emotional expressions.* Speech Communication, vol. 46 (3-4), pages 252–267, 2005.

[Bänziger 04] Bänziger, T. *Communication vocale des émotions : perception de léxpression vocale et attributions émotionnelles.* Thèse de doctorat, Faculté de Psychologie, Université de Genève, Suisse, 2004.

[Boë 00] Boë, L.-J. *Forensic voice identification in France.* Speech Communication, vol. 31, pages 205–224, 2000.

[Boersma and Weenink 09] Boersma, P. and Weenink, D. *Praat : doing phonetics by computer (Version 5.1.05) [Computer program].* From http ://www.praat.org/, Retrieved May 1, 2009.

[Boersma 93] Boersma, P. *Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.* Institute of Phonetics Sciences, University of Amsterdam, vol. 17, pages 97–110, 1993.

[Bogert et al. 63] Bogert, B., Healy, M. and Tukey, J. Symposium in time series analysis, chapitre The quefrency analysis of time series for echoes : cepstrum, pseudo-autovariance, cross-cepstrum and spahe cracking, pages 209–243. John Wiley & Sons, 1963.

[Breazeal and Aryananda 02] Breazeal, C. and Aryananda, L. *Recognition of affective communicative intent in Robot-Directed speech.* Autonomous Robots, vol. 12, pages pp. 83–104, 2002.

[Breazeal 99] Breazeal, C. *Robot in Society : Friend or Appliance ?* In Autonomous Agents Workshop on Emotion-Based Agent Architectures, Seattle, WA, USA, 1999.

[Brendel et al. 10] Brendel, M., Zaccarelli, R. and Devillers, L. *Building a system for emotions detection from speech to control an affective avatar.* In LREC, Valetta, Malta, 2010.

[Burkhardt et al. 05] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. and WeissI, B. *A database of german emotional speech.* In Interspeech, pages 1517–1520, Lisbon, Portugal, 2005.

[Burkhardt et al. 10] Burkhardt, F., Eckert, M., Johannsen, W. and Stegmann, J. *A Database of Age and Gender Annotated Telephone Speech.* In LREC, Valetta, Malta, 2010.

[Callejas and Lopez-Cozar 08] Callejas, Z. and Lopez-Cozar, R. *On the Use of Kappa Coefficients to Measure the Reliability of the Annotation of Non-acted Emotions.* In IEEE Workshop on Perception and Interactive Technologies for Speech-Based Systems : Perception in Multimodal Dialogue Systems,, 2008.

[Campbell 04a] Campbell, N. *Acounting for voice quality variation.* In Speech Prosody, pages 217–220, Nara, Japan, 2004.

[Campbell 04b] Campbell, N. *Perception of Affect in Speech - towards an Automatic Processing of Paralinguistic Information in Spoken Conversation.* In International Conference on Spoken Language Processing, Jeju Island, Korea, 2004.

[Campione et al. 98] Campione, E., Flachaire, E., Hirst, D. and Véronis, J. *Evaluation de modèles d'étiquetage automatique de l'intonation.* In Journées d'Etude sur la Parole (JEP), numéro 22, pages 99–102, Martigny, Suisse, 1998.

[Caridakis et al. 10] Caridakis, G., Karpouzis, K., ManolisWallace, Kessous, L. and Amir, N. *Multimodal user's affective state analysis in naturalistic interaction.* Journal of Multimodal User Interfaces, vol. 3, pages 49–66, 2010.

[Castellano et al. 10] Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A. and McOwan, P. *Affect recognition for interactive companions : challenges and design in real world scenarios.* Journal of Multimodal User Interfaces, vol. 3, pages 89–98, 2010.

[Chang and Lin 11] Chang, C.-C. and Lin, C.-J. *LIBSVM : a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, vol. 2 (3), pages 1–27, 2011.

[Chastagnol and Devillers 12] Chastagnol, C. and Devillers, L. *Personality traits detection using a parallelized modified SFFS algorithm.* In Interspeech, Portland, Oregon, USA, 2012.

[Chung 00] Chung, S. *L'expression et la perception de l'émotion extraite de la parole spontanée : évidences du coréen et de l'anglais.* Thèse de doctorat, Institut de Linguistique et Phonétique Générales et Appliquées - Paris 3, 2000.

[Clavel and Richard 10] Clavel, C. and Richard, G. Reconnaissance acoustique des émotions, chapitre 5, pages 11–49. 2010.

[Clavel et al. 07] Clavel, C., Devillers, L., Richard, G., Vidrascu, I. and Ehrette, T. *Abnormal situations detection and analysis through fear-type acoustic manifestations.* In ICASSP, volume IV, pages 21–24, Honolulu, HI, U.S.A., 2007.

[Clavel 07] Clavel, C. *Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales.* Thèse de doctorat, Ecole Nationale Supérieur des Télecommunications, 2007.

[CNRS 08] CNRS. *Sécurité, identification, vie privée, ce que va changer le numérique.* In Journal du CNRS, octobre 2008.

[Cohen 60] Cohen, J. *A coefficient of agreement for nominal scales.* Educational and Psychological Measures, vol. 20, pages 37–46, 1960.

[Cowie et al. 00] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M. and Schröder, M. *Feeltrace : an instrument for recording perceived emotion in real time.* In ISCA Workshop on Speech and Emotion, pages 19–24, Newcastle, Northern Ireland, U.K., 2000.

