# Analyse et détection des émotions verbales dans les interactions orales

Laurence Vidrascu

**Université Paris-Sud 11**
Faculté des Sciences d'Orsay
91405 ORSAY CEDEX

**LIMSI-CNRS**
BP 133
F-91403 ORSAY CEDEX

Thèse pour obtenir le grade de Docteur de l'Université Paris 11
Discipline : Informatique

Présentée et soutenue publiquement par
Laurence Vidrascu

# Analyse et détection des émotions verbales dans les interactions orales

Soutenu publiquement le 20 décembre 2007 devant le jury composé de

| | |
|---|---|
| Laurence Devillers | Directeur |
| Jean-Paul Haton | Rapporteur |
| Anton Batliner | Rapporteur |
| Nick Campbell | Examinateur |
| Lori Lamel | Examinateur |
| Joseph Mariani | Examinateur |

# Remerciements

Je tiens tout d'abord à remercier mon directeur de thèse, Laurence Devillers, pour son encadrement et ses nombreux conseils ainsi que Lori Lamel qui a encadré ma thèse pendant 2 ans pour ses encouragements et son aide pour l'utilisation des outils du groupe.
Mes remerciements vont également aux membres de mon jury : Jean Paul Haton, Anton Batliner, Nick Campbell et Joseph Mariani.
Merci à toute le groupe TLP et aux autres doctorants, en particulier Bianca (et Emma) pour son aide incommensurable et son amitié, Cécile pour ses relectures et son soutien, Daniel pour avoir relié ma thèse et Anne-Laure pour tous ses conseils.
Merci enfin à tous ceux, famille et amis, qui m'ont soutenue et supportée, mention spéciale à ma soeur et à son cachougué.

# Table des matières

# I Introduction

# INTRODUCTION GENERALE

La présente thèse a pour sujet la détection automatique des émotions dans la voix. Longtemps dédaigné par la communauté scientifique, le domaine des émotions est aujourd'hui en plein essor. Les avatars pouvant exprimer une émotion, comme ceux disponibles sur Yahoo Messenger par exemple, se multiplient. De même on voit de plus en plus de gadgets du type lapin « nabaztag[1] » qui exprime des émotions[2] et à qui on peut envoyer des messages en utilisant une voix plus ou moins « en forme » ou stressée. De façon moins ludique, on commence aussi à s'intéresser aux émotions dans le domaine de l'éducation avec par exemple pour objectif des tuteurs virtuels dont la stratégie évoluerait suivant que la personne qui interagit avec eux est intéressée, ennuyée ou frustrée. La prise en compte des émotions peut également servir pour les centres d'appels, où la satisfaction du client est primordiale. Concernant ce dernier point, la détection des émotions peut avoir plusieurs intérêts. Tout d'abord une grande quantité de données est actuellement enregistrée et il peut être intéressant de détecter automatiquement les portions de dialogue correspondant à de la satisfaction ou à de l'énervement afin de les analyser a posteriori et de modifier les stratégies (pour des agents humains comme pour des agents virtuels), le but final étant de ne pas perdre de client. Ensuite, la détection des émotions a également était envisagée afin de superviser les interactions et d'intervenir en cas de problème[3] (là aussi avec des agents humains ou des systèmes de dialogue). Les premiers outils de « quality monitoring » utilisent à la fois la reconnaissance de la parole et des indices acoustiques (voix superposées, silences, hésitations, temps d'interaction, etc.) pour inférer de la non-satisfaction d'un appelant. Un premier système de ce type a d'ailleurs été commercialisé en 2006 par les laboratoires NICE[4] avec un module « emotion detection » visant à détecter un « évènement émotionnel » mais aucune évaluation de ce module n'a été effectuée.

Indépendamment de ces applications émergentes, le domaine des émotions est particulièrement intéressant par son aspect pluridisciplinaire (psychologie, physiologie, neurologie, traitement de la parole, traitement du signal, réalité virtuelle). Il a motivé la création d'un réseau d'excellence, HUMAINE (Human-Machine Interaction Network on Emotion), dans lequel le LIMSI est

---

[1] www.nabaztag.com

[2] En activant une fonction « humeur », le lapin prend la parole à des moments aléatoires et ses intonations sont souvent assez marquées.

[3] Bar Veinstein, de NICE Systems : "Des recherches montrent que si vous répondez à un consommateur dans les 24 heures après qu'il ait eu une mauvaise expérience avec l'un de vos produits, vous avez de grande chance de regagner ce consommateur et de le fidéliser"

[4] http://nicesystem.ru/news/newsletter/6_07/analyze.php

impliqué, et qui réunissait des experts issus de plusieurs disciplines dans le but de partager les différentes expertises afin de progresser vers des systèmes 'orientés émotions'. J'ai d'ailleurs participé à CEICES (Combining Efforts for Improving Automatic Classificationof Emotional User States), une collaboration de plusieurs sites de HUMAINE, dont l'objectif était de se pencher sur la classification des états émotionnels exprimés vocalement.

Dans ce manuscrit, nous nous concentrons sur la communication vocale des émotions. Nous avons choisi de travailler sur des données téléphoniques provenant de centres d'appel car elles sont particulièrement adaptées à ce type de travaux, l'émotion s'exprimant uniquement par la voix. En contrepartie, la qualité du signal n'est pas toujours optimale par rapport à des données non téléphoniques et la bande passante est réduite.

## Les émotions dans la voix

Depuis une quinzaine d'années, de plus en plus de chercheurs se sont intéressés à l'étude des émotions dans la voix avec souvent la même manière de procéder : à partir de données étiquetées en émotion, un ensemble d'indices est extrait et des méthodes de fouille de donnée sont utilisées pour reconnaître les émotions. Dans la plupart des études, la notre y compris, le terme émotion sera utilisé au sens large pour signifier état affectif. En 2003, deux états de l'art [Scherer 2003], [Juslin et Laukka 2003] ont été fait sur plus d'une centaine d'études avec de nombreuses critiques :

*"most of the studies in this area lack theoretical and analytical rigor, and some of the most serious shortcomings are the following: using actor portrayed emotion utterances, as opposed to naturally occurring emotional vocalizations; not systematically controlling important variables such as the number of speakers, the type of emotions studied, the instructions for portrayal, and the verbal material used." [Kappas et al. 1991 p213]*

En effet la plupart des études, bien que visant des applications réelles, s'appuyaient sur des données jouées par des acteurs (dont le nombre était d'ailleurs souvent restreint). Elles ne précisaient pas systématiquement si une validation du jeu des acteurs avait été effectuée. (Est-ce que la colère produite par l'acteur est vraiment perçue comme de la colère ?). De plus le nombre d'émotions étudié était souvent limité, avec une majorité d'études essayant de discriminer entre 2 ou 3 classes d'émotions assez larges (positif, neutre, négatif).

Le peu d'études réalisées avec des données spontanées semblait indiquer que les performances avec des données actées ne reflétaient pas du tout ce qui serait obtenu avec des données réelles. En effet, les données réelles, comparées aux données actées, sont souvent moins intenses et beaucoup plus complexes avec en plus des mélanges qui ne sont pas toujours qualifiables avec une étiquette émotionnelle simple. Il était également difficile d'avoir une idée du nombre

maximum d'émotions pouvant être discriminées. Le domaine étant assez récent, un flou existait également sur la manière d'évaluer les performances et les performances maximales que l'on pouvait imaginer obtenir avec des indices et des algorithmes idéaux.

Nous avons essayé de répondre aux différentes critiques en travaillant sur des données spontanées particulièrement riches avec une grande diversité de locuteurs (âge, sexe, contexte, accent) et un large éventail d'états émotionnels.

Dès lors que l'on travaille avec des données réelles, plusieurs questions se posent :

- comment annoter les données pour rendre compte de leur richesse et de leur complexité ?
    - tout d'abord qu'est ce qui est annoté ? En général on choisit une unité statique comme le tour de parole, mais pourrait-on envisager un traitement plus dynamique ?
    - comment former des annotateurs experts combien d'annotateurs faut-il et comment valider les annotations ?
- de nombreuses théories existent sur les mélanges d'émotions, mais peu d'études empiriques ont été effectuées. Comment les étudier dans la pratique ? Est-ce que tout le monde les perçoit ? Peut-on typer les différents mélanges ?
- quels sont les indices les plus pertinents pour discriminer les émotions ? Existe-il un profil vocal pour les émotions « de base » comme par exemple la colère ? Est-il possible de tous les obtenir de manière automatique ? comment les combiner ? y a-t-il des types indices émergents pour reconnaître les émotions ?
- comment gérer la grande variabilité d'émotions/voix ?
- combien de classes d'émotions peut-on discriminer ?
- les modèles obtenus sont-ils généralisables sur des données comparables ? Sur d'autres types de données ? Et dans des langues différentes ?

Travaillant dans le groupe traitement de la parole du LIMSI, d'autres questions se posaient en arrière plan et des perspectives s'ouvrent. Les données émotionnelles affectent-t-elles les performances de reconnaissance de la parole. Pourrait-on envisager dans le long terme d'ajouter un module émotion au système de reconnaissance de la parole ?

## Plan de thèse

Le manuscrit sera divisé en 3 parties. La première partie dressera un état de l'art à la fois théorique et technique (chapitre 1) sur les émotions. La deuxième partie rendra compte des difficultés à travailler sur des données spontanées : leur collection (chapitre 2), leur annotation (chapitre 3) et leur analyse (chapitre 4). Enfin, la dernière partie traitera de la modélisation des émotions. Pour détecter des émotions dans des données spontanées, il faut combiner de nombreux indices de différentes natures, ce qui sera détaillé dans le chapitre 5. Les systèmes de détection, leur portabilité et leur universalité seront décrits dans le chapitre 6. Nos conclusions et perspectives sont élaborées dans le chapitre 7.

# Chapitre 1

# Etat de l'art

Résumé

Qu'appelle-t-on émotion dans nos travaux ? Quelles sont les différentes théories sur les émotions ? Existe-t-il un nombre fini d'émotions discrètes ou est ce que ce que nous appelons émotions discrètes sont en fait des zones dans des espaces à plusieurs dimensions sans véritable frontière ?

Dans ce chapitre, nous évoquons d'abord brièvement le problème de la définition des émotions et des différents états affectifs, ainsi que l'intérêt de les étudier. Nous présentons ensuite le modèle de Brunswik adapté par Scherer qui modélise la communication verbale des émotions. La question se pose alors de savoir si un humain peut juger efficacement de l'émotion d'un autre humain, ce qui est une hypothèse des travaux sur la détection d'émotions.

Nous décrivons ensuite les principales théories sur la représentation des émotions : une représentation sur des axes abstraits, la théorie d'un nombre fini d'émotions de base et la théorie d'évaluation, en nous penchant plus particulièrement sur celle de Klaus Scherer pour les émotions vocales. Enfin, après une présentation des différentes problématiques rencontrées lorsqu'on s'intéresse à la détection des émotions dans la voix, nous donnerons un état de l'art de différentes études au commencement de ma thèse et les défis posés dans cette thèse.

*What is meant by emotion and what are the different theories about what an emotion is? Is the assumption that there are distinct discrete labels theoretically correct?*

*In this chapter, we start by briefly tackling the issues of the definition of an emotion and the reasons for studying them. Once we have defined what we mean by "emotion", the question arises whether a human (or a machine) can perceive accurately his own or other people's emotion. In order to answer that, we present Scherer's adaptation of Brunswik model, which models how emotions are conveyed and report of several perceptual tests. We then briefly describe the main theories on how to represent emotions: discrete labels, continuous dimensions and the appraisal theory. Finally we give several issues in relation to the study of vocal emotions as well as a brief state of the art.*

# 1. ETAT DE L'ART

## 1.1. Théorie des émotions

Avant d'entrer dans le vif du sujet, il est nécessaire de rappeler les principales problématiques liées à la définition des émotions.

### 1.1.1. Définitions et fonction

#### *Qu'est ce qu'une émotion ?*

Dès la période classique de l'antiquité grecque, des philosophes tels que Platon et Aristote se sont intéressés aux émotions : Platon les considéraient comme une perturbation de la raison, tandis qu'Aristote déclarait dans *Rhétorique* (livre 2 1378a20 1380a4), *"J'entends par état affectif, l'appétit, la colère, la crainte, l'audace, l'envie, la joie, l'amitié, la haine, le regret de ce qui a plu, la jalousie, la pitié, bref toutes les inclinaisons accompagnées de plaisir ou de peine".*

Les scientifiques n'arrivent pas à s'accorder sur une réponse à la question "Qu'est ce qu'une émotion ? ", célèbre titre de l'article de William James[1]. Et comme le remarquent Fehr et Russel, *« Everyone knows what an emotion is, until asked to give a definition. Then it seems, no one knows »* [Fehr et Russell 1984]. Kleinginna & Kleinginna ont fait une liste des définitions existantes dans [Kleinginna et Kleinginna 1981] et ont tenté d'en extraire des caractéristiques communes. Devillers dresse un état de l'art de cette problématique dans [Devillers 2006].

Une liste non exhaustive de définitions que j'ai pu rencontrer est donnée en Annexe1. Les membres du réseau d'excellence humaine citent souvent Scherer [Scherer et al. 2004], qui définit l'émotion comme :

> *"Episodes of massive, synchronized recruitment of mental and somatic resources allowing to adapt or cope with a stimulus event subjectively appraised as being highly pertinent to the needs, goals and values of the individuals"*

[1] Pour James, les émotions sont des réactions physiologiques : lorsqu'on est dans la forêt et qu'un ours apparaît, nos os tremblent à cause de l'ours et on éprouve de la peur parce qu'on sent nos os trembler (et non pas à cause de l'ours).

## *Vocabulaire des différents états affectifs*

Le terme « émotion » peut être confondu ou opposé à d'autres états affectifs[1].

Au niveau du réseau d'excellence humaine, on se réfère souvent aux définitions de Scherer [Scherer 2003], chaque état étant décrit en termes de différentes évaluations comme par exemple évaluation de l'intensité ou de la durée. Le Tableau 1-1 résume les principaux états affectifs.

| | *Impact sur le comportement* | *Intensité* | *Rapidité du changement* | *Brièveté* | *Focus sur un événement* | *Elicitation d'évaluations* | *Synchronisation* |
|---|---|---|---|---|---|---|---|
| *Emotion*<br>ex : colère, joie, tristesse, peur | ++ | ++ | ++ | ++ | ++ | ++ | ++ |
| *Positions entre personnes*<br>(Interpersonal Stances)<br>ex : distant, froid, méprisant, chaleureux | + | + | ++ | + | + | | |
| *Humeurs* (Moods)<br>Ex : joyeux, irritable, déprimé, de bonne humeur | | + | + | + | | | |
| *Attitudes* (attitudes)<br>ex: amour, haine, amitié, désir | + | + | | | | | |
| *Dispositions affectives*<br>(Affect dispositions)<br>ex : nerveux, anxieux, morose, hostile | + | | | | | | |

**Tableau 1-1.** *Etats affectifs (adapté de [Scherer 2003]).*

Dans le cadre de ce travail et comme pour une majorité des études en « affective computing » (sciences affectives), le terme émotion sera utilisé au sens large et inclura tout état affectif, notion sur laquelle les scientifiques s'accordent[2].

[1] Pour une discussion sur la différence entre les différents états affectifs, on peut se reporter à la discussion "How Are Emotions Distinguished from Moods, Temperament, and Other Related Affective Constructs" [Ekman et Davidson 1994] où différents auteurs donnent leurs définitions de termes tels que « moods » (humeur), « emotion states/traits » (état émotionnel ), episodes, sentiments, personality, dispositions, temperament. Une liste de termes est également explicitée par Cowie dans [Cowie 2007].

[2] Cowie dans [Cowie 2007] cite la définition de AlleyDog.com « fancy way to say "feelings". »

## *Pourquoi s'intéresser aux émotions ?*

L'émotion a un impact sur notre jugement[1] [Clore 1994] et notre raisonnement[Damasio 1994]. Elle influe également sur l'attention, la motivation, la mémoire, la résolution de problèmes ou la prise de décision.

Lors de la réunion plénière Humaine de 2007[2], Paul Ekman a décrit un outil qu'il a mis en place afin de déterminer l'émotion d'une personne (peur, colère, …) par l'observation des « unités d'action[3]» du visage. Il rapportait que les gens amélioraient leur reconnaissance des émotions, que cette amélioration perdurait plusieurs mois après l'apprentissage et que certaines personnes avaient constaté une amélioration de leurs relations avec les autres, suite à cet apprentissage. R. Picard a fait des observations similaires, particulièrement lors d'études[4] avec des individus ayant du mal à reconnaître les émotions comme les autistes. Des expériences sur des simulateurs de voiture [Nass et al. 2005] ont mis en évidence le fait qu'une personne de bonne humeur réagissait mieux à une voix de bonne humeur (moins d'accidents). Par contre, une personne stressée allait mieux réagir à une voix plus sobre et une voie joyeuse allait au contraire l'irriter encore plus et augmenter le nombre d'accidents. Dans toutes les études où on compare deux versions d'un outil l'une avec un module de traitement sur les états affectifs, aussi basique soit-il, et un autre sans ce module, l'utilisateur va systématiquement préférer la version affective et cela va souvent se répercuter sur ses performances. De même, dans le domaine de l'éducation, un tuteur virtuel qui adapterait sa stratégie à l'état émotif d'un élève pourrait lui permettre de progresser plus rapidement et avec plus de plaisir. Au MIT également, des études en cours[5] proposent l'intégration des biosenseurs à des produits du type ipod ou téléphone portable afin par exemple d'adapter la musique de l'ipod à l'humeur du sujet ou de prévenir les rechutes d'anciens toxicomanes en détectant les signaux physiologiques de manque.

---

[1] Clore cite une expérience de Martin [Martin 1986] où des sujets, après avoir effectué une tâche qui induisait une réponse émotionnelle, lisaient une description ambiguë d'une personne et devaient ensuite juger cette personne. Martin a mis en évidence le fait que leur jugement était biaisé par leur première expérience affective.

[2] http://emotion-research.net/ws/plenary-2007/

[3] Ekman et Friesen ont codé les modifications de l'expression du visage en « FACS » (Facial Action Coding System), une « unité d'action » décrivant l'effet d'un muscle sur un trait du visage.

[4] http://affect.media.mit.edu/projects.php?id=1935

[5] http://www.media.mit.edu/research/ResearchPubWeb.pl?ID=30

### 1.1.2. Les émotions dans les interactions sociales : le modèle de Brunswik, encodage et décodage des émotions

Notre recherche s'appuie sur la base théorique du modèle de perception de Brunswik, modèle développé pour l'étude perceptive de la vision et appliqué à différents types d'analyse de jugement. Ce modèle a été adapté par Klaus Scherer [Scherer et al. 2003] comme paradigme pour la recherche sur la communication vocale des émotions. Scherer insiste sur la distinction entre l'expression (**ou encodage**) de l'émotion par le locuteur, la transmission du son et le décryptage par le receveur **(décodage)**.

Dans son modèle (voir Figure 1-1), les états internes d'un locuteur s'expriment par des modifications physiologiques (respiration, 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 interne 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. La partie gauche du modèle correspond à l'encodage, la droite au décodage.

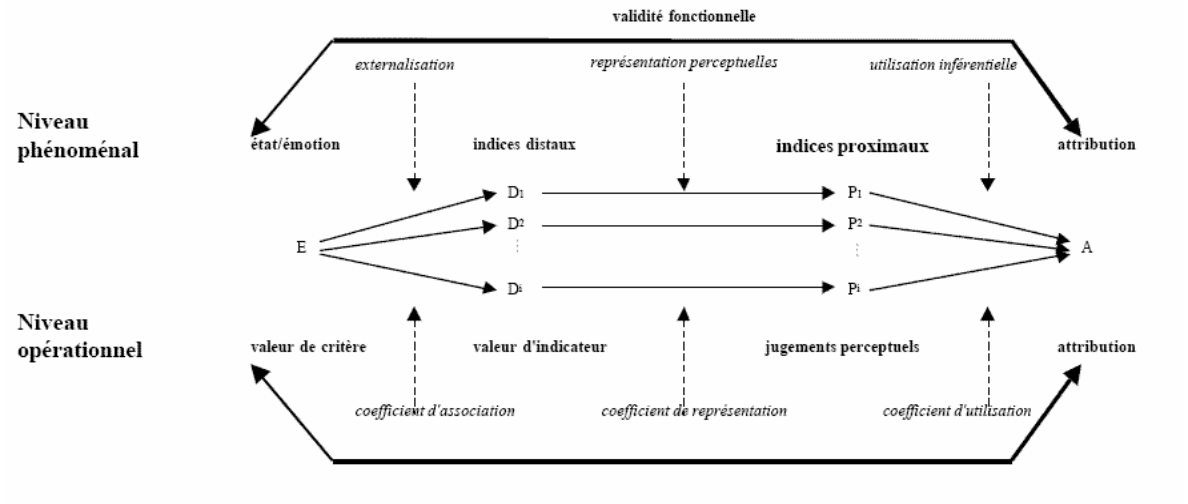



**Figure 1-1:** *le modèle de Brunswik adapté par Scherer.*

Une illustration est donnée dans [Scherer 2003] pour le cas de la fréquence fondamentale du signal.

> *«… the fundamental frequency of a speech wave constitutes the distal characteristics that gives rise to the pattern of vibration along the basilar membrane, and, in turn, the pattern of excitation along the inner hair cells, the consequent excitation of the auditory neurons, and finally, its representation in the auditory cortex. Either phase in the input, transduction and coding process could be considered a proximal representation of the distal stimulus"*

Même si les indices proximaux sont censés refléter les indices distaux, ils peuvent être modifiés ou déformés par la transmission du son (distance[13], bruit) et les caractéristiques structurelles de l'organe perceptif (plus de détails dans [Scherer 2003]).

Nos travaux portent sur la partie décodage du modèle, nous utilisons les caractéristiques de la voix pour inférer l'émotion de l'émetteur (exprimée volontairement ou non).

Comme le remarque Ortony dans [Ortony et al. 1988], de même qu'il n'y a aucun moyen de prouver qu'une personne est en train de percevoir une couleur précise, il n'y a pas de mesure objective connue pour établir qu'une personne est en train d'éprouver une émotion spécifique.

En pratique, une des méthodologies les plus utilisées est le "self-report" d'une émotion : on demande à une personne par exemple de se remémorer un épisode émotionnel ou on lui pose des questions après une expérience en lui demandant de décrire les émotions qu'elle a éprouvées. Cependant, même en supposant qu'il soit possible de questionner le locuteur, comme le remarque [Cornelius 1996 p13] :

> *Studies of emotion employing self-report methodologies assume, of course that people are able and willing to tell researchers what the researchers wants to know about their emotion. This, it turns out is a somewhat dodgy assumption to make and is one that has occasioned a great deal of controversy*

En effet, le temps peut avoir un impact sur les souvenirs ou la formulation même de la question pourrait d'ailleurs biaiser sa réponse. La personne pourra également amplifier ou inventer des émotions afin de satisfaire l'expérimentateur [Schachter et Singer 1962] [14]. Plutchik [Plutchik et Kellerman] donne d'ailleurs une liste de raisons pour lesquelles les reports verbaux ne décrivent pas nécessairement l'état émotionnel (voir Tableau 1-2).

---

[13] Si par exemple le receveur est situé physiquement loin de l'encodeur, il va devoir produire un signal plus intense, ce qui aura des répercussions sur les indices acoustiques.

[14] Dans les expériences de [Schachter et Singer 1962], des étudiants étaient mis dans des conditions supposées induire de l'exaltation (elation) et de la colère et devaient évaluer leurs sentiments de "joie" et "colère" sur des échelles. Dans la condition "colère", les sujets se sont plus notés comme « content» que comme « en colère ». A la fin de l'expérience, il s'est avéré que les sujets éprouvaient plus de colère que de joie, mais avaient peur de le dire car on leur avait promis 2 points de plus à leur examen final s'ils faisaient l'expérience.

| |
|---|
| Un observateur peut assumer de manière erronée qu'aucune émotion n'existe car aucune n'a été reportée. |
| La demande de rapport de l'émotion immédiate de quelqu'un pose le problème que le processus d'observation peut modifier l'objet étudié[15]. |
| Les rapports verbaux peuvent être des distorsions ou des vérités partielles pour des raisons conscientes ou inconscientes.[16]<br>En général ils sont rétrospectifs et dépendent donc de la mémoire. Les souvenirs peuvent être atténués, déformés ou réprimés par exemple.<br>On peut délibérément tromper une autre personne. |
| Les émotions pures sont rarement expérimentées. Typiquement une situation va générer des émotions complexes plus difficiles à décrire. Les rapports verbaux dépendent de l'histoire d'un individu et de sa facilité avec les mots. L'ambiguïté inhérente du langage pose également le problème du véritable sens du terme émotionnel |

**Tableau 1-2.** *Des arguments contre le « self report » des émotions (adapté de [Plutchik et Kellerman p4]).*

Même en considérant le « self-report » comme valide, il n'est pas toujours possible de demander aux locuteurs de verbaliser leurs émotions, surtout pour des données réelles du type enregistrements de conversations téléphoniques.

Des juges humains peuvent-ils reconnaître les émotions, et en particulier à partir du seul canal audio ? De nombreuses études ont tenté de donner une preuve affirmative empirique à travers des tâches types où des acteurs ou professionnels expriment différentes émotions que des juges/annotateurs essaient de reconnaître. Scherer [Scherer 1989] a passé en revue une trentaine de ces études qui s'accordaient sur des taux de discrimination plus de cinq fois supérieures aux taux du hasard. Mais ces études comprenaient de nombreux biais parmi lesquels le nombre restreint d'émotions prises en compte, avec souvent peu d'émotions positives et un manque de variabilité dans l'expression de ces émotions. Banse et Scherer ont essayé de traiter ce problème [Banse et Scherer 1996] en utilisant un large ensemble de stimuli avec 14 émotions, parfois de même classe du type colère chaude, colère froide, honte, exprimées par 12 professionnels et ont obtenu un taux de reconnaissance de 48%. Le taux variait selon l'émotion à reconnaître avec certaines expressions comme la colère chaude et l'ennui très bien reconnus alors que d'autres comme la honte étaient très mal reconnues malgré un profil acoustique distinct.

Il est également important de différencier des catégories de juges ou annotateurs, ce qui est rarement fait de façon claire dans la majorité des études en détection des émotions. Nous nous

[15] Lorsqu'on dit par exemple « Je ne suis pas en colère »

[16] Dans une étude sur des passagers aériens reportant des bagages perdus, [Scherer et Ceschi] décrivent comment certains passagers vont consciemment ou non décrire leurs sentiments différemment de leur véritable expérience, soit pour projeter une image stoïque en essayant d'apparaître impavide après la perte de leur bagage, soit au contraire en exagérant leur irritation afin de produire le comportement stéréotypé « normalement » attendu.

référons aux catégories définies en ISO standard 8566-2 pour définir les juges naïfs et experts [Soren et Zacharov 2006]. Un juge naïf est instruit de la procédure à suivre mais n'est pas entraîné pour la faire. La dénomination d'expert selon cette norme nécessite un apprentissage des juges, une évaluation de leur potentiel et une sélection des juges.

Kappas *et al.* [Kappas et al. 1991] soulignent également les différences de perception suivant que l'annotateur connaît ou non le locuteur. Une personne peut avoir naturellement une voix très tendue ou aigue qui entraînerait une mauvaise perception de son état émotionnel.

Dans une étude sur des données naturelles d'interactions dans un aéroport international entre des passagers dont les valises ont été perdues et les agents d'un aéroport, [Scherer et Ceschi 2000] ont comparé l'auto-annotation des états émotionnels des passagers (5 classes : Colère/Irritation, Inquiétude/Stress, « Bonne humeur », Résignation/Tristesse, Indifférence), leur annotation par les agents avec qui ils ont interagi et celle par des juges (étudiants en psychologie) disposant de la vidéo et de l'audio. Ils ont trouvé peu de corrélations entre le « self report » et l'annotation par les agents et juges. Les classes « Bonne humeur » et Inquiétude/Stress étaient bien corrélées, Résignation/Tristesse et Indifférence étaient corrélées au niveau du hasard. Même en admettant que les passagers aient été honnêtes dans le rapport de leur état interne, ils ont pu contrôler leur comportement et paroles lors de l'interaction avec l'agent pour masquer leur colère et au contraire délibérément non contrôler leur stress afin de susciter de l'empathie. Leur conclusion finale est que malgré la difficulté de la tâche, il est possible d'étudier des phénomènes émotionnels dans des conditions réalistes « sur le terrain ». Hess s'est intéressé à l'effet auditoire et avance que les expressions émotionnelles pouvait être comprises comme « des communications d'intentions, modulées par la présence des autres et indépendantes de l'état émotionnel concomitant » [Hess 2006]. Elle cite les travaux de Fridlund [Fridlund 1991] qui a montré que l'affichage des expressions faciales négatives et positives subit une augmentation en présence d'un public réel ou imaginé.

## 1.1.3. La représentation des émotions

### *Quatre courants théoriques sur les émotions*

Il existe quatre courants théoriques principaux en recherche émotionnelle, largement documentés et résumés dans le tableau suivant traduit de [Cornelius 1996], dont vont découler les différentes représentations.

| Tradition | Idée principale | Référence | Recherche plus contemporaine |
|---|---|---|---|
| Darwinienne | Les émotions ont des fonctions adaptatives qui sont universelles | [Darwin 1872] | [Ekman et Fridlung 1987] |
| Jamesienne | Emotion=réponse physiologique | [James 1884] | [Levenson et al. 1990] |
| Cognitive | Les émotions résultent d'un processus d'évaluations (appraisals) | [Arnold 1960] | [Smith et Lazarus 1993] |
| Constructivisme social | Les émotions sont le produit de constructions sociales | [Averill 1980] | [Smith et Kleinman 1989] |

**Tableau 1-3.** *Quatre théories des émotions en psychologie (d'après [Cornelius 1996] p12).*

### *Dimensions abstraites*

En 1957, le psychologue américain Osgood [Osgood et al. 1957], dans le but de décrire l'espace sémantique, a utilisé le *Roget's International Thesaurus* pour aider à la construction d'une cinquantaine d'échelles bi-polaires fondées sur des opposés sémantiques tels que "good-bad", "large-small", "beautiful-ugly", "hard-soft", "sweet-sour", "strong-weak" etc. Le résultat des recherches d'Osgood sur l'espace sémantique est l'existence de 3 dimensions universelles mesurables sous-jacentes aux dimensions émotionnelles : Evaluation (pleasant to unpleasant), Potency (in control to out of control) et Activity (calm to excited) aussi appelées EPA. Depuis lors, de nombreuses études internationales ont validé la réalité de cet espace sémantique et sa validité inter culturelle. Selon Osgood, les dimensions sont adaptées aux études inter cultures parce qu'il est difficile de traduire les étiquettes émotions quand on passe d'un langage à un autre (certaines émotions existent d'ailleurs dans certains langages et pas dans d'autres[17]).

[17] Wierzbiecka cite par exemple le mot russe *toska* (mélancolie tourment, angoisse), *zalet* (« to lovingly pity someone ») ou le concept Ifaluk *fago* (qui peut signifier simultanément tristesse/compassion/amour) [Wierzbicka 1999 p8] ou le mot allemand *Schadenfreude* (joie provoquée par le malheur d'autrui).

Le philosophe Spinoza a été probablement le premier à décrire les émotions à partir de 3 dimensions au dix-huitième siècle. Les émotions peuvent être plaisantes ou non plaisantes, fortes ou faibles et plus ou moins persistantes. Pour Wundt, les 3 axes plaisir/non plaisir, stress/relaxation, excitation/calme suffisent à placer de manière distincte tous les états émotionnels [Wundt 1896]. Schlosberg a proposé le modèle "circumplex" avec toutes les émotions placées sur la circonférence d'un cercle [Schlosberg 1941]. L'activation juge l'énergie avec une gradation allant de passif à actif. La valence va du déplaisir au plaisir. Ce modèle a eu une grande influence, bien qu'il ait été critiqué ([Lazarus 1991], [Larsen et al. 1992]) parce qu'il ne permettait pas de faire la différence entre certaines émotions : par exemple, la peur et la colère sont toutes les deux déplaisantes et très actives. De nombreuses études ont été réalisées depuis, le plus souvent avec deux axes ([Cowie et al. 2000], [Cacioppo et al. 2000], [Lang et al. 1997], [Carver 2001]) ou trois axes ([Russell et Mehrabian 1977], [Osgood et al. 1957], [Smith et Ellsworth 1985]). Les dimensions les plus fréquemment introduites [Ortony et al. 1988 p6] sont l'arousal et la valence (l'arousal pouvant être vue comme l'Activation d'Osgood, et la valence, comme un mélange de Potency et Activity). Une troisième dimension est le contrôle qui évalue l'aptitude d'un individu à gérer une situation. Une autre dimension est l'intensité de l'émotion.

La consistance des quatre dimensions « valence », « potency », « activation » et « unpredictability » a été prouvée pour le hollandais, le français, l'anglais et le chinois[Roesch et al. 2006].

Bernston a indiqué, lors de l'école d'été de HUMAINE en 2006, que pour le moment, les dimensions n'étaient pas suffisantes lorsqu'on cherche à représenter les émotions pour des situations du type « Je viens de gagner 2 dollars, mais j'aurai pu en gagner 10 » où la personne ressent à la fois de la joie et de la déception.

Il n'y a pas de contradictions entre les dimensions et des étiquettes discrètes [Ekman et Davidson 1994]. Ces deux représentations s'emploient dans des buts différents. Albrecht *et al.* remarquent d'ailleurs qu'il est possible de faire un mapping entre une catégorie d'émotion et l'espace dimensionnel[Albrecht et al. 2005]. Le contraire n'est pas possible. Aucun ensemble de dimensions ne permet cependant de capturer de manière adéquate les différences entre les émotions discrètes.

## *Théorie des émotions de base*

Selon cette théorie, les émotions ne sont pas fondamentalement similaires et simplement différentiables par une position sur différents axes, mais il existe un nombre fini d'**émotions de base** ou **émotions primaires**, chacune correspondant à un pattern/prototype bien défini (expressions comportementale, manifestations physiologique[18], antécédents …) et découlant à la base d'une fonction vitale[19].

Selon plusieurs de ces théories [Lazarus 1991], les émotions de base ont évolué de manière à s'adapter à certaines difficultés caractéristiques de la vie comme la concurrence (colère), le danger (peur), la perte (tristesse). Ekman[20] [Ekman 1992] donne neuf caractéristiques permettant de distinguer les émotions de base (voir Tableau 1-4). Il a prouvé, par des études sur les expressions du visage le caractère universel de six de ces émotions (Colère, Peur, Tristesse, Dégoût, Joie, Surprise).

| |
|---|
| elles sont universelles (expériences sur les expressions du visage reconnues par différentes populations) |
| elles existent chez d'autres primates et animaux[21] |
| elles s'appuient sur un contexte physiologique spécifique |
| elles se manifestent dans des contextes semblables. Il donne l'exemple de la perte d'un proche, qui est souvent l'antécédent de la tristesse. |
| il y a congruence/cohérence entre l'expérience émotionnelle et son expression |
| le déclenchement est rapide |
| la durée est brève |
| il y a un mécanisme d'évaluation qui peut être automatique |
| elles peuvent se manifester de manière inopportune. Parce que leur déclenchement est rapide, il n'est pas toujours possible de les inhiber. |

**Tableau 1-4.** *Les neuf caractéristiques des émotions de base selon [Ekman 1992].*

Ces émotions de base caractérisent des familles d'émotions [Ortony et al. 1988] avec des variations d'intensité à l'intérieur d'une même famille. Brenner [Brenner 1980] en donne un exemple pour la peur:

[18] Plusieurs études ont prouvé des "patterns" (patrons) distinctifs dans l'activité système nerveux autonome (ANS) pour la colère, la peur et le dégoût [Levenson et al. 1991]

[19] Pour Averill [Averill 1994] une émotion sera vitale pour la survie d'une espèce (point de vue biologique), d'une société (critère social) ou de soi-même (critère psychologique). Par conséquent, elles sont universelles, observables chez certains primates et héréditaires.

[20] La majorité des travaux de modélisation des émotions du visage sont basés sur les émotions de base définies par Ekman.