[Cowie et al. 01] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, G., Votsis, N., Kollias, S., Fellen, W. and Taylor, J. *Emotion recognition in human-computer interaction.* IEEE Signal Processing Magazine, vol. 18 (1), pages 32–80, 2001.

[Cowie et al. 05] Cowie, R., Douglas-Cowie, E. and Cox, C. *Beyond emotion archetypes : databases for emotion modelling using neural networks.* Neural Networks, vol. 18, pages 371–388, 2005.

[Cox 08] Cox, S. *Voice stress analysis stirs controversy and debate.* IEEE Signal Processing Society, Speech and Language Technical Comitte e-newsletter, 2008.

[Craggs and Woods 04] Craggs, R. and Woods, M. *A categorical annotation scheme for emotion in the linguistic content of dialogue.* In Affective Dialogue Systems, Proceedings of a Tutorial and Research Workshop, Kloster Irsee, pages 89–100, 2004.

[d'Alessandro and Mertens 95] d'Alessandro, C. and Mertens, P. *Automatic pitch contour stylization using a model of tonal perception.* Computer Speech and Language, vol. 9, n°3, pages 257–288, 1995.

[Damasio 94] Damasio, A. Descartes' error : Emotion, reason and the human brain. New York : Grosset/Putnam, 1994.

[de Cheveigné and Kawahara 02] de Cheveigné, A. and Kawahara, H. *YIN, a fundamental frequency estimator for speech and music.* Journal of the Acoustical Society of America, vol. 111(4), pages 1917–1930, 2002.

[Degottex et al. 08] Degottex, G., Bianco, E. and Rodet, X. *Usual to particuliar phonatory studied with high-speed videoendoscopy.* In International Conference on Voice Physiology and Biomechanics, pages 19–26, Tampere, Finland, 2008.

[Degottex et al. 10] Degottex, G., Roebel, A. and Rodet, X. *Function of phase-distortion for glottal model estimation.* In ICASSP, Prague, Rep. Tchèque, 2010.

[Degottex 10] Degottex, G. *Glottal source and vocal-tract separation.* Thèse de doctorat, Université Paris VI - Pierre et Marie Curie, 2010.

[Delaborde et al. 10] Delaborde, A., Tahon, M., Barras, C. and Devillers, L. *Affective Links in a Child-Robot Interaction.* In LREC, Valetta, Malta, 2010.

[Dellandréa et al. 03] Dellandréa, E., Makris, P. and Vincent, N. *Wavelets and zipf law for audio signal analysis.* In Signal Processing and its Applications, volume 2, pages 483–486, 2003.

[Dellandrea et al. 02] Dellandrea, E., Makris, P., Boiron, M. and Vincent, N. *A medical acoustic signal analysis method based on Zipf law.* In Digital Signal Processing, volume 2, pages 615–618, Santorini, Greece, 2002.

[Dellandrea et al. 03] Dellandrea, E., Makris, P., Boiron, M. and Vincent, N. *Analyse de signaux sonores par les lois de Zipf et Zipf Inverse.* In GRETSI, Paris, France, 2003.

[Dempster et al. 77] Dempster, A., Laird, N. and Rubin, D. *Maximum likelihood from incomplete fata via the em algorithm.* Journal of the Royal Statistical Society, vol. 39(1), pages 1–38, 1977.

[Devillers and Vidrascu 06] Devillers, L. and Vidrascu, L. *Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs.* In Interspeech, Pittsburgh, PA, USA, 2006.

[Devillers and Vidrascu 07] Devillers, L. and Vidrascu, L. *Positive and negative emotional states behind laughs in spontaneous spoken dialogs.* In Interdisciplinary Workshop on The Phonetics of Laughter, pages 37–40, Saarbrücken, Germany, 2007.

[Devillers et al. 04] Devillers, L., Vasilescu, I. and Vidrascu, L. *Anger and fear detection in recorded conversations.* In Speech Prosody, Nara, Japon, 2004.

[Devillers et al. 05a] Devillers, L., Abrilian, S. and Martin, J.-C. *Representing reali-life emotions in audiovisual data with non-basic emotional patterns and context features.* In ACII, Beijing, China, 2005.

[Devillers et al. 05b] Devillers, L., Vidrascu, L. and Lamel, L. *Challenges in real-life emotion annotation and machine leraning based detection.* Journal of Neural Networks, Special Issue on Emotion and Brain, vol. 18 (4), pages 407–422, 2005.

[Devillers et al. 06] Devillers, L., Cowie, R., Martin, J.-C., Douglas-Cowie, E., Abrilian, S. and McRorie, M. *Real-life emotions in French and English TV video clips : an integrated annotation protocol combining contiuous and discrete approaches.* In LREC, Genoa, Italy, 2006.

[DiCristo 03] DiCristo, A. *De la métrique et du rythme dans la parole ordinaire : l'exemple du français.* Semen, vol. 16, 2003.

[Ding et al. 12] Ding, N., Sethu, V., Epps, J. and Ambikairajah, E. *Speaker variability in emotion recognition - an adaptation based approach.* In ICASSP, Kyoto, Japan, 2012.

[Douglas-Cowie et al. 03] Douglas-Cowie, E., Campbell, N., Cowie, R. and Roach, P. *Emotional speech : Towards a new generation of databases.* Speech Communication, vol. 40, pages 33–60, 2003.

[Douglas-Cowie et al. 07] Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., claude Martin, J., Devillers, L., Abrilian, S., Batliner, A., Amir, N. and Karpouzis, K. *The HUMAINE Database : Addressing the Collection and Annotation of Naturalistic and. In Induced Emotional Data.* Lecture Notes in Computer Science, Affective Computing and Intelligent Interaction, vol. 4638, pages 488–500, 2007.

[Doval and Rodet 93] Doval, B. and Rodet, X. *Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs.* In ICASSP, 1993.

[Dowek et al. 09] Dowek, G., Guiraud, D., Kirchner, C., Métayer, D. L. and Oudeyer, P.-Y. *Rapport sur la création d'un comité d'éthique en Sciences et Technologies du Numérique.* Rapport technique, INRIA, mai 2009.

[Dumouchel et al. 09] Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R. and Boufaden, N. *Cepstral and Long-Term Features for*

Emotion Recognition. In Interspeech, pages 344–347, Brighton, U.K., 2009.

[Engberg et al. 97] Engberg, I. S., Hansen, A. V., Andersen, O. and Dalsgaard, P. *Design, recording and verification of a danish emotional speech database.* In Eurospeech, pages 1695–1698, Rhodes, Greece, 1997.

[Erickson et al. 06] Erickson, D., Yoshida, K., Menezes, C., Fujino, A., Mochida, T. and Shibuya, Y. *Exploratory study of some acoustic and articulatory characteristics of sad speech.* Phonetica, vol. 63 (1), no. 63, pages 1–25, 2006.

[Eyben et al. 10] Eyben, F., Batliner, A., Schuller, B., Seppi, D. and Steidl, S. *Cross-corpus classification of realistic emotions : some pilot experiments.* In LREC, Workshop on EMOTION : Corpora for Research on Emotion and Affect, pages 77–82, Valetta, Malta, 2010. ELRA.

[Fant et al. 85] Fant, G., Liljencrants, J. and Lin, Q.-G. *A four-parameter model of glottal fow.* STL-QPSR, vol. 26 (4), pages 1–13, 1985.

[Farner et al. 09] Farner, S., Röbel, A. and Rodet, X. *Natural transformation of type and nature of the voice for extending vocal repertoire in high-fidelity applications.* In Conference of the Audio Engineering Society (AES), New-York, NY, USA, 2009.

[Fernandez and Picard 03] Fernandez, R. and Picard, R. W. *Modeling drivers' speech under stress.* Speech Communication, vol. 40, pages 145–159, 2003.

[Galves et al. 02] Galves, A., Garcia, J., Duarte, D. and Galves, C. *Sonority as a basis for rhythmic class discrimination.* In Speech Prosody, Aix-en-Provence, France, 2002.

[Garnier et al. 04] Garnier, M., Dubois, D., Poitevineau, J., Henrich, N. and Castellengo, M. *Etude perceptive et acoustique de la qualité vocale dans le chant lyrique.* In International Conference on Voice Physiology and Biomechanics, Marseille, France, 2004.

[Gauvain and Lee 94] Gauvain, J.-L. and Lee, C. *Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains.* IEEE Transaction on Speech and Audio Processing, vol. 2 (2), pages 291–298, 1994.

[Gendrot 04] Gendrot, C. *Rôle de la qualité de la voix dans la simulation des émotion : une étude perceptive et physiologique.* 2004.

[Goldman 11] Goldman, J.-P. *EasyAlign : an automatic phonetic alignment tool under Praat.* In Interspeech, Firenze, Italy, 2011.

[Grabe and Low 02] Grabe, E. and Low, E. L. *Durational variability in speech and the rhythm class hypothesis.* Papers in Laboratory Phonology, vol. 7, pages 515–546, 2002.

[Grichkovtsova et al. 08] Grichkovtsova, I., Morel, M. and Lacheret-Dujour, A. *Identification des émotions en voix naturelle et synthétique : paradigme d'ancrage.* In JEP, Avignon, France, 2008.

[Grichkovtsova et al. 09] Grichkovtsova, I., Morel, M. and Lacheret, A. The role of prosody in affective speech, chapitre Perception of affective prosody in natural and synthesized speech : which methodological approach ? Peter Lang AG, Bern, 2009., 2009.

[Grimm et al. 08] Grimm, M., Kroschel, K. and Narayanan, S. *The Vera am Mittag German audio-visual emotionnal speech database.* In IEEE, editeur, International Conference on Mulitmedia and Expo (ICME), pages 865–868, Hannover, Germany, 2008.

[Halioua 07] Halioua, B. Le procès des médecins de nuremberg : L'irruption de l'éthique médicale moderne. Espace Ethique, 2007.

[Hall et al. 09] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. *The WEKA Data Mining Software : An Update.* SIGKDD Explorations, vol. 11 (1), 2009.

[Hansen and Bou-Ghazale 97] Hansen, J. H. L. and Bou-Ghazale, S. E. *Getting Started with SUSAS : A Speech Under Simulated and Actual Stress Database.* In Eurospeech, 1997.

[Harnsberger et al. 08] Harnsberger, J., Shrivastav, R., Brown, W., Rothman, H. and Hollien, H. *Speaking Rate and Fundamental Frequency as Speech Cues to Perceived Age.* Journal of Voice, vol. 22 (1), pages 58–69, 2008.

[He et al. 10] He, L., Lech, M., Maddage, N. and Allen, N. *Stress and Emotion recognition using log-gabor filter analysis of speech spectrograms.* ACII, 2010.

[Hermansky 90] Hermansky, H. *Perceptual linear predictive (PLP) analysis for speech.* Journal of the Acoustical Society of America, vol. 87, pages 1738–1752, 1990.