[21][Hebb 1972] cité par [Plutchik 1984 p6] : "The dog is definitely capable of jealousy and occasionally, in some dogs, there are signs of sulking. In the chimpanzee, however, we have the full picture of human anger in its three main forms : anger, sulking, and the temper tantrum"

> *"[A]nxiety is unpleasure accompanied by an expectation that something bad is going to happen […] Under the broad heading of anxiety, however, different terms are often used to indicate variations both in the intensity of the unpleasure that an anxious person experiences and in the nature of the conscious and unconscious ideas associated with it. If the danger is perceived to be acute or imminent, we are likely to label the affect "fear". If the unpleasure is intense we use the word "panic". If the unpleasure is mild and the danger is slight, uncertain, or distant, we may well speak of worry or uneasiness."*

Pour Brenner, pour définir les affects et les différencier, il suffit souvent de (a) spécifier si l'affect correspond à une expérience de plaisir ou de déplaisir et donner son intensité et (b) faire un lien avec l'idée qui lui est associée.

| | |
|---|---|
| *Anger* | L: a demeaning offence against me and mine<br>S: something interferes with the person's attainment of certain goals; a person perceives something as harming him in some way; the angry person makes the perception that the harm is illegitimate, situation is contrary to what ought to be |
| *Anxiety* | L: facing uncertain, existential threat<br>B: unpleasure accompanied by an expectation that something unpleasurable is going to happen |
| *Fright* | L: an immediate, concrete and overwhelming physical danger |
| *Fear* | S: interpretation of events as potentially dangerous or threatening to self |
| *Guilt* | L: having transgressed a moral imperative |
| *Shame* | L: failing to live up to an ego ideal |
| *Sadness* | L: having experienced an irrevocable loss<br>B: unpleasure connected with ideas that something bad already happened. |
| *Embarrassment* | [Parrott et Harré 1996]: Expression of the judgement that other people will think that something about us or something we have done is improper in the context. In displaying embarrassment we express a kind of apology for the real or imagined fault |
| *Envy* | L: wanting what someone else has |
| *Jealousy* | L: resenting a third party for the loss of, or a threat to, another's affection or favour |
| *Disgust* | L: taking in or being too close to an indigestible object or (metaphorically speaking) idea |
| *Happiness* | L: making reasonable progress toward the realisation of a goal<br>B: feeling of pleasure in connection with an experience or fantasy of instinctual gratification |
| *Pride* | L: enhancement of one's ego-identity by taking credit for a valued object or achievement, either one's own or that of someone or group with whom one identifies |
| *Relief* | L: a distressing goal-incongruent condition that has changed for the better or gone away |
| *Hope* | L: fearing the worst but wanting better |
| *Love* | L: desiring or participating in affection, usually but not necessarily reciprocated |
| *Compassion* | L: being moved by another's suffering and wanting to help |

**Tableau 1-5.** *Emotions & their core relational theme (d'après L : [Lazarus 1998] ,B : [Brenner p345] S :Shaver et al).*

L'ensemble minimal d'émotions primaires varie suivant les chercheurs (voir Tableau 1-6 inspiré de Ortony). Cependant, les émotions joie, peur, colère, tristesse, dégoût et surprise se retrouvent dans une majorité des études[22].

| | Joie | Tristesse | Dégoût | Peur | Colère | Surprise | Aversion | Courage | Découragement | Désir | Désespoir | Haine | Amour | Espoir | Intérêt | Mépris | Culpabilité | Honte | Rage | Terreur | Anxiété | Chagrin (grief) | Emerveillement | Bonheur | Confiance | Apathie | Peine (sorrow) | Détresse |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Darwin | + | + | + | + | + | | | | | | | | | | | | | | | | | | | | | | | |
| Arnold (1960) | | + | | + | + | | + | + | + | + | + | + | + | + | | | | | | | | | | | | | | |
| Izard (1971) | + | + | + | + | + | + | | | | | | | | | + | + | + | + | | | | | | | | | | |
| Plutchik (1980) | + | + | + | + | + | + | | | | | | | | | | | | | | | | | | | + | + | | |
| Tomkins (1980) | + | | + | + | + | + | | | | | | | | | + | + | | + | | | | | | | | | | + |
| Ekman, Friesen & Ellsworth (1982) | + | + | + | + | + | + | | | | | | | | | | | | | | | | | | | | | | |
| Gray (1982) | + | | | | | | | | | | | | | | | | | | + | + | + | | | | | | | |
| James (1884) | | | | + | | | | | | | | | + | | | | | | + | | | + | | | | | | |
| Oatley & Johnson Laird (1984) | | + | + | | + | | | | | | | | | | | | | | | | + | | | + | | | | |
| Frijda (1986) | | | | | | + | | | | + | | | | | + | | | | | | | | + | + | | + | | |

**Tableau 1-6.** *Les émotions de base, d'après [Ortony et Turner 1990].*

[22] Dans une étude [Fehr et Russell 1984], réalisée auprès de 200 personne à qui il était demandé de lister en une minute un maximum d'items de la catégorie 'EMOTION', les émotions citées par plus de 40% des sujets étaient « Bonheur », « Colère », « Tristesse », « Amour », « Peur », « Haine » et « Joie »

## *Les émotions complexes*

Dans la vie réelle, les émotions pures ne sont pas fréquentes. Quel est le lien entre les émotions de base et toutes les autres émotions ? La plupart des théories ont une approche combinatoire. Plutchik parle de "mixed states", de "dyads" ou de "triads" d'émotions primaires [Plutchik 1962]. De même Averill parle de "compounds", émotions composées de plusieurs émotions primaires[Averill 1975]. Les manières dont les émotions primaires se mélangent sont variées : pour Ekman et Plutchik, les deux émotions de bases sont observables dans le mélange[23]. Ekman suggère cependant que les mélanges peuvent refléter une séquence très rapide de deux émotions de base. Pour Averill seule la résultante du mélange est observable[24].
Izard [Izard 1972] a défini des "patterns" d'émotions comme au moins deux émotions fondamentales exprimées simultanément ou très rapidement l'une après l'autre.
*"Discrete emotions retain their essential genotypical characteristics when they occur in patterns"* [Izard 1972 p24] Pour lui par exemple l'anxiété est un mélange de Peur avec une autre émotion discrète.

Averill [Averill] décrit les émotions conflictuelles ("conflictive émotions") :

> *"Conflictive emotions are […] like conversion reactions in the Freudian sense […]. In the case of a conversion reaction, the individual wishes to engage in some behavior that conflicts with personal norms or standards […] [S]tandard conflictive emotions can be viewed as conversion-like phenomena on a socialcultural as opposed to an individual level of analysis"*

Brenner [Brenner 1980] sur « les mixed affects » :

> *"[T]here are also affects characterized by a mixture of feelings of pleasure and unpleasure as well as by ideas that include various combinations of good and bad expectations. Such affects are rather the rule than the exception."*

[23] Ortony dans [Ortony et al. 1988] donne l'image d'un mélange de sucre et de sel, les 2 goûts étant alors perceptibles.
[24] De même que ni l'oxygène, ni l'hydrogène ne sont observables dans l'eau [Ortony et al. 1988].

## Représentation des émotions de base et des émotions complexes avec le modèle de Plutchik

Plutchik a cherché comment représenter l'ensemble des émotions. Pour lui [Plutchik 1984], il existe un ensemble d'émotions de base. Il en identifie huit prototypiques : peur/terreur, colère/rage, tristesse/chagrin (grief), acceptation/confiance, dégout/répugnance (loathing), espoir/anticipation et surprise/étonnement (astonishment). Afin de tenir compte de toutes les étiquettes émotionnelles, il lui a fallu trouver un moyen d'organiser les émotions les unes par rapport aux autres. Tout d'abord, les émotions varient en intensité, ensuite certaines émotions sont plus proches que d'autres. Par exemple, la honte et la culpabilité sont plus proches l'une de l'autre que le dégoût et la joie. Enfin, il voit également des polarités dans les émotions avec des émotions opposées comme joie/tristesse. Plutchik [Plutchik 1984 p200] utilise la métaphore d'une palette de couleurs pour faire une distinction entre des émotions fondamentales ou primaires et d'autres dérivées ou secondaires : à partir d'une base de trois couleurs primaires et des variations d'intensité, toutes les couleurs observables dans la nature peuvent être représentées ; le même principe peut s'appliquer aux émotions. Il modélise les relations entre les huit émotions de base par un "solide émotion" représenté Figure 1-2.

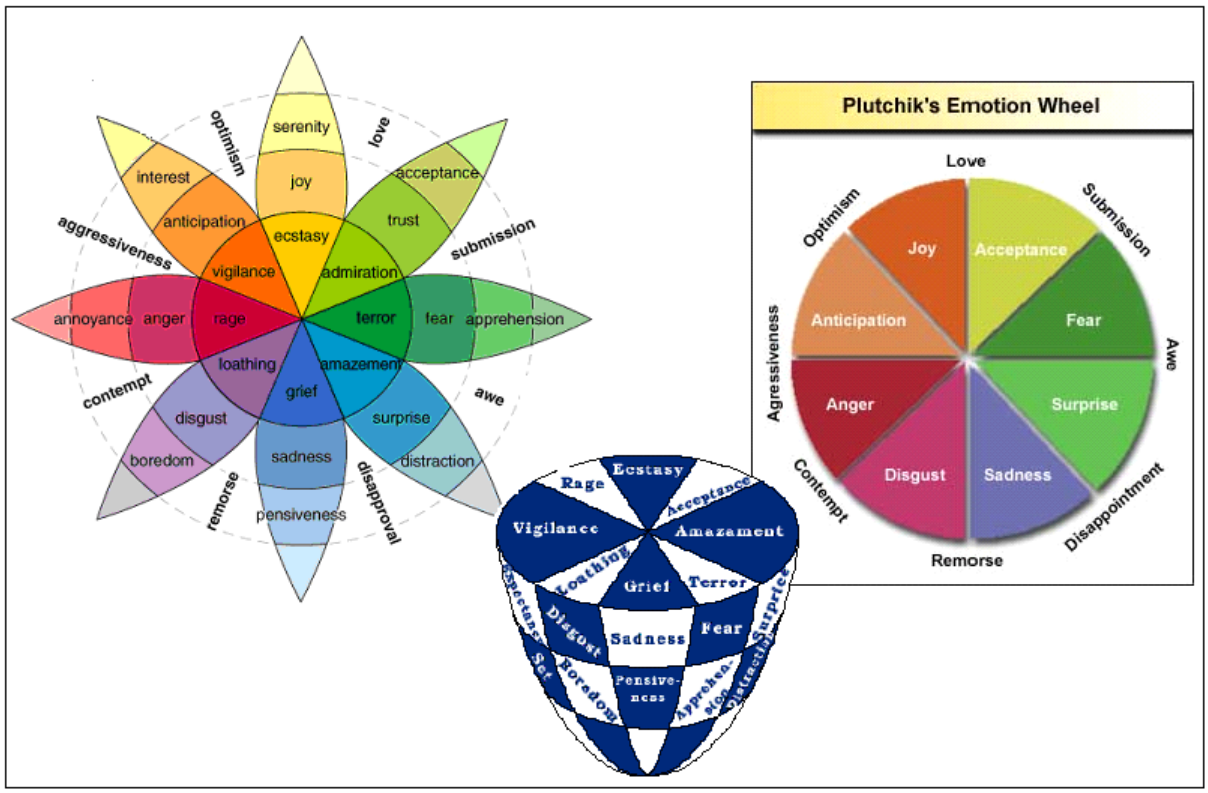



**Figure 1-2.** *"Solide émotion" de Plutchik. (de [Plutchik 1984]).*

Les émotions les moins intenses sont en bas du solide. Elles deviennent de plus en plus intenses et de plus en plus différenciées quand on va vers le haut. Chaque "tranche" représente une émotion de base. Les émotions complexes se situent au niveau des frontières entre deux tranches. Par exemple, l'amour est un mélange d'acceptation et de sérénité.

## Une étude sur les mélanges d'émotions dans des données réelles : Lost luggage

Malgré les nombreuses théories sur les mélanges d'émotions, peu d'études ont été réalisées sur le sujet. Scherer a cherché des méthodes pour les étudier à travers 3 tâches [Scherer 1998]. En particulier, il a filmé et enregistré 112 passagers rapportant la perte de leurs bagages dans un aéroport international, puis les a interviewés en leur demandant d'évaluer leur état affectif, avant et après l'entretien avec l'agent de l'aéroport, sur une échelle de 1 à 5 pour les catégories d'émotion Colère/Irrité=*Colère*, Résigné/Triste=*Résignation*, Indifférent, Worry/Stressé=*Worry* et de bonne humeur. Il a d'abord essayé de regrouper les différents mélanges en cluster, mais le nombre de classes obtenues était trop important pour pouvoir être analysé. Il a alors regroupé les classes « indifférent » et « de bonne humeur » en une classe « good spirit» et a ensuite divisé les résultats en état émotionnel « dominant » lorsqu'une émotion était ressentie avec plus d'intensité que les autres et « blend » *Colère/Worry*, *Résignation/Worry*, *Colère/Résignation* sinon. Il a analysé l'évolution des « blends » au cours de l'interaction, mais sa conclusion générale était que les réponses émotionnelles indiquées par les passagers étaient trop riches et complexes pour pouvoir être étudiées.

## *Modèle d'évaluation (appraisal)*

Le mot "appraisal" a été employé pour la première fois[25] par Magda Arnold [Arnold 1960]. Elle soutenait que l'on évalue en permanence l'impact des changements de notre environnement sur notre bien-être et que de ces évaluations jouent un rôle dans l'apparition et la différentiation des émotions. Pour elle, l'émotion naît quand un évènement est jugé comme nuisible ou bénéfique et des émotions différentes apparaissent parce que des événements sont jugés de différentes manières.

Différentes théories existent quant aux dimensions d'évaluations les plus importantes. Lazarus a passé en revue les plus importantes[26] [Lazarus 1998 p358]. Parmi les dimensions sur lesquels nombreux chercheurs s'accordent, on trouve

- une *composante motivationnelle* : pour éprouver une émotion, il faut avoir un but.
- La *valence* de l'émotion qui dépend souvent des conditions favorables ou non à la réalisation du but
- une dimension reliée à la *responsabilité de soi et des autres* : le fait qu'un tort ou un bénéfice nous soit attribué conduit a des sentiments de fierté, honte ou culpabilité alors que la responsabilité en bien ou mal des autres conduit à des sentiments de colère par exemple.

Une seule évaluation ne peut pas justifier un état émotionnel. Par exemple pour la colère :

> *"Not only does the subject have to feel thwarted, his/her self esteem has to have been demeaned, responsibility has to have been attributed, and the responsible person has to have been presumed in control of his/her actions" [Lazarus 1998 p358]*

[25] Le premier théoricien cognitiviste des émotions est en fait Aristote qui, de manière étonnement moderne, définit entre autres la colère dans *Rhétorique* (Rethorique II 1378a), comme "un désir de vengeance accompagné d'une peine provoquée par ce qui semble un dédain injuste […] Pour la colère par exemple, en quel habitus y est-on porté ; contre quelles personnes se met on habituellement en colère et à quels sujets"

[26] Théories de Frijda, Lazarus, Reisenzein, Roseman, Scherer, Smith & Ellsworth …

**Le modèle des processus-composantes de Scherer pour les émotions vocales**

Dans le cas des émotions vocales, le modèle de référence est celui de Scherer [Scherer et al. 2003]. Selon Scherer, les émotions vocales, de par leur aspect dynamique et changeant imposent de s'éloigner de la conception statique et figée des émotions (émotions de base), en prenant en compte le contexte. Son modèle stipule que la réaction émotionnelle est le résultat d'une séquence de processus d'évaluations de l'événement inducteur de l'émotion (voir Tableau 1-7). Cette séquence d'évaluations, dénommée « séquence de traitement de la stimulation » (stimulus evaluation checks ou SEC) dans le modèle de Scherer est récursive et se fait en boucle.

| **Séquence de traitement de la stimulation** | |
|---|---|
| **Nouveauté** | |
| Soudaineté<br>Familiarité<br>Prévisibilité | *Quelle l'est importance de l'évènement ? Est il connu ou au contraire nouveau ?* |
| **Agrément intrinsèque** | |
| **Rapports aux buts** | |
| Pertinence<br>Degré de certitude de la prédiction des conséquences<br>Attente<br>Opportunité<br>Urgence | *L'événement va-t-il à l'encontre de mes buts ou au contraire les favorise-t-il ?* |
| **Potentiel de maitrise** | |
| Causalité : interne<br>Causalité : externe<br>Contrôle<br>Puissance<br>Ajustement | *Est ce que l'individu peut maîtriser la situation ?* |
| **Accord avec les standards** | |
| Standards externes<br>Standards internes | *La réaction sera différente suivant que l'évènement sera jugé comme moral ou non.* |

**Tableau 1-7**. *Critères d'évaluation des séquences de traitement dans le modèle de Scherer (extrait de [Scherer et Sangsue 2006 p20]) .*

Les "étiquettes de bases" peuvent être détaillées/explicitées par ce processus d'évaluation. Par exemple une situation peu contrôlable va entraîner une réaction émotionnelle du type peur.

Scherer [Scherer 1986] a étudié des répercussions physiologiques, de qualité vocale et sur certains paramètres acoustiques (F0, formants) de la voix des résultats de ces évaluations qui sont

détaillées dans le **Tableau 1-8**. Il différencie notamment la colère froide de la colère chaude. En outre, il observe que plusieurs manifestations de la même "émotion de base" résultent en fait d'évaluations très différentes ce qui conduit à des manifestations très variées, et s'oppose à l'idée que chaque émotion fondamentale correspondrait à un pattern bien défini. Il donne en exemple une étude de Frick [Frick 1986] sur deux types de colère, une liée à la frustration et l'autre à l'agression qui se manifestent différemment sur le plan acoustique et sont différentiables perceptivement.

| Criterion | ENJ/HAP | ELA/JOY | DISP/DISG | CON/SCO | SAD/DEJ | DESPAIR | ANX/WOR |
|---|---|---|---|---|---|---|---|
| Relevance | | | | | | | |
| Novelty | | | | | | | |
| Suddenness | Low | High/med | Open | Open | Low | High | Low |
| Familiarity | Open | Open | Low | Open | Low | Very low | Open |
| Predictability | Medium | Low | Low | Open | Open | Low | Open |
| Intrinsic pleasantness | High | Open | Very low | Open | Open | Open | Open |
| Goal/need relevance | Medium | High | Low | Low | High | High | Medium |
| Implication | | | | | | | |
| Cause : agent | Open | Open | Open | Other | Open | Oth/nat | Oth/nat |
| Cause : motive | Intent | Cha/int | Open | Intent | Cha/neg | Cha/neg | Open |
| Outcome probability | Very high | Very high | Very high | High | Very high | Very high | Medium |
| Discrepancy from expectation | Consonant | Open | Open | Open | Open | Dissonant | Open |
| Conclusivenes | Conducive | Vcon | Open | Open | Obstruct | Obstruct | Obstruct |
| Urgency | Very low | Low | Medium | Low | Low | High | Medium |
| Coping potential | | | | | | | |
| Control | Open | Open | Open | High | Very low | Very low | Open |
| Power | Open | Open | Open | Low | Very low | Very low | Low |
| Adjustment | High | Medium | Open | High | Medium | Very low | Medium |
| Normative significance | | | | | | | |
| Internal Standards | open | Open | Open | Very low | Open | Open | Open |
| External Standards | Open | Open | Open | Very low | Open | Open | Open |

| Criterion | FEAR | IRR/COA | RAG/HOA | BOR/IND | SHAME | GUILT | PRIDE |
|---|---|---|---|---|---|---|---|
| Relevance | | | | | | | |
| Novelty | | | | | | | |
| Suddenness | High | Low | High | Very low | Low | Open | Open |
| Familiarity | Low | Open | Low | High | Open | Open | Open |
| Predictability | Low | Medium | Low | Very high | Open | Open | Open |
| Intrinsic pleasantness | Low | Open | Open | Open | Open | Open | Open |
| Implication | | | | | | | |
| Cause : agent | Oth/nat | Open | Other | Open | Self | Self | Self |
| Cause : motive | Open | Int/neg | Intent | Open | Int/neg | Intent | Intent |
| Outcome probability | High | Very high | Very high | Very high | Very high | Very high | Very high |
| Discrepancy from expectation | Dissonant | Open | Dissonant | Consonant | Open | Open | Open |
| Conclusivenes | Obstruct | Obstruct | Obstruct | Open | Open | High | High |
| Urgency | Very high | Medium | High | Low | High | Medium | Low |
| Coping potential | | | | | | | |
| Control | Open | High | High | Medium | Open | Open | Open |
| Power | Very low | Medium | High | Medium | Open | Open | Open |
| Adjustment | Low | High | High | High | Medium | Medium | High |
| Normative significance | | | | | | | |
| Internal Standards | Open | Open | Open | Open | Very low | Very low | Very high |
| External Standards | Open | Low | Low | Open | Open | Very low | High |

**Tableau 1-8.** *Les évaluations prédites pour les émotions les plus étudiées. ENJ/HAP, contentement/bonheur; ELA/JOY, joie; DISP/DISG, déplaisir/dégout ;CON/SCO, mépris; SAD/DEJ, tristesse/abattement; IRR/COA, irritation/colère froide; RAG/HOA, rage/colère chaude; BOR/IND, ennui/indifférence; de[Sander et al. 2005 p. 326].*

# 1.2. La détection des émotions dans la voix

## 1.2.1. Méthodologie pour construire un système de détection des émotions

La méthodologie pour classifier les émotions est illustrée dans la Figure 1-3 et chaque étape sera détaillée dans les différents chapitres de la thèse.

(a) Choisir un corpus de données (voir chapitre 2). On peut demander à des acteurs de jouer des émotions particulières, construire des systèmes pour induire des émotions ou utiliser des données plus ou moins naturelles et spontanées.

(b) Etiquetage de ces données : si les données sont actées, il est nécessaire de vérifier que l'émotion a été bien jouée ; si au contraire on travaille sur des données plus naturelles, il faut mettre en place un protocole pour annoter les données. L'unité d'annotation pourra aller du mot au tour de parole et plusieurs stratégies dérivées des différents modèles d'émotions seront décrites. A la fin de cette étape, on sélectionne un sous corpus de données étiquetées pour entraîner des modèles computationnels.

(c) Extraction d'indices : elle peut se faire à différents niveaux (partie voisée, mot, au tour de parole, à intervalle régulier ou non,…). Une multitude d'indices peuvent être extraits au niveau local ou global, le plus souvent linguistiques et paralinguistiques, mais pouvant également inclure des indices contextuels, des actes de dialogue …

(d) Classification : elle suit souvent une phase de sélection des meilleurs attributs. De nombreux algorithmes peuvent être utilisés ou mélangés.

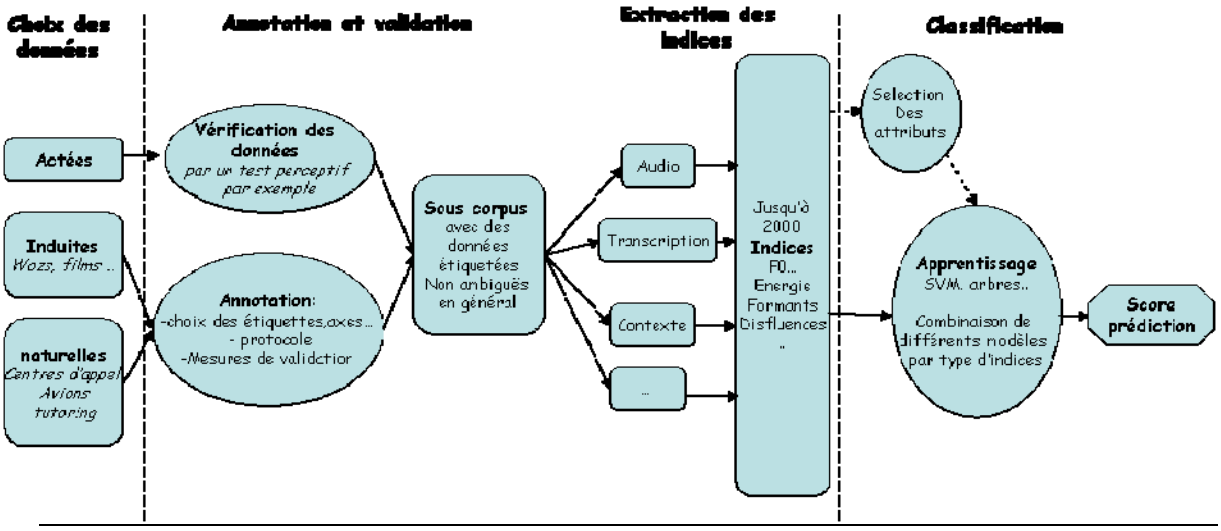



**Figure 1-3.** *Méthodologie pour construire un système de détection des émotions.*

### 1.2.2. Perception des émotions : Les performances humaines

D'après plusieurs études, parmi lesquelles celles de Scherer [Scherer 1986], pour la détection automatique des émotions, il faut viser des taux de reconnaissance de l'ordre de ceux des humains.

Scherer a évalué les performances humaines pour la reconnaissance de 6 émotions (Colère, Peur, Joie, Tristesse, Dégoût, Etat Neutre) dans la voix et le visage en s'appuyant sur un ensemble d'études comparables réalisées avec des acteurs [Scherer 2003]. Les taux globaux de reconnaissance pour la voix se situent entre 55% et 65% avec de grandes variations suivant les émotions étudiées, la colère et la tristesse étant les mieux reconnues avec des scores compris entre 70 et 80%, puis la peur et la joie autour de 60% et enfin le dégoût à 31%. De même Lee *et al.* arrivent à un taux de reconnaissance de 68.7% avec 4 juges ayant à distinguer 4 émotions jouées par des acteurs professionnels : Colère, Joie, Etat Neutre, Tristesse et constatent principalement des confusions Colère/Joie et Neutre/Tristesse [Lee et al. 2004] ; Petrushin constate 63.5% d'accord entre 23 observateurs naifs pour les 5 états Normal, Joie, Colère, Tristesse, Peur avec moins de variations pour la peur et la colère [Petrushin 1999].

Oudeyer [Oudeyer 2003] cite deux expériences où des japonais et des américains devaient reconnaître les émotions dans des stimuli sans information sémantique prononcés par des locuteurs des deux langues. Il avait peu de différences de performances entre les 2 langues, mais les scores de reconnaissance étaient de l'ordre de 60%. Il souligne lui aussi le fait qu'on ne peut s'attendre à une reconnaissance parfaite des émotions, mais qu'il est réalisable de viser des performances de l'ordre des performances humaines.

### 1.2.3. Etat de l'art des systèmes de détection sur les émotions dans la voix

Quel est l'état de l'art actuel des systèmes de détections des émotions ? A cause de la difficulté de la tâche de catégorisation et d'annotation et souvent le manque de données, la plupart des études se sont focalisées sur un ensemble minimal d'émotions comprenant des émotions : positives et négatives [Lee et al. 2001] ou émotion vs. état neutre [Batliner et al. 2003]. Certains autres chercheurs considèrent des comportements ou attitudes dépendantes de la tâche ; stressé vs. non stressé [Petrushin 1999], [Narayanan 2002], [Fernandez et Picard 2003]; frustration/colère vs. neutre/amusé [Ang et al. 2002], ou colérique, maternel, emphatique et un état neutre [Steidl et al. 2005].

Dans la communauté scientifique, les modèles les plus souvent utilisés pour la détection des émotions sont les SVM (Support Vector Machine détaillés p.127), les GMMs (Mélange de Gaussiennes), les kNN (K plus proches voisins), les arbres de décision. Les résultats obtenus entre ces différents classificateurs sont souvent comparables [Batliner et al. 2006] et il n'y a pas de consensus sur le choix d'un algorithme précis ou sur les meilleurs paramètres. Ceux ci semblent en effet être dépendants des données.
Plusieurs études rapportant les expériences de détection automatique sont listées dans le Tableau 1-9. Elles donnent une idée (non exhaustive) des tendances actuelles. Extrêmement peu d'études sont menées avec des données issues de corpus enregistrés dans des contextes réels.
La majorité des travaux sur la détection d'émotion de base porte sur des données jouées par des acteurs (peu de locuteurs, nombre d'échantillons restreint, peu de variabilité, données prototypiques). Sur un ensemble de plus 100 études, Juslin *et al.* en répertorient 87% sur des données actées[Juslin et Laukka 2003]. Pour faire des systèmes utilisables il faut pourtant travailler sur des données naturelles en situation (contexte).

Il est important de noter que les performances sont difficilement comparables car elles varient par exemple en fonction :

- Du type de données (émotions en général plus prototypiques pour les données actées d'où de meilleures performances. *Cf.* p33)
- De l'unité utilisée pour les annotations [Batliner et al. 2007].
- Du choix des tours sélectionnés pour entraîner et tester les modèles. Batliner donne l'exemple d'un vote majoritaire avec 5 annotateurs. On peut choisir de garder les segments/mots/tours pour lesquelles 3 annotateurs sur 5 sont d'accord, ou ceux pour lesquels 4 sur 5 sont d'accords et les performances seront supérieures dans le deuxième cas de figure.
- De la manière dont les paramètres sont obtenus. Corriger manuellement la F0 par exemple donne de meilleurs résultats. Nous décrirons d'ailleurs dans le chapitre 6 p151 un protocole pour nommer et décrire les paramètres qui permet de mieux les comparer et d'en avoir une description plus transparente.

Juslin et Laukka [Juslin et Laukka 2003] proposent d'utiliser un score prenant en compte le nombre de classes et les biais liés au jugement du décodeur.

| Référence Auteur | Style de corpus | Taille du Corpus | Etiquettes Emotions | Type de paramètres | Algorithme Modèle | Taux de Détection |
|---|---|---|---|---|---|---|
| [Dellaert et al. 1996] | Acteurs | 1000 Tours (5 acteurs) | Joie, Peur, Colère, Neutre | Prosodiques (Pitch contour) | MLB, KR, kNN | 60-65% (acté, 4 classes) |
| [Petrushin 1999] | Acteurs | 700 Tours | Colère, Tristesse, Peur, joie, neutre | Prosodiques | NNs | 70% (acté, 5 classes) |
| | Acteurs (non professionnels) | 56 appels (15 - 90s) | Calme, Agitation | Spectraux (F1, F2, F3) | | 77% (acté, 2 classes) |
| [Batliner et al. 2003] | Acteurs | 96 Tours (E) | Emotion (E), Non émotion (NE) | Prosodiques Spectraux Part-of-speech Dialogiques | NNs | 95% (acté, 2 classes) |
| | Parole lue | 50 Tours(E), 50 (NE). | | | | 79% (lu, 2 classes) |
| | WOZ Vermobil | 2395 Tours (20 dial.) | | | | 73% (Woz, 2 classes) |
| [Ang et al. 2002] | DHM Communicator | 21kTours (~3500 Tours Emotion) | Frustration, Agacement, Autre | Prosodiques | CART (Arbre de décision) | 75% (DHM, 2 classes) |
| | | | | Langage 3-gram | | 60-65% (DHM, 2 classes) |
| [Lee et Narayanan 2004] | DHM Callcenters (real-life) | 1187 appels | Négatif, Non Négatif | Prosodiques | LDC | 80% |
| | Speech-Works | 7200 Tours | | Lexicaux | | 88% (DHM, 2 classes) |
| | | | | Pros + Lex | | 93,5% |
| [Shafran et al. 2003] | DHM AT&T How May I Help You | 5147 Tours | Négatif, Non Négatif | Spectral (MFCC) F0 | HMM | 76 % (DHM, 2 classes) |
| | | | | Lexical | SVM | 81 % (DHM, 2 classes) |
| [Forbes-Riley et Litman 2004] | DHH Machine-Mediated | 385 Tours (90 Neg, 15 Pos, 280 Neu) | Positif, Négatif, Neutre | Prosodiques Lexicaux Dialogiques | AdaBoost | 84% (DHH-M, 3 classes) |
| [Steidl et al. 2005] | DHM AIBO | ~6000 mots (51 enfants) | Colère, Maternel, Emphatique, Neutre | Prosodiques Part-of-speech | NNs | 60% (DHM,4 Classes) |

**Tableau 1-9.** *Tableau récapitulatif d'études sur la détection des émotions : référence de l'auteur, style de corpus de travail (acté, Woz, DHH : dialogue Homme-Homme, DHM : dialogue Homme-Machine), size (Tours de parole) et nombre de locuteurs, Les représentations des émotions. Corpora Emotion labels, Type d'indices (Spectraux, Prosodique (Fréquence Fondamentale, Energie, Débit), Disfluences, Lexiques, Langage (n-gram), Syntax/semantic (Etiquettes : Part-of-Speechs) et enfin Dialogique), modèle d'apprentissage (MLB: Maximum Likelihood Bayes Classifier, KR: « Kernel Regression », LDC: « Linear discriminant classifier, » kNN: k Nearest-Neighbors, SVM: Support Vector Machine, HMM: Hidden Markov Model, NNs: Neural Networks, decision trees, Adaboost, etc), et finalement le taux de détection.*

## 1.3. Conclusion de l'état de l'art

Bien que différents courants existent quant à la définition précise d'une émotion et qu'on lui substitue de plus en plus le terme « état affectif », on peut extraire de leur étude des connaissances robustes[27]. On s'accorde aussi sur le fait que les humains et a fortiori les machines sont capables d'inférer de l'état émotionnel d'un autre humain (même si avec plus ou moins de finesse et de manière non infaillible).

Les mêmes problématiques se posent pour toutes les études sur la détection des émotions orales : quel matériel utiliser ? Comment annoter les données émotionnelles ? Quels indices extraire ? Quel méthode ou algorithme sont les plus appropriés pour la détection ?

Au commencement de la thèse, la majorité des travaux sur la détection d'émotion abordaient déjà la difficulté d'extraire des indices pertinents et de trouver des algorithmes d'apprentissage efficaces, mais peu se focalisaient sur les étapes préalables à l'obtention de données étiquetées en émotion. En effet la majorité des études portait sur des données jouées par des acteurs (Sur un ensemble de plus 100 études, Juslin et Laukka en ont répertorié 87% sur des données actées [Juslin et Laukka 2003]) et toutes les critiques convergeaient sur le fait qu'il fallait travailler sur des données spontanées pour pouvoir un jour construire des systèmes affectivement intelligents. La première étape est donc de collecter des données riches et en quantité suffisantes, ce qui est décrit en détails dans le chapitre 2.

[27] Pour illustrer ce point Rosalind Picard cite McCarthy [Picard 1997 p21]:« we can't define Mt Everest precisely _whether or not a particular rock or piece of ice is or isn't part of it ; but it is true, without qualification, that Edmund Hillaryand Tenzing Normay climbed it in 1953. In other words, we can base solid facts on structures that are themselves imprecisely defined"

# II Travailler sur des données spontanées

# Chapitre 2

# Les corpus émotionnels


Résumé

Dans la première partie de la thèse, nous avons longuement insisté sur l'importance du choix des données avant de pouvoir faire des expériences de détection. Plusieurs types de données peuvent être utilisés : données « actées » jouées par des acteurs, données induites ou données naturelles. Pourquoi choisir des données réelles ? Quels sont les avantages et inconvénients des différents types de données ?

Dans ce chapitre, nous présentons également les données LIMSI, et en particulier 2 corpus, sur lesquels nous nous sommes appuyés pour nos travaux. Toutes sont des données réelles issues de centres d'appel. Le premier Corpus contient des appels provenant de transactions boursières. Il a été analysé et annoté préalablement au début de ma thèse et contenait 13% de données émotionnelles, souvent peu intenses. La plupart des expériences ont été effectuées sur le deuxième corpus, CEMO pour *Corpus Emotion*, issu d'un centre d'appel médical dont j'ai suivi le traitement. Vingt heures de données ont été manuellement transcrites et le signal a été aligné phonétiquement à la transcription. Le corpus contient près de 30 % de données émotionnelles. Pour chaque dialogue, des métadonnées sur les locuteurs du type âge, sexe, accent, type de voix ont également été annotées.

*In the first chapter of the thesis, we have pointed out the importance of choosing the right type of data in order to study emotions. When studying emotion in speech, we can make use of different data types: acted speech, induced data or natural data. Why choose real-life data? In a first part, we will present the advantages and drawbacks of the different types of data. Then we will describe the LIMSI corpora that we based our work on. All data were recorded in French call centers. The first corpus contains data recorded in a stock exchange customer service and was annotated and analysed prior to my phd. It contains 13% emotional speech, mostly not very intense. Most of our experiments were made on the CEMO corpus (20 hours) recorded in a medical call center and I co-supervised its processing. It is a lot richer emotionally with 30% of emotional data. It was transcribed manually and for each dialog, metadata on the age, gender, accent and voice type was also annotated.*

# 2.LES CORPUS EMOTIONNELS

## 2.1. Quel matériel ? Les différents types de corpus[28] : avantages et inconvénients

La première problématique lorsqu'on étudie les émotions est de se procurer ou de créer un corpus de données. Plusieurs types de données peuvent être utilisées et leurs avantages et inconvénients seront discutés dans ce chapitre. La plupart des expériences sur les émotions ont été effectuées sur des données « actées » enregistrées par des acteurs avec souvent peu de classes émotionnelles (Neutre, Négatif, Positif ou les « émotions de base » de Ekman (*cf.* Chapitre 1).