[Hirsh 12] Hirsh, E. *Fondements de l'éthique biomédicale, textes de références.* Université Paris-Sud, Département de recherche en éthique, 2012.

[Hirst and Espesser 93] Hirst, D. and Espesser, R. *Automatic modelling of fundamental frequency using quadratic spline function.* Travaux de l'Institut de Phonétique d'Aix, vol. 15, pages 75–85, 1993.

[Hirst et al. 00] Hirst, D., Cristo, A. D. and Espesser, R. Levels of representation and levels of analysis for the description of intonation systems, chapitre Prosody : Theory and Experiment. Kluwer Academic Press, 2000.

[Hogg and Ledolter 87] Hogg, R. V. and Ledolter, J. Engineering statistics. McMillan, 1987.

[Jarina et al. 02] Jarina, R., O'Connor, N., Marlow, S. and Murphy, N. *Rhythm detection for speech-music discrimination in MPEG compressed domain.* In International Conference on Digital Signal Processing (DSP/IEEE), Santorini, Greece, 2002.

[John and Srivastava] John, O. P. and Srivastava, S. Handbook of personality : Theory and research, chapitre The Big-Five Trait Taxonomy : History, Measurement, and Theoretical Perspectives, pages 102–138. New York : Guilford.

[Kennedy and Ellis 04] Kennedy, L. S. and Ellis, D. P. *Laughter detection in meetings.* In NIST ICASSP Meeting Recognition Workshop, pages 11–14, Montreal, Canada, 2004.

[Kessous et al. 10] Kessous, L., Castellano, G. and Caridakis, G. *Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis.* Journal of Multimodal User Interfaces, vol. 3, pages 33–48, 2010.

[Kienast and Sendlmeier 00] Kienast, M. and Sendlmeier, W.-F. *Acoustical analysis of spectral and temporal changes in emotional speech.* In ISCA ITRW on Speech and Emotion, pages 92–97, Belfast, Ireland, 2000.

[Kim et al. 08] Kim, S., Hirst, D., Cho, H. and Chung, H.-Y. L. M. *Korean MULTEXT : a korean prosody corpus.* In Speech Prosody, Campinas, Brazil, 2008.

[Kipp 01] Kipp, M. *Anvil - A Generic Annotation Tool for Multimodal Dialogue.* In Eurospeech, numéro 7, pages 1367–1370, Aalborg, Germany, 2001.

[Kreiman and Gerratt 03] Kreiman, J. and Gerratt, B. R. *Jitter, shimmer, and noise in pathological voice quality perception.* In VOQUAL'03, pages 57–62, Genève, Suisse, August 2003.

[Kreiman et al. 03] Kreiman, J., Gabelman, B. and Gerratt, B. *Perception of vocal tremor.* Journal of Speech, Language, and Hearing Research, vol. 46, pages 203–214, 2003.

[Lacheret-Dujour and Victorri 02] Lacheret-Dujour, A. and Victorri, B. *La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques.* In JEP, Nancy, France, 2002.

[Lanchantin et al. 08] Lanchantin, P., Morris, A. C., Rodet, X. and Veaux, C. *Automatic phoneme segmentation with relaxed textual constraints.* In LREC, Marrakech, Marocco, 2008.

[Ledoux 89] Ledoux, J. Cognitive-emotional interactions in the brain, volume 3, chapitre Cognition and Emotion, pages 267–289. 1989.

[Lee et al. 04] Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S. and Narayanan, S. *Emotion recognition based on phonem classes.* In ICSLP, pages 889–892, Jeju Island, Korea, 2004.

[Leggetter and Woodland 95] Leggetter, C. and Woodland, P. *Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Model.* Computer Speech and Language, vol. 9, pages 171–185, 1995.

[Leung et al. 08] Leung, C.-C., Ferras, M., , Barras, C. and Gauvain, J.-L. *Comparing prosodic models for speaker recognition.* In Interspeech, pages 1945–1948, Brisbane, Australia, 2008.

[Li et al. 05] Li, D., Yang, Y., Wu, Z. and Wu, T. *Emotion-state conversion for speaker recognition.* In ACII, volume LNCS 3784, pages 403–410, Beijing, China, 2005.

[Li et al. 10] Li, M., Jung, C.-S. and Han, K. J. *Combining five acoustic level modeling methods for automatic speaker age and gender recognition.* In Interspeech, Makuhari, Japan, 2010.

[Mallat 00] Mallat, S. Une exploration des signaux en ondelettes. Les Editions de l'Ecole Polytechnique, 2000.

[Marchi et al. 12] Marchi, E., Batliner, A. and Schuller, B. *Speech, emotion, age, language, task and typicality : trying to disentangle performance and future relevance.* In Workshop on Wide Spectrum Social Signal Processing (ASE/IEEE International Conference on Social Computing), Amsterdam, Netherlands, 2012.

[Mariani et al. 09] Mariani, J., Besnier, J.-M., Bordé, J., Cornu, J.-M., Farge, M., Ganascia, J.-G., Haton, J.-P. and Serverin, E. *Pour une éthique de la recherche en Sciences et Technologies de l'Information et de la Communication (STIC).* Rapport technique, Comité d'éthique du CNRS, 2009.

[Martin 87] Martin, P. *Prosodic and rhythmic structures in French.* Linguistics, vol. 25, pages 925–949, 1987.

[Martin 05] Martin, P. *WinPitch LTL, un logiciel multimédia d'enseignement de la prosodie.* Apprentissage des langues

et systèmes d'information et de communication (Alsic), vol. 8 (2), pages 95–108, 2005.

[Martin 06] Martin, P. *Intonation du français : parole spontanée et parole lue*. EFE, vol. XV, pages 133–162, 2006.

[Mehrabian 96] Mehrabian, A. *Analysis of the Big-five personality factors in terms of the PAD Temperament Model*. Australian Journal of Psychology, vol. 48, pages 86–92, 1996.

[Mertens 04] Mertens, P. *The Prosogram : Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model*. In Speech Prosody, Campinas, Brazil, 2004.

[Milgram 63] Milgram, S. *Behavioral study of obedience*. Journal of Abnormal and Social Psychology, vol. 67, pages 371–378, 1963.

[Milgram 74] Milgram, S. Obedience to authority. Harper and Row, 1974.

[Miwa et al. 03] Miwa, H. I., Ito, D. T. K. and Takanishi, A. H. *Introduction of the Need Model for Humanoid Robots to Generate Active Behavior*. In IEEE/RSJ International Conference on Intelligent Robots and Systems, volume 2, pages 1400–1406, 2003.

[Mohammadi et al. 10] Mohammadi, G., Mortillaro, M. and Vinciarelli, A. *The voice of personality : Mapping nonverbal vocal behavior into trait attributions*. In International Workshop on Social Signal Processing (SSPW), pages 17–20, Florence, Italy, 2010.

[Montacie and Chollet 87] Montacie, G. and Chollet, G. *Système de références pour l'évaluation d'applications et la caractérisation de base de données en reconnaissance automatique de la parole*. In JEP, 1987.

[Mori 70] Mori, M. *The uncanny valley*. Energy, vol. 7 (4), pages 33–35, 1970.

[Mower et al. 09] Mower, E., Metallinou, A., Lee, C.-C., Kazemzadeh, A., Busso, C., Lee, S. and Narayanan, S. *Interpreting ambiguous emotional expressions*. In ACII, volume 978 (1), pages 4244–4799, Amsterdam, The Netherlands, 2009.

[Obin et al. 08a] Obin, N., Goldman, J.-P., Avanzi, M. and Lacheret-Dujour, A. *Comparaison de trois outils de détection automatique de proéminences en français parlé*. In JEP, Avignon, France, 2008.

[Obin et al. 08b] Obin, N., Rodet, X. and Lacheret-Dujour, A. *Un modèle de durées des syllabes fondé sur leurs propriétés*

intrinsèques et les varations locales de débit. In JEP, Avignon, France, 2008.

[Osgood et al. 75] Osgood, C., W.H. and Miron, M. Cross-cultural universals of affective meaning. University of Illinois Press, Urban, 1975.

[Patel et al. 06] Patel, A. D., Iversen, J. R. and Rosenberg, J. C. *Comparing the rhythm and melody of speech and music : the case of British English and French.* Journal of Acoustic Society of America, vol. 119(5), pages 3034–3047, 2006.

[Peeters 04] Peeters, G. *A large set of audio features for sound description (similarity and classification) in the CUIDADO project.* Ircam, 2004.

[Pelecanos and Sridharan 01] Pelecanos, J. and Sridharan, S. *Feature warping for robust speaker verification.* In Workshop on Speaker Recognition (ISCA), Crete, Greece, 2001.

[Pierrehumbert 80] Pierrehumbert, J. B. *The Phonology and Phonetics of English Intonation.* Thèse de doctorat, MIT, 1980.

[Plaisant et al. 10] Plaisant, O., Courtois, R., Réveillère, C., Mendelsohn, G. and John, O. *Validation par analyse factorielle du Big Five Inventory français (BFI-Fr). Analyse convergente avec le NEO-PI-R.* Annales Médico-psychologiques, revue psychiatrique, vol. 168 (7), pages 97–106, 2010.

[Plutchik 84] Plutchik, R. Approaches to emotion, chapitre A General Psychoevolutionary Theory. Erlbaum, Hillsdale, NJ, 1984.

[Polzehl et al. 10] Polzehl, T., Möller, S. and Metze, F. *Automatically assessing personality from speech.* In IEEE International Conference on Semantic Computing (ICSC), pages 134–140, Pittsburgh, PA, 2010.

[Pudil et al. 02] Pudil, P., Novovicova, J. and Somol, P. *Feature selection toolbox software package.* Pattern Recognition Letters, vol. 23 (4), pages 487–492, 2002.

[Ramus et al. 99] Ramus, F., Nespor, M. and Mehler, J. *Correlates of linguistic rhythm in the speech signal.* Cognition, vol. 73 (3), pages 265–292, 1999.

[Raskin and Hall 79] Raskin, R. and Hall, C. *Interpersonal and intrapsychic adaptiveness of trait self-enhancement : A mixed blessing ?* Journal of Personality and Social Psychology, vol. 74, pages 1197–1208, 1979.

[Ray et al. 08] Ray, C., Mondada, F. and Siegwart, R. *What do people expect from robots ?* In IEEE/RSJ International Confe-

rence on Intelligent Robots and Systems, pages 3816–3821, Nice, France, 2008.

[Röbel and Rodet 05] Röbel, A. and Rodet, X. *Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation*. In Digital Audio Effects (DAFx'05), Madrid, Spain, September 20-22 2005.

[Reynolds and Rose 95] Reynolds, D. A. and Rose, R. C. *Robust text-independent speaker identification using Gaussian mixture speaker models*. Transaction on Speech and Audio Processing (IEEE), vol. 3, pages 72–83, 1995.