### *Données actées*

Les premières études et la plupart des études actuelles ont été réalisées sur des données actées [Dellaert et al. 1996]. Ces données présentent plusieurs avantages. Elles ne soulèvent pas de problèmes éthiques, elles sont faciles à collecter et elles permettent de s'appuyer sur une grande quantité de données étiquetées pour chaque classe d'émotions (en pratique ce n'est cependant pas souvent le cas). De plus, elles rendent possible la comparaison de segments avec un contenu linguistique identique, ce qui permet d'attribuer les différences de perception aux seuls indices acoustiques. Cependant, elles s'avèrent insuffisantes pour représenter la réalité et le manque de contexte et le nombre réduit d'acteurs (souvent moins de 10) font que ces corpus de données contiennent moins de variabilité que les corpus de données spontanées. En effet, les acteurs utilisent souvent des stéréotypes caractéristiques de l'émotion et qui sont très différents de la véritable expression de celle-ci. Ils auront tendance à accentuer les codes sociaux de communication (effet « pull »)[29], mais l'effet "push", normalement associé à la relation physiologique émotionnelle, sera absent. Notons cependant que dans le cas des données GEMEP par exemple, différents scénarios permettant d'éliciter des émotions ont été donnés aux acteurs afin qu'ils jouent des émotions moins prototypiques.

[28] Nous entendons par corpus est un ensemble de données recueillies pour un sujet d'étude. Un corpus est souvent un recueil de données annotées.

[29] Par exemple [Williams et Stevens 1972] ont comparé le commentaire radio de la catastrophe aéronautique de Hindenburg et un acteur le simulant et ont trouvé une augmentation de la plage et du médian de la F0 mais beaucoup plus prononcé chez l'acteur.

De plus, il n'y a pas toujours d'évaluation de la qualité des émotions exprimées par des acteurs. Si lorsqu'on demande à un acteur d'exprimer de la colère, la phrase est étiquetée automatiquement *Colère* sans aucune validation, alors on ne pourra pas obtenir de conclusions valides. Cependant, des tests perceptifs sont réalisés dans beaucoup d'études, à la suite desquels on ne garde que les données validées. Pour ce faire, des techniques statistiques sont présentées dans [Banse et Scherer 1996] par exemple.

Dans un livre récapitulant différentes problématiques rencontrées au cours de ses recherches depuis les années 1940, Lazarus [Lazarus 1998 p161], après avoir étudié différents types de données actées ou induites a choisi de s'appuyer sur des données naturelles :

> *"I was now convinced I needed to find another way of studying stress, emotion, and coping in daily life, and it should be in the field rather than in the laboratory".*

De même, Scherer *et al.* arrivent à la conclusion qu'on ne peut pas généraliser les résultats obtenus sur des données actées à des données naturelles [Scherer et al. 1991].

Batliner *et al.* ont comparé des expériences effectuées sur des données actées, induites (magicien d'Oz[30]) et réelles (en interaction homme_machine) et ont montré que des modèles performants pour des données actées ne l'étaient pas pour les données réelles : les scores de bonne détection étaient inversement proportionnels au naturel des données [Batliner et al. 2003]. Des résultats similaires ont été présentés [Vogt et Andre 2005] avec là encore des performances bien plus élevées sur des données actées que sur des données naturelles. Par ailleurs, ils ont montré que ce ne sont pas les mêmes indices qui sont les plus pertinents pour les différents types de données.

De plus, les performances obtenues avec des données actées sont largement supérieures à celles obtenues avec des données naturelles [Vogt et Andre 2005] et les modèles entraînés sur des données actées ont de très mauvaises performances sur les données réelles [Batliner et al. 2003].

## *Données de fiction (film, théâtre)*

Une solution pour obtenir des données avec des émotions authentiques en quantité suffisante et sans contraintes de confidentialité est d'utiliser la fiction, par exemple en sélectionnant des données jouées par des acteurs professionnels. La mise en situation de l'acteur pourra permettre d'améliorer le réalisme des émotions jouées. Les inconvénients de l'utilisation de ce type de données sont qu'elles sont souvent accompagnées de bruitages et restent susceptibles de ne pas

[30] L'expérience de magicien d'Oz est une expérience dans laquelle les sujets interagissent avec un système informatique qu'ils croient autonome, mais qui est en fait totalement ou partiellement contrôlé par un humain (http://fr.wikipedia.org/wiki/Magicien_d'Oz_(exp%C3%A9rience))

refléter des comportements réels. Clavel, dans ses travaux sur les manifestations de type peur [Clavel 2007], a sélectionné des séquences de films en anglais en appliquant un critère de crédibilité pour construire son corpus. Elle a décrit en détails les avantages et inconvénients de ce type de corpus.

## *Données induites : Magicien d'Oz et expériences contrôlées*

Plusieurs techniques ont été mises en place pour induire des émotions: hypnose [Grossberg et Wilson 1968], présentation de films ([Gross et Levenson 1995], [Philippot 1993][31]), images ou jeux induisant une réponse émotionnelle, réalisation d'une tâche difficile à effectuer en peu de temps pour induire du stress, expériences de magicien d'Oz [Batliner et al. 2003].
Les émotions induites sont souvent de faible intensité. De plus, les mêmes protocoles d'induction n'induisent pas nécessairement des états émotionnels identiques. Un autre biais est le contrôle de l'environnement sur l'expression des individus [Hochschild 1979].

## *Données naturelles*

Des données naturelles de toutes sortes ont été enregistrées : pilotes d'avion, séances thérapeutiques[32], télé réalité.
La qualité de l'enregistrement est souvent assez mauvaise, la quantité de données émotionnelles assez faible et il y a en général peu de parole par locuteur. En plus, il n'est pas toujours évident de connaître l'émotion exprimée par le locuteur. Le matériel d'enregistrement peut aussi devenir un obstacle à l'aspect naturel des données. Par exemple, les personnes enregistrées à la télévision peuvent être en «display» (affichage de certaines émotions liées aux interactions sociales [Hess 2006]), ce qui peut remettre en cause la validité des données[33]. L'utilisation des centres d'appels est une alternative intéressante, en particulier lorsqu'on s'intéresse uniquement aux émotions dans la voix. L'enregistrement imperceptible permet d'obtenir des données spontanées. De plus, avec les données téléphoniques, l'émotion doit s'exprimer par la voix sans possibilité de conflits avec d'autres modalités comme les actions, gestes ou expressions du visage. Cependant, le contenu est majoritairement assez faible émotionnellement, souvent de l'ordre de 10%.

---

[31] Pierre Philippot dans ses expériences a choisi des séquences de film visant à éliciter des émotions spécifiques et montre que les films sont des bons moyens pour éliciter les émotions voulues.
[32] Lazarus [Lazarus 1991] a enregistré des entretiens de 10-15 minutes auprès de 61 patients d'un hôpital à la veille d'une opération (hernie, vessie, thyroïde) et a établi des liens entre leur degré de stress et leur durée de rétablissement (pas d'analyse acoustique).
[33] Dans les cas d'émotion « intense» ou lorsque filmé longtemps, ces effets peuvent disparaître.

Le réseau d'excellence HUMAINE recense les principales bases de données plus ou moins naturelles utilisées par les différents collaborateurs. (http://emotion-research.net/wiki/Databases). Les données naturelles sont données dans le Tableau 2-1 ci-dessous.

Afin de pouvoir comparer les résultats des études et prendre en compte les différences entre individus, il est encouragé d'utiliser des corpus contenant des personnalités et des manifestations d'une même émotion variés ou au moins contrôlés [Kappas et al. 1991].

| Identifier | Emotional content | Emotion elicitation methods | Size | Language |
|---|---|---|---|---|
| Reading-Leeds database | Range of full blown emotions | Natural: Unscripted interviews on radio/television | Around 4 ½ hours material | English |
| France et al. | Depression, suicidal state, neutrality | Natural: therapy sessions & phone conversations. | 115 subjects: 48 females 67 males. | English |
| CREST database | Wide range of emotional states and emotion-related attitudes | Natural: volunteers record their domestic and social spoken interactions for extended periods throughout the day | Target - 1000 hrs over 5 years | English Japanese Chinese |
| Stock Exchange Customer Service (Devillers & Vasilescu) | Mainly negative - fear, anger, stress | Natural: call center human-human interactions | Unspecified | French |
| AIBO | Joyful, surprised, emphatic, helpless, touchy (irritated), angry, motherese, bored, reprimanding, neutral | Human machine: interaction with robot | 51 german children, 51.393 words English 30 children, 5.822 words | German |

**Tableau 2-1.** *Des données naturelles.*

## *Ethique : consentement conscient et confidentialité…*

Lors d'expériences physiologiques sur la peur et la colère, Ax [Ax 1953] a créé l'illusion qu'il y avait un problème grave avec l'équipement auquel les sujets étaient attachés en simulant une panique des expérimentateurs accompagnée de bruits sinistres et d'étincelles. Il rapporte les réactions des sujets :

> *« One woman kept pleading, "Please take the wires off. Oh! Please help me." Another said during the interview that she had prayed to be spared during the fear episode. A man said, "Well, everybody has to go sometime. I thought this might be my time." »*

De telles expériences désagréables ou douloureuses pour le sujet ne seraient plus reproductibles aujourd'hui pour des raisons éthiques[34]. Le réseau d'excellence humaine est particulièrement concerné par tous les problèmes liés à l'éthique, à la fois dans la manière de récolter et de traiter les données, mais aussi dans l'utilisation qui en sera faite, ce qui a donné lieu à un rapport (deliverable ) « Science and society : Ethics ». De même, à la session plénière HUMAINE de 2007, la fondatrice du groupe « affective computing » du MIT Rosalind Picard qui présentait un ordinateur capable de faire des signes de désapprobation lorsque l'utilisateur adopte une mauvaise posture se positionnait contre ce type d'utilisation.

Lors des élections aux conseils de quartier d'Issy-les-Moulineaux en 2005, la mairie avait fait appel à un avatar réalisé par la Cantoche[35] pour inciter les gens à aller voter, mais quels sont les dangers de ce type d'applications ? Le public tend à avoir foi en la machine et à oublier qu'elle fonctionne à partir des décisions parfois subjectives d'un programmeur humain.

[34] Dès 1975, la question a été soulevée par Osgood dans [Osgood et al. 1975 p28]: "Collection of data relating to subjective culture always involve potential misuse as well as potential invasion of the privacy of the individual … two salient issues: first, the degree of informed consent that the tested individual should exercise; second the degree of confidentiality that should be maintained in the use of the collected information".

[35] http://cantoche.com/fr~Avatars.html

## 2.2. Données LIMSI : des centres d'appels

En réponse à la problématique soulevée par les deux états de l'art [Scherer 2003] [Juslin et Laukka 2003] « Comment obtenir des données et quel type de données utiliser ? », nous avons choisi les centres d'appel. Une critique de l'utilisation de données réelles est que le locuteur, parce qu'il a conscience d'être filmé/enregistré, va être « en display » et réagir de telle sorte à avoir un certain effet sur une audience. Pour des données téléphoniques, les appelants ne sont pas vraiment en display car cachés derrière leur téléphone et en général sans public. Dans le cas des « hotline » où le motif de l'appel est souvent de se renseigner ou de résoudre un problème rapidement, ils ne sont pas non plus focalisés en permanence sur la manière dont on va les juger et réagissent de manière spontanée. Ce ne sera plus forcément le cas en interaction avec un système vocal automatique. Par exemple, aux Etats-Unis, une stratégie des appelants est d'exagérer volontairement leur parler afin d'être redirigé le plus rapidement possible vers un opérateur humain.
Quant aux réserves concernant l'annotation de ce type de données (le locuteur ne peut pas être interrogé sur son état émotionnel), notre stratégie sera détaillée dans le chapitre Annotation.
Une dernière critique des données réelles est qu'il est difficile d'avoir assez de données par locuteur et encore moins des données permettant de comparer plusieurs individus à cause des grandes variations d'expression des émotions à la fois dans le contenu linguistique et acoustique. Le domaine d'application pour un centre d'appel étant fini, il est possible d'obtenir des réactions similaires, même si évidement non identiques et en utilisant des étiquettes les plus fines possible pour décrire le corpus, on diminuera les variations par classe. De plus, pour le corpus CEMO détaillé ci-dessous, grâce à l'enregistrement d'une grande quantité de données, il y a des cas de segments de contenu lexical identique et exprimé avec différentes émotions. Enfin, le nombre d'agents est assez petit donc nous avons beaucoup de données par agent, même si de par leur rôle ils sont contraints de contrôler leurs émotions et de suivre des normes sociales (compassion…).

Les données utilisées pour les expériences proviennent de deux corpus LIMSI de dialogues oraux naturels enregistrés dans des centres d'appels :

- le premier corpus (**transactions boursières**) préalablement transcrit et annoté, a été utilisé pour les premières expériences de détection. Les problèmes rencontrés sur ce corpus ont servi de base à la mise en forme des protocoles de transcription et d'annotation du second corpus.
- le **corpus CEMO (**pour Corpus EMOtion) sur lequel ont été effectuées la plupart des expériences.

L'utilisation des données s'est faite dans le respect des conventions éthiques assurant l'anonymat des appelants, le caractère privé des informations personnelles et la non diffusion du corpus et des annotations.

### 2.2.1. Corpus de transactions boursières

Le Corpus 1 est composé de dialogues réels provenant d'un centre d'appel gérant les portefeuilles boursiers des comptes client par téléphone. Le but de cet enregistrement était indépendant du travail d'étude sur les émotions et était disponible dans le cadre du projet Amities [Hardy et al. 2002].

Le service peut être contacté via une connexion Internet ou en appelant directement un agent. Bien qu'une grande partie des appels soient liés à des problèmes d'utilisation du web (informations générales, requêtes compliquées, transactions, confirmations, problèmes de connections), certains appelants préfèrent simplement interagir avec un agent humain. Les dialogues ont été transcrits orthographiquement. Les normes de transcription sont données dans http://www.dcs.shef.ac.uk/nlp/amities/. Le corpus contient 100 dialogues en français entre un agent et un client (avec en tout 4 agents différents), soit 6200 tours de parole. Il a été enregistré sur seulement un canal et il y a donc beaucoup de recouvrements qui n'ont pas été transcrits (environ 20% du corpus).

### 2.2.2. CEMO

Le corpus CEMO contient des enregistrements de conversations réelles entre des agents et des appelants obtenus à la suite d'une convention entre un centre médical et le LIMSI-CNRS. Le service, dont le rôle est de donner des conseils médicaux, peut être contacté 24h sur 24 et 7 jours sur 7. Lors d'une interaction, un agent va utiliser une stratégie précise et prédéfinie afin d'obtenir un certain nombre d'informations de la manière la plus efficace possible. Son rôle est de déterminer le sujet de l'appel et d'obtenir assez de détails sur les circonstances de l'appel pour évaluer son degré d'urgence et prendre une décision. Les principaux motifs d'appel sont les situations d'urgence, les demandes de conseil médical et les demandes d'informations (numéro d'un docteur …). La décision prise pourra être d'envoyer une ambulance, de rediriger l'appelant vers les urgences sociales ou psychiatriques, ou de conseiller l'appelant, par exemple en lui enjoignant d'aller à l'hôpital ou d'appeler son médecin. L'appelant peut être le patient ou un tiers (famille, ami, collègue, voisin, etc.) Dans les cas d'appels urgents, l'appelant va souvent exprimer du

stress, de la douleur, de la peur, voire de la panique. L'étude a été faite sur un sous-ensemble de 20 heures, soit 688 dialogues (7 agents et 784 appelants distincts).

Bien que le corpus CEMO ait été enregistré sur 2 canaux, seul le canal correspondant à l'agent est propre. L'autre contient environ 10% de recouvrements qui n'ont pas été transcrits. Ces recouvrements ont été exclus de l'étude bien qu'ils puissent être corrélés à la parole émotionnelle : ils sont coûteux à transcrire et il est difficile d'en extraire des paramètres acoustiques sans erreurs. Le Tableau 2-2 résume les principales caractéristiques des 2 corpus.

| | **Corpus transactions boursières** | **CEMO** |
|---|---|---|
| #agents | 4 (3H, 1F) | 7 (3H, 4F) |
| #clients | 100 dialogues (91H, 9F) | 688 dialogues (271H, 513F) |
| #tours/dialogue | Moyenne : 50 | Moyenne : 48 |
| #mots distincts | 3k | 9.2k |
| #total de mots | 44k | 143k |

**Tableau 2-2.** *Caractéristiques des deux corpus : Corpus 1: 100 dialogues agent-client d'environ 3 heures (H: homme, F: femme), Corpus 2: 688 dialogues agent-client d'environ 20h (H : homme, F : femme) Dans 96 dialogues, des tiers interagissent.*

# 2.3. Transcription du corpus CEMO

## 2.3.1. Protocoles

Les protocoles de transcription sont similaires à ceux utilisés pour la transcription des dialogues oraux dans le projet FP5-Amities, très proches des normes de transcription LDC (www.ldc.upenn.edu). Des marqueurs ont été ajoutés pour indiquer les entités nommées, mais également les éléments non verbaux (*cf.* Tableau 2-4) tels que respirations, rire, pleurs, raclements de gorge et autres bruits (bruits de bouche), qui peuvent être signifiants de l'état émotionnel [Schröder 2000]. Les silences et les signaux inintelligibles sont aussi marqués. Un manuel de transcription a été rédigé.

## 2.3.2. Outils et vitesse de transcription

Les appelants et agents avaient été enregistrés sur deux canaux différents et afin d'avoir une transcription la plus riche possible, les deux canaux ont d'abord été transcrits séparément à l'aide de l'outil Transcriber[36] [Barras et al. 2000], puis sous emacs en groupant les 2 canaux afin d'accélérer le traitement.

## 2.3.3. Caractéristiques du Corpus

Des caractéristiques du corpus sont données dans les tableaux ci-dessous, avec en particulier la fréquence des principaux états affectifs.

| | |
|---|---|
| #mots distincts | 9.2 k |
| #total mots | 238 k |
| %Parole Inintelligible (PI) | 0,4% |
| #mots/tour | 6,9 |

**Tableau 2-3.** *Caractéristiques du corpus.*

| | |
|---|---|
| # rires | 159 |
| # pleurs | 244 |
| # « heu » | 7347 |
| # bruits de bouche | 4500 |
| # respiration | 243 |

**Tableau 2-4.** *Marqueurs affectifs indiqués par la transcription sur les 20 heures.*

[36] http://trans.sourceforge.net/

## 2.4. Métadonnées

Au niveau de chaque dialogue, des informations existent

sur les appelants (âge, sexe, relation avec le patient, accent), le patient (âge, sexe), le motif de l'appel et son issu, ainsi que sur les conditions de l'appel (type de téléphone, lieu d'où l'appel est passé).

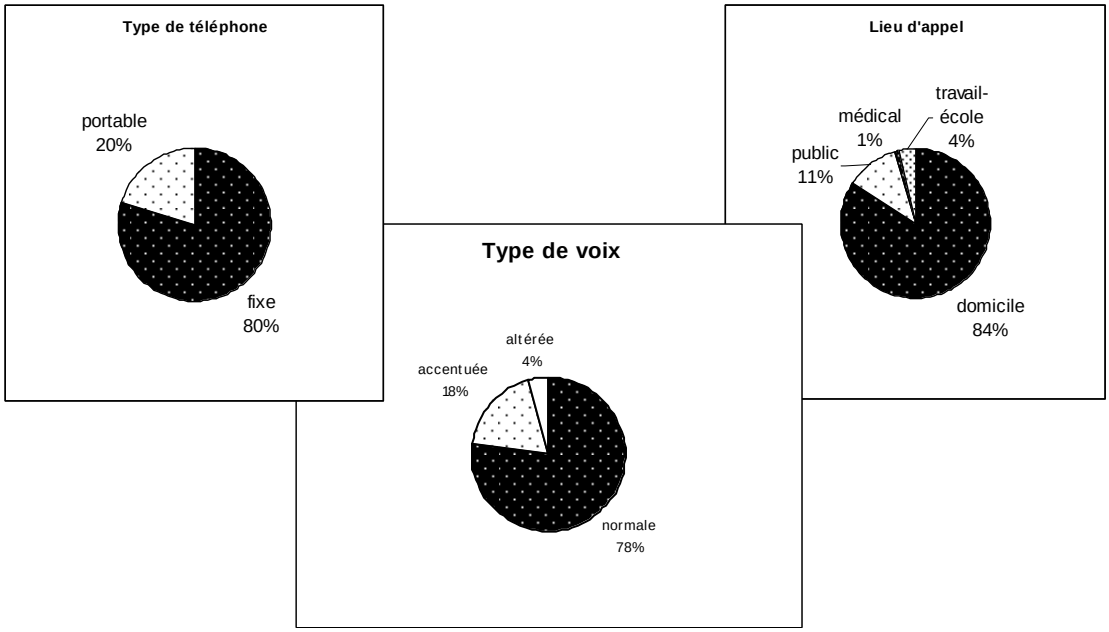



**Figure 2-1.** *Métadonnées liées à l'acoustique. (Gauche) Type de téléphone; (Droite) Lieu d'appels (Bas) Type de voix normale, accentuée (accents étrangers et régionaux) et altérée.*

La majorité des appelants (59,5%) sont des femmes adultes (Figure 2-2). La majorité des patients sont également des femmes (59 % des cas). Le patient appelle directement dans 30 % des cas. Sinon, l'appel est fait par un tiers, qui peut être plus ou moins proche du patient.

Ce corpus est extrêmement intéressant pour comprendre le rôle du contexte dans la perception des émotions. Nos modèles de détection (voir chapitre 5 p 137) ont pris en compte comme indices contextuels le rôle dans le dialogue : agent vs. appelant, et le sexe, mais il faudrait étudier d'autres indices contextuels comme l'âge par exemple.

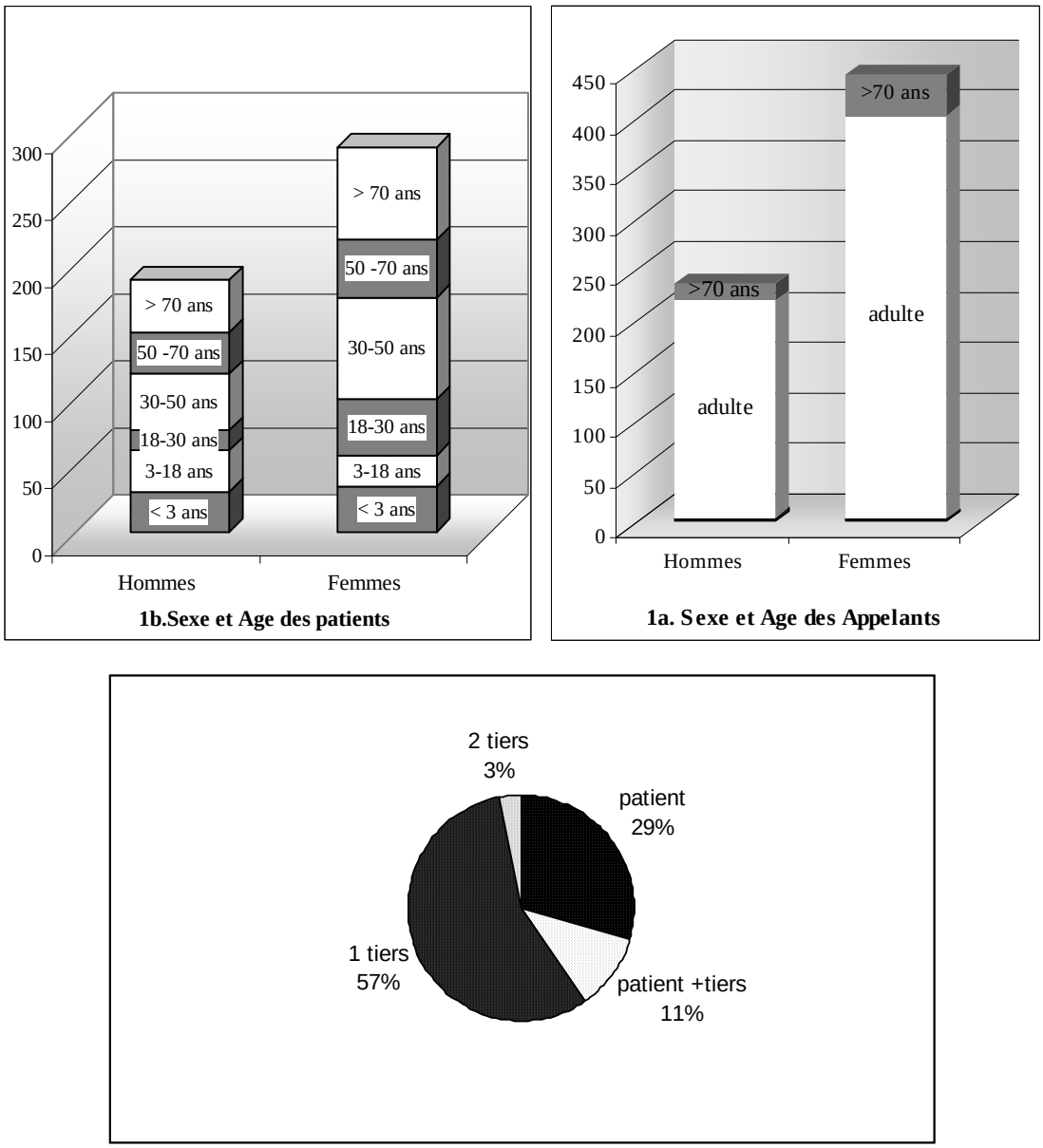



**Figure 2-2.** *Métadonnées.* ***1a*** *Age et Sexe des locuteurs et appelants;* ***1b*** *Répartition des appelants,* ***1c*** *Répartition des appels avec seulement le patient ou 1 tiers, 15% des appels on plus de 2 interlocuteurs (11% patient+tiers, 3% 2 tiers).*

## 2.5. Conclusion

Dans ce chapitre, nous avons justifié notre décision de travailler sur des données réelles et avons présenté en détails les 2 corpus qui ont été principalement utilisés pour nos expériences, et en particulier le corpus CEMO très riche et avec une grande variabilité tant dans les contextes que dans les locuteurs. La difficulté est alors de trouver une palette d'émotions et un protocole d'annotations décrivant au mieux ces données.

# Chapitre 3

# Annotation des émotions


Résumé

Dans les chapitres précédents, nous avons justifié notre choix de travailler sur des données réels et l'intérêt des données provenant de centres d'appels lorsqu'on étudie les émotions vocales. Une fois ces données collectées, l'étape suivante est de les annoter.

Ce Chapitre décrit notre schéma d'annotation pour le corpus CEMO. Au préalable, différentes problématiques liées à l'annotation sont présentées. Qu'est ce qui est annoté ? Quelles annotations choisir (étiquettes discrètes, axes...) ? Combien d'annotateurs sont nécessaires et comment les entraîner et les évaluer ? Comment valider le protocole d'annotation ?

Une expérience sur le Corpus de transactions boursières a mis en évidence la présence de mélanges d'émotions dans nos données, ce qui nous a conduits à la mise en place d'un protocole original, permettant d'annoter deux émotions par segment émotionnel, l'émotion *Majeur*, principale et l'émotion *Mineur* en arrière plan. Un segment émotionnel, qui peut être inférieur au tour de parole, est introduit. Le choix des étiquettes émotionnels est expliqué. Des dimensions (valence, contrôle) et des informations dialogiques (répétitions, ironie) ont été annotés en plus de ces étiquettes. Un protocole d'annotation comprenant une phase de calibration et détaillant toutes ces annotations a été mis en place.

L'annotation, effectuée par deux experts (un homme et une femme), a été validée par des mesures de cohérence inter-annotateur (coefficient kappa), intra-annotateur (expérience de ré-annotation) et par un test perceptif.

Enfin, nous introduisons un vecteur émotion par segment émotionnel, résultant de la combinaison de plusieurs annotations.

*In the previous chapter, we have stressed the importance of choosing appropriate data for emotion detection experiments, especially if the application is natural spontaneous speech and we demonstrated the pertinence of call-centers. With a 20-hour corpus manually transcribed, the next step is to add emotion labels. This chapter describes our annotation protocol. Several issues must be raised before annotating: the unit to be used, the kind of annotation such as labels or axis, the number of annotators as well as the method to train and evaluate them.*

*An experiment on the stock option corpus revealed the presence of emotion blends. This lead to an original annotation protocol, enabling to annotate two emotions per emotion segment. The main emotion was denoted as "Major" and the background one was denoted as "Minor". The emotion unit can be shorter than the speaker turn. Abstract dimensions, as well as dialogic information (repetitions, irony) were annotated in addition to emotion labels. An annotation protocol including a calibration phase was established. The annotation was done by two experts and was validated. Finally, we introduce an Emotion Vector per segment, which is the result of a combination of several annotations.*

# 3.ANNOTATION DES EMOTIONS

L'un des défis de mon travail de thèse était de mettre en place un protocole d'annotation adapté à la complexité du corpus étudié car il n'existe pas actuellement de normes pour décrire les émotions et leurs annotations. L'élaboration d'un langage standardisé pour représenter et annoter les émotions dans le cadre d'applications informatiques est d'ailleurs un axe des recherches actuelles. Une première tentative avait été effectuée par le réseau d'excellence HUMAINE avec EARL (Emotion Annotation and Representation Language : http://emotion-research.net/earl), maintenant un W3C[37] incubator group (Groupe d'incubation sur les émotions) dans lequel le LIMSI est impliqué [Schröder et al. 2007].

## 3.1. Problématiques liées à l'annotation

### 3.1.1. Choix d'une unité de dialogue

Le premier problème avant de pouvoir annoter des données spontanées est ce choisir une unité [Batliner et al. 1998]. Même s'il n'existe pas de consensus sur l'unité de parole, les études utilisent :

- Le mot : Batliner *et al.* considèrent que c'est la meilleure unité car la plus petite qui soit. Elle permet si on le souhaite de passer à une unité plus grande [Batliner et al. 2003].
- La parole d'un même locuteur sans interruption par un autre (ou entre 2 souffles), ce qui constitue un tour de parole [Traum et Heeman 1997]
- Une unité intermédiaire entre le mot et le tour de parole, souvent appelée chunk, qui peut être définie de manière plus ou moins rigoureuse : N mots, unité émotionnelle…
- Une unité qui définit un seul acte de dialogue [Batliner et al. 2003]

Dans notre cas, on cherche à la fois à détecter les émotions et à prédire le commencement de troubles dans le dialogue. Si une unité trop longue, comme parfois le tour de parole[38], est choisie, elle pourra contenir une séquence de plusieurs émotions ou une partie « Neutre» et une autre partie plus riche en émotion. Pour une unité trop petite, certains paramètres, comme le débit par exemple, ne pourront pas forcément être calculés. Certaines recherches font en parallèle des prédictions sur différentes unités (mot et tour de parole par exemple) et combinent ces prédictions.

[37] W3C : organisme en charge des standards du web.
[38] Par exemple dans le corpus CEMO, la taille moyenne des tours de parole est de 2 secondes environ, mais près de 600 tours ont une durée supérieure à 10 secondes.

Schuller *et al.* ont montré que le choix de l'unité avait une incidence sur les scores de détection[Schuller et al. 2007b].

## 3.1.2. Choix des axes/étiquettes

### *Utilisation d'étiquettes discrètes*

Comme indiqué dans l'état de l'art, la majorité des études se font sur peu d'étiquettes (Neutre/Négatif par exemple) ou sur un nombre assez petit d'émotions primaires jouées par des acteurs ou induites volontairement. Des psychologues comme Kappas *et al* [Kappas et al. 1991] insistent sur l'importance de distinguer les différentes formes d'une émotion donnée (ex irritation vs. colère chaude).

#### Comment choisir les étiquettes ? Choix libre ou ensemble restreint ?

Le plus souvent on demande à un juge de faire un choix forcé parmi une liste d'étiquettes. Les réponses peuvent alors être influencées par la liste des choix proposés [Ekman et Davidson 1994].
Quelques expériences ont également été effectuées en choisissant une étiquette libre ([Greasley et al. 2000], [Devillers et al. 2002]). Avec le choix libre, il est nécessaire de bien formuler les consignes des annotateurs en précisant bien au juge de désigner l'émotion que le locuteur veut exprimer. Il y a des risques de réponses complètement hors propos et les différentes réponses peuvent être difficiles à classifier.
Greasley *et al.* ont comparé des annotations avec un choix d'étiquettes libres et un choix forcé entre 5 étiquettes de base sur les mêmes données [Greasley et al. 2000]. Pour ce faire, ils ont sélectionné 89 échantillons de parole émotionnelle extraits d'émissions de télévision et de radio et les ont fait annoter par 28 étudiants en psychologie avec à la fois un choix libre et un choix forcé parmi les 5 émotions de base peur, colère, tristesse, dégoût, joie. Ils sont arrivés à la conclusion que lorsque des juges s'accordaient sur des étiquettes libres, celles-ci étaient consistantes avec les étiquettes standards (à l'exception du dégoût), mais avec des variations dans le degré d'émotion (plus d'une quinzaine de termes utilisés pour désigner une étiquette standard). Par contre, ils ont constaté que si des données actées peuvent être facilement étiquetées avec des étiquettes standard (grand pourcentage d'accord parmi les annotateurs), ce n'est pas le cas des données naturelles. Pour 46% des échantillons, ils n'ont pas trouvé d'accord significatif entre les annotations utilisant

les émotions de base. Ils ont explicité plusieurs cas[39] où des mélanges de plusieurs émotions de base sont clairement perçus, en séparant les émotions éprouvées en réaction à une personne (agent), à un événement et à un objet. **Pour décrire et a fortiori pour annoter des données naturelles, une seule étiquette discrète n'est pas suffisante**. Nous l'avons également constaté avec le corpus de données boursières (*cf.* p 56).

Pour des expériences visant à développer une application spécifique comme celles décrites dans [Clavel 2007] ou [Liscombe 2006] les étiquettes sont choisies en fonction d'un but ou d'une application précis. Il convient cependant de ne pas travailler sur une tâche trop limitée afin de pouvoir généraliser les résultats[40].

## *Utilisation des axes abstraits*

On peut aussi demander aux annotateurs d'évaluer le stimulus sur une ou plusieurs échelles continues. Cowie *et al.* ont créé 'feeltrace', un instrument permettant de rendre compte de l'aspect dynamique de l'épisode émotionnel[Cowie et al. 2000]. Ils utilisent les deux axes activation et valence dans un espace continu à 2 dimensions, représenté par un cercle sur l'écran d'ordinateur (*cf.* Figure 3-1) et l'utilisateur déplace le curseur tout en écoutant l'extrait de parole.

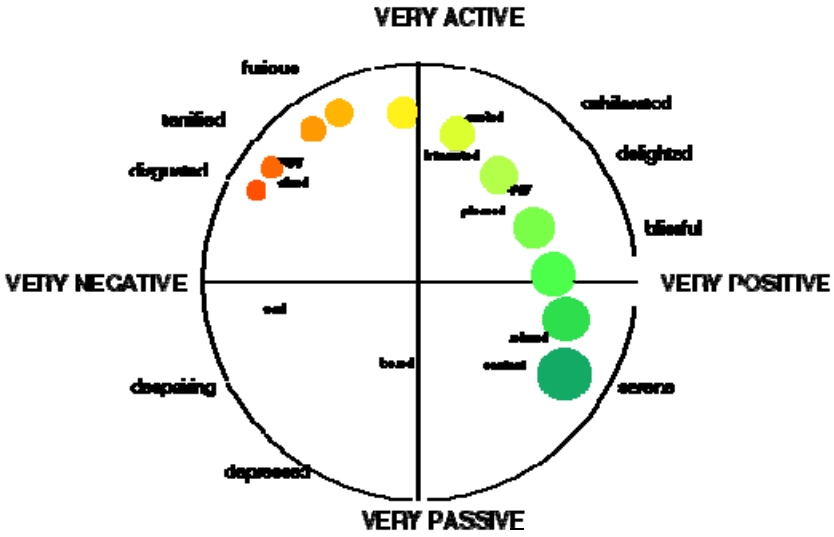



**Figure 3-1.** *Exemple d'affichage de Feeltrace, extrait de [Cowie et al. 2000].*

[39] Ils décrivent 2 cas en particulier ; celui d'une mère parlant du meurtre de sa fille en disant qu'elle ne pourra jamais pardonner aux meurtriers qui exprime à la fois son désespoir face à la situation et sa haine envers les meurtriers et celui d'une femme décrivant sa sœur qui après avoir gagner à la loterie a abandonné sa famille et exprime à la fois de la tristesse devant la situation et du dégoût et de la colère envers sa sœur.