[Reynolds et al. 00] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B. *Speaker verification using adapted gaussian mixture models*. Digital Signal Processing, vol. 10, pages 19–41, 2000.

[Ringeval et al. 08] Ringeval, F., Sztaho, D., Chetouani, M. and Vicsi, K. *Automatic Prosodic disorders analysis for impaired communication children*. In Workshop on Child, Computer and Interaction (WOCCI), ICME (IEEE), numéro 1st, 2008.

[Rollet et al. 09] Rollet, N., Delaborde, A. and Devillers, L. *Protocol CINEMO : the use of fiction for collecting emotional data in naturalistic controlled oriented context*. In ACII, Amsterdam, the Netherlands, 2009.

[Rong et al. 09] Rong, J., Li, G. and Chen, Y.-P. P. *Acoustic feature selection for automatic emotion recognition from speech*. Information Processing and Management, vol. 45, pages 315–328, 2009.

[Roubeau et al. 09] Roubeau, B., Henrich, N. and Castellengo, M. *Laryngeal Vibratory Mechanisms : The Notion of Vocal Register Revisited*. Journal of Voice, vol. 23 (4), pages 425–438, 2009.

[Ruiz et al. 08] Ruiz, R., de Hugues, P. P. and Legros, C. *Analysing cockpit and laboratory recordings to determine fatigue levels in pilots' vocies*. Journal of the Acoustical Society of America, vol. vol. 123 :3070, Issue N°5, 2008.

[Ruiz et al. 10] Ruiz, R., de Hugues, P. P. and Legros, C. *Advanced voice analysis of pilots to detect fatigue and sleep inertia*. Acta Acustica United with Acustica, vol. 96, n°3, pages 567–579, 2010.

[Russel 97] Russel, J. The psychology of facial expression, chapitre Reading emotions from and into faces : resurrecting a dimensional-contextual perspective, pages 295–320. Cambridge University Press, Cambridge, UK, 1997.

[Sacks 09] Sacks, O. Musicophilia, la musique, le cerveau et nous. 2009.

[Said et al. 10] Said, C. P., Moore, C. D., Norman, K. A., Haxby, J. V. and Todorov, A. *Graded representations of emotional expressions in the left superior temporal sulcus.* Frontiers in Systems Neurosciences, vol. 4, article 6, pages 1 – 8, 2010.

[Scherer and Peper 01] Scherer, K. R. and Peper, M. *Psychological theories of emotion and neuropsychological research.* Handbook of Neuropsychology, vol. 5 (Emotional behavior and its disorders), pages 17–48, 2001.

[Scherer et al. 80] Scherer, U., Helfrich, H. and Scherer, K. R. *Internal push or external pull ? Determinants of paralinguistic behavior.* Oxford - New York : Pergamon, 1980.

[Scherer et al. 03] Scherer, K., Johnstone, T. and Klasmeyer, G. Handbook of affective sciences, chapitre Vocal expression of emotion, pages 433–456. Numeéro 23. Oxford University Press, Oxford, New-York., 2003.

[Scherer 86] Scherer, K. R. *Vocal affect expressions : A Review and a Model for Future Research.* Psychological Bulletin, vol. 99 (2), pages 143–165, 1986.

[Scherer 94] Scherer, K. R. Affect bursts, chapitre Emotions, pages 161–193. Hillsdale, NJ : Lawrence Erlbaum, 1994.

[Scherer 99] Scherer, K. R. Handbook of cognition and emotion, chapitre Appraisal Theory, pages 637–663. T. Dalgleish and M. J. Powe, 1999.

[Schröder 03] Schröder, M. *Experimental study of affect bursts.* Speech Communication - Special session on speech and emotion, vol. vol. 40, Issue 1-2, 2003.

[Schuller and Devillers 10] Schuller, B. and Devillers, L. *Incremental acoustic valence recognition : an inter-corpus perspective on features, matching and performance in a gating paradigm.* In Interspeech, Makuhari, Chiba, Japan, 26 - 30 sept 2010.

[Schuller et al. 09a] Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G. and Wendemuth, A. *Acoustic emotion recognition : a benchmark comparison of performances.* In Automatic Speech Recognition and Understanding Workshop (ASRU/IEEE), pages 552–557, Merano, 2009.

[Schuller et al. 09b] Schuller, B., Steidl, S. and Batliner, A. *The INTERSPEECH 2009 Emotion Challenge.* In Interspeech, Brighton, U.K., 2009.

[Schuller et al. 10a] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. and Narayanan, S. *The INTERSPEECH 2010 Paralinguistic Challenge.* In Interspeech, pages 2830–2833, Makuhari, Chiba, Japan, 26 - 30 sept 2010.

[Schuller et al. 10b] Schuller, B., Zaccarelli, R., Rollet, N., and Devillers, L. *CINEMO - A French spoken language resource for complex emotions : facts and baselines.* In LREC, Valetta, Malta, 2010.

[Schuller et al. 11a] Schuller, B., Batliner, A., Steidl, S. and Seppi, D. *Recognising realistic emotions and affect in speech : state of the art and lessons learnt from the first challenge.* Speech Communication, Special Issue on "Sensing Emotion and Affect - Facing Realism in Speech Processing", vol. 53 (9/10), pages 1062–1087, 2011.

[Schuller et al. 11b] Schuller, B., Steidl, S., Batliner, A., Schiel, F. and Krajewski, J. *The INTERSPEECH 2011 Speaker State Challenge.* In Interspeech, Firenze, Italy, 2011.

[Schuller et al. 12a] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. and Narayanan, S. S. *Paralinguistics in Speech and Language, State-of-the-Art and the Challenge.* Computer Speech and Language (CSL), Special Issue on Paralinguistics in Naturalistic Speech and Language, 2012.

[Schuller et al. 12b] Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G. and Weiss, B. *The INTERSPEECH 2012 Speaker Trait Challenge.* In Interspeech, Portland, Oregon, USA, 2012.

[Shilker 09] Shilker, T. S. *Analysis of affective expression in speech.* Thèse de doctorat, Cambridge University, Computer Laboratory, 2009.

[Simon 02] Simon, A.-C. *Le rôle de la prosodie dans le repérage des unités textuelles minimales.* Cahiers de linguistique française, vol. 23, pages 99–125, 2002.

[Smith and Honing 08] Smith, L. M. and Honing, H. *Time-frequency representation of musical rhythm by continuous wavelets.* Journal of Mathematics and Music, vol. 2 (2), 2008.

[Steidl et al. 09] Steidl, S., Batliner, A., Schuller, B. and Seppi, D. *The hinterland of emotions : facing the open-microphone challenge.* In ACII, pages 4244–4799, 2009.

[Sturmel et al. 09] Sturmel, N., d'Alessandro, C. and Rigaud, F. *Glottal closure instant detection using lines of maximum amplitudes (loma) of the wavelet transform.* In ICASSP, Taipei, Taiwan, 2009.

[Sturmel 11] Sturmel, N. *Analyse de la qualité vocale appliquée à la parole expressive.* Thèse de doctorat, Université Paris-Sud, LIMSI, 2011.

[Sun 02] Sun, X. *Pitch determination and voice quality analysis using subharmonic to harmonic ratio.* In ICASSP, volume I, pages 333–336, Orlando, Florida, USA, 2002.

[t Hart 81] t Hart, J. *Differential sensitivity to pitch distance, particularly in speech.* Journal of the Acoustical Society of America, vol. 69 (3), pages 811–821, 1981.

[Tahon and Devillers 10] Tahon, M. and Devillers, L. *Acoustic measures characterizing anger across corpora collected in artificial or natural context.* In Speech Prosody, Chicago, USA, 2010.

[Tahon et al. 10] Tahon, M., Delaborde, A., Barras, C. and Devillers, L. *A corpus for identification of speakers and their emotions.* In LREC, Valetta, Malta, 2010.

[Tahon et al. 11] Tahon, M., Delaborde, A. and Devillers, L. *Real-life Emotion Detection from Speech in Human-Robot Interaction : Experiments across Diverse Corpora with Child and Adult Voices.* In Interspeech, Firenze, Italia, 2011.

[Tahon et al. 12a] Tahon, M., Degottex, G. and Devillers, L. *Usual voice quality features for emotionnal valence detection.* In Speech Prosody, Shanghai, China, 2012.

[Tahon et al. 12b] Tahon, M., Delaborde, A. and Devillers, L. *Corpus of children voices for mid-level social markers and affect bursts analysis.* In LREC, Istanbul, Turckey, 2012.

[Tamburini and Wagner 07] Tamburini, F. and Wagner, P. *On automatic prominence detection for German.* In Interspeech, Antwerp, Belgium, 2007.

[Tamburini 06] Tamburini, F. *Reliable Prominence Identification in English Spontaneous Speech.* In Speech Prosody, Dresden, Germany, 2006.

[Tao and Kang 05] Tao, J. and Kang, Y. *Features Importance Analysis for Emotional Speech Classification.* In ACII, volume 13784/2005, pages 449–457. Lecture Notes in Computer Sciences, 2005.

[Tato 99] Tato, R. *Emotion Recognition in Speech Signal.* Thèse de doctorat, University of Manchester, 1999.

[Trouvain 01] Trouvain, J. *Phonetics aspects of "speech-laughs".* In Orality and gestuality (ORAGE 2001), pages 634–639, 2001.

[Truong and van Leeuwen 07a] Truong, K. P. and van Leeuwen, D. A. *Evaluating automatic laughter segmentation in meetings using acoustic and acoustic-phonetic features.* In Interdisciplinary Workshop on The Phonetics of Laughter, Saarbrucken, Germany, 2007.

[Truong and van Leeuwen 07b] Truong, K. P. and van Leeuwen, D. A. *An "open-set' detection evaluation methodology for automatic emotion recognition in speech.* In International workshop on Paralinguistic Speech - between models and data, Saarbrücken, Germany, 2007.

[Vaudable et al. 10] Vaudable, C., Rollet, N. and Devillers, L. *Annotation of affective interaction in real-life dialogs collected in a call-center.* In LREC, Workshop on Emotion and Affect, Valetta, Malta, 2010.

[Vaudable 12] Vaudable, C. *Analyse et reconnaissance des émotions lors de conversations de centre d'appels.* Thèse de doctorat, Université Paris-Sud, LIMSI, 2012.

[Ververidis and Kotropoulos 03] Ververidis, D. and Kotropoulos, C. *A review of emotionnal speech databases.* In Penhellenic Conf. on Informatics (PCI), numéro 9th, pages 560–574, Thessaloniki, Greece, 2003.

[Vidrascu and Devillers 05] Vidrascu, L. and Devillers, L. *Annotation and detection of blended emotions in real Human-Human dialogs recorded in a Call center.* In ICME, Amsterdam, The Netherlands, 2005.