[40] Par exemple, Yacoub [Yacoub et al. 2003] ont développé un système qui discriminait la colère et le neutre dans le but de détecter les troubles de la communication. Ce système a ensuite été testé sur des émotions positives qui ont été classifiées comme de la colère.

Grimm *et al.* ont utilisé le « Self Assessment Manikins » pour annoter par tour de parole la Valence (positif vs. négatif), l'Activation (niveau d'excitation haut vs. bas) et la Dominance (force apparente du locuteur 'fort vs. faible') [Grimm et al. 2007].

## 3.1.3. Combien d'annotateurs ?

Le nombre minimum d'annotateurs est de deux, trois si on veut faire un vote majoritaire. Batliner en a utilisé cinq pour annoter les données AIBO [Batliner et al. 2004]. Un test perceptif de Abrilian *et al.* a montré que vers 10 annotateurs avec 20 étiquettes, la courbe des annotations se stabilise ([Abrilian et al. 2006]). Dans leur étude, un des résultats du test perceptif était que les annotations données par les hommes et par les femmes étaient différentes.

## 3.1.4. Validation des annotations

### 3.1.4.1. Cas d'étiquettes discrètes

Avant de pouvoir utiliser les annotations effectuées, il est indispensable d'évaluer leur fiabilité. On va donc chercher à estimer un taux d'accord réel entre plusieurs juges pour des jugements qualitatifs dans des cas où il n'existe pas de référence.

Une première mesure serait de calculer le **pourcentage de fois où les juges sont d'accord**, mais cette mesure est biaisée [Wagner 1993], en particulier lorsqu'une catégorie domine, ce qui est souvent le cas pour les émotions. Si on prend le cas des données réelles ou le pourcentage de neutre est typiquement supérieur à 70%, on aura facilement un pourcentage d'accord très haut, même si les émotions moins représentées sont assez confondues. Par exemple dans le cas fictif du Tableau 3-1, le pourcentage de fois où les juges sont d'accord est supérieur à 75% alors que les 3 émotions Peur, Colère et Tristesse ne sont pas bien différenciées.

| Juge1/Juge2 | Neutre | Peur | Colère | Tristesse |
|---|---|---|---|---|
| Neutre | 7100 | **900** | 50 | 85 |
| Peur | **1000** | 500 | **100** | **90** |
| Colère | 15 | 15 | 50 | 5 |
| Tristesse | 50 | 20 | 0 | 20 |

**Tableau 3-1.** *Exemple de matrice d'inter annotation. Les chiffres sont fictifs.*

De plus l'accord observé entre un ou plusieurs jugements aura toujours une composante aléatoire et une composante réelle [Bergeri et al. 2002].

La plupart des mesures d'inter-annotation existantes utilisent une évaluation de cette composante aléatoire. Plusieurs manières de la définir ont été proposées, ce qui a donné lieu à plusieurs mesures d'interannotation dont le coefficient Kappa, le Pi de Scott et le « S index » de Bennet *et al.* (*cf.* Tableau 3-2 voir [Zwick 1988] pour des définitions et références).

Les études sur les émotions annotées avec des étiquettes discrètes (données nominales), qui ont reporté des mesures de fiabilité, ont le plus souvent utilisé le **coefficient Kappa** [Cohen 1960] défini ci-dessous.

## *Mesure de l'inter-annotation à l'aide du coefficient Kappa dans le cas de deux juges*

Intéressons nous d'abord au cas de deux juges pouvant choisir entre k catégories. Pour plusieurs juges, on pourra comparer les juges deux à deux ou discuter/moyenner les résultats ou encore adapter les formules (voir pour le cas du coefficient Kappa avec N juges [Fleiss 1971]).

Pour 2 juges, les résultats de l'annotation peuvent être représentés par une matrice (agreement matrix) k x k, dont la diagonale représente les cas d'accord entre les 2 juges (comme dans l'exemple fictif du Tableau 3-1).

Un coefficient d'inter-annotation est alors :

$$A = \frac{P_{obs} - P_{\exp}}{1 - P_{\exp}}$$

avec :

- $P_{obs}$ la proportion d'accord observée :

$$P_{obs} = \sum_{i=1}^{k} pii$$ avec *pii* la proportion de cas de la i$^{ème}$ diagonale de la matrice.

- $P_{\exp}$ la proportion d'accord aléatoire (exp : « expected by chance »), qui se calcule différemment suivant les coefficients (voir Tableau 3-2**).**

On essaie de corriger $P_{obs}$ en lui soustrayant $P_{\exp}$ correspondant aux cas de la diagonale qui sont dus au hasard ; le numérateur est divisé par (1- $P_{\exp}$), l'accord maximum lorsqu'on retire le hasard.

| **Coefficient** | **Définition** |
|---|---|
| Kappa (Cohen) | $\sum_{i=1}^{k} p_{i+} p_{+i}$ |
| Pi (Scott) | $P_{obs} = \sum_{i=1}^{k} (\frac{p_{i+} p_{+i}}{2})^2$ |
| S (Bennet, Alpert et >Goldstein) | $P_{obs} = 1/Kappa$ |

**Tableau 3-2.** *Pexp reproduit de [Zwick 1988] avec pi+ la somme des proportions de la ligne i de la matrice et p+i la somme des proportions de la colonne i.*

Le coefficient d'inter-annotation est un nombre réel compris entre -1 et 1. L'accord est d'autant plus élevé que sa valeur est proche de 1. Il est maximal quand les deux jugements sont les mêmes : tous les exemples sont sur la diagonale de la matrice de confusion.

– Il vaut 0 lorsque les deux jugements sont indépendants

– Il vaut -1 lorsque les juges sont en total désaccord.

Landis et Koch [Landis et Koch 1977] ont proposé une échelle de degré d'accord pour le coefficient Kappa selon la valeur du coefficient pour la biologie :

| | |
|---|---|
| Excellent | >0.81 |
| Bon | 0.80-0.61 |
| Modéré | 0.6-0.21 |
| Mauvais | 0.20-0.0 |
| Très mauvais | <0 |

**Tableau 3-3.** *Degré d'accord suivant la valeur du coefficient kappa*

.

Cette évaluation pourra varier selon les domaines[41] et un accord « modéré » dans cette échelle pourra être considéré comme « bon » pour un autre domaine comme la psychiatrie ou il y a plus d'incertitude. [Bergeri et al. 2002]. Dans le cas du Tableau 3-1, le kappa est de 0.25.

Le kappa s'accompagne normalement de son degré de signification, calculé à partir de la déviation standard de kappa : $\sigma_k \cong \sqrt{\frac{pobs(1-P_{obs})}{N(1-P_{\exp})^2}}$ [Cohen 1960]

Cohen [Cohen 1968] a également présenté une variante du coefficient Kappa prenant en compte les différences de distances entre les catégories. (Par exemple, une confusion Agacement/Impatience sera moins grave qu'une confusion Agacement/Amusement). Cela implique cependant de donner des poids a priori à chaque cellule de la matrice.

D'autres stratégies existent pour mesurer les corrélations entre les évaluations de deux juges pour des données ordinales [Howell 1999 p 550-553] ou pour des variables continues (coefficient alpha de Cronbach [Cronbach 1951])

#### 3.1.4.2. Mesures de validation pour les axes

Grimm *et al.* [Grimm et Kroschel 2007] ont regardé la déviation standard pour chaque axe avec 6 juges pour une expérience et 17 pour une autre et ont fait des mesures de corrélation avec le coefficient de Pearson.

[41] En biologie, on mesure souvent un diagnostic positif ou négatif pour plusieurs techniques médicales.

## 3.2. Annotation du corpus CEMO

Avec des données réelles enregistrées, il est impossible de demander au locuteur de nous renseigner sur ses émotions et on se place donc nécessairement au niveau du décodage. Il s'agissait tout d'abord de **choisir de manière rigoureuse une palette d'émotions ou d'axes** décrivant les données. Il nous fallait également concilier notre volonté d'utiliser des étiquettes fines pour une meilleure analyse des données (*cf.* [Kappas et al. 1991]) et la nécessité d'avoir un nombre suffisant d'échantillons par classe pour pouvoir ensuite entraîner des systèmes de reconnaissance.

Se posait ensuite le problème de la **rigueur et la validité des annotations**. Plutôt que de faire appel à un grand nombre d'annotateurs naïfs pour annoter le corpus et pour des raisons pratiques, seules deux personnes expertes ont annoté les données, un homme et une femme, afin de tenir compte des différences de sexe.

Le terme « Emotion » sera utilisé au sens large en accordance avec le réseau d'excellence Humaine. L'émotion « Neutre », également sujet de controverse désignera un état avec un faible niveau d'activation affective.

Des expériences sur les données du Corpus 1, annoté avant ma thèse, ont aidé à mettre en place le protocole d'annotation.

### 3.2.1. Expérience tirée des travaux sur le Corpus de transactions boursières

#### *Une première annotation du Corpus financier antérieure au commencement du travail de thèse*

L'annotation initiale du corpus financier [Devillers et al. 2002] a été développée avec, pour des raisons pratiques, un nombre restreint d'étiquettes émotion. 5000 phrases ont été conservées et annotées par 2 personnes différentes avec un choix de 5 états émotionnels : *Neutre*, *Colère*, *Peur*, *Satisfaction* et *Excuse*. Une troisième personne tranchait en cas de désaccord (3% des phrases étaient ambiguës et cette indécision était le plus souvent entre un état *Neutre* et une émotion). « *Neutre* » est l'état de référence, la plupart des segments étant peu émotionnels.

Des tests perceptifs ([Devillers et al. 2003b] : 20 juges naïfs dans chacun des tests : avec et sans écoute du signal de parole) ont révélé que dans ce corpus, les émotions classiques *Colère* et *Peur* correspondaient plus à de l'énervement (« *elle me l'a déjà expliqué mais c'est quand même mal foutu quoi* ») ou de l'inquiétude (« *parce que bon euh et et il faut attendre euh* »). La classe *Peur* correspond souvent à la

peur de perdre de l'argent. *Satisfaction* («*parfait* », « *c'est très gentil* ») et *Excuse* (« *vraiment désolé demain normalement ça devrait aller* ») sont plus des attitudes que des émotions, mais sont des étiquettes adaptées à la description du corpus.

Les segments non neutres constituent 13% du corpus. Le nombre de tours de parole par segment est donné dans le Tableau 3-4**.**

| | **Peur** | **Colère** | **Excuse** | **satisfaction** | **Neutre** | **Total** |
|---|---|---|---|---|---|---|
| **Agent** | 34 | 19 | 48 | 106 | 2423 | 2630 |
| **Client** | 158 | 234 | 3 | 62 | 1913 | 2370 |
| **Total** | 192 | 253 | 51 | 168 | 4336 | 5000 |

**Tableau 3-4.** *Nombre de fichiers pour chaque état émotionnel dans le corpus de données boursières.*

## *Des confusions entre les classes Peur et Colère dans le Corpus 1*

Avant de décider du protocole d'annotation pour le corpus CEMO, des premières expériences de classification ont été réalisées sur le Corpus 1 à l'aide d'indices paralinguistiques. Alors que de bons résultats étaient obtenus pour la classification Neutre/Négatif (voir), les performances de classification Peur/Colère étaient proches du hasard avec les seuls indices acoustiques (47 indices à l'époque) et de l'ordre de 60% de bonne détection après ajout d'indices liés aux disfluences. Après réécoute des signaux, il est apparu que pour certains segments on pouvait percevoir à la fois de la peur et de la colère. Cela peut s'expliquer par le fait que les deux émotions peuvent être mélangées pour cette tâche financière. Les clients sont en colère car ils ont peur de perdre de l'argent. Ce lien entre la peur et la colère est d'ailleurs mentionné dans [Lazarus 1998 p159].

## *Expérience de ré-annotation des segments négatifs avec possibilité de choisir 2 étiquettes par segment*

Une expérience de ré-annotation des segments négatifs (445 segments *Peur* et *Colère*) a été menée en contexte afin de vérifier ces ambiguïtés avec deux annotateurs différents de ceux ayant fait la première annotation. Huit étiquettes étaient utilisées pour cette expérience (*cf.*3.2.2) : Neutre, Tristesse, Peur, Colère, Embarras, Autre Négatif, Empathie, Autre Positifs. Un annotateur avait la possibilité de choisir une deuxième étiquette s'il percevait un mélange d'émotions. L'émotion principale était qualifiée de *Majeure* ; si une deuxième émotion était perçue, elle était qualifiée de *Mineure.*

Les 2 annotateurs ont perçu la même émotion *Majeure* dans 64% des cas et 13% des segments étaient ambigus (pas d'étiquette commune entre les 2 annotateurs).

## *Combinaison des nouvelles annotations en un vecteur par segment*

Chaque annotation a été transformée en un vecteur Emotion [Peur, Colère] avec un poids de 2 pour le *Majeur*, et un poids de 1 pour le *Mineur* (Voir paragraphe 3.3). Les vecteurs correspondant à chaque annotateur ont ensuite été moyennés afin d'avoir un vecteur par segment émotionnel. 4 classes de segments se déduisent de ces vecteurs : *Peur* (Peur>0; Colère=0), *Colère* (Peur=0; Colère>0), *Mélange* (Peur>0; Colère>0) et *Autre* (Peur=0; Colère=0). La distribution de ces classes est représentée dans la Figure 3-2.

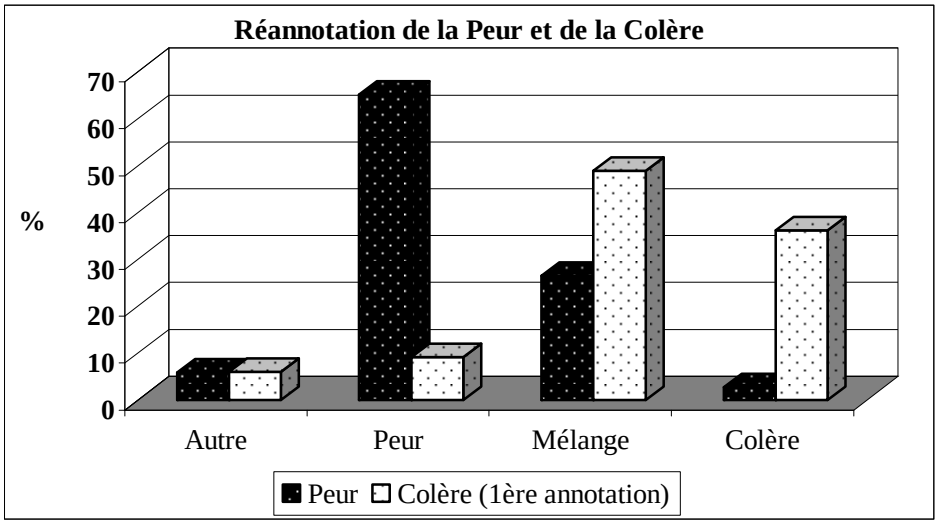



**Figure 3-2.** *Répartition des segments annotés précédemment Peur et Colère après la ré-annotation.*

## *Consistance entre la première et la deuxième annotation*

Nous avons d'abord vérifié que les deux annotations étaient consistantes. Pour un segment préalablement annoté Peur, nous avons considéré que les deux annotations étaient équivalentes si le champ *Peur* du vecteur était strictement positif et supérieur ou égal au champ *Colère* (idem pour *Colère*). C'est le cas pour 78% des segments.[42]

40% des segments ont été réannotés comme mélange. Cela peut expliquer les faibles scores de classification Peur/Colère et souligne la nécessité d'avoir des classes le plus distinctes possibles pour entraîner des systèmes en vue de faire de la classification.

[42] Par exemple pour un segment Colère par la première annotation, si l'annotateur 1 le perçoit comme Peur et l'annotateur 2 comme Colère, les 2 annotations ne sont pas cohérentes. Par contre si l'annotateur 1 perçoit de la colère et l'annotateur 2 perçoit de la Peur en *Majeur* et de la Colère en *Mineur*, les deux annotations sont cohérentes.

Le fait d'utiliser plusieurs étiquettes lors de l'annotation est ainsi un moyen d'identifier les données complexes et de ne pas les utiliser lors de l'entraînement de classifieurs. Cependant cela pose problème à la fois pour mesurer la validité des annotations et pour choisir l'étiquette émotion attribuée à un segment. Traditionnellement (*cf.* 3.1.3), N annotateurs jugent les données et une étiquette est attribuée à un segment, en général par vote majoritaire. Des mesures d'accord permettent de vérifier la robustesse de l'annotation et les segments sur lesquels les N annotations sont trop différentes sont exclus lors des expériences de détection. Accepter d'avoir plusieurs étiquettes par segment émotionnel multiplie le nombre de classes possibles et implique de mettre en œuvre des méthodes pour valider l'annotation. Il faut ensuite établir des règles pour combiner N annotations complexes afin d'avoir une annotation (étiquette, vecteur ?) par segment.

## 3.2.2. Annotation du corpus CEMO

Nous avons adopté une palette discrète d'émotions verbales pour annoter les émotions du corpus[Plutchik 1984] [Cowie 2000], ce qui permet de mélanger les catégories verbales afin d'obtenir une description des émotions complexes.

L'annotation utilise à la fois des **dimensions abstraites** (intensité et contrôle) et des **étiquettes**, avec **une ou deux étiquettes par segment** (Majeur et Mineur). Ces étiquettes sont hiérarchisées en partant des étiquettes les plus fines jusqu'à des étiquettes plus larges.

Toutes les annotations ont été faites en contexte : l'annotateur écoutait l'ensemble du dialogue et en particulier les tours précédents le segment a annoter.

Nous ne considérons pas qu'il y ait une bonne annotation par segment, mais plusieurs perceptions différentes.

### *Le segment émotion*

Le tour de parole est segmenté par groupe de souffle en plusieurs segments si nécessaire. L'annotation se fait par défaut au niveau du tour de parole (environ 32900 tours de parole au total), mais pour gérer l'aspect dynamique, chaque annotateur avait la possibilité de couper le tour en plusieurs segments émotionnels s'il percevait 2 émotions différentes séquentiellement[43]. Cette coupure se faisait cependant au niveau des séparateurs syntaxiques définis par les transcripteurs.

Nous avons ainsi créé des unités émotionnelles, qui peuvent être inférieures aux tours de parole bien que le tour de parole reste l'unité majoritairement utilisée. Les bruits ont été également retirés, ainsi que les échos (550 bruits ou échos) afin d'avoir des données le plus propre possible.

### *Quelles étiquettes ?*

Notre objectif était de choisir un ensemble d'étiquettes adaptées à nos données et comparables à celles d'autres études. Une liste de 52 termes émotionnels pertinents à des interfaces du futur sensibles aux émotions, établie par Roddy Cowie lors de l'école d'été de HUMAINE à Belfast. (http://emotion-research.net/ws/summerschool1), a servi de référence pour le choix des étiquettes (voir Figure 3-3 ).

[43] Des coupures similaires, dépendant de l'appréciation de 3 juges avaient également étaient effectuées par [Greasley et al. 2000] afin d'obtenir des segments comprenant un seul état émotionnel pour un test perceptif.

| | | | | |
|---|---|---|---|---|
| Admiration | Contempt | Fear | Jealousy | Shame |
| Affection | Cruelty | Friendliness | Mockery | Shock |
| Amusement | Despair | Greed | Neutrality | Stress |
| Annoyance | Determination | Guilt | Panic | Surprise |
| Anxiety | Disagreeableness | Happiness | Pleasure | Sympathy |
| Approval | Disppointment | Hopeful | Relaxation | Wariness |
| Boredom | Disapproval | Hot anger | Relief | Weariness |
| Calm | Disgust | Hurt | Resentment | Worry |
| Cold anger | Distraction | Impatience | Sadness | Confidence |
| Coldness | Embarrassment | Indifference | Satisfaction | Excitement |
| Interest | Serenity | | | |

**Figure 3-3.** *Liste de termes émotionnels pertinents pour des interfaces du futur sensibles aux émotions, établie par Cowie.*

Cinq personnes familières avec le corpus CEMO ont évalué pour chaque émotion de la liste son degré de pertinence avec le corpus sur une échelle de 0 à 3. Après un vote majoritaire, nous avons abouti à une liste de 18 termes émotionnels : Anxiété, Stress, Peur, Panique, Agacement, Impatience, Colère froide, Colère chaude, Déception, Tristesse, Désespoir, Douleur, Embarras, Soulagement, Intérêt, Amusement, Surprise et Neutre. A ces termes ont été ajoutés pendant une phase de calibration (*cf.* p61) les étiquettes Désarroi (ne sait pas quoi faire), Résignation et Compassion. En cas de difficulté à reconnaître l'émotion, les annotateurs pouvaient également utiliser les étiquettes *Positif*, *Négatif* ou *Unknown* (« je ne sais pas »), mais ces étiquettes ont été rarement utilisées : pour 1% des segments pour un annotateur (soit environ 450 segments sur 34280 au total) et 0.1% (4 segments) des segments pour l'autre.

## *Une hiérarchie en méta-catégories*

Des ensembles de différents niveaux de granularité peuvent être dérivés de cette liste afin de tenir compte des proximités entre les différents termes (l'irritation est plus proche de la colère que de la surprise) et d'avoir assez d'instances par catégories pour pouvoir plus tard construire des modèles. Au plus haut niveau, on retrouve la séparation entre les émotions négatives et positives.
Les catégories sont identifiées par un des termes de la catégorie, comme défini par Ortony ([Ortony et Turner 1990 p8]):

> *"For some categories of emotions, a language like English provides a relatively large number of tokens, thus reducing the need for metaphorical descriptions of emotional quality. In such cases, it becomes necessary to identify one of the words in the category as the unmarked form or category label. […] It may be helpful to think of the word "fear" as a relatively neutral word for an emotion type, fear. "*

| Valence | Classe large (7 classes) | Etiquettes fines (20 classes + Neutral) |
|---|---|---|
| Négatif | Peur | Peur, Anxiété, Stress, Panique |
| | Colère | Agacement, Impatience, Colère froide, Colère chaude |
| | Tristesse | Tristesse, Déception, Résignation, Désarroi, Embarras Désespoir |
| | Douleur | Douleur |
| Négatif ou Positive | Surprise | Surprise |
| Positive | Positif | Intérêt, Compassion, Amusement, Soulagement |
| Neutre | Neutre | Neutre |

**Tableau 3-5.** *Hiérarchie des classes d'émotion.*

Ce groupement a également été effectué par vote majoritaire entre 5 personnes (les mêmes qui ont choisi les étiquettes) en s'adaptant aux instances du corpus CEMO. Par exemple, le stress est dans le corpus beaucoup plus proche de la peur que de la colère. Parce qu'il y a peu de manifestations positives dans le corpus, une seule classe *Positif* regroupe toutes les autres. Elle pourrait être divisée en *Empathie* (Intérêt, Compassion), *Soulagement* et *Autres positifs*

Pour les autres émotions, nos sous-catégories correspondent à celles définies par les psychologues [Shaver et al. 2001]. Dans une étude sur la structure hiérarchique des termes émotionnels en anglais, ils ont remarqué que même si les gens ont des difficultés à définir certains termes émotionnels, ils s'accordent facilement sur des catégories regroupant les différents termes. Il a été demandé à cent étudiants en psychologie de regrouper une centaine de concepts émotionnels (extraits du « semantic atlas of Emotional concepts » [Averill 1975] en catégories. Les résultats de l'analyse par clustering hiérarchique sont donnés Tableau 3-6 et sont cohérents avec nos sous catégories.

| Cluster | Nom du sous Cluster sélectionné empiriquement | Sous-cluster |
|---|---|---|
| **Love** | Affection | Adoration, affection, love, fondness, liking, attraction, caring, tenderness, **compassion**, sentimentality |
| | Lust | Arousal, desire, lust, passion, infatuation |
| | Longing | Longing |
| **Joy** | Cheerfulness | **Amusement**, bliss, cheerfulness, gaiety, glee, jolliness, joviality, joy, delight, enjoyment, gladness, happiness, jubilation, elation, satisfaction, ecstasy, euphoria |
| | Zest | Enthusiasm, zeal, zest, excitement, thrill, exhilaration |
| | Contentment | Contentment, pleasure |
| | Pride | Pride, triumph |
| | Optimism | Eagerness, hope, optimism |
| | Enthrallment | Enthrallment, rapture |
| | Relief | **Relief** |
| **Surprise** | Surprise | Amazement, surprise, astonishment |
| **Anger** | Irritation | Aggravation, irritation, agitation, **annoyance**, grouchiness, grumpiness |
| | Exasperation | Exasperation, frustration |
| | Rage | **Anger**, rage, outrage, fury, wrath, hostility, ferocity, bitterness, hate, loathing, scorn, spite, vengefulness, dislike, resentment |
| | Disgust | Disgust, revulsion, contempt |
| | Envy | Envy, jealousy |
| | Torment | Torment |
| **Sadness** | Suffering | Agony, suffering, hurt, anguish |
| | Sadness | Depression, **despair**, hopelessness, gloom, glumness, **sadness**, unhappiness, grief, sorrow, woe, misery, melancholy |
| | Disappointment | **Dismay**, **disappointment**, displeasure |
| | Shame | Guilt, shame, regret, remorse |
| | Neglect | Alienation, isolation, neglect, loneliness, rejection, homesickness, defeat, dejection, insecurity, **embarrassment**, humiliation, insult |
| | Sympathy | Pity, sympathy |
| **Fear** | Horror | Alarm, shock, **fear**, fright, horror, terror, **panic**, hysteria, mortification |
| | Nervousness | **Anxiety**, nervousness, tenseness, uneasiness, apprehension, worry, distress, dread |

**Tableau 3-6.** *Résultats d'une analyse par clustering hiérarchique de 135 noms d'émotion (d'après* [Averill 1975]).

## *Deux étiquettes possibles par segment*

Afin de pouvoir rentre compte des émotions complexes, les annotateurs avaient la possibilité de choisir deux étiquettes par segment, l'étiquette *Majeur* (comme définie page 53) et *Mineur*. Nous adopterons la notation «Emotion1/Emotion2» pour décrire une annotation avec comme étiquette *Majeur* «Emotion1» et comme étiquette *Mineur* «Emotion 2».

L'annotateur 1 a utilisé une étiquette *Mineure* pour 31% des segments non neutre contre 17% pour l'annotateur 2. Les émotions mixtes sont de plusieurs types et seront décrites plus en détail dans le Chapitre 4 (p75) :

- Hésitation entre deux étiquettes d'une même grande classe. Par exemple si l'émotion perçue est entre l'impatience et la colère chaude, elle sera annotée «Impatience/Colère chaude»
- Perception quasi simultanément de deux émotions différentes. Exemple : «Anxiété/Agacement»
- Annotation de la surprise : la surprise étant la seule émotion dont on ne peut déduire la valence, il était demandé aux annotateurs percevant de la surprise d'indiquer en émotion mineure sa valence (entre positif ou négatif). Dans 80% des cas, cette valence a été annotée comme *Unknown*.

## *Autres annotations*

Pour des raisons de coût, nous avons considéré que la valence pouvait se déduire des étiquettes (d'où la nécessité de l'indiquer en *Mineur* pour la Surprise). Cette décision a été confirmée par un test perceptif où les sujets devaient annoter à la fois la valence et les étiquettes de segments émotionnels du corpus (voir p80 pour une description du test perceptif). Sur 1600 segments annotés par 44 sujets d'origine à la fois française et étrangère, il n'y a que 4% des cas où la valence perçue par les sujets ne correspond pas à l'étiquette du *Majeur*.

Parce qu'il y a souvent des confusions entre activation (passif, normal, actif) et intensité, l'intensité seulement est jugée sur une échelle de 1 à 5 (faible à fort). Nous avons ajouté une autre dimension, le contrôle (est ce que le locuteur semble contrôler son émotion?), différente de l'axe Puissance/Contrôle défini par Osgood [Osgood et al. 1975]. Il a été annoté sur un axe de -3 à +3. Il permet notamment de représenter des nuances très intéressantes, comme la simulation d'un état émotionnel.

Certaines annotations dialogiques, comme les répétitions (de soi ou de l'interlocuteur), le mensonge et l'ironie sont également annotées.

Toutes les annotations sont résumées dans la Figure 3-4 ci-dessous :

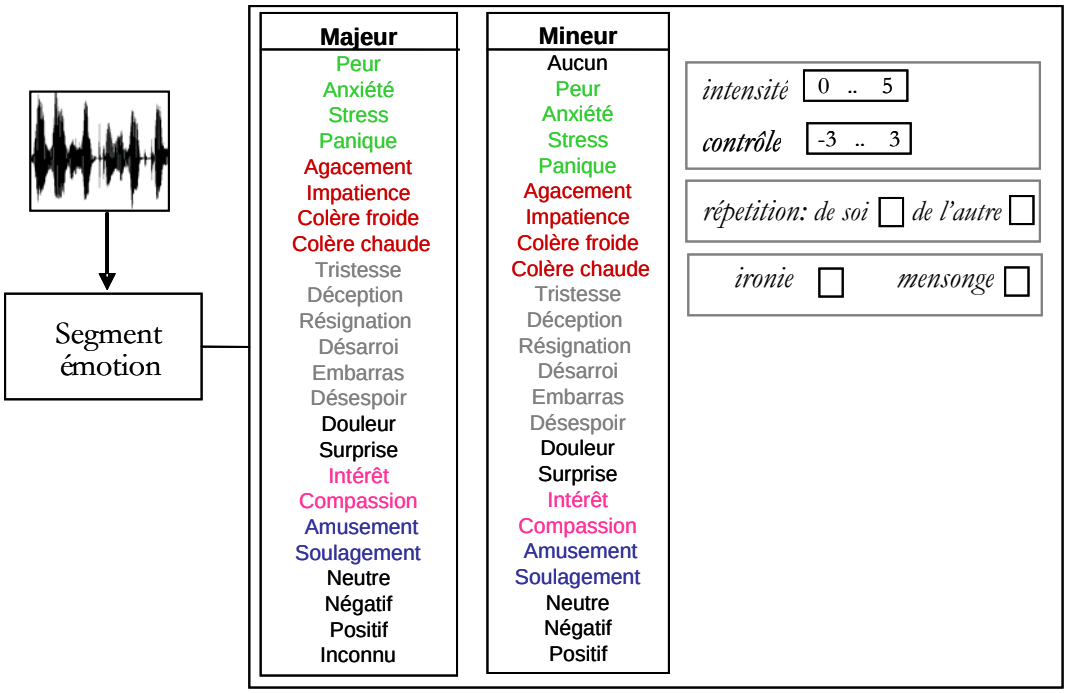
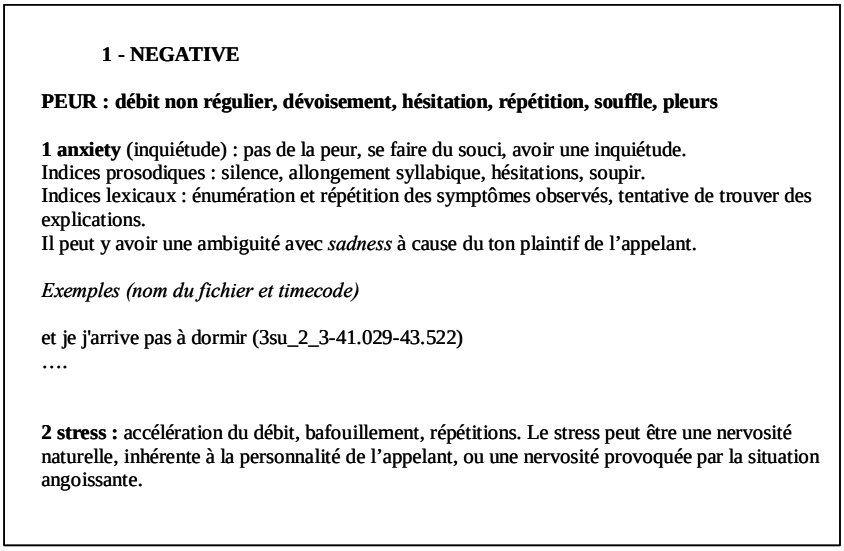



**Figure 3-4.** *Le schéma d'annotation : récapitulatif, l'annotation est faite en contexte, chaque tour pouvant être coupé en segment. Pour chaque segment sont annotés : une ou deux étiquettes, ainsi que l'intensité et le contrôle. L'annotateur peut aussi indiquer si la personne répète ce qu'elle a déjà dit ou ce que son interlocuteur a dit et si elle perçoit de l'ironie ou du mensonge.*

## *Phase de calibration préalable à l'annotation des données*

Le protocole d'annotation nécessite une phase d'apprentissage et de calibration pour les échelles d'évaluation. De plus les émotions n'étant pas stéréotypées dans des données naturelles, il n'existe pas nécessairement d'étiquette décrivant rigoureusement un type de comportement et nous voulions nous assurer que les mêmes comportements seraient décrits par la même étiquette par chaque annotateur. En effet, des expériences sur des données naturelles comme celle de Scherer *et al.* ont montré que des annotateurs différents pouvaient utiliser de manière consistante deux étiquettes différentes pour décrire un même comportement [Scherer et Ceschi 2000].

Afin d'effectuer cette vérification, il faut tout d'abord extraire un sous-ensemble de dialogues du corpus. On peut les sélectionner de manière aléatoire en aveugle ou au contraire choisir des dialogues comportant des phénomènes intéressants. Une quinzaine de dialogues (certains choisis spécifiquement, d'autres sélectionnés au hasard) ont été annotés par 4 personnes durant cette phase de calibration. Cela a mené à la mise en place d'un guide d'annotation (voir Figure 3-5) avec une définition et des exemples pour chaque étiquette, ainsi qu'une décision de frontière entre un état avec un faible degré d'émotion (étiquette Neutre) et une Emotion.



**1 - NEGATIVE**

**PEUR : débit non régulier, dévoisement, hésitation, répétition, souffle, pleurs**

**1 anxiety** (inquiétude) : pas de la peur, se faire du souci, avoir une inquiétude.
Indices prosodiques : silence, allongement syllabique, hésitations, soupir.
Indices lexicaux : énumération et répétition des symptômes observés, tentative de trouver des explications.
Il peut y avoir une ambiguité avec *sadness* à cause du ton plaintif de l'appelant.

*Exemples (nom du fichier et timecode)*

et je j'arrive pas à dormir (3su_2_3-41.029-43.522)
….

**2 stress :** accélération du débit, bafouillement, répétitions. Le stress peut être une nervosité naturelle, inhérente à la personnalité de l'appelant, ou une nervosité provoquée par la situation angoissante.

**Figure 3-5.** *Un extrait du protocole d'annotation.*

## *Outil d'annotation : le logiciel transcriber*

Les annotations ont été effectuées avec le logiciel Transcriber, déjà utilisé pour la transcription [Barras et al. 2000] qui a été enrichi d'une dtd[44] émotion (voir Figure 3-6).

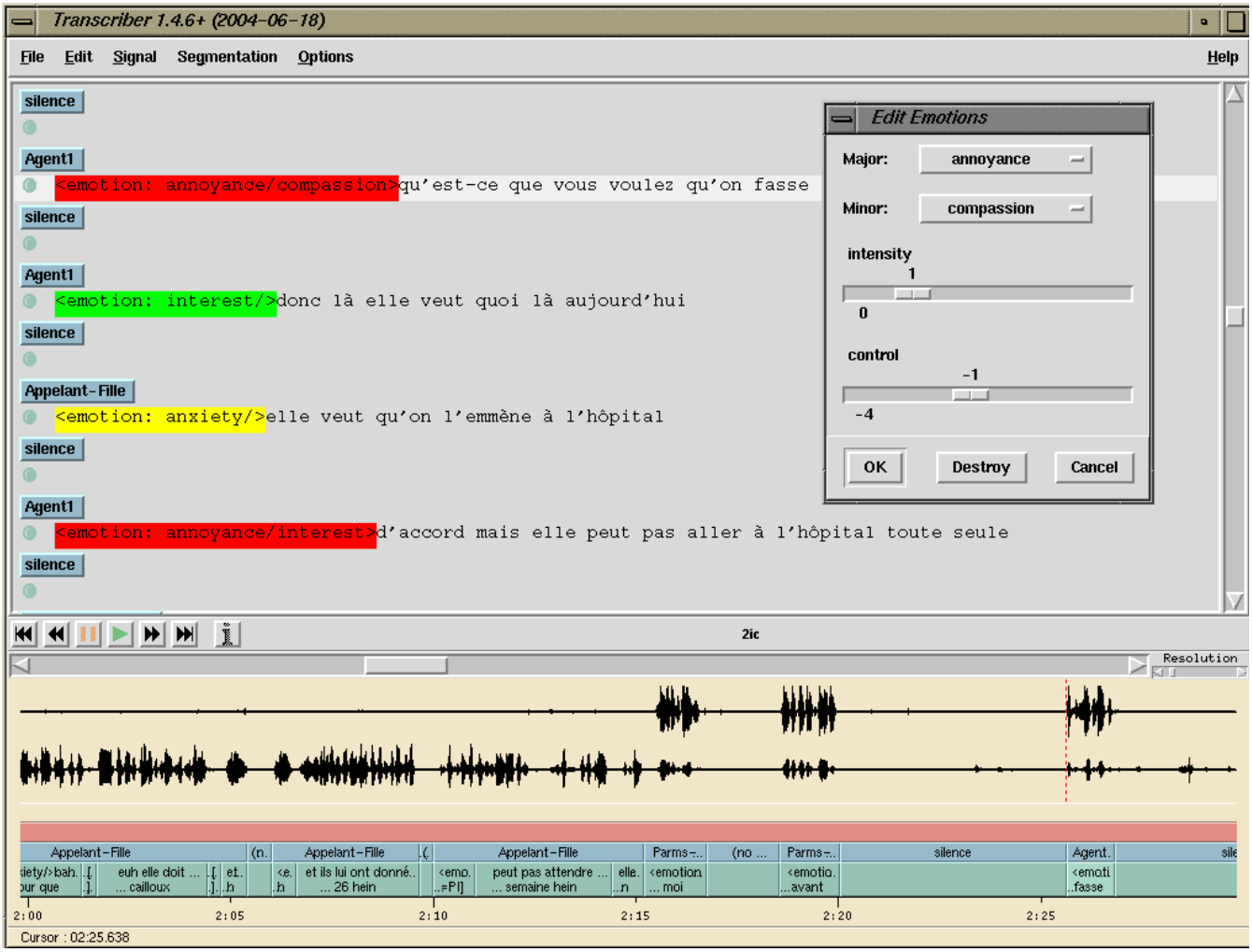



**Figure 3-6.** *Logiciel transcriber avec une dtd émotion utilisée pour l'annotation. L'extrait se situe à la fin d'un dialogue assez long entre un agent et la fille d'une patiente qui appelle pour la deuxième fois en quelques jours. La fois précédente, une ambulance avait été envoyée, mais la situation avait été considérée comme non critique et la patiente avait été ramenée chez elle. L'agent n'arrivant pas à déterminer précisément le motif de l'appel est un peu agacé par la situation, malgré sa compassion pour la patiente.*

Les tours « Neutre » ne sont pas annotés, sauf si passage d'un état émotionnel à un état Neutre au milieu d'un tour. L'annotation se présente ensuite sous forme de balises.

[44]dtd :Document Type Definition

### 3.2.3. Validation

Comment comparer et valider N annotations avec des unités parfois différentes et éventuellement plusieurs étiquettes émotion ?

Les annotateurs ayant chacun la possibilité de couper le tour de parole, la première nécessité est de choisir la taille du segment émotionnel final, puis d'appliquer des mesures de validité.

Comme indiqué page 49, plusieurs mesures d'**inter-annotation** existent, comme par exemple le coefficient Kappa dans le cas de données ordinales. Ces mesures se font sur des segments avec une seule étiquette. De plus, elles s'appuient sur le principe d'existence d'une annotation correcte vers laquelle les annotateurs doivent idéalement converger. Nous partons du principe qu'il n'y a pas systématiquement une «bonne» annotation, mais qu'il peut y avoir différentes perceptions possibles et plus particulièrement lorsqu'on se penche sur des étiquettes très fines et complexes. Cependant, si deux annotateurs ont une même perception, ils doivent utiliser la même étiquette. La proportion d'émotions complexes étant assez faible, nous nous attendons toutefois globalement à une certaine convergence, que nous voulons pouvoir comparer à d'autres études semblable : c'est pourquoi nous avons calculé le **coefficient kappa**.

Par ailleurs, s'il y a des segments perçus de la même manière par un grand nombre d'auditeurs naïfs, notre annotation doit refléter cette perception. Pour le vérifier, nous avons réalisé un **test perceptif** en utilisant à la fois des segments ou les jugements des deux annotateurs experts convergeaient et d'autres où il n'y avait pas d'accord.

Un bon annotateur devra en tout cas être cohérent avec lui-même (**intra-cohérence**), ce qui est plus facile à mesurer.

## *Segment émotionnel :*

Chaque annotateur a choisi son segment émotionnel et l'a annoté. Finalement, 1,4% (466) des tours de parole ont été coupés par l'annotateur 1 et 1,6% (395) par l'annotateur 2. Ces coupures ne concernent pas nécessairement les mêmes tours de parole et dans le cas où le même tour est coupé, ce n'est pas forcément au même endroit. (*cf.* exemples Figure 3-7 et Figure 3-8). Dans le cas de la Figure 3-7, les 2 annotateurs ont perçu une progression séquentielle de l'agacement vers la peur, mais pas au même moment.

| | | | **Anno1** | **Anno2** |
|---|---|---|---|---|
| APPELANT | t1 | [Parole] | Agacement | Agacement |
| | t2 | [Bruit de bouche] | Anxiété/ Stress | |
| | t3 | [Parole] | | |
| | t4 | [Silence] | | Anxiété |
| | t5 | [Parole] | | |
| | T6 | [Silence] | | |
| | t7 | [Parole] | | |

**Figure 3-7.** *Exemple de tour de parole coupé différemment par les 2 annotateurs. t1…t7 sont les time-codes correspondant au début des données transcrites à droite.*

Différentes stratégies sont envisageables pour combiner les choix des deux annotateurs. Dans le cas de la Figure 3-7, on pourrait couper le tour à t2 quand le premier changement est perçu, choisir de conserver le tour de parole comme unité en cas de désaccord ou demander à une troisième personne de trancher. Nous avons choisi de garder le segment le plus petit, afin d'avoir des émotions le plus « pures » possibles lorsqu'on entraînera des systèmes. Au final, il y aura donc un segment étiqueté *Agacement* de t1 à t2, suivi d'une phase de transition de t2 à t4 qui sera étiqueté avec la combinaison des annotations *Agacement* et *Anxiété/Stress* et enfin un segment correspondant à de la peur de t4 à t8. Cela peut amener à avoir des segments de taille trop petite pour être étudiés. Pour remédier à ce problème, il faudrait une phase de correction d'erreurs d'annotations et de « synchronisation » des segments. Par exemple dans le cas de la Figure 3-8, si un annotateur coupe le tour en t2 et l'autre en t3, le tour sera coupé en 3 segments avec le deuxième segment de taille très petite, composé d'un bruit de bouche annoté *Soulagement* par un annotateur et *Agacement* par l'autre. Cependant, ces cas sont très rares (une dizaine de cas) et une grande majorité des segments de très petite taille est due à des bruits ou échos en milieu de tour de parole. Pour être plus rigoureux et pouvoir s'adapter à un nombre plus grands d'annotateurs, il faudrait néanmoins ajouter une phase au protocole pour définir de façon consensuelle les unités émotionnelles et ensuite les annoter.

| | | | **Anno1** | **Anno2** |
|---|---|---|---|---|
| Appelant | T1 | Ah d'accord | Soulagement | Soulagement |
| | T2 | [Silence] | | |
| | T3 | [bruit de bouche] | | Agacement |
| | T4 | c'est pas xxxxxxx | Anxiété | |

**Figure 3-8.** *Exemple 2 : tour de parole coupé différemment par les 2 annotateurs.*

### 3.2.4. Cohérence inter-annotateur : le coefficient kappa

Le calcul du Kappa suppose que les classes d'émotions soient indépendantes. Or pour des étiquettes fines, la distance entre les classes *Agacement* et *Surprise* et plus grande qu'entre *Agacement* et *Colère* par exemple. Nous pourrions utiliser le kappa pondéré, mais comment choisir la pondération ?

Afin de pouvoir se comparer à des études similaires et de ne pas avoir à gérer les problèmes de proximité entre les classes fines, le kappa a été calculé sur les grandes classes. Reste le problème que le Kappa se calcule normalement avec une étiquette par classe.

#### *Comment gérer les doubles étiquettes ?*

Une tentative de solution a été apportée par Rosenberg *et al.* [Rosenberg et Binkowski 2004], mais les valeurs de kappa deviennent alors très faibles et il n'y a aucune valeur canonique pour juger du résultat et nous ne savons pas comment analyser le coefficient résultant. Si on ne regarde que les classes larges, les annotations du type «2 étiquettes de degré différent» (*Agacement/Colère chaude*) ou «mélange de deux émotions d'une même grande classe» (*Agacement/Impatience*) deviennent équivalentes à une seule étiquette «*Colère*». Idem dans le cas de la surprise, où le *Mineur* doit seulement préciser la valence par une étiquette «*Positif*» ou «*Négatif*».

On peut alors regarder uniquement le *Majeur*, considérant que les étiquettes doubles sont moins fréquentes que les simples et que le *Majeur* correspond à l'émotion principale perçue. Dans ce cas cependant, des annotations du type Annotateur1: *Agacement/Anxiété*; Annotateur2 : *Anxiété/Agacement* seront considérées comme différentes.

Il aurait été possible de faire des règles d'accord comme par exemple de considérer comme égales (ou demi-égales) deux annotations inversées "Majeur/Mineur» =«Mineur/Majeur". Ce type de règles peut s'avérer difficile avec beaucoup d'annotateurs.

Finalement, pour les données CEMO, le Kappa est de **0,57 pour les clients** et de **0,35 pour les agents** lorsqu'on ne regarde que le *Majeur*. La plupart des désaccords sont entre un état neutre et un état émotionnel. Ces valeurs sont du même ordre que celles trouvées dans d'autres études (0.48 pour [Grimm et al. 2007] avec4 émotions actées et 4 juges). Le Kappa est légèrement supérieur lorsqu'on utilise des règles d'accord entre 2 annotations.

### 3.2.5. Cohérence intra-annotateur : ré-annotation

Nous avons voulu évaluer la cohérence dans le temps des annotateurs. Pour ce faire, des ensembles de dialogues ont été réannotés à différents moments (après un mois par exemple). La fiabilité des annotations semble se stabiliser à 85% (*cf.* Tableau 3-7).

| | **Dec-Fev** | **Jan-Fev** | **Mar-Avr** | **Avr-Mai** |
|---|---|---|---|---|
| **Agent** | 76.4 (369 seg.) | 82.9 (287 seg.) | 86.1 (495 seg.) | 85.7 (405 seg.) |
| | 66.5 (369 seg.) | 80.8 (279 seg.) | 86.8 (499 seg.) | 87.6 (412 seg.) |
| **Client** | 73.9 (356 seg.) | 83.9 (255 seg.) | 83.4 (499 seg.) | 84.2 (442 seg.) |
| | 78.5 (350 seg.) | 76.5 (254 seg.) | 81.4 (505 seg.) | 85.8 (450 seg.) |

**Tableau 3-7.** *Mesure de fiabilité d'un annotateur : % accord entre deux annotations par un même annotateur à deux moments différents. Dec-Fev signifie une première annotation en décembre et une deuxième en février, (14 dialogues), Jan-Fev première annotation en janvier,deuxième en février (11 dialogues), Mar-Avr (16 dialogues), Avr-Mai (16 dialogues). Les 2 lignes pour agent et client correspondent aux 2 annotateurs.*

### 3.2.6. Test perceptif

Un test perceptif a été réalisé auprès de 40 sujets à la fois sur des segments simples où les 2 annotateurs étaient d'accords et sur des segments où ils n'étaient pas d'accord (voir Chapitre 4 p80 pour une description du test perceptif). Ce test a validé à la fois l'existence d'émotions doubles et l'expertise des annotateurs. Il n'a pas mis en évidence de cas où l'ensemble des sujets convergeait sur des segments pour lesquels les annotateurs experts ne percevaient pas la même émotion. (Toutefois, il aurait fallu tester plus de segments). Les sujets pouvaient ajouter une étiquette libre et le test n'a pas révélé de failles dans notre ensemble d'étiquettes (*cf.* p 86).

## 3.3. Combiner les annotations : un vecteur émotion

Comme chaque segment était annoté par plusieurs annotateurs et qu'on pouvait lui assigner une ou deux étiquettes, il a été nécessaire de créer un « mapping » afin de pouvoir effectuer un apprentissage.

Une annotation sera considérée comme un vecteur (Majeur, Mineur) et N annotations seront combinées en un vecteur *émotion* (voir exemple Figure 3-9 ) : la taille du vecteur émotion est le nombre d'émotions étudiées. Différents poids sont attribués au Majeur ($w_M$) et au Mineur ($w_m$) et les différentes annotations sont sommées et moyennées. Cette représentation sous forme de vecteur d'émotions a été également utilisée représenter des émotions complexes dans des données audio-visuelles [Devillers et al. 2005a].

**Segment annotation :**

**Annotateur 1 :** *Majeur* Agacement, *Mineur* Intérêt

**Annotateur 2 :** *Majeur* Stress, *Mineur* Agacement

**Conversion en un vecteur émotion :**

-> (wM/W **Agacement**, wm/W **Stress**, wm/W **Interêt)**

Pour wM=2, wm=1, W=6

-> (0. 5 Enervement, 0.33 Stress, 0,17 Intérêt).

**Figure 3-9.** *Exemple de création d'un vecteur d'émotions pondérées.*

Si on souhaite mettre une étiquette finale à un segment (analyse, détection), elle se déduit des champs les plus grands du vecteur *émotion*. A peu près 50% du corpus a ainsi été étiqueté comme neutre.

Les étiquettes émotions sont différentes pour les agents et les appelants. Globalement, les plus fréquentes dans la classe *Positive* sont le *Soulagement*, l'*Intérêt* et la *Compassion*, ceux de la classe *Négative* sont le *Stress*, l'*Anxiété*, l'*Agacement*, l'*Impatience* et l'*Embarras*.

La proportion des étiquettes les plus fréquentes pour les agents et les clients est donnée dans le Tableau 3-8 ci-dessous.

| Client | Neutre | Anxiété | Stress | Soulagement | Douleur | Autre |
|---|---|---|---|---|---|---|
| 10810 seg. | 67.6% | 17,7% | 6.5% | 2.7% | 1.1% | 4.5% |
| Agent | Neutre | Intérêt | Compassion | Agacement | Surprise | Autre |
| 11207 seg. | 89.2% | 6.1% | 1.9% | 1.7% | 0.6% | 0.6% |

**Tableau 3-8.** *Répartition des étiquettes fines (5 meilleures classes) avec le même Majeur. (688 dialogues), Autre donne le pourcentage de segments annotés avec les 19 étiquettes restantes.*

## 3.4. Clustering sur les annotations utilisant un algorithme divisif

La matrice de confusion des annotations permet également de donner une idée des classes émotions qui sont le plus différenciées par les annotateurs. Elle peut être représentée graphiquement sous forme de clustering hiérarchique aussi appelé dendrogramme [Kaufman et Rousseeuw 1990] : les données sont représentées sous la forme d'un arbre binaire dans lequel la distance verticale entre deux feuilles est fonction de leur distance dans la matrice de confusion. Deux manières existent pour construire l'arbre :

- méthode descendante (division) : on commence avec un groupe contenant toutes les données et on le divise à chaque itération en utilisant des mesures de distance.
- méthode ascendante (agglomération) : on commence avec chaque classe de données dans un groupe séparé et les données les plus proches sont regroupées à chaque itération.

Les dendrogrammes ont été construits pour les agents et les appelants à l'aide du logiciel libre R[45], en utilisant la méthode ascendante Agnès avec la distance euclidienne (des figures similaires étaient obtenues avec d'autres distances et d'autres méthodes)

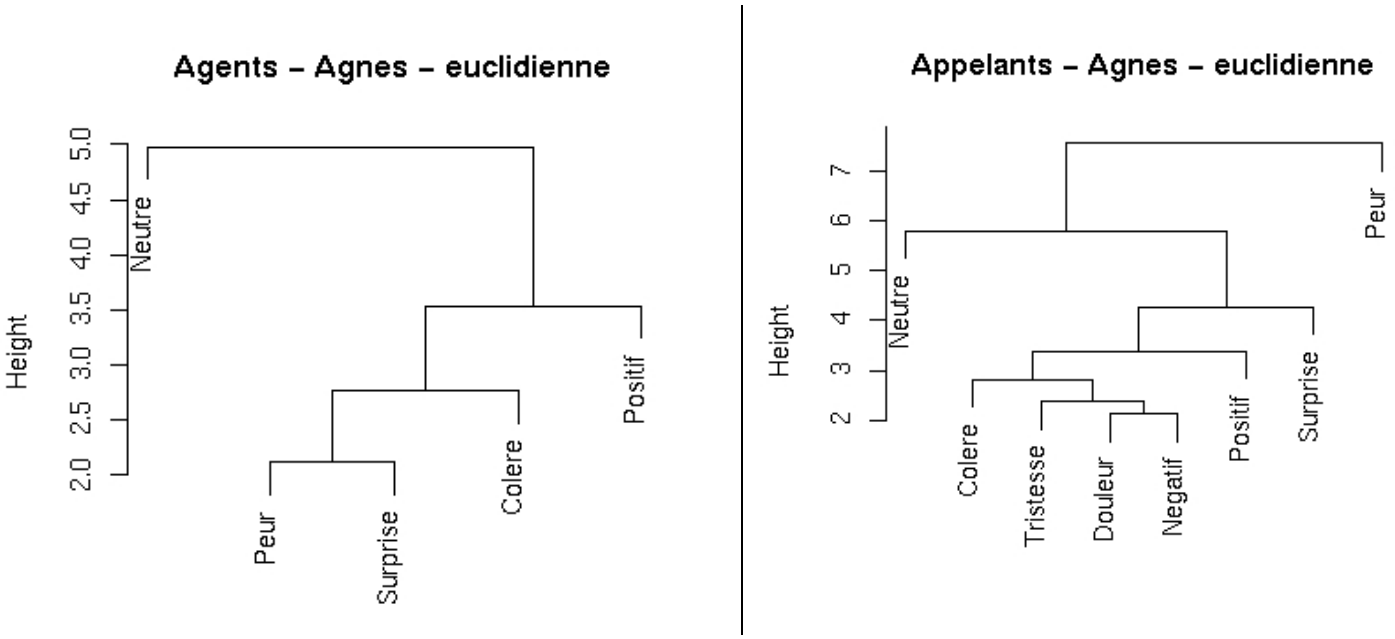



**Figure 3-10.** *Dendrogrammes issus du clustering agglomératif utilisant une distance euclidienne.*

[45] www.r-project.org

Finalement, les classes les mieux différenciées sont celles les plus représentées avec pour les agents un groupe *Neutre* et un groupe *Emotion* qui peut être séparé entre *Positif* et *Autre émotion* et pour les appelants un groupe *Peur*, un groupe *Neutre* et un groupe avec les autres émotions.

## 3.5. Conclusion

En conclusion nous avons proposé un schéma d'annotation et un protocole de validation de nos étiquettes émotions. Nous n'avons pas encore exploité toutes les annotations de ce corpus.

Un des résultats de ma thèse est l'observation et la représentation d'émotions complexes dans les données spontanées. Des analyses perceptives ont été menées pour valider la présence d'émotions complexes, nous les décrivons dans le chapitre 4.

# Chapitre 4

# Analyse des mélanges d'émotions dans le corpus CEMO

Résumé

Nous disposons désormais d'un ensemble de données annotées de manière fiable avec un vecteur émotion par segment. Nous pouvons alors nous pencher plus en détails sur les données qui seront utilisé pour une majorité de nos travaux. Dans ce chapitre, nous donnons d'abord un aperçu de la variété du corpus. Nous allons ensuite nous focaliser sur les mélanges d'émotions, et en particulier les mélanges positifs et négatifs peu souvent étudiés, à travers deux expériences perceptives. Comment valider leur présence ? Ces mélanges peuvent-ils être perçus hors contexte ? Quels types d'indices permettent de les percevoir ? Une expérience perceptive sur les types d'indices permettant de percevoir les mélanges émotions sera décrite. Cette expérience a été généralisée sur d'autres catégories de mélanges et validée avec des annotateurs naïfs.

*We are now armed with a 20-hour corpus precisely annotated, each segment being described by an emotion vector. In this chapter we focus on emotion mixtures, what has seldom been done and look more closely at the data that will be used for most of our experiments. We also describe 2 perceptive tests, one investigating the different type of cues that enable expert annotators to perceive 2 emotions at the same time, and one focusing on how naïve judges perceive these mixtures in conditions close to those of the detection system (data removed from its original context). Those tests also serve as a validation of our annotation protocol.*

# 4. ANALYSE DES MELANGES D'EMOTIONS DANS LE CORPUS CEMO

## 4.1. Distribution des émotions

L'émotion exprimée va dépendre de nombreux facteurs tels que l'âge du sujet, son sexe, son rôle ou sa relation avec le patient. De plus, des manifestations très variées vont être désignées par la même étiquette émotion. La Figure 4-1 indique la répartition des émotions pour les agents, pour les 20 heures de données qui ont été annotées. Comme il a déjà été indiqué, les émotions exprimées sont principalement négatives, pour l'appelant comme pour les agents. Si l'on exclut les segments ambigus (à la limite entre un état neutre et une émotion), près de la moitié du corpus contient des données émotionnelles, principalement de la peur. En poussant l'analyse un peu plus loin, on pourrait distinguer des profils types suivant le « type » d'appelant (sexe, relation avec le patient, âge ?) ou entre les différents agents [Devillers et al. 2004]. Il y a en effet des différences de comportement suivant la personne qui interagit et Campbell montre par exemple qu'il y aura des différences significatives de certains paramètres acoustiques suivant que quelqu'un interagit avec quelqu'un de sa famille, ses amis ou une personne moins proche [Campbell et Mokhtari 2003].

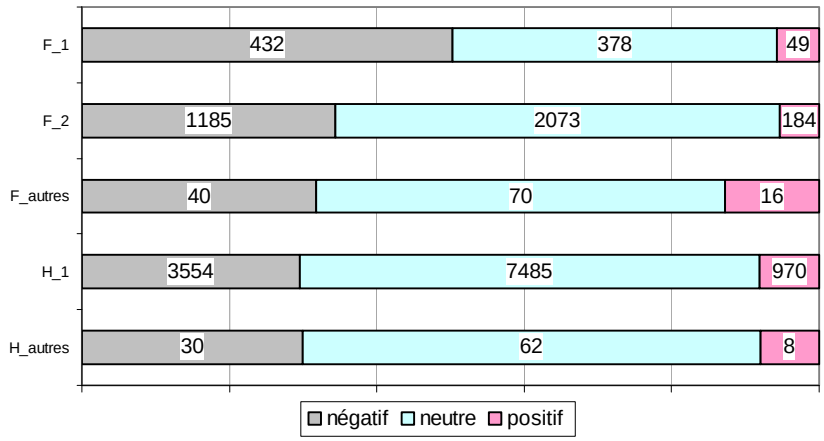



**Figure 4-1.** *Répartition des émotions entre positif, négatif et neutre pour les agents. Dans les données récoltées, 3 agents interviennent beaucoup F_1, F_2 et H_1. Les nombres indiquent le nombre de segment pour chaque cas.*

Hess reporte des différences dans l'expression des émotions selon le sexe, les femmes exprimant en général plus de peur que les hommes et exprimant certaines émotions différemment, par exemple en pleurant plus quand elles se mettent en colère[Hess 2006]. Sur notre échantillon de 20h de conversation, les hommes semblent moins émotionnels que les femmes lorsque le patient n'est pas un proche et ils expriment un peu plus de colère.

# 4.2. Les mélanges d'émotions

## 4.2.1. Différents cas dans le corpus CEMO

Comme indiqué dans le chapitre Annotation, les annotateurs ont respectivement annoté 31% et 17% de mélanges. Une typologie des mélanges a été dérivée des différents cas.

On définit les mélanges d'émotion comme *ambigus* quand les émotions appartiennent à la même grande classe (i.e. Colère froide/Agacement/Colère chaude), *conflictuels* si elles appartiennent à des classes différentes de valence différente (i.e. Anxiété/Amusement) et *non conflictuels* si elles appartiennent à des classes différentes de même valence (i.e. Anxiété/Agacement). L'étiquette *Surprise* étant à part, les émotions mixtes contenant de la surprise seront étudiées séparément. Les proportions des différents types de mélanges annotés sont indiquées Figure 4-2.

Les étiquettes ambiguës seront le plus souvent utilisées lorsqu'un annotateur ne trouve pas d'étiquette correspondant exactement à l'émotion qu'il perçoit. Sa stratégie sera alors de donner deux étiquettes proches. Dans les autres cas d'émotions mixtes, il perçoit des émotions différentes simultanément. Elles sont d'ailleurs souvent causes d'erreurs lorsqu'elles sont utilisées pour créer ou tester des modèles[Vidrascu et Devillers 2005c] ).

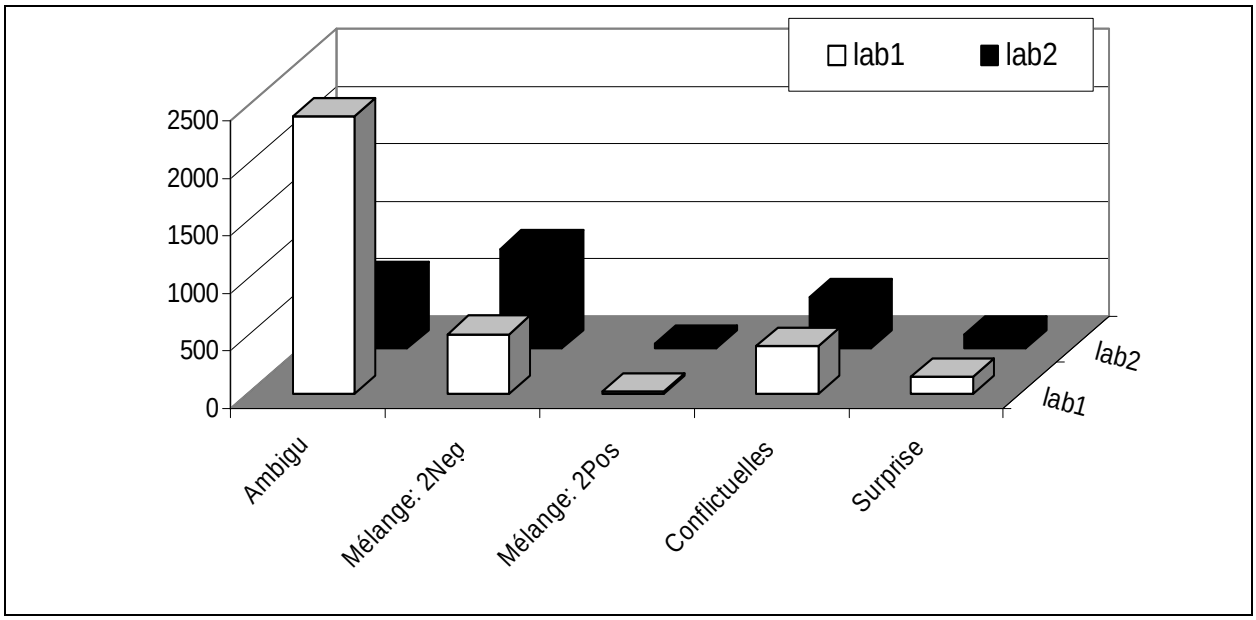



**Figure 4-2.** *Répartition des mélanges d'émotion pour chaque annotateur. lab1 and lab2 sont les 2 annotateurs; Mélange: 2Pos signifie que les 2 étiquettes sont choisies dans des classes différentes d'émotions positives ('Amusement', 'Soulagement', 'Compassion/Intérêt' ); Mélange: 2Neg signifie que les 2 étiquettes sont choisies dans 2 classes négatives différentes ('Peur', 'Colère', 'Tristesse' et 'Douleur' ).*

Nous parlons d'émotions conflictuelles lorsqu'un segment est annoté à la fois par une étiquette positive et par une étiquette négative. Elles sont différentes selon leur position dans le dialogue (entre les agents et les appelants). Un exemple typique dans notre corpus sera pour un agent d'éprouver de la compassion envers un appelant teintée par de l'agacement (ou d'essayer de masquer son irritation par une voix compatissante): « *je comprends bien Madame mais j'ai pas de solution miracle* ». Pareillement, un appelant pourra être réconforté par l'agent et exprimer son soulagement en même temps que du stress ou de l'anxiété après une intervention de l'agent « *Hum d'accord là j(e) là je me sens beaucoup [parole inintelligible] parce que j'ai les renseignements parce que ».*

Les deux émotions peuvent être séquentielles, mais elles peuvent aussi être exprimées en même temps.

### 4.2.2. Différents indices : Une étude sur les « émotions conflictuelles »

#### *Objectif et protocole*

Une étude a été conduite [Vidrascu et Devillers 2005b] afin de valider les mélanges d'émotions positives et négatives et de les étudier plus précisément. L'objectif était à la fois de voir si une des deux émotions était vraiment dominante et si les types d'indices permettant de percevoir les 2 émotions étaient différents (par exemple une émotion est-elle perçue grâce à des indices lexicaux et l'autre grâce à des indices prosodiques ?). Le focus de l'étude étant le type d'indices utilisés, l'expérience a été réalisée par des personnes « expertes » connaissant le corpus et familières avec les différents mélanges d'émotions et les différents types d'indices.

30 segments (20 appelants, 10 agents) où chacun des deux annotateurs « experts » avaient perçu un mélange positif/négatif[46] ont été réannotés par 3 personnes (dont les 2 annotateurs). En plus de choisir une ou deux étiquettes émotions par segments, il leur était demandé de préciser les indices qui avaient motivé leur choix pour chaque étiquette, en choisissant une ou plusieurs des catégories :

- *Indices lexicaux* : mots et syntaxe
- *Indices prosodiques* : rythme, mélodie, platitude
- *Disfluences* : pauses vides, hésitations (« euh »), répétitions
- *Contexte* : segment précédent, rôle (appelant, agent, témoin …)

#### *Résultats*

Tout d'abord, la deuxième annotation était cohérente avec la première : chacun des 3 annotateurs ont perçu un mélange d'émotions conflictuelles pour la majorité des segments (sauf pour 6 segments sur les 30 où seulement 2 annotateurs sur 3 ont perçu un mélange conflictuel) avec globalement les mêmes classes que pour la première annotation47 (*cf.* exemple Figure 4-3).

Afin de sélectionner les segments où une des 2 émotions « dominait », une quatrième personne a annoté les segments avec les étiquettes *Positive* et *Négative*.

[46] Les 2 annotateurs n'étaient pas forcément d'accord sur les étiquettes fines ni sur l'émotion Majeur entre la positive et la négative.

[47] Malgré que la première annotation a été faite en contexte et pas la deuxième

Pour chaque segment, les 4 annotations ont été mélangées en un vecteur [Négatif, Positif] avec un poids de 2 pour le *Majeur* et de 1 pour le *Mineur (*voir p53 pour la définition de *Majeur* et *Mineur).* Une étiquette POS/NEG (étiquette avec le plus grand poids gagnant positive et deuxième étiquette négative), NEG/POS (cas contraire) se déduisait ensuite de ce vecteur (voir Figure 4-3).

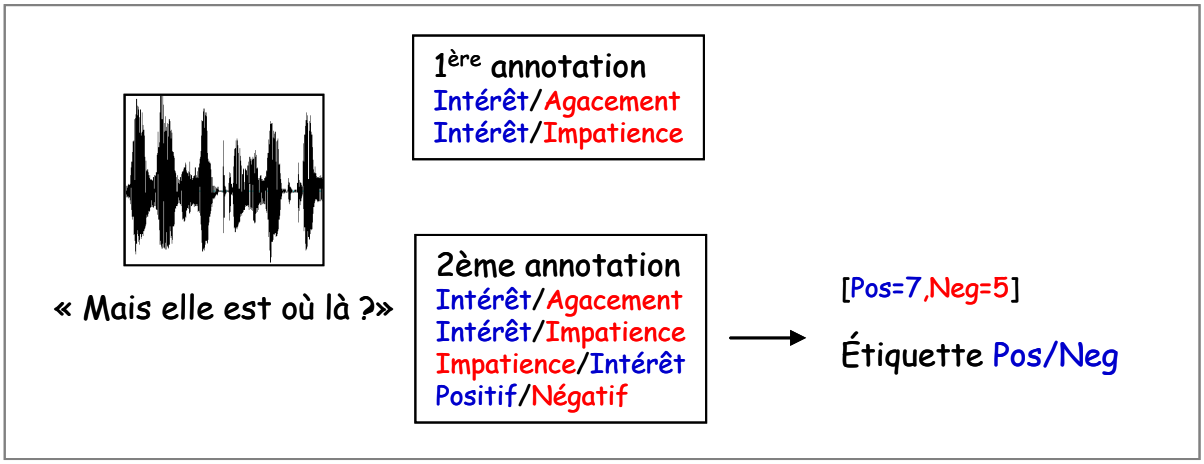



**Figure 4-3.** *Exemple de segment réannotés.*

Les indices de contexte et disfluences étant très peu utilisés, seuls les indices prosodiques et lexicaux ont été pris en compte pour l'étude. Une technique de vote majoritaire (accord entre au moins deux juges) a été utilisée pour sélectionner les types d'indices perçus par segment émotionnel.

La Figure 4-4 indique pour chaque classe POS/NEG et NEG/POS les pourcentages de segments pour lesquels des indices prosodiques ou lexicaux ont été perçus respectivement pour les émotions *Majeur* et *Mineur*. Par exemple des indices prosodiques sont utilisés pour percevoir le *Majeur* dans plus de 90% des segments. La figure révèle des répartitions similaires pour les *Majeur* et *Mineur* des deux classes POS/NEG et NEG/POS. Les données étaient assez équilibrées avec respectivement 14 et 15 segments pour les classes POS/NEG et NEG/POS. Chaque type d'indice (lexical et prosodique) est deux fois plus sélectionné pour l'étiquette *Majeur* que *Mineur* quelle que soit la classe. De plus, le *Mineur* est seulement perçu grâce à un indice, soit lexical (10%), soit prosodique (70%), jamais les 2 à la fois. Une autre observation intéressante est que les indices lexicaux ne sont jamais sélectionnés à la fois pour la perception des émotions *Majeur* et *Mineur*, ce qui est par contre souvent le cas pour les indices prosodiques.

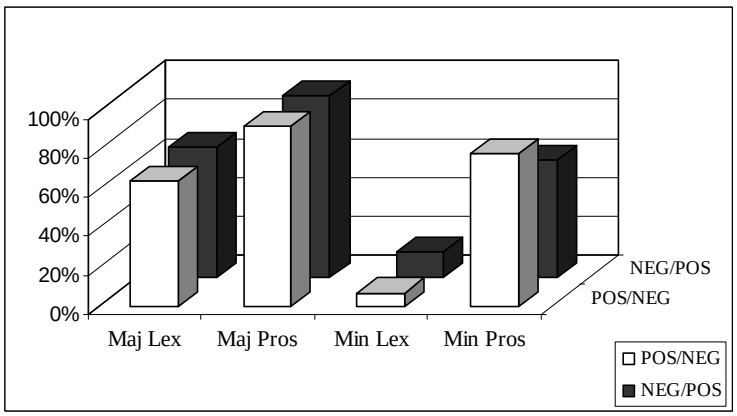



**Figure 4-4.** *Répartition des indices lexicaux et prosodique entre le Majeur et le Mineur pour les « émotions conflictuelles », (appelants et agents).*

Bien que d'autres tests sur plus de segments et avec plus d'annotateurs soient nécessaires pour confirmer certaines des observations, cette expérience a apporté *une validation à l'existence dans le corpus de mélanges* ***d'émotions*****, de valences opposées et a mis en évidence la pertinence des indices lexicaux et prosodiques**. Les types d'indices ne sont pas nécessairement différents pour les 2 émotions et les indices prosodiques peuvent permettre de percevoir les deux émotions. L'émotion perçue grâce au plus grand nombre de types d'indices est analysée comme dominante.

### 4.2.3. Test perceptif sur les émotions complexes

#### *Objectif et protocole*

Le premier objectif de ce test [Vidrascu et Devillers 2006] était de valider le protocole d'annotation :

- **l'expertise des 2 annotateurs** : 2 annotateurs sont-ils suffisants pour annoter le corpus de manière fiable ? En quoi les 2 annotateurs sont-ils experts ? Un objectif était de confronter les annotations des deux experts à celles d'un « grand nombre » de juges naïfs. Nous avons donc sélectionné un sous ensemble de segments, majoritairement où les 2 experts s'accordaient, en prenant soin de ne pas inclure de données confidentielles. Quelques segments pour lesquels les 2 experts divergeaient ont également été sélectionnés (une convergence des perceptions des juges naïfs sur ces segments aurait remis en cause l'annotation des données). Pour le traitement automatique des émotions, le contexte n'est pas pris en compte : les segments étaient donc présentés hors contexte.
- **Le choix des étiquettes** : est ce que les étiquettes proposées sont pertinentes pour la description du corpus ? La distinction entre des étiquettes fines du type *Agacement* et *Impatience* ou I*ntérêt* et *Compassion* est-elle utile malgré les confusions dans certains cas.

**La valence :** était-il erroné de considérer que la valence de l'émotion pouvait être déduite de l'étiquette émotion ?

Le test perceptif se voulait également un complément aux études précédentes sur les mélanges d'émotions.

- Ces mélanges sont ils perçus hors contexte ?
- Dans le cas affirmatif, sont ils plutôt perçus comme simultanés ou séquentiels ?
- Quels types d'indices sont pertinents pour leur perception ?

Une quarantaine de stimuli ont été sélectionnés parmi lesquels 14 segments « simples » (annoté par une seule même étiquette par les deux annotateurs), 11 mélanges non conflictuels, 13 mélanges conflictuels et 3 segments « complexes » pour lesquels les annotateurs ne s'accordaient pas. Les mélanges conflictuels correspondaient aux cas prototypiques d'agents exprimant à la fois de la compassion et de l'irritation, d'appelants exprimant de la peur et le soulagement de savoir qu'on allait les aider, ou encore de l'embarras et du « self amusement », comme par exemple une baby-sitter, qui lorsqu'on lui demande le numéro de l'appartement

répond avec un rire embarrassé «*je vais juste prendre mon agenda parce que je sais même plus quel numéro on est euh ne quittez pas* »

Les stimuli étaient présentés hors contexte à 44 sujets : 34 français natifs (13 femmes et 21 hommes) et 10 non natifs et pouvaient être rejoués à volonté. Le sujet devait évaluer la valence du stimulus (du très négatif au très positif) sur une échelle allant de -3 à +3. Il devait ensuite choisir une étiquette pour l'émotion perçue dans la liste : Neutre, Anxiété, Stress, Peur, Panique, Agacement, Impatience, Colère froide, Colère chaude, Embarras, Déception, Tristesse, Désarroi, Résignation, Désespoir, Surprise, Soulagement, Intérêt, Compassion, Amusement, Douleur. Des définitions des différents termes lui étaient fournies et il avait la possibilité d'interroger l'expérimentateur pendant le déroulement du test. L'intensité et le contrôle pour cette émotion étaient ensuite évalués sur une échelle de 1 à 5 et le sujet devait donner le type d'indices qui lui permettaient de percevoir l'émotion (lexical, prosodique, disfluences …). S'il percevait une deuxième émotion, il devait la choisir dans la même liste et répondre aux mêmes questions en précisant en plus si les 2 émotions étaient perçues séquentiellement ou simultanément. Enfin, il pouvait donner le nom de l'émotion perçue si elle n'était pas dans la liste[48]. L'interface a été réalisée en tcl/Tk.

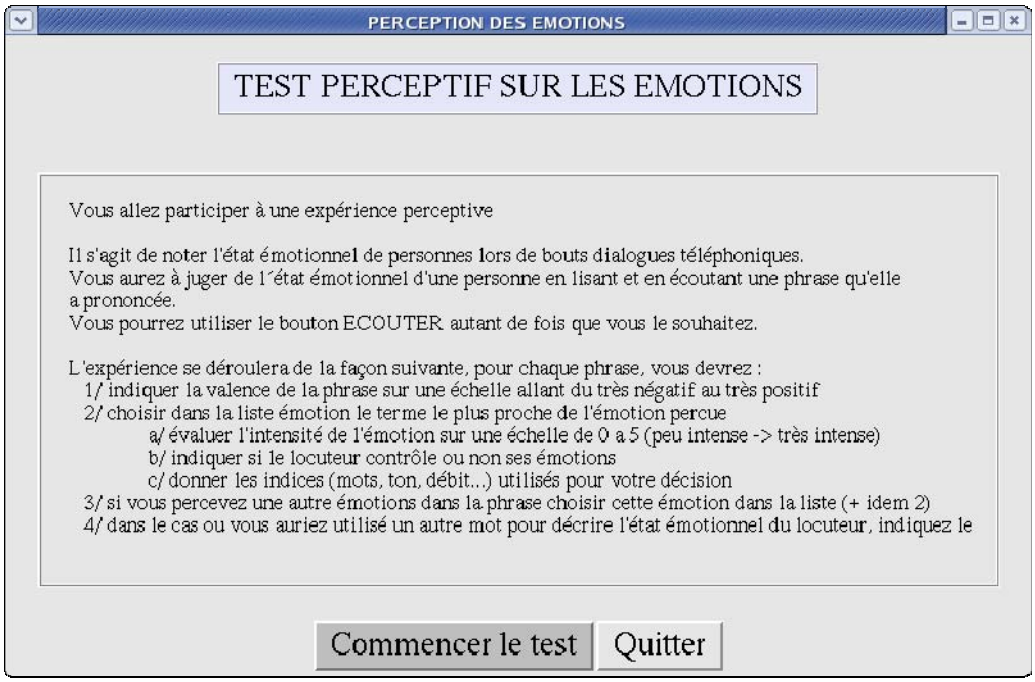



**Figure 4-5.** *Introduction et instructions du test perceptif.*

[48] Pour avoir des annotations libres, [Greasley et al. 2000] avaient procédé en demandant aux sujets d'utiliser des mots se référant à « comment une personne se sent ». Il aurait été judicieux de faire de même pour éviter des réponses du type « remerciement » ou « excuse ».

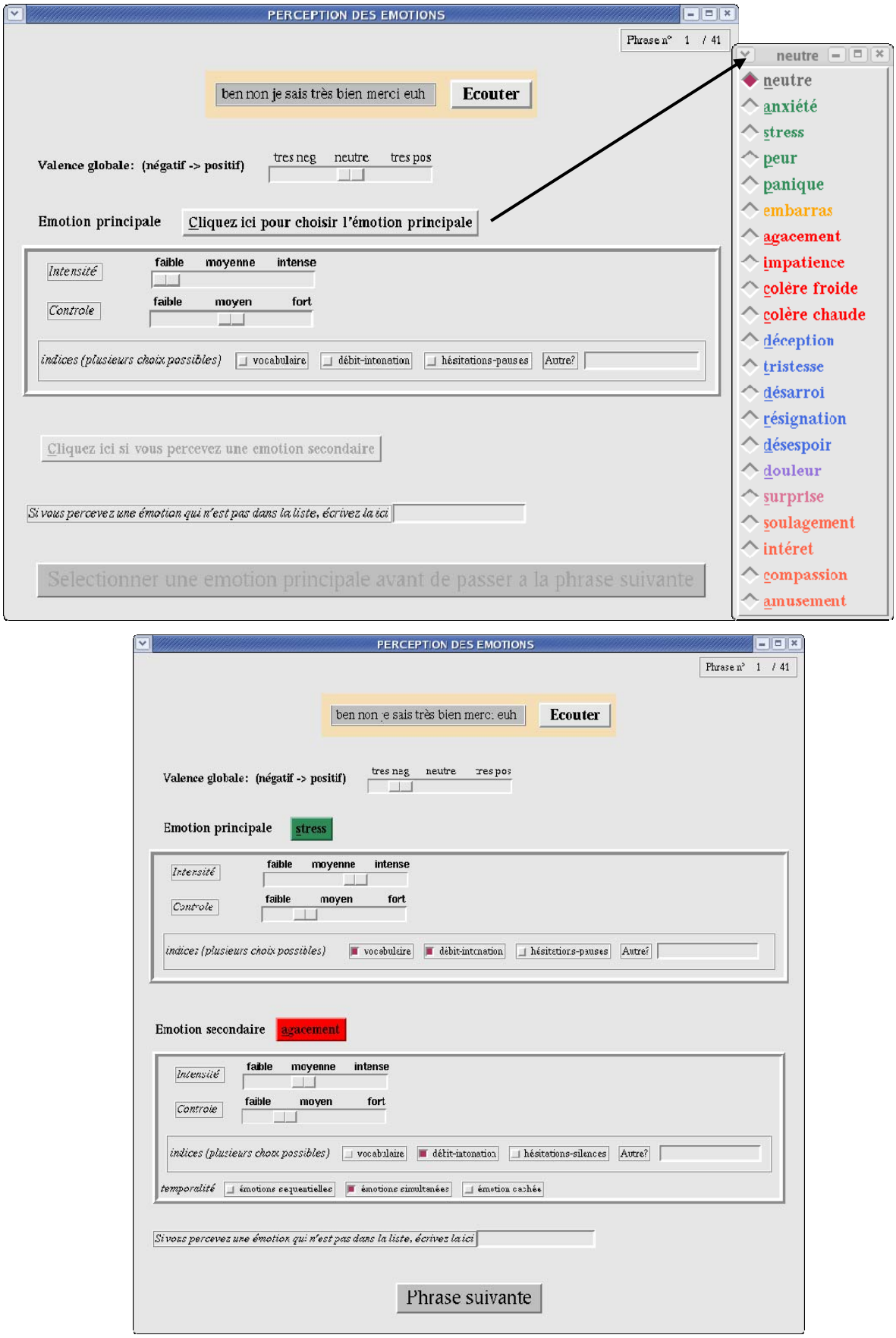



**Figure 4-6.** *Interface du test perceptif.*

## *Résultats*

Les non natifs ayant trouvé la tâche trop difficile à cause de l'absence de contexte et du grand nombre d'étiquettes, ils n'ont pas été pris en compte pour l'étude sur les mélanges d'émotions[1].

### Résultats par Sujet

Même sans contexte, tous les locuteurs français ont perçu des mélanges d'émotions et à part deux sujets, ils ont également perçu des mélanges d'émotions positives et négatives. Cependant un sujet percevait en moyenne un mélange d'émotions pour 9 segments, parfois d'ailleurs sur des segments simples, ce qui est peu. Le Tableau 4-1 donne pour chaque sous ensemble (émotion simple, mélange non conflictuel, mélange conflictuel) les pourcentages de mélanges annotés par les juges naïfs.

| Annotaté comme -> | **Simple/ambigu** | **Non conflictuel** | **Conflictuel** |
|---|---|---|---|
| **Simple** (14 segts) | 87% | 7% | 6% |
| **Non conflictuel** (11 segts) | 76% | 19% | 5% |
| **Conflictue**l (13 segts) | 71% | 10% | 18% |

**Tableau 4-1.** *Pourcentages d'émotions « simples » et complexes des 33 sujets français ayant effectué le test perceptif.*

Bien que tous les locuteurs aient perçu des mélanges d'émotions, 70% des stimuli complexes ont été annotés comme simples, les femmes percevant plus d'émotions conflictuelles que les hommes pour cette étude. Parallèlement, 15% des stimuli étaient jugés comme complexes alors qu'ils étaient étiquetés comme simple. Ces mauvais résultats montrent la difficulté pour des annotateurs naifs de percevoir les émotions complexes exprimées dans ces stimuli sans contexte.

### Résultats par vecteur

Même si les annotations individuelles des sujets ne correspondaient pas toujours à celle des experts, le vecteur combinant les annotations des sujets semble correspondre au vecteur combinant les annotations des experts[2]. En effet, les annotations des sujets ont été regroupées en un vecteur (Neutre, Peur, Colère, Tristesse, Compassion-Intérêt, Soulagement) par stimulus avec

[1] Par contre leurs performances étaient comparables à celles des natifs pour la reconnaissance des émotions « simples »

[2] Ce type de résultat se retrouve chez [Scherer et Ceschi] qui évaluent la fiabilité de 31 juges ensemble et séparément et obtiennent une très bonne fiabilité pour l'ensemble des juges, mais une fiabilité assez faible par juge en moyenne et précisent que ce phénomène est très fréquent dans les « rating studies using lay observers »

un poids de 1 pour le Majeur et le Mineur. Ce vecteur a été comparé à celui des 2 experts. On a d'abord regardé le plus grand coefficient de chaque vecteur : il y avait 85% d'accord entre sujets naïfs et experts. En considérant les 2 plus grands coefficients, il y avait accord pour 18 des 24 émotions complexes. Par exemple, un segment annoté Peur/Tristesse par les experts était réannoté en Peur par 50% des sujets, Tristesse par 30% des Sujets et Peur/Tristesse par 5%. Les cas de désaccord entre expert et naïfs étaient souvent expliqués par l'absence de contexte. Dans le cas de la Figure 4-7, l'intonation et la répétition de « oh ma pauvre » faisaient que les sujets percevaient de la compassion quand en contexte, le locuteur exprimait clairement une émotion négative. D'ailleurs, l'annotation de la valence du stimulus est négative et en contradiction avec l'étiquette. Seulement 2 sujets sur 34 ont choisi une valence strictement positive pour la phrase contre 25 sur 34 avec une valence strictement négative.

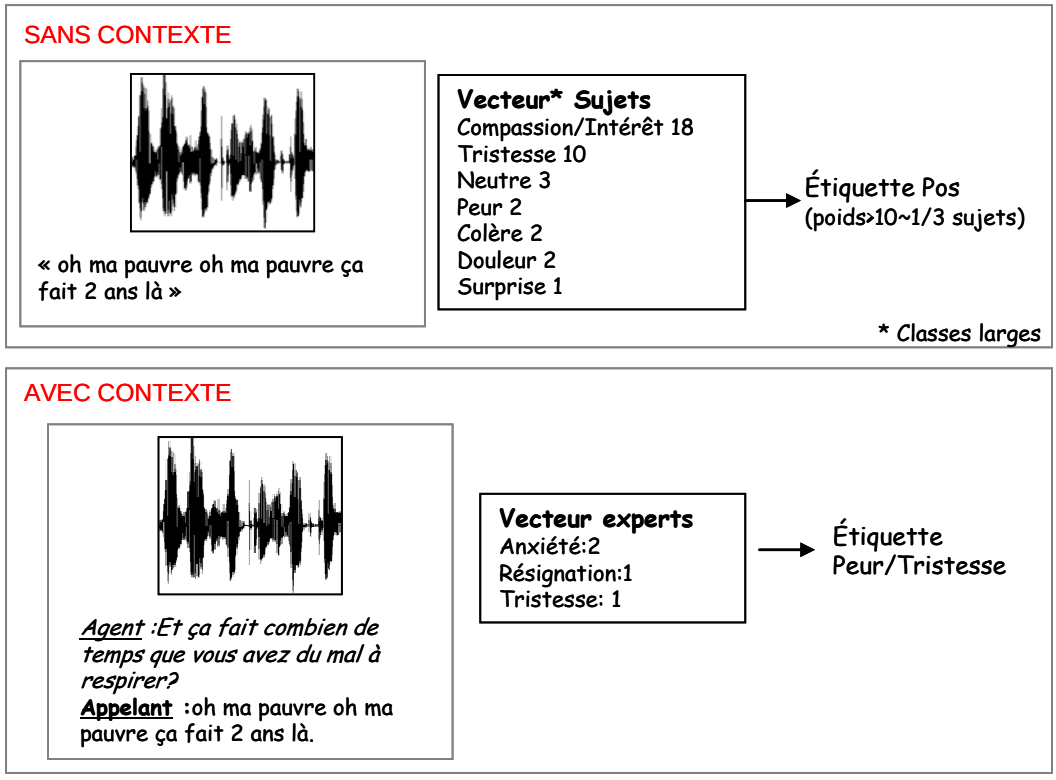



**Figure 4-7.** *Le rôle du contexte dans les différences entre les annotations. Les annotations des sujets sont regroupées en un vecteur émotion (étiquettes larges) avec un poids de 1 par étiquette. Pour déduire l'étiquette finale, on a choisi de ne garder que celles choisies par plus de 1/3 des sujets (poids>10).*

Pour les stimuli sur lesquels les experts n'étaient pas d'accord, aucun consensus n'a été trouvé parmi les sujets du test.

## Comparaison des différentes annotations avec celles d'un SVM

Dans le cas de la détection automatique des émotions, le système n'aura pas accès au contexte (ce pourquoi les sujets du test n'y avaient pas accès non plus) et en plus il n'aura pas accès à l'information lexicale. Nous avons profité des annotations du test perceptif pour comparer les étiquettes attribuées aux 41 stimuli par les annotateurs experts, les sujets naïfs et le système de détection (voir Chapitre 6 pour une description de ce système). Les 3 disposent de différents niveaux de connaissances, comme montré Tableau 4-2.

| | **paralinguistique** | **lexical** | **contexte** |
|---|---|---|---|
| **Annotateurs experts** | X | X | X |
| **Sujets naïfs** | X | X | |
| **Détection automatique** | X | | |

**Tableau 4-2.** *Différents niveaux d'information.*

Un « modèle » paralinguistique a été créé pour les agents avec les quatre émotions Neutre, Colère, Compassion et Surprise et un pour les appelants avec les 4 émotions Peur, Colère, Tristesse et Soulagement. Les prédictions pour les 41 segments (qui n'avaient pas été utilisés pour construire le modèle !) ont été comparées à celles des experts et des naïfs. Les pourcentages de même détection entre experts, naïfs et système automatique sont donnés dans le Tableau 4-3 ci-dessous.

| | |
|---|---|
| Experts=naïfs=automatique | 61 % |
| experts=naïfs | 85 % |
| expert=automatique | 66 % |

**Tableau 4-3.** *Pourcentage d'accord en ne considérant que le plus grand coefficient des vecteurs, expert : annotation initiale, naïf : annotation des sujets du test perceptif, automatique : détection automatique.*

Bien que le système ait été entraîné avec les annotations des experts, le pourcentage d'accord entre le système et les annotateurs est loin derrière celui entre experts et naïfs. Il faudrait apporter des informations sur le contexte et le lexique et les combiner aux indices paralinguistiques pour améliorer la détection.

## Validation des étiquettes

Les réponses données pour l'étiquette libre sont regroupées dans le Tableau 4-4. La formulation de l'énoncé étant trop ouverte, certaines réponses ne correspondent pas vraiment à des états affectifs (comme par exemple fermeté). La plupart des classes où une autre étiquette a été donnée correspondent à celles qui ont été redéfinies pour s'adapter aux données. Comme il a déjà été indiqué dans la description des données, l'étiquette *Soulagement* s'appliquait typiquement à un appelant en fin d'interaction à qui de l'aide allait être apportée et s'exprimait lexicalement souvent par « merci », d'où des étiquettes type « gratitude », « remerciement ».

De même, « Désarroi » avait été ajouté à la liste d'étiquettes pour décrire l'émotion exprimée par des locuteurs éprouvant un sentiment d'impuissance et se manifestant lexicalement par des phrases du type « je ne sais pas quoi faire ». La plupart du temps, les étiquettes désignent une même émotion ou sont hors sujet…

| | Choix libre |
|---|---|
| Intérêt<br>Compassion | Curiosité<br>politesse, rassurante |
| Soulagement | Reconnaissance (x5), remerciement, gratitude (x6) |
| Autre Positif | bonne humeur, satisfaction (x4), plaisir |
| Peur<br>Anxiété | crainte, suppliant<br>nervosité (x2) |
| Embarras<br>Désarroi<br>Résignation | Gêne (x4), honte, excuse, vexé (x2)<br>impuissance x2, incertain, indécision, confusion<br>indifférence, désintérêt (x2), ennui (x2), distraction, fatalisme |
| Agacement | exaspération (x3), énervement<br>Indignation |
| Surprise | étonnement (x2) |
| Autre | Réflexion<br>ironie (x6), autodérision<br>incrédulité (x2), fermeté |

**Tableau 4-4.** *Les résultats du choix libre pour l'émotion perçue.*

Un test de Khi-2 a révélé une différence significative (Khi-2>60) dans l'emploi d'étiquettes qui semblent parfois confondues comme *Agacement* vs. *Impatience*, *Compassion* vs. *Intérêt* ou *Stress* vs. *Anxiété*. Il était donc pertinent de conserver ces étiquettes.

## La valence

Dans le protocole d'annotation, nous avions considéré qu'il n'était pas nécessaire d'annoter la valence, car elle pouvait se déduire de l'étiquette émotion (à l'exception de la *Surprise*, d'où la directive de donner en *Mineur* une étiquette précisant sa valence). L'annotation explicite de la valence pour le test perceptif était un moyen de vérifier cette hypothèse. Le nombre de cas où la valence ne se déduit pas de l'étiquette est donné Tableau 4-5. Pour les segments non complexes, il y a moins de 5% des cas pour lesquels la valence ne se déduit pas de l'annotation. Le nombre élevé d' « erreurs » pour la compassion peut s'expliquer par des cas du type celui décrit Figure 4-7. Les sujets choisissent une étiquette de valence opposée à celle des experts (à cause sans doute de l'absence de contexte), mais perçoivent toutefois des indices qui les poussent à annoter la valence différemment. De même pour l'embarras, une grande partie des segments annotés *Embarras* par les sujets avaient été annotés comme conflictuels par les experts. De façon non surprenante, les « erreurs » sont d'autant moins nombreuses que les émotions sont fortes (aucune « erreur » pour la panique ou la colère chaude).

| Majeur | Tous les segments | Sans Mineur |
|---|---|---|
| Compassion | 41% (63 segts) | 17% (35 segts) |
| Intérêt | 10% (121 segts) | 5% (79 segts) |
| Soulagement | 3% (90 segts) | 3% (66 segts) |
| Amusement | 5% (132 segts) | 6% (69 segts) |
| **Total Positif** | **11,6% (406 segts)** | **6,4% (249 segts)** |
| Impatience | 7% (102 segts) | 7% (70 segts) |
| Agacement | 1% (258 segts) | 1% (189 segts) |
| ***Colère chaude*** | ***0% (45 segts)*** | ***0% (32 segts)*** |
| Colère froide | 3% (69 segts) | 4% (49 segts) |
| Anxiété | 6% (98 segts) | 7% (70 segts) |
| Stress | 5% (56 segts) | 6% (35 segts) |
| Peur | 5% (21 segts) | 0% (10 segts) |
| ***Panique*** | ***0% (35 segts)*** | ***0% (23 segts)*** |
| Embarras | 16% (116 segts) | 17% (70 segts) |
| Désarroi | 6% (86 segts) | 5% (60 segts) |
| Résignation | 7% (105 segts) | 4% (78 segts) |
| ***Tristesse*** | ***0% (41 segts)*** | ***0% (32 segts)*** |
| Désespoir | 2% (44 segts) | 4% (24 segts) |
| Déception | 8% (76 segts) | 4% (56 segts) |
| ***Douleur*** | ***0% (10 segts)*** | ***0% (10 segts)*** |
| **Total Négatif** | **5,1% (1162 segts)** | **4,6% (808 segts)** |
| **Total** | **6,8% (1568 segts)** | **5% (1057 segts)** |

**Tableau 4-5.** *Pourcentage de cas où la valence est en contradiction avec les étiquettes émotions par émotion. Pour « Tous les segments », la valence est comparée à celle de l'émotion Majeur et pour « Sans Mineur », on ne regarde que les segments annotés avec une seule étiquette. Le nombre total de segments est indiqué entre parenthèses).*

### Les indices

Différentes catégories d'indices ont motivé les annotations, avec beaucoup d'indices prosodiques. Ces indices montrent la richesse de ce corpus et la grande diversité des expressions émotionnelles dans la parole conversationnelle spontanée.

| Prosodie, qualité vocale | « Affect burst » | Lexical | "dialogique" |
|---|---|---|---|
| intensité (x3)<br>mode impératif<br>intonation aigue (x2)<br>ton sec (x2)<br>tonalite de la voix<br>tremblement dans la voix | respiration (x2)<br>expiration<br>souffle (x4)<br>soupir (x3)<br>« Oooh, pfff »<br>rire (x12)<br>rire nerveux | Manque de cohérence<br>emploi du "oui" à la fin de la phrase<br>"mais enfin"<br>"merci" (x2)<br>"très bien" au lieu de "bien" | grammaire, sémantique dans un contexte téléphonique.<br>intérêt simulé ?<br>rire faux<br>laconisme<br>répétition (x3) |

**Figure 4-8.** *Résultats du choix libre d'indices ayant motivé les annotations.*

## 4.3. Conclusions

Dans ce chapitre, nous avons présenté et typé différents cas de mélange d'émotions dans le corpus CEMO. Ces mélanges peuvent être perçus même sans contexte et des indices acoustiques peuvent être perçus en même temps pour plusieurs émotions. Un test perceptif a validé l'annotation des experts, le choix des étiquettes et la décision de déduire la valence des étiquettes lors de l'annotation. Même si individuellement les sujets naïfs ne percevaient sans contexte qu'une seule émotion d'un mélange, lorsqu'on regroupe leurs annotations en un vecteur, on retrouve l'annotation des deux experts.

Ces mélanges sont très intéressants à étudier, mais sont susceptibles de causer des confusions si on les utilise pour entraîner des systèmes. Dans les parties qui suivre, ils ne seront donc pas utilisés.

# III Modélisation

# Chapitre 5

# Les paramètres

Résumé

Dans les chapitres précédents, nous avons décrit la collection et l'annotation en émotion de nos données. Avant de pouvoir commencer les expériences de détection, la première étape est de trouver un ensemble d'indices pertinents et de les extraire. Quels indices extraire du signal pour identifier les émotions ? Existe-t-il un profil acoustique par émotion ? Alors que les juges humains parviennent relativement bien à reconnaître les émotions dans la voix, les chercheurs n'ont pas encore réussi à s'accorder de manière précise sur les paramètres acoustiques corrélés à ces émotions. Scherer et Juslin ont comparé différentes études et résumé les divergences et les convergences sur le comportement des paramètres les plus souvent étudiés. Dans ce chapitre, nous présentons tout d'abord le modèle de Fónagy sur la transmission et le décodage du message oral. Nous décrivons ensuite brièvement la production de la parole et les différentes mesures qui en découlent, et en particulier celles utilisées dans les expériences sur les émotions, ainsi que les conclusions de Juslin et Scherer sur certains des paramètres les plus étudiés. Nous listons alors les paramètres que nous avons extraits en distinguant ceux extraits de manière purement automatique et ceux nécessitant d'avoir une transcription des données. Nous nous penchons ensuite sur le problème de l'extraction de la fréquence fondamentale. Nous nous intéressons également à la normalisation des paramètres prosodiques. Nous comparons ensuite les tendances dans les données CEMO avec les conclusions de Scherer et Juslin. Nous regarderons brièvement les triangles vocaliques par émotions.

*Now that we have collected emotional data and annotated it, the next step before detection experiments is to extract a set of cues. How to find the most relevant ones? Whereas human judges are pretty accurate for recognizing emotions from the voice, scientists haven't been able to agree on an accurate set of relevant parameters. Juslin and Scherer have compared several studies on vocal emotion detection and compared the trends of the most studies cues.*
*In this chapter, we first present Fónagy's model on how an oral message is transmitted and decoded. We then briefly describe speech production and different measures than can be made, especially for the purpose of emotion detection. We list the different types of cues that are used for emotion detection and distinguish between the "blind" ones that can be extracted automatically and those that require human processing.*

# 5.LES PARAMETRES

## 5.1. Etat de l'art des paramètres utilisés

Dans la communication orale des émotions, on distingue en général deux types d'informations : les informations paralinguistiques et les informations linguistiques. (*cf.* 5.1.1, modèle de Fónagy). Ces deux types d'informations vont servir à la compréhension et à l'interprétation d'un message et la perception d'une émotion découlera de leur interprétation. Pour détecter automatiquement les émotions, on va chercher à approcher les différentes caractéristiques perçues, comme l'intensité, le timbre ou la mélodie par des mesures physiques afin d'extraire des indices pertinents. Comprendre comment la parole est produite (*cf.* 5.1.2. la production de la parole) peut nous guider dans le choix des mesures à effectuer.

Scherer a exprimé le paradoxe suivant [Scherer 1986][1]: alors que les humains parviennent facilement à décoder les émotions dans la voix, les scientifiques n'arrivent pas à se mettre d'accord sur un ensemble de paramètres qui identifieraient correctement les émotions. Les indices les plus pertinents varient selon les études et les émotions que l'on cherche à discriminer (*cf.* p101). C'est pourquoi la stratégie utilisée est d'extraire le plus possible d'indices[2], jusqu'à des milliers [Schuller et al. 2006]. Ces indices sont déduits de tous les types d'informations de la parole émotionnelle avec majoritairement des indices paralinguistiques et linguistiques, mais aussi des indices contextuels et dialogiques. Ils sont souvent redondants et des algorithmes d'optimisation sont généralement appliqués pour réduire leur nombre et sélectionner les plus pertinents.

### 5.1.1. Le modèle de Fónagy

D'après le modèle de Fónagy [Fónagy 1983 p14], le message oral est transmis par deux actes successifs d'encodage : un encodage linguistique, qui transforme un message global en une séquence de phonèmes; et un deuxième codage, le code paralinguistique[3], qui correspond à la manière dont les sons vont être exprimés. Deux messages opposés peuvent ainsi être transmis simultanément comme dans l'exemple donné par Fónagy [Fónagy 1983 p14] :

---

[1] "*Whereas judges seem to be rather accurate in decoding emotional meaning from vocal cues, researchers in psychoacoustics and psychophonetics have so far been unable to identify a set of vocal indicators that reliably differentiate a number of discrete emotions* "

[2] Les mots indices ou paramètres pourrons être utilisés (cues, parameters, features en anglais).

[3] On entend par paralinguistique, les informations de type acoustiques, prosodiques et les manifestations non verbales de type marqueurs affectifs. Ces marqueurs sont des éléments très brefs qui sont extrêmement porteurs de connaissance sur l'état émotionnel de la personne.

> *«on pourrait imaginer la source transmettant au Premier Encodeur un message global, de remercier par exemple l'interlocuteur de sa gentillesse. Ce message serait décomposé en une séquence de phonèmes par le Premier Encodeur:/o :vuz ɛt vremã trop emabl*[1]*/. La source confiera en même temps un message, de nature différente et de caractère opposé au Deuxième Encodeur : d'exprimer un sentiment de haine, de mépris. Le Deuxième encodeur transformera, conformément à cette instruction, la séquence de sons que vient de lui présenter le premier encodeur, d'une telle manière à ce que la haine et le mépris soient clairement exprimés – par la compression des cordes vocales et des bandes ventriculaires, el allongeant les consonnes, […] »*

## 5.1.2. La production de la parole

Il n'existe pas d'organe spécifique destiné à la production de la parole et pour produire un signal, l'homme va utiliser son système respiratoire et son système digestif [Marchal 1980]. L'ensemble du système vocal se compose des poumons et du conduit trachéobronchique, du larynx et du conduit vocal, formé par le pharynx et les cavités orales et nasales (voir Figure 5-1 ci dessous).

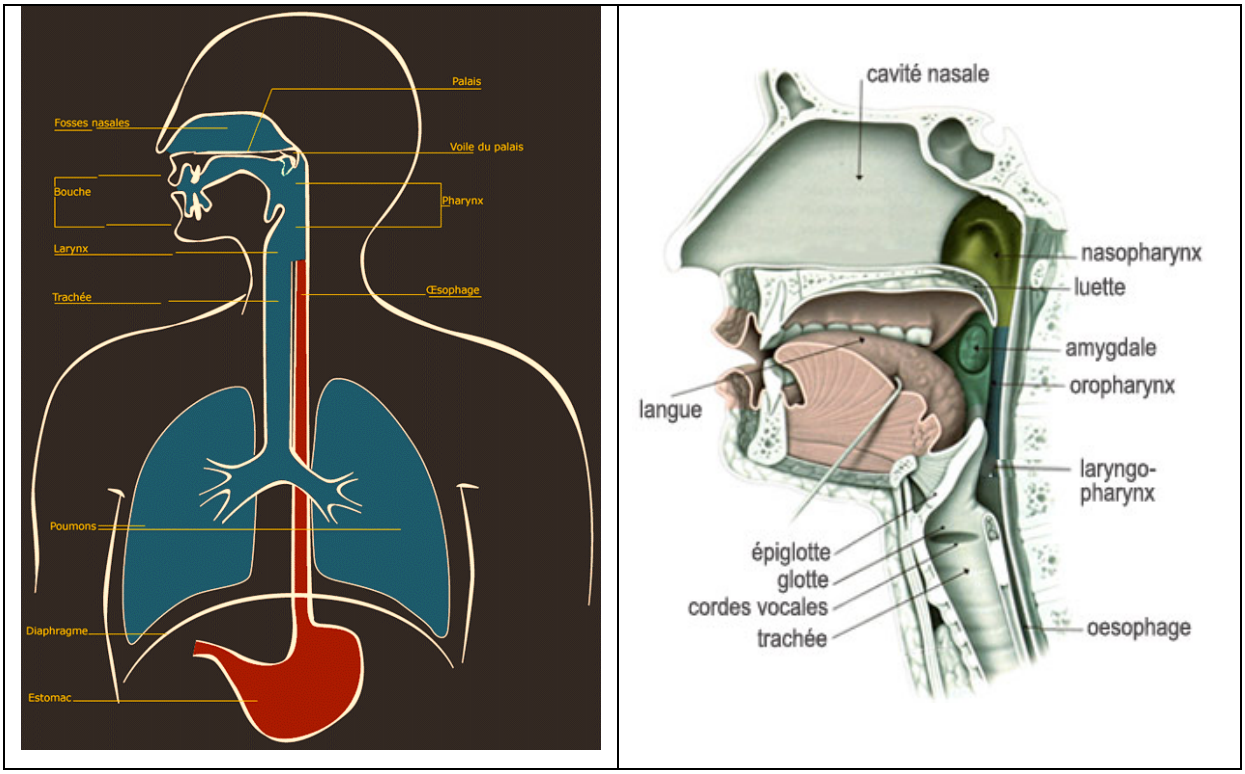



**Figure 5-1.** *L'appareil phonatoire.*

(*http://catalogue.ircam.fr/sites/Voix/decrire/appareil.html* ;
*http://lecerveau.mcgill.ca/flash/capsules/outil_bleu21.html*)

[1] « Oh vous êtes vraiment trop aimable »

Les sons sont produits par la modification du courant d'air de l'expiration en provenance des poumons. L'air passe d'abord à travers la trachée artère. Au sommet de celle-ci se trouve le larynx où la pression de l'air est modulée avant d'être appliquée au conduit vocal. Le son lui-même est produit au niveau des cordes vocales[1]. Il est fonction d'une part de la pression sous-glottique (pression due à l'obstacle des cordes vocales à l'air en provenance des poumons) et d'autre part de la masse effective des cordes vocales. Les sons résultants d'une vibration périodique des cordes vocales sont dits « voisés ». A l'opposé, lorsque l'air passe librement dans la glotte[2], les sons résultants sont dits « non voisés » ou « sourds ».

Les traits acoustiques du signal sont liés à sa production [Boite et al. 1999] :

- l'intensité du son est liée à la pression de l'air en amont du larynx (sous-glottique)
- sa fréquence correspond à la fréquence d'ouverture/fermeture des cordes vocales (déterminée par la tension des muscles qui la contrôlent)
- son spectre résulte du filtrage du signal glottique par le conduit vocal.
- le pharynx, la bouche et éventuellement les lèvres constituent une cavité de résonance qui peut être assimilé à un tube de diamètre uniforme (de 17 à 20 cm de longueur) fermé à un bout (la glotte) et ouvert à l'autre (les lèvres). Les fréquences renforcées par le phénomène de résonance sont appelées **formants**[3].

[1] Les cordes vocales sont deux lèvres symétriques placées en travers du larynx. Elles peuvent fermer complètement le larynx et en s'écartant elles déterminent une ouverture triangulaire appelée glotte.

[2] Cas également de la respiration et des chuchotements

[3] Le formant est défini rigoureusement [Liénard 1977] comme « un maximum de la fonction de transfert du conduit vocal »

## 5.1.3. Les indices extraits pour la détection des émotions

Les émotions vont se manifester par des variations des paramètres ou des perturbations par rapport à leur valeur standard. Nous allons tout d'abord décrire les principaux types d'indices paralinguistiques et linguistiques utilisés pour la détection des émotions. Parmi eux, on pourra différencier ceux qui peuvent être obtenus de manière purement automatique et ceux qui nécessitent un prétraitement ou une connaissance humaine (transcription orthographique des données par exemple).

### 5.1.3.1. Le niveau paralinguistique

Les différents types de paramètres paralinguistiques qui seront décrits ci-dessous sont :

(a) les paramètres prosodiques : fréquence fondamentale, énergie, durée et qualité vocale
(b) la microprosodie : jitter, shimmer, HNR (Harmonic to Noise Ratio), NHR (Noise to Karmonic ratio)
(c) les paramètres spectraux : formants et leur bande passante et MFCCs
(d) les disfluences prosodiques
(e) les marqueurs affectifs acoustiques

#### *(a) Les paramètres prosodiques*

Le terme **prosodie** se confond souvent avec celui d'« intonation ». La prosodie concerne le suprasegmental[1] et englobe des phénomènes tels que accentuation, variations de hauteur, de durée et d'intensité. C'est ce qui donne un ton naturel et cohérent à la parole. . Elle est représentée dans le diagramme donné Figure 5-2 [Hirst et Di Cristo p5].

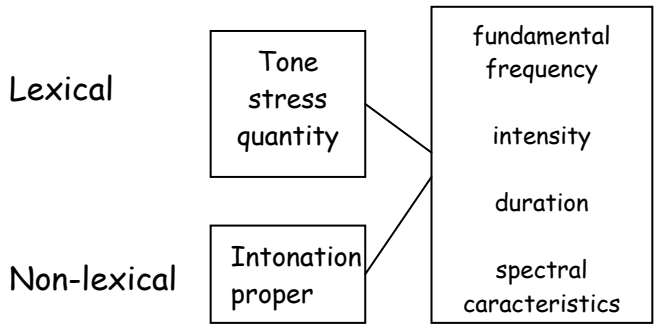



**Figure 5-2.** *La prosodie selon [Hirst et Di Cristo]*

[1] Terme créé par Hockett en 1942, défini dans le Dictionnaire de linguistique comme « un trait suprasegmental, ou trait prosodique, est une caractéristique qui affecte un segment plus long que le phonème » [Rossi et al. 1981 p10].

Elle intervient par exemple dans la distinction des questions et des réponses ou pour lever les ambiguïtés du langage parlé par l'insistance sur certains mots. Elle permet aussi d'exprimer des attitudes et des états émotionnels

Au niveau physique, la prosodie se réfère aux variations d'un ensemble de paramètres acoustiques parmi lesquels la fréquence fondamentale, l'intensité et la durée.

## La fréquence fondamentale (F0)

.

La fréquence fondamentale (ou pitch) est un indice de mesure globale de la voix, qui correspond à la fréquence de vibration des cordes vocales. Elle est calculée sur les parties voisées[1] du signal et peut être exprimée en plusieurs unités : en Hertz, échelle Mel ou Bark.

Elle dépend de facteurs spécifiques aux locuteurs comme le sexe (pour une femme, la F0 moyenne est de 250 Hz alors que pour un homme elle est estimée à 150 Hz), l'âge, la langue maternelle ou l'accent. La F0 moyenne apporte une mesure globale de la hauteur de la voix (aiguë, grave ...)

La F0 s'avère être un paramètre très important pour la reconnaissance des émotions, et a été étudiée depuis les années 60 [Liberman et Michaels 1962]. La plage de la F0 et le contour de la F0 sont des paramètres typiques pour discriminer certaines émotions.

## L'intensité du signal

Elle apporte une mesure globale de la force sonore de la voix (faible ou forte). Elle se mesure généralement en décibel (dB). Pour une voix triste ou neutre, l'intensité sera beaucoup moins forte que pour une voix colérique. L'intensité du signal est un paramètre difficile à normaliser, notamment au téléphone. Une voix faible peut-être proche du téléphone et une voix forte loin du téléphone. F0 et intensité sont corrélés.

## La durée et le rythme (tempo)

La durée correspond au temps d'émission et est liée à la notion de rythme et aux silences. Elle englobe des variables paralinguistiques comme les longueurs (phrase, mot, syllabe, phonème, partie voisée…) ou les débits (mots/syllabes/phonèmes par unité de temps), mais aussi les pauses et

[1] Signal périodique ou quasi périodique

silences. Une mesure de débit fréquemment calculé dans les études sur les émotions est l'inverse de la longueur moyenne des parties voisées.

### La qualité vocale

Certains ajoutent à la prosodie une quatrième dimension, la qualité de la voix (timbre, voix rauque, chuchotée, grinçante, voix de fausset...), due à des caractéristiques laryngales ou supralaryngales. Un indice lié à la qualité vocale est le NAQ **« Normalized Amplitude Quotient »,** défini par Campbell *et al.* [Campbell et Mokhtari 2003]. Il permet d'avoir une mesure sur l'onde de débit glottal. Il peut être considéré comme une normalisation du « temps de déclinaison » et s'exprime comme le rapport de l'amplitude crête à crête de l'onde de débit glottique et du pic négatif maximal de sa dérivée, normalisé par la période fondamentale. Il est cependant difficilement utilisable sur des données réelles car il nécessite une prise de son parfaite.

## *(b) La microprosodie*[2]

### Le coefficient Shimmer[1]

Le shimmer mesure les variations d'amplitude entre deux cycles d'oscillation :

- le *shimmer moyen* représente la moyenne des rapports d'amplitudes entre deux cycles d'oscillation consécutifs
- le *shimmer factor* relativise le shimmer moyen en divisant par l'amplitude moyenne
- l'*APQ* (Amplitude Perturbation quotient) mesure la moyenne des variations d'amplitude sur 11 périodes consécutives, le tout rapporté à l'amplitude moyenne du signal observé

### Le coefficient Jitter[2]

C'est un indice représentatif de la perturbation à court terme de la fréquence fondamentale, qui se traduit par des variations de fréquence entre chaque cycle d'oscillation. Il peut être intéressant de le mesurer pour des phrases où la fréquence est normale puis s'accélère brutalement (pour des émotions comme la peur, le stress ou le désespoir par exemple). Plusieurs mesures existent :

- le *jitter absolu moyen* est la moyenne de la différence de F0 en valeur absolu, entre deux cycles de vibration consécutifs
- le *jitter factor* permet de relativiser le jitter moyen en le comparant à la F0 moyenne

[1] http://www.fon.hum.uva.nl/praat/manual/Voice_3__Shimmer.html
[2] http://www.fon.hum.uva.nl/praat/manual/Voice_2__Jitter.html

- le *jitter ratio* mesure la moyenne des variations de période entre deux cycles de vibration consécutifs et relativise cette valeur par la période moyenne du signal observé
- le *RAP (Relative Average Perturbation)* mesure la moyenne des variations de trois périodes consécutives rapportée à la période moyenne du signal observé

## Le pourcentage de fenêtres non voisées dans le segment[1]

Le signal sonore peut être divisé en plusieurs fenêtres (trames) voisées ou non voisées (pas de signal F0). Le pourcentage des fenêtres non voisées dans la phrase est révélateur de la quantité pauses sur cette phrase. Ainsi, une phrase prononcée à un rythme normal contiendra beaucoup plus de pauses (donc plus de fenêtres non voisées) qu'une phrase prononcée avec un rythme élevé, comme dans le cas de la colère ou de la peur.

**Fundamental Frequency**

| | |
|---|---|
| Mean | mean (Hz) = $\frac{1}{N}\sum_{i=1}^{N} F0_i$ |
| std dev | standard deviation (Hz)= $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(F0_i - F0_{moy}\right)^2}$ |
| c.v | coefficient of variation (%) = $100 \times \frac{\text{standard deviation}}{mean}$ |
| Cycle to Cycle perturbation | Mean Jitter (Hz) = $\frac{1}{N-1}\sum_{i=1}^{N-1}\left\|F0_i - F0_{i+1}\right\|$ |
| Percentage | Jitter factor (%) = $100 \times \frac{\text{mean jitter (Hz)}}{F0_{mean}}$ |
| Per 1000 | Jitter ratio (‰) = $1000 \times \frac{\text{mean jitter (ms)}}{T0_{moy}}$ |
| Average perturbation | Relative Average Perturbation<br>RAP= $\frac{\frac{1}{N-2}\sum_{i=2}^{N-1}\left\|\frac{T0_{i-1}+T0_i+T0_{i+1}}{3} - T0_i\right\|}{T0_{moy}}$ |

**Intensity/amplitude**

| | |
|---|---|
| Mean | mean (dB SPL) = $\frac{1}{N}\sum_{i=1}^{N} Intensity_i$ |
| std dev | standard dev. (dB SPL) = $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Intensity_i - Intensity_{moy}\right)^2}$ |
| c.v | coefficient of variation (%) = $100 \times \frac{\text{standard deviation}}{mean}$ |
| Cycle to Cycle perturbation | $shimmer_{dB} = \frac{1}{N-1}\sum_{i=1}^{N-1}\left\|20.\log\left(\frac{A_i}{A_{i+1}}\right)\right\|$ |
| percentage | Shimmer factor = $100 \times \frac{\text{shimmer}}{20.\log\left(A_{mean}\right)}$ |
| Average perturbation | Amplitude Perturbation Quotient<br>APQ = $\frac{\frac{1}{N-10}\sum_{i=6}^{N-5}\left\|\frac{A_{i-5}+\ldots+A_i+\ldots+A_{i+5}}{11} - A_i\right\|}{A_{mean}}$ |

**Figure 5-3.** *Les paramètres acoustiques (extrait de http://aune.lpl.univ-aix.fr/~ghio/doc/Doc-VoiceParameters.pdf)*

[1] http://www.fon.hum.uva.nl/praat/manual/Voice_1__Voice_breaks.html

### *(c) Les coefficients spectraux*

#### Les formants et leurs largeurs de bande

Les formants correspondent à des pics d'énergie. Les trois premiers formants et surtout les différences entre les formants peuvent être des indices de comportements affectifs. On ajoute en général aussi les largeurs de bande.

#### Les coefficients MFCCs (Mel Frequency Cepstral Coefficients)

Ils sont caractéristiques des résonances du conduit vocal à un instant donné. Il est d'usage de les calculer également sur une fenêtre temporelle et de calculer leurs dérivées premières et secondes.

### *(d) Les disfluences prosodiques*

Elles existent au niveau linguistique et acoustique et sont souvent difficiles à classifier. Elles désignent toutes les ruptures dans le signal : répétitions de mots, silences, hésitations. Elles sont souvent déduites de la transcription, mais il n'est pas exclu de pouvoir les détecter automatiquement.

### *(e) Les marqueurs affectifs acoustiques*

Les indices non verbaux appelés aussi marqueurs affectifs pour "affect bursts" (rires, pleurs, toux, raclements de gorge, interjections, etc.) ont souvent un haut pouvoir de discrimination des émotions [Polzin et Waibel 1998] [Schröder 2000]. Parmi les marqueurs affectifs, les rires[1][Devillers et Vidrascu 2007], respirations et pleurs sont de plus en plus étudiés.

## 5.1.3.2. Le niveau linguistique

### *Le contenu lexical*

Particulièrement pour des données enregistrées au téléphone, le niveau linguistique peut apporter des informations pour la détection des émotions. Il sera plus ou moins utile selon les émotions que

[1] Il y a d'ailleurs eu un workshop interdisciplinaire dessus http://www.coli.uni-saarland.de/conf/laughter-07/

l'on cherche à discriminer. Par exemple dans le corpus CEMO, la détection du soulagement peut être attribuée à certaines marques lexicales spécifiques comme « merci ». Les émotions négatives peuvent aussi être liées à certains termes, comme « problème » ou à des formes négatives « ne pas ». Dans nos données cependant, les expressions de la peur sont souvent plus syntaxiques que lexicales à travers des répétitions, des reformulations etc.

## *Le contexte dialogique*

- Les annotations émotionnelles peuvent être corrélées avec les actes dialogiques[1].

Ce type d'annotations est moins fréquent. Dans les travaux de Devillers et.al, les actes dialogiques ont été annotés (adaptés d'après DAMSL standard dialogs acts annotation) Des mesures de corrélation ont montré que les émotions négatives Colère et Peur sont susceptibles de générer plus fréquemment des Assertion, Réassertions, Requêtes et Répétitions, tandis que les émotions positives comme la Satisfaction et le Neutre sont corrélées avec l'Acceptation [Devillers et al. 2002]. Dans les travaux de Lee *et al.* [Lee et Narayanan 2004], l'utilisation de cinq actes de dialogue (du type rejection, répétition) en plus d'indices lexicaux et prosodiques a amélioré les scores de détection, ils ont également contribué à améliorer les scores de détection pour Batliner *et al.* [Batliner et al. 2003]. De même, dans les travaux de Liscombe *et al.* [Liscombe et al. 2005], 10 actes de dialogues ont été annotés ainsi que des informations sur les émotions des deux tours précédents (prédite ou réelle) et ces indices ont augmenté la reconnaissance.

[1] Notion introduite par [Austin 1962] qui correspond à une « unité de contexte » dans le dialogue. Le fait de dire quelque chose revient à faire une action et ces actions peuvent être typées, par exemple asserion, rejection, répetition.

### 5.1.4. Les variations des paramètres suivant les états émotionnels dans la littérature

Scherer [Scherer 2003] a résumé les effets des émotions les plus fréquemment étudiées sur certains paramètres en s'appuyant sur les résultats empiriques d'une trentaine d'études des soixante dernières années. La plupart de ces études ont été effectuées sur des données actées. Une synthèse des résultats empiriques est donnée dans le Tableau 5-1 ci-dessous.

| Paramètres acoustiques | Stress | joie | Irritation/ Colère f | Colère /Rage | Tristesse/ Abattement | Affliction/ Desespoir | Peur/ Anxiété | Peur/ Panique | Ennui |
|---|---|---|---|---|---|---|---|---|---|
| **Débit et Fluency** | | | | | | | | | |
| Nombre de syllabes par seconde | > | >= | | > | < | > | > | >> | < |
| Durée des syllabes | < | <= | | <> | > | | | < | > |
| Durée des voyelles accentuées | >= | >= | | > | >= | | | < | >= |
| Nombre et durée des pauses | < | < | | < | > | | = | <> | > |
| Durée relative des segments voisés | | | | > | | | | <> | |
| Durée relative des segments non voisés | | | | < | | | | <> | |
| **F0 et Prosodie** | | | | | | | | | |
| Moyenne F0 | > | > | <> | <> | <> | > | > | >> | <= |
| F0: 5ème percentile | > | > | | < | <= | | | > | <= |
| déviation standard de F0 | > | > | < | >> | < | > | < | >> | < |
| Plage F0 | > | > | < | >> | < | > | | <>> | <= |
| Fréquence des syllabes accentuées | > | >= | | > | < | | | | |
| Gradient of F0 rising and falling | > | > | | > | < | | | <> | <= |
| F0 final fall: range and gradient | > | > | | > | < | | | <> | <= |
| **Effort Vocal et Type de Phonation** | | | | | | | | | |
| Intensité moyenne (dB) | > | >= | | > | <= | > | = | > | <= |
| déviation standard de l'Intensité | > | > | > | > | < | > | > | > | < |
| pente spectrale (spectral slope) | < | < | | < | > | | | <> | > |
| Laryngalisation | | = | | = | > | | | > | = |
| Jitter | > | >= | | >= | > | | > | > | = |
| Shimmer | | >= | | >= | | | | > | = |
| HNR | | > | | > | < | | | < | <= |
| **Formants** | | | | | | | | | |
| précision des Formants | ? | >= | > | > | < | > | > | <> | <= |
| Bande passante des formants | < | | | < | > | | | | >= |
| F1 (M) | | < | > | > | <> | > | > | > | > |
| F2 (M) | | | < | < | < | < | < | < | < |
| F1 (bw) | | > | << | << | <> | << | < | << | < |

**Tableau 5-1.** *Synthèse des résultats empiriques pour l'effet des émotions sur les paramètres vocaux (extrait [Scherer et al. 2003], [Juslin et Laukka 2003],,[Juslin et Scherer 2005] ) < "plus petit/ lent/ plat/étroit"; > "plus grand/haut/rapide/pentu/large" ;=égal au "Neutre"; <> : Des études ont reporté à la fois des résultats plus grand et plus petits. Les résultats surlignés en gris concernent les données naturelles ou induites.*

Ces résultats concordent globalement avec ceux de Juslin [Juslin et Laukka 2003] obtenus en comparant 104 études, dont 12 effectuées sur des données naturelles. La colère par exemple

s'exprime vocalement par un accroissement de la F0 moyenne et de son intensité ainsi que par la variabilité de la plage de F0.

## 5.2. Paramètres extraits sur nos corpus

Dans nos données, différentes phrases avec un même contenu lexical pourront s'exprimer de multiples manières avec différentes courbes de F0 ou des différences dans la longueur des phonèmes qui peuvent être autant d'indices sur l'émotion exprimée. Un aperçu de ces variations est donné dans les spectrogrammes ci-dessous pour le même segment « je sais pas » exprimé d'abord par un même locuteur de manière neutre, puis agacée (Figure 5-4 ); puis par différents locuteurs avec quelques cas d'émotions intenses. La F0 est représentée en bleu, l'énergie en jaune et la durée du segment en rouge en haut à droite des spectrogrammes. Les formants sont visibles par les zones les plus noires.

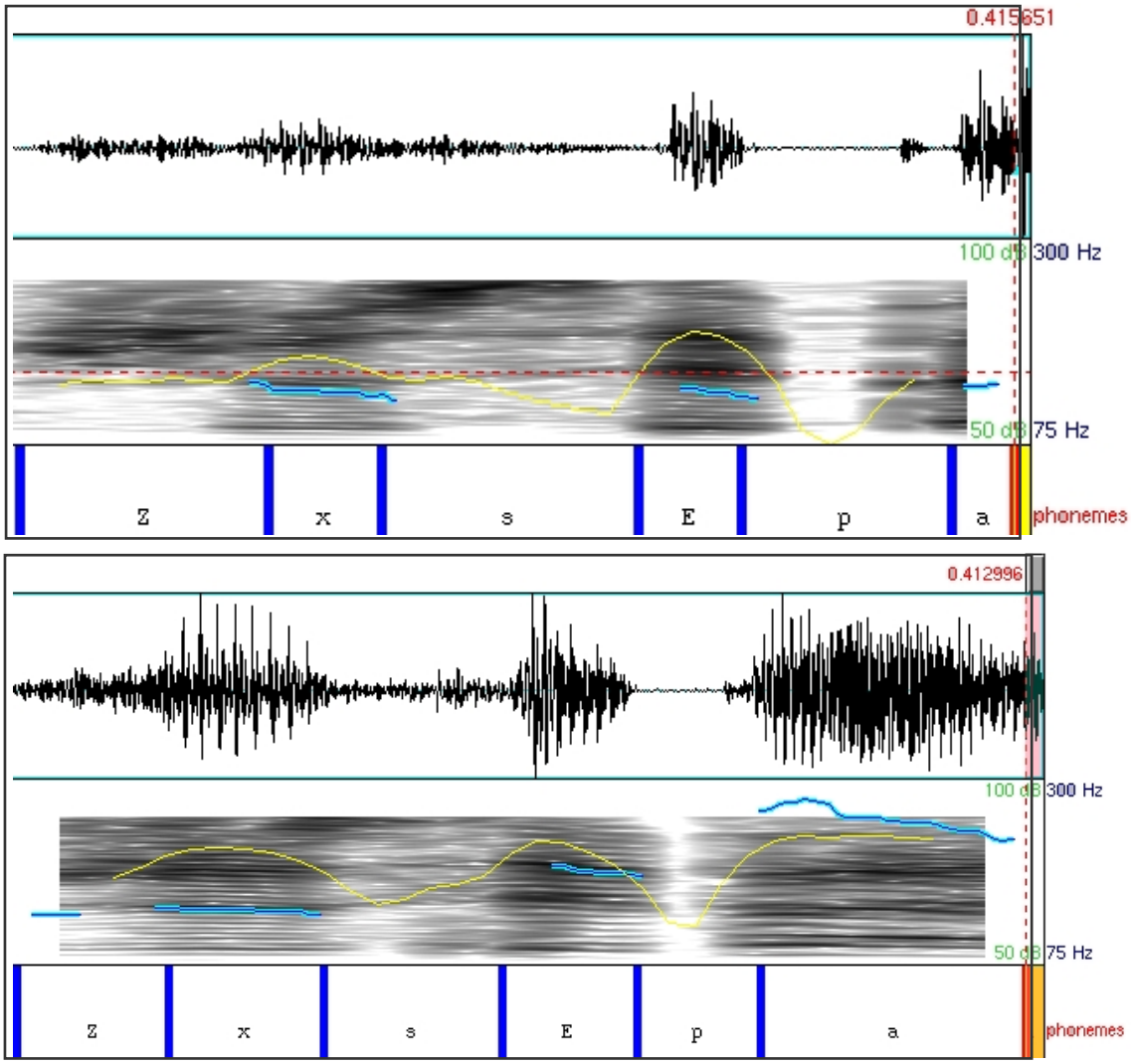



**Figure 5-4.** *Le même contenu lexical « Je sais pas » et le même locuteur de manière neutre puis agacée.*

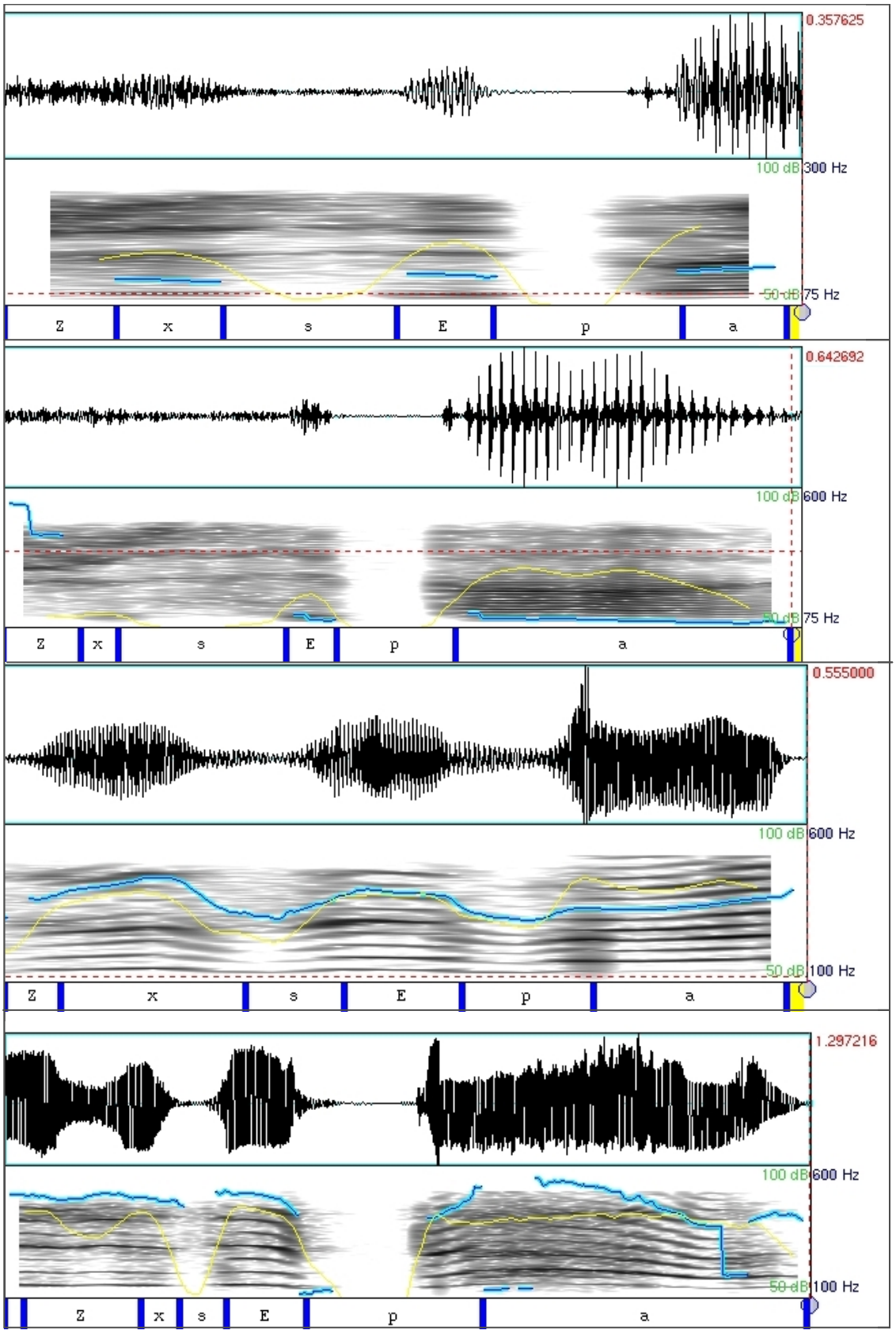



**Figure 5-5.** *« Je sais pas » : plusieurs locuteurs, plusieurs émotions (neutre, stress, désespoir, désespoir ).*

Comme il n'existe pas de consensus sur une liste de paramètres pertinents et que le choix de ces paramètres semble dépendre des données, notre stratégie est d'en extraire le plus possible, même si la plupart sont redondants et d'utiliser des méthodes de fouille de données pour sélectionner les meilleurs.

## 5.2.1. Paramètres extraits de manière automatique

### 5.2.1.1. Paramètres prosodiques

Nous avons utilisé le logiciel Praat [Boersma et Weenink 2005] pour extraire les mesures de F0, d'intensité et de voisement des segments.

Praat utilise pour la détection de la F0 un algorithme robuste de détection de la périodicité travaillant dans le « lag auto-correlation domain » (Boersma, 1993). Cet algorithme est particulièrement bien adapté pour des conditions de bruit (parole téléphonique) et permet de détecter des phénomènes acoustiques particuliers. Les différences homme (F0 moyenne autour de 150 Hz), femme (F0 moyennes autour de 250Hz) et enfants ont été prises en compte lors de l'extraction de la F0.

Finalement, la F0 est extraite sur les segments voisés toutes les 10 ms. Nous avons considéré les segments courts (<40ms) comme des erreurs de détection (*cf.* Figure 5-6 ) et les avons éliminés.

Avec Praat, il est également possible de pénaliser les sauts d'octave, les silences trop grands ou le trop grand nombre de segments voisés/non voisés. La difficulté est toujours de trouver le compromis entre pénaliser les erreurs et effacer des indices pertinents pour les émotions (saut d'octave par exemple).

Les paramètres déduits de l'extraction de la F0 sont :

- F0 maximum, F0 minimum, plage[1] F0, médian F0, premier quartile F0, troisième quartile F0, moyenne F0, déviation standard F0
- variation maximale de F0 entre 2 segments voisés adjacents (voir max ΔF0Inter Figure 5-6)
- variation maximale de F0 à l'intérieur d'un segment voisé (voir max ΔF0Intra Figure 5-6)
- le maximum et la moyenne de la pente de la F0, du coefficient de régression et de l'erreur moyenne quadratique par segment voisée
- des paramètres de durée :
    - position sur l'axe des temps où F0 est maximum (resp. minimum)

[1] Plage=maximum-minimum

- ratio entre les parties voisées et non voisées ($\frac{\sum_i \Delta tvoi_i}{\sum_i \Delta tunv_i}$)
- débit (inverse de la longueur moyenne des parties voisées.

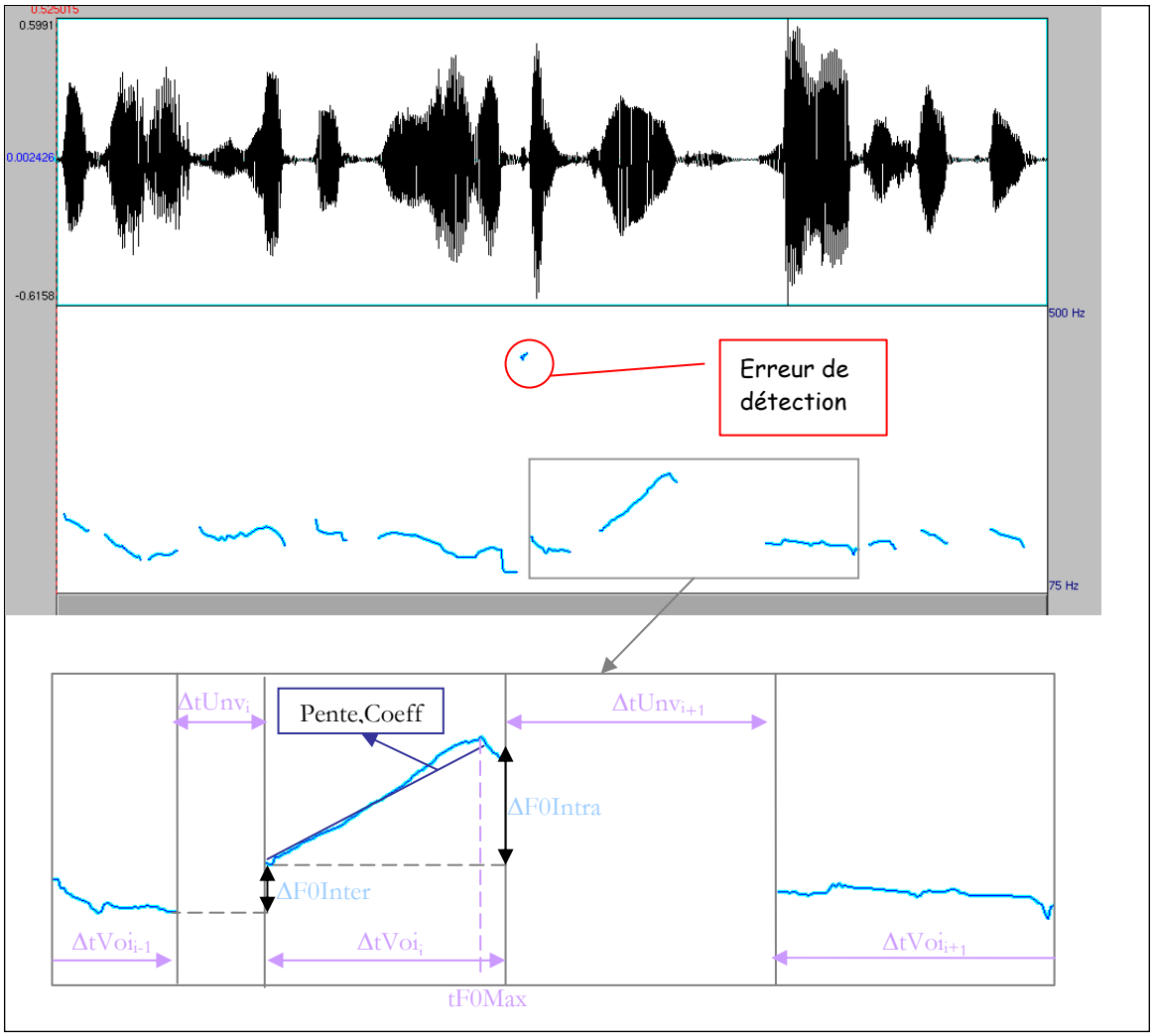



**Figure 5-6.** *Exemple d'extraction de F0 avec Praat : la courbe de la F0 est indiquée en bleu et des informations sont données sur les différents traitements effectués.*

Des paramètres similaires sont calculés pour l'énergie.

### 5.2.1.2. Paramètres spectraux

Les 3 premiers formants et leurs bandes passantes ont été extraits avec Praat toutes les 10ms en prenant compte des différences hommes femmes et suivant l'algorithme de Burg [Childers 1978; Teukolsky et al. 1992] qui ne tient pas compte des formants en dessous de 50Hz.

Seules les valeurs extraites sur les parties voisées ont été conservées et des paramètres (minimum, maximum, moyenne, médian, premier et troisième quartile, déviation standard, plage) ont été extraits pour chaque formant et bande passante, ainsi que pour les différences (F2-F1) et (F3-F2).
Les paramètres cepstraux sont des paramètres standard pour les systèmes de transcription [Gauvain 2002]. Ils ont été extraits toutes les 10ms en utilisant une fenêtre de 30ms sur une bande 0-8kHz. Nous avons calculé les maxima et minima des 15 coefficients cepstraux, ainsi que des coefficients Δ et ΔΔ.

### 5.2.1.3. Microprosodie

Le jitter, shimmer, NHR, HNR, ont été extraits par Praat au niveau du segment.
Un exemple de variation de F0 (tremolo, voix tremblante) est donné dans la Figure 5-7.

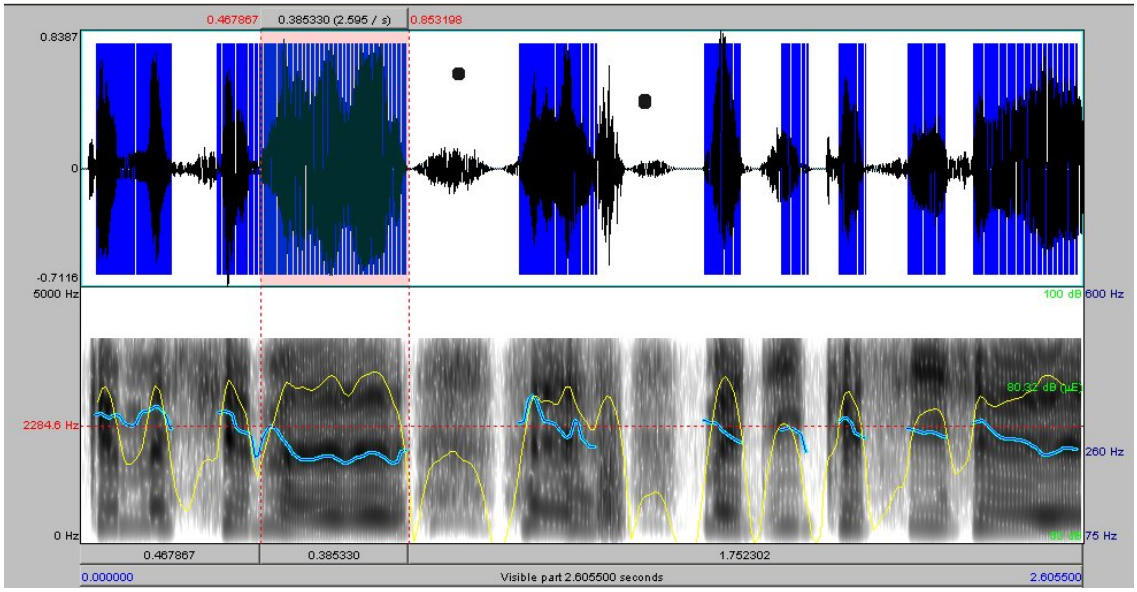



**Figure 5-7.** *Exemple de voix tremblante (variation de F0), extrait annoté détresse/désespoir/tristesse.*

L'énergie, les paramètres spectraux et les formants ont seulement été extraits sur les parties voisées (i.e. parties où Praat détecte la F0). Certains signaux, comme les voix chuchotées en particulier (Figure 5-8) ont très peu d'indices.

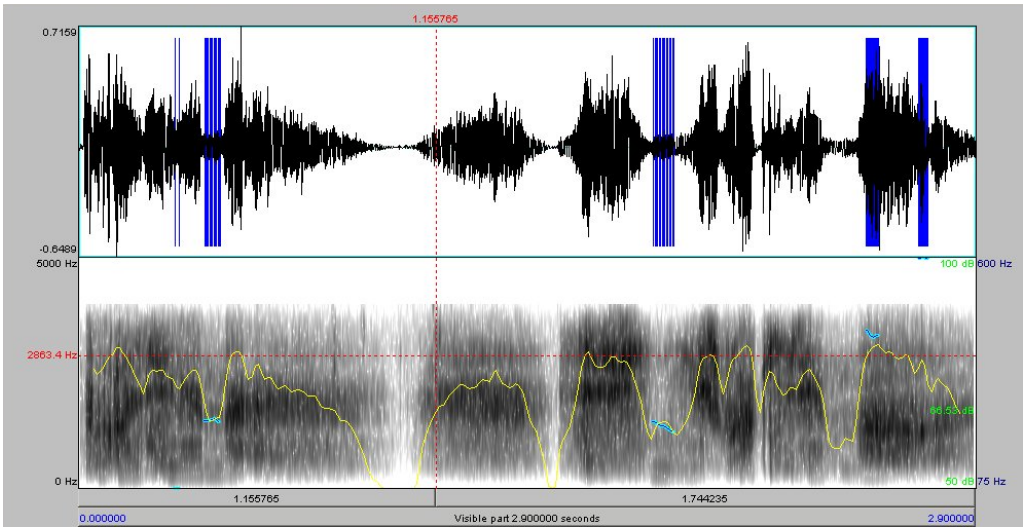



**Figure 5-8.** *Exemple d'une voix chuchotée avec très peu d'indices.*

## *En résumé*

La Figure 5-9 récapitule les différents types de paramètres acoustiques extraits automatiquement et leurs traitements.

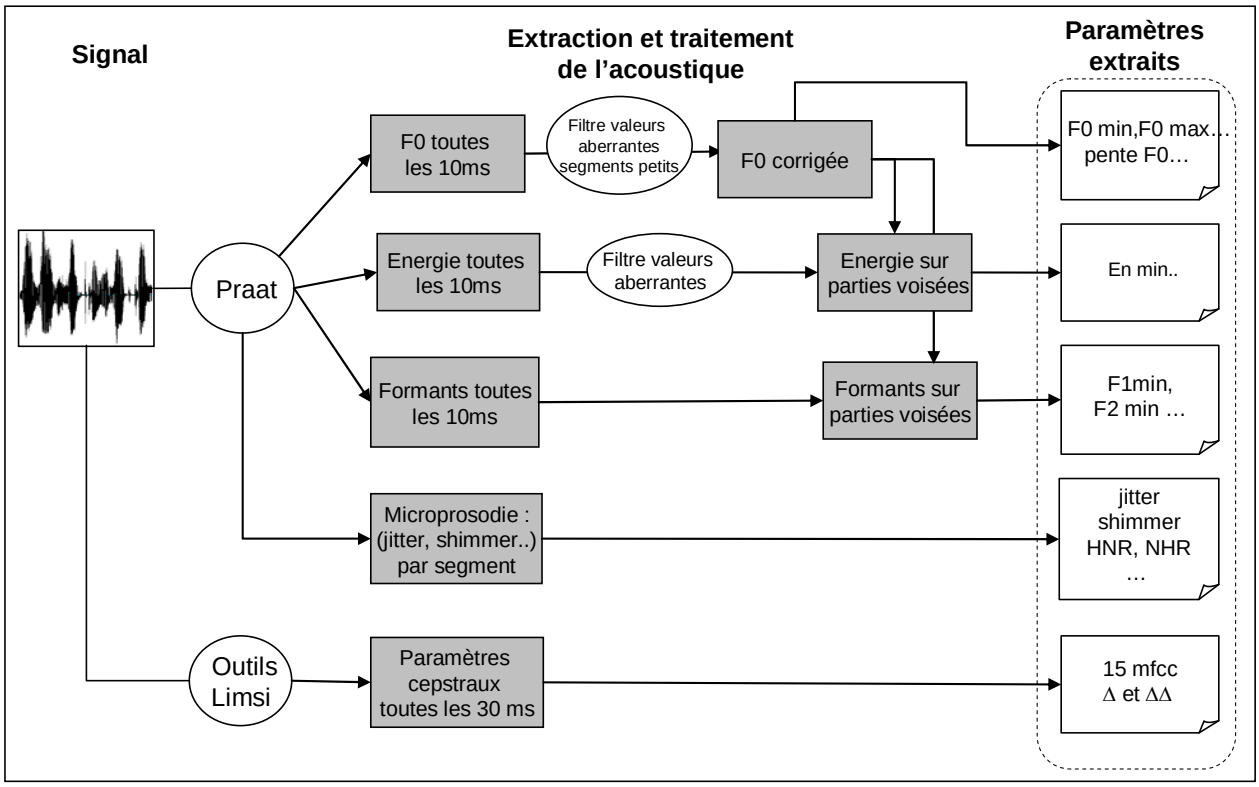



**Figure 5-9.** *Résumé des paramètres acoustiques extraits « automatiquement ».*

### 5.2.2. Paramètres déduits de la transcription manuelle et de l'alignement phonémique

*Alignement Phonémique*

Le corpus a été segmenté en phonèmes[1] en utilisant des modèles acoustiques indépendants du contexte, mis au point au Limsi pour des conversations téléphoniques.

La procédure (*cf.* Figure 5-10), fondée sur l'alignement dynamique de modèles de Markov cachés à densité continue, indépendants du contexte, est décrite dans [Adda-Decker 1999]. Elle nécessite une transcription orthographique fine de la parole, avec tous les phénomènes de disfluences que cela comporte : les lapsus, mots tronqués, hésitations… ainsi qu'un dictionnaire contenant les différentes prononciations possibles (transcription phonétique basée sur 36 phonèmes) de tous les mots du lexique (121k mots). A partir des données audio, de leur transcription manuelle et du dictionnaire de prononciations, le décodeur produit la séquence de phonèmes réalisée la plus probable et leur association temporelle. Les résultats produits par le décodeur dépendent bien sûr du degré de finesse avec lequel a été élaboré le dictionnaire, des modèles acoustiques et plus généralement des paramètres du système.

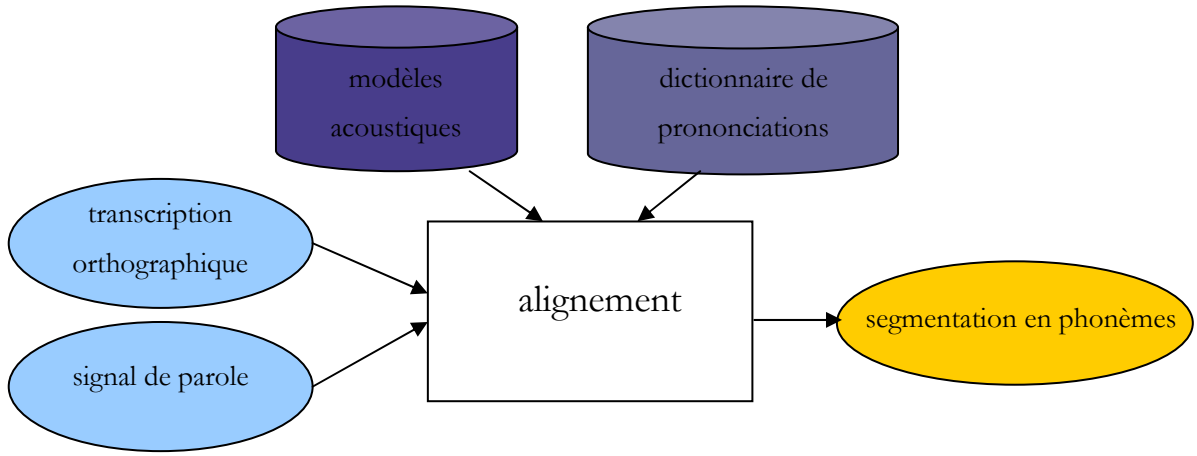



**Figure 5-10.** *L'alignement phonémique.*

Ont été extraits de la transcription et de l'alignement phonémique :

- *des marqueurs affectifs* : nombre de rires, de pleurs, de respirations, de mots inintelligibles, de mots tronqués, nombre de mots par segment et débit (#nombre mots/longueur du signal)
- des *disfluences* : nombre d'hésitations « euh » et leur durée

[1] Le phonème est la plus petite unité linguistique (36 en français).

- des informations sur *les durées des phonèmes*[1]. En particulier, nous avons regardé la durée moyenne et maximum des phonèmes et le débit phonémique (nombre de phonèmes divisés par la longueur de la phrase). Ces mêmes paramètres ont été calculés pour les voyelles seulement.

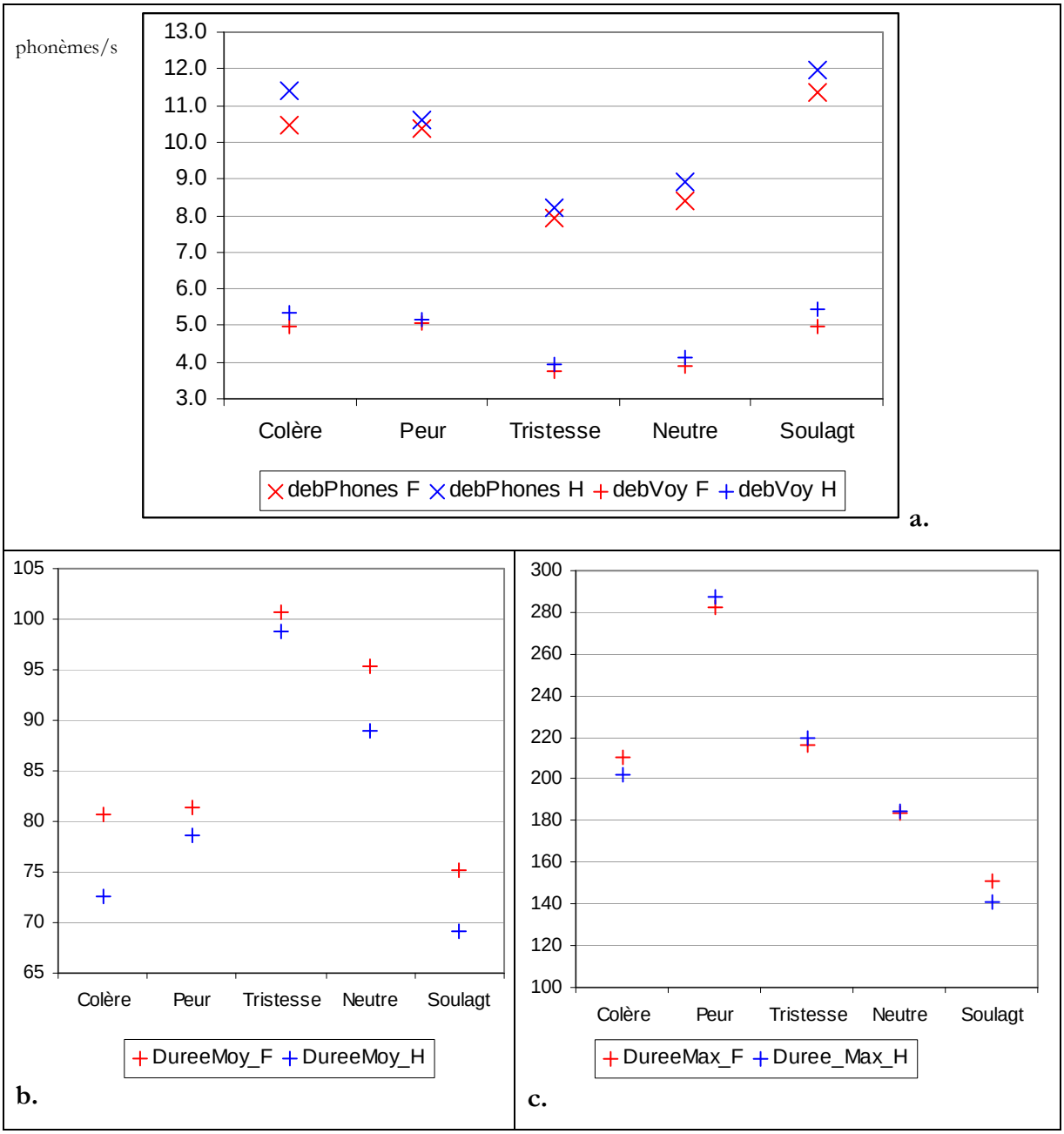



**Figure 5-11.** *Quelques paramètres issus de l'alignement phonémique pour les classes émotionnelles Peur/Colère/Tristesse/Neutre/Soulagement ; a. : débit phonémique et #voyelles/durée du segment pour les 5 émotions en regardant les hommes et les femmes séparément ;, b. : durée moyenne des phonèmes, c. : durée maximum des phonèmes.*

[1] Les fréquences des formants pour les voyelles n'ont pas été ajoutées aux paramètres, faute de temps

La Figure 5-11 donne un aperçu de la variation de certains paramètres obtenus après alignement phonémique pour les 5 classes émotionnelles Peur, Colère, Soulagement et Neutre.

Dans une étude sur les données boursières [Devillers et al. 2004], nous avions étudié les durées des silences et hésitations. Les résultats montraient que les silences étaient plus présents et plus longs chez les appelants que chez les agents (le rôle de l'agent empêche la manifestation de silences) et plus nombreux pour les émotions négatives que neutre ou positives. De même l'étude montrait la corrélation entre le nombre de « euh » et les segments étiquetés « peur/inquiétude ».

En résumé, le nombre d'indices par type est donné dans le Tableau 5-2.

| Type de paramètre | Description | # indices |
|---|---|---|
| coefficients MFCC | minimum et maximum des 15 coefficients cepstraux, coefficients Δ et ΔΔ | 90 |
| paramètres déduits de l'extraction de la F0 | F0 : min, médian, premier et troisième quartile, maximum, moyenne, déviation standard, plage au niveau du segment, pente (moyenne et max) pour le segment voisé. Coefficient de régression et son erreur quadratique moyenne (calculé sur les parties voidées), variation maximale de F0 entre 2 segments voisés adjacents. (inter-segment) et pour chaque segment voisé (intra-segment), position sur l'axe de temps où est maximum (resp. minimum), ratio du nombre de segments voisés et non voisés. Durées: débit (inverse de la durée moyenne des parties voisées), nombre et longueur des silences (portions non voisées entre 200-800 ms). | 25 |
| Paramètres spectraux (extraits sur les parties voisées du signal et normalisés) | formants et leurs bandes passantes, différence entre le troisième et le second formant, différence entre le second et le premier formant : min, max, moyenne, déviation standard, plage. | 48 |
| Energie (Normalisée) | min, max, moyenne, déviation standard et plage au niveau du segment. pente (moyenne et max) sur les parties voisées, coefficient de régression et erreur quadratique moyenne. | 20 |
| Microprosodie | jitter, shimmer, NHR, HNR | 14 |
| *Trans1* : indices extraits de la transcription | inspiration, expiration, bruits de bouche, rires, pleurs, nombre de mots tronqués et de paroles inintelligible, nombre de mots, débit (#mots/durée du segment). Disfluences : nombre de "euh" | 11 |
| *Trans2 :* Durées obtenues après alignement phonémique | durée moyenne et maximum des phonèmes, débit phonémique (#phonèmes/ durée du segment), longueur (max et moyenne) des hésitations. | 11 |

**Tableau 5-2.** *Résumé des différents paramètres paralinguistiques extraits.*

### 5.2.3. Normalisation des paramètres prosodiques

La normalisation des paramètres est indispensable étant donné que certains paramètres dépendent des locuteurs. Par exemple, la F0 moyenne dépend du locuteur et elle est d'environ 150 Hz pour les hommes, 250 Hz pour les femmes et 350 Hz pour les enfants.

Plusieurs méthodes existent pour normaliser les paramètres. La difficulté est de lisser les différences entre locuteurs sans effacer les variations causées par les émotions. Pour un paramètre P, les différentes possibilités de normalisation sont les suivantes :

- *Z-Norme* : normalisation par rapport à la moyenne et à la déviation standard :

$$P_{ZNorme} = \frac{P - moyenne_{Loc}}{Sd_{Loc}}$$

avec $moyenne_{Loc}$ et $Sd_{Loc}$ les moyennes et déviation standards de P pour un locuteur.

- Méthode logarithmique [Wrede et Shriberg 2003]: Pnorm=log(P/Pmin) avec

$$P_{Norme} = \log \frac{P}{P\min_{Loc}}$$

avec $P_{min}$ tel que 3% des données par locuteur soient inférieures à $P_{min}$.

- Normalisation de Nearey [Adank 2003] (est censer éliminer les différences dues aux différentes longueurs du conduit vocal)

$$P_{Nea} = \frac{\log P}{Moyenne(\log(P_i))}$$

Un exemple de segment normalisé avec les différentes méthodes est donné Figure 5-12. La normalisation ne semble pas lisser la courbe.

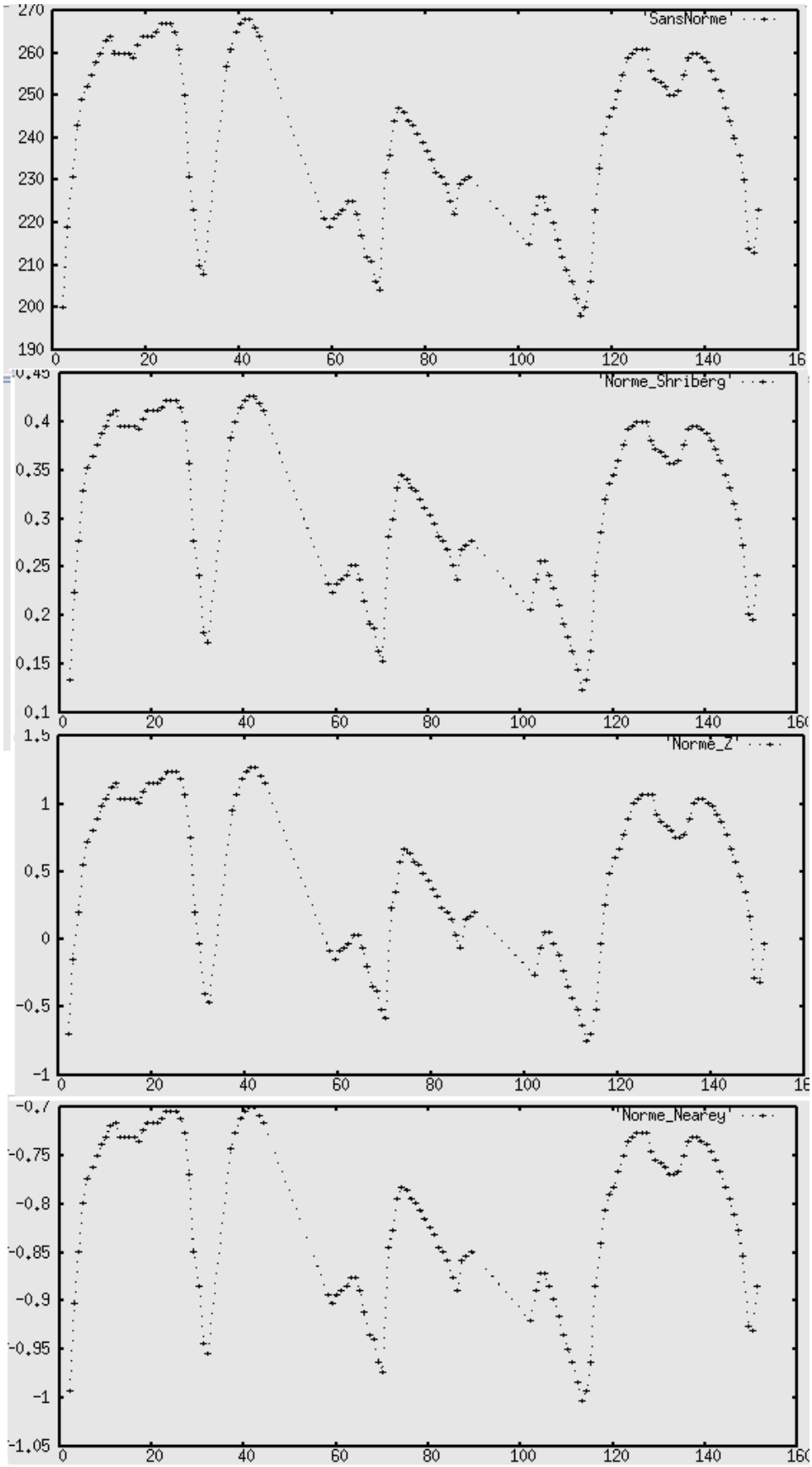



**Figure 5-12.** *Comparaison entre les courbes de F0 sans normalisation, en utilisant la Z-norme, la normalisation de Shriberg et celle de Nearey.*

### 5.2.4. Tendances des paramètres comparées à celles de Scherer

En séparant simplement hommes et femmes et en regardant les paramètres prosodiques tous locuteurs du même sexe confondus, on n'observe pas vraiment de tendances. (Il y a plus de variations inter-locuteurs qu'inter-émotions). Nous avons donc comparé différents paramètres sur des données normalisées avec la Znorme. N'ayant pas effectué d'alignement syllabique des données, nous avons considéré que les durées syllabiques devaient évoluer à peu près pareillement que les durées phonémiques.

Les résultats pour les paramètres calculés sont représentés dans le Tableau 5-3. Si pour la Peur et la Colère les paramètres se comportent à peu près conformément à l'étude de Scherer, ce n'est pas toujours le cas de la tristesse. Il peut y avoir en particulier, une grande variation de F0. Pour nos données la déviation standard de F0 pour la tristesse était légèrement inférieure à celle du Neutre pour les femmes et légèrement supérieure pour les hommes.

Les paramètres jitter, shimmer et HNR peuvent s'obtenir de plusieurs façons et n'étaient pas normalisés dans notre étude ce qui peut expliquer les différences. Avec des données réelles où les émotions peuvent s'exprimer avec une grande variabilité, on n'observera donc pas nécessairement les mêmes tendances que pour des données plus contrôlées.

| Paramètres acoustiques | Stress | Colère/ Rage | Tristesse | Peur /Panique |
|---|---|---|---|---|
| **Débit et Fluency** | | | | |
| Nombre de phonèmes par seconde | > | <> | < | > |
| Durée des phonèmes | < | <> | > | < |
| Nombre et durée des pauses | < | ≤ | > | <> |
| **F0 et Prosodie** | | | | |
| moyenne F0 | > | > | < | ≥ |
| déviation standard de F0 | > | > | ≤ | > |
| Plage F0 | > | > | ≤ | <> |
| **Effort Vocal et Type de Phonation** | | | | |
| Intensité moyenne (dB) | > | > | <= | |
| Jitter | | >= | | > |
| Shimmer | | >= | | ≥ |
| HNR | | ≥ | ≤ | ≤ |

**Tableau 5-3.** *Comparaison entre la review de Scherer (cf.Tableau 5-1) et les données CEMO. Les conclusions partagées sont surlignées en jaune et celles différentes barrées en rouge.*

### 5.2.5. Triangles vocaliques

Le triangle des voyelles montre que « les zones privilégiées des fréquences phonatoires, renforcées par les différents résonateurs pharyngés, varient l'une en fonction de l'autre suivant la voyelle prononcée.

Nous avons tracé les triangles vocaliques en suivant la méthode de Vieru *et al.*[Vieru-Dimulescu et Boula de Mareüil 2006]. Les valeurs des 2 premiers formants des voyelles n'ont pas été ajouté au vecteur de paramètres, mais il serait intéressant de les évaluer.

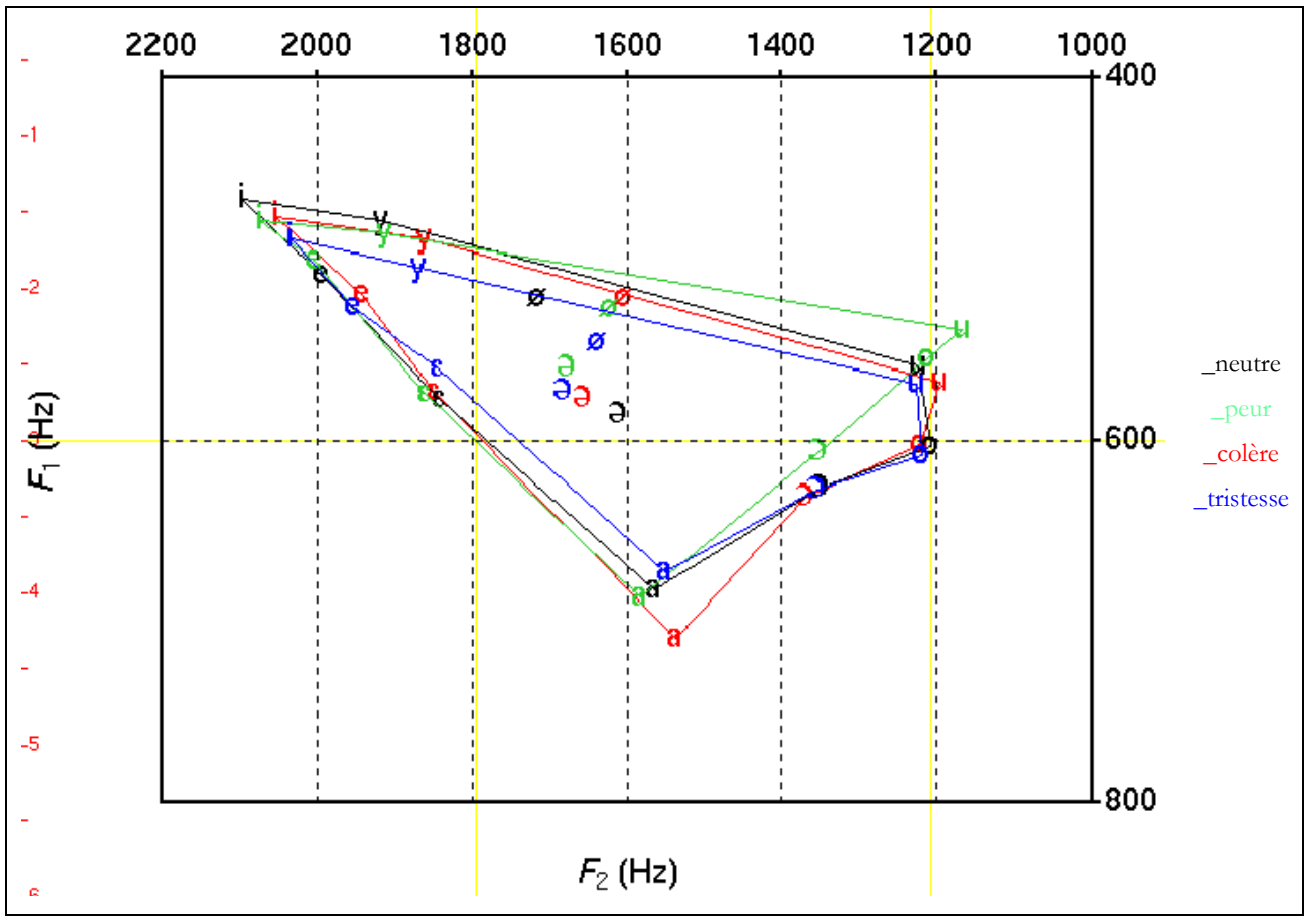



**Figure 5-13.** *Triangle vocalique des femmes pour les émotions Neutre/Peur/Colère/Tristesse (normalisation de Nearey.*

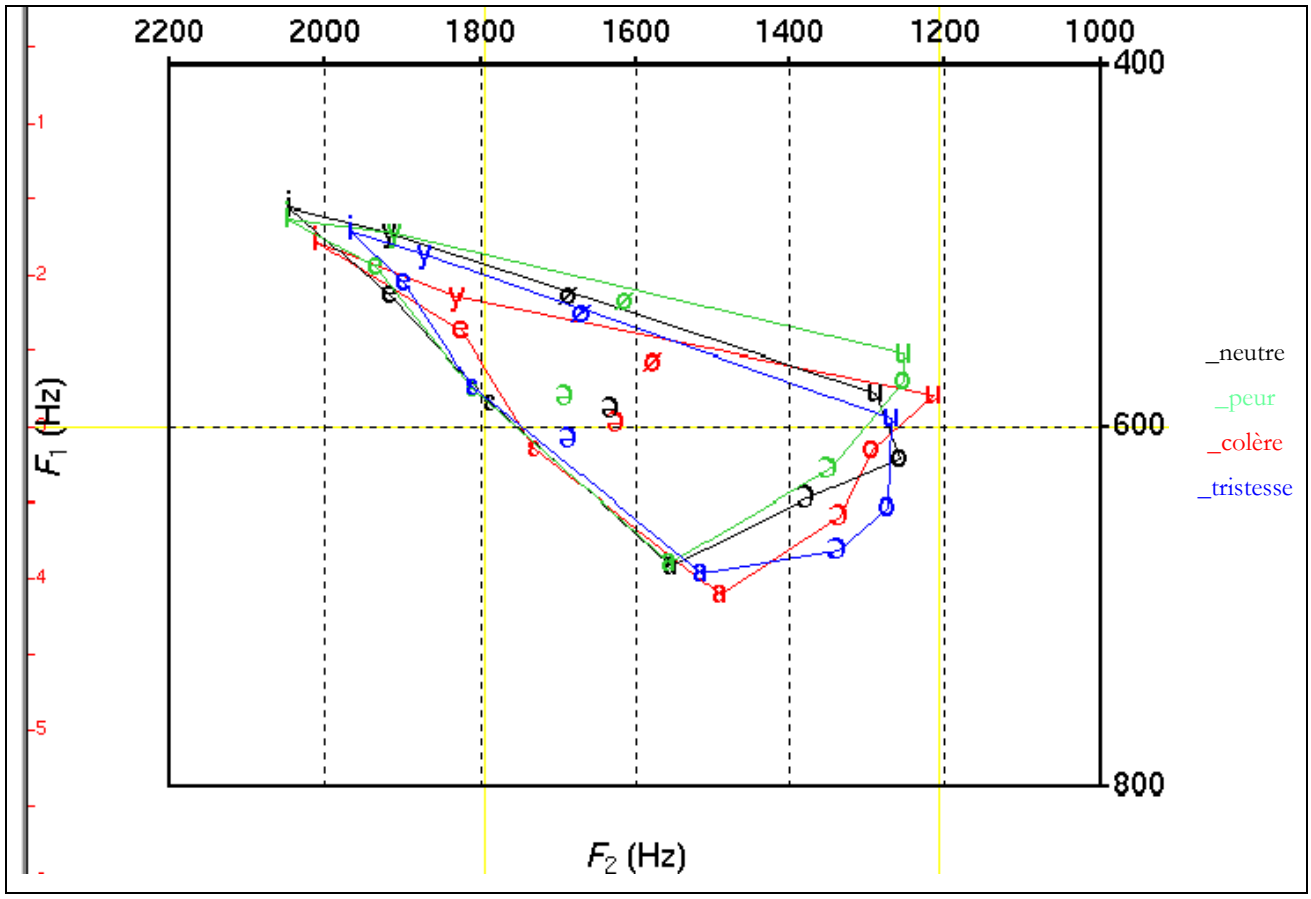



**Figure 5-14.** *Triangle vocalique des hommes pour les émotions Neutre/Peur/Colère/Tristesse (normalisation de Nearey).*

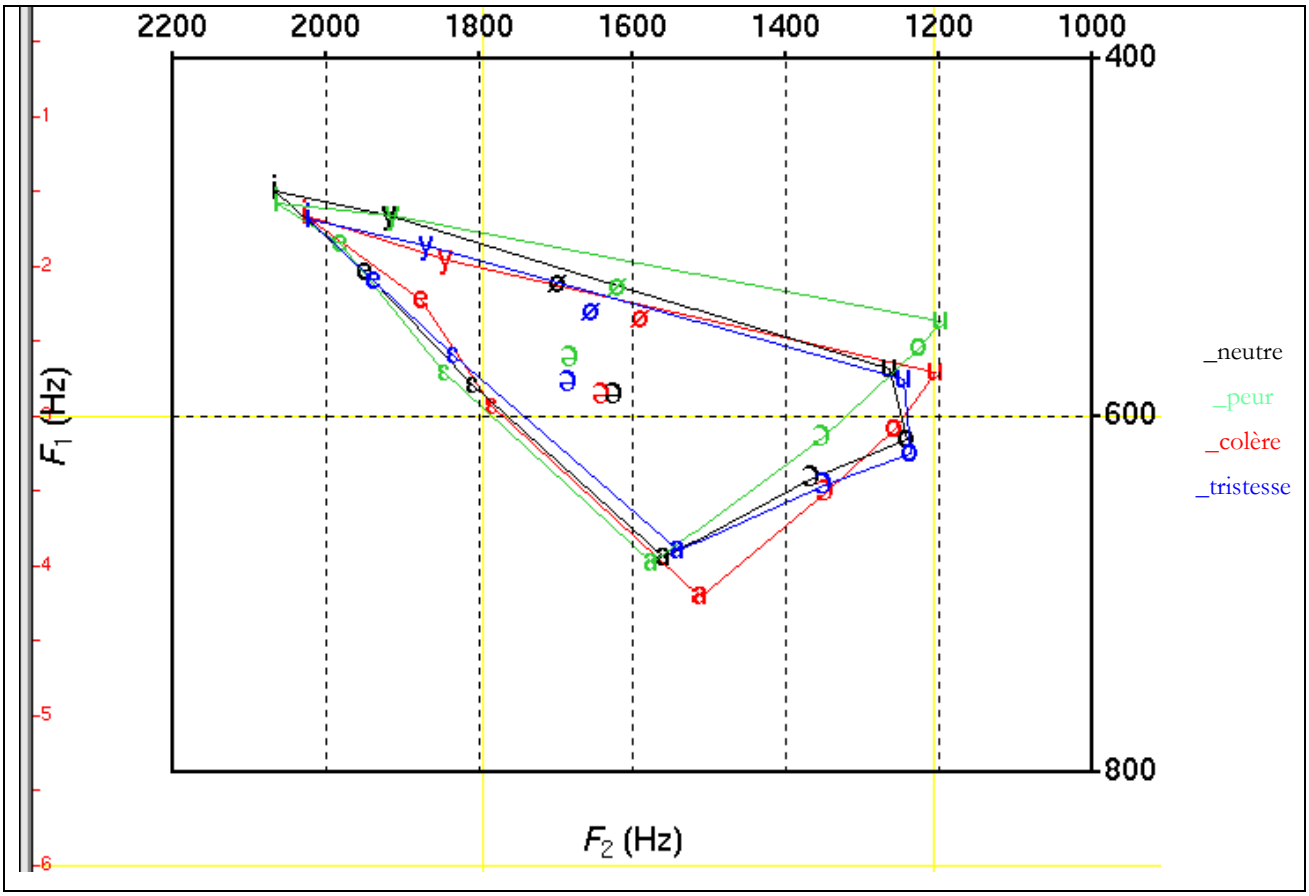



**Figure 5-15.***Triangle vocalique pour les émotions Neutre/Peur/Colère/Tristesse (normalisation de Nearey).*

## 5.3. Conclusion

Dans ce chapitre, nous avons décrit une multitude d'indices différents pouvant être extrait pour la détection des émotions avec en particulier des indices prosodiques, spectraux pouvant être extraits automatiquement, et que nous qualifierons de « blind » dans la suite de cette thèse et d'autres nécessitant un traitement manuel, bien qu'il n'est pas exclu qu'ils puissent être exclus automatiquement dans les années à venir. Nous avons décrit en détail les mesures déduites de ces indices avec au total plus d'une centaine de paramètres extraits pour chaque segment, certains très locaux comme par exemple le maximum et d'autres globaux comme la moyenne. Nous avons également insisté sur la nécessité de normaliser ces paramètres en présentant plusieurs méthodes de normalisation. Certains sont très redondantes et une stratégie usuelle sera de leurs appliquer des algorithmes de sélection.

Quels sont les plus pertinents ? Sont-ils tous nécessaires ? En quoi leur combinaison pourra-t-elle améliorer la détection ? C'est ce qui est abordé dans le prochain chapitre.

# Chapitre 6

# Apprentissage pour la détection des émotions


Résumé

Ce chapitre porte sur l'apprentissage de modèles computationnels pour la détection des émotions. Plusieurs questions se posent lorsqu'on veut utiliser la fouille de données pour la détection des émotions. Quel algorithme utiliser ? Combien de classes d'émotions peut-on envisager de discriminer ? Comment optimiser les résultats (pré-traitement des données, choix de l'algorithme apprentissage, normalisation, sélection et combinaison des paramètres, …) ? Y-a-t-il des différences entre les différents rôles (agent vs. appelant, homme vs. femmes) Notre méthodologie est-elle transposable sur d'autres types de données ? Quelles performances les classifieurs entrainés sur un corpus ont-ils sur différentes données ? La méthodologie développée sur les données CEMO peut elle être appliquée à d'autres données et les « modèles » construits sur CEMO peuvent-ils être testés sur d'autres données ?

Dans ce chapitre, nous décrivons d'abord les algorithmes que nous avons utilisés, ainsi que différentes considérations méthodologiques sur la préparation des données. Nous présentons ensuite les expériences réalisées sur les corpus LIMSI (données boursières et CEMO) avec plusieurs axes de recherche :

- variation du nombre de classes émotion
- ajout de contexte : différences agent/appelant et hommes/femmes
- importance des différents types d'indices
- mélange d'indices linguistiques et paralinguistiques

Nous présentons ensuite les résultats obtenus sur les données AIBO dans le cadre de la collaboration CEICES (Combining Efforts for Improving Automatic Classification.of Emotional User States) dans le réseau d'excellence FP6-HUMAINE visant entre autre à pallier le manque de méthode standard d'évaluation et l'absence de corpus de référence.

Enfin nous testons la portabilité des classifieurs entrainés sur les données CEMO sur d'autres données collectées dans les mêmes conditions acoustiques (centre d'appel sur une autre tâche) et sur des données actées prototypiques collectées par l'université de Genève.

*With a 20h corpus rich in emotions, how many emotion classes could be discriminated? What are the best algorithms and the most relevant parameters? How to optimize the results? Is there a difference between the speaker roles (agent vs. caller), between the genders? Can we use our method for other type of data? And how well do classifiers trained