[Vidrascu 07] Vidrascu, L. *Analyse et détection des émotions verbales dans les interactions orales.* Thèse de doctorat, Université Paris-Sud, LIMSI, 2007.

[Villavicencio et al. 09] Villavicencio, F., Röbel, A. and Rodet, X. *Applying improved spectral modeling for high quality voice conversion.* In ICASSP, pages 4285–4288, Taipei, Taiwan, 2009.

[Vinciarelli et al. 08] Vinciarelli, A., Pantic, M., Bourlard, H. and Pentland, A. *Social Signals, their Function, and Automatic Analysis : A Survey.* In Conference on Multimodal Interfaces (ACM),, pages 61–68, Chania, Greece, 2008.

[Vinciarelli et al. 10] Vinciarelli, A., Salamin, H., Mohammadi, G. and Truong, K. Toward autonomous, adaptive, and context-aware multimodal interfaces : Theoretical and practical issues, volume Vol. 6456, chapitre More than words : inference of socially relevant information from nonverbal vocal cues in speech, pages 24–34. Lecture Notes in Computer Science, 2010.

[Wigging 96] Wigging, J. The five-factor models of personality. Guilford, 1996.

[Wöllmer et al. 11] Wöllmer, M., Weninger, F., Steidl, S., Batliner, A. and Shculler, B. *Speech-based Non-prototypical Affect Recognition for Child-Robot Interaction in Reverberated Environments.* In Interspeech, Firenze, Italy, 2011.

[Wu et al. 10] Wu, D., Parsons, T. D. and Narayanan, S. *Acoustic feature analysis in speech emotion primitives estimation.* In Interspeech, Makuhari, Chiba, Japan, 2010.

[Wundt 13] Wundt, W. Die psychologie im kampf ums dasein. Leipzig : Kröner, 1913.

[Xiao 08] Xiao, Z. *Classification of emotions in audio signals.* Thèse de doctorat, Ecole Centrale de Lyon, 2008.

[Xie and Niyogi 06] Xie, Z. and Niyogi, P. *Robust acoustic-based syllable detection.* In Interspeech, Pittsburgh, PA, USA, September Pittsburgh, PA, USA, September Pittsburgh, PA, USA, 2006. ISCA.

[Yang 04] Yang, H.-R. *Real-time Musical Rythm detection.* Thèse de doctorat, Tiatung University, China, 2004.

[Yildirim et al. 04] Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S. and Narayanan, S. *An acoustic study of emotions expressed in speech.* In ICSLP, Jeju Island, Korea, 2004.

[Zellner-Keller 04] Zellner-Keller, B. *Prosodic styles and Personality styles, are the two intercorrelated ?* In Speech Prosody, Nara, Japan, 2004.

[Zhou et al. 01] Zhou, G., Hansen, J. H. L. and Kaiser, J. F. *Nonlinear Feature Based Classification of Speech under Stress.* Transaction on Speech and Audio Processing, vol. 9 (3), pages 201–216, 2001.

Table des figures

Liste des tableaux

Publications pendant la thèse

Conférences internationales avec comité de lecture

Tahon Marie, Gilles Degottex, Laurence Devillers, "Usual voice quality features and glottal features for emotional valence detection", Speech Prosody, Shanghai, China, 2012.

Tahon Marie, Agnes Delaborde, Laurence Devillers, "Corpus of children voices for mi-level social markers and affect burst analysis", LREC, Istanbul, Turkey, 2012.

Tahon Marie, Agnes Delaborde, Laurence Devillers, "Real-life emotion detection from speech in Human-Robot interaction : experiments across diverse corpora with child and adult voices", Interspeech, Firenze, Italy, 2011.

Tahon Marie, Devillers Laurence, "Acoustic measures characterizing anger across corpora collected in artificial or natural context", Speech Prosody, Chicago, USA, 2010.

Tahon Marie, Agnès Delaborde, Claude Barras and Laurence Devillers, "A corpus for identification of speakers and their emotions", Workshop Emotion, LREC, Malta, 2010.

Agnès Delaborde, Tahon Marie, Claude Barras and Laurence Devillers, "Corpus NAO-children : affective links in a Child-Robot interaction, Workshop Emotion, LREC, Malta, 2010.

Agnes Delaborde, Marie Tahon, Claude Barras, and Laurence Devillers, "A Wizard-of-Oz Game for Collecting Emotional Audio Data in a Children-Robot Interaction". In Proc. of the International Workshop on Affective-aware Virtual Agents and Social Robots, ICMI-MLMI, Boston, USA, 2009.

Présentations scientifiques sans comité de lecture

Tahon Marie, "Reconnaissance du locuteur et de son état émotionnel à partir d'un signal acoustique verbal lors d'une interaction homme-robot", GT Affect, Compagnons Artificiels et Interaction (ACAI), 6 avril 2012, Lip6, Paris.

Tahon Marie, "Étude de descripteurs acoustiques pour la reconnaissance des émotions et l'identification du locuteur ", Journée des Doctorants, projet ROMEO, 18 nov. 2010, Telecom-Paris, Paris.

Tahon Marie, "Proposition de mesure du degré de naturel dans un corpus émotionnel à partir des manifestations acoustiques de la colère ", Journées Jeunes Chercheurs en Audition, Acoustique Musicale et Signal Audio (JJCAAS), 17-19 nov. 2010, Ircam, Paris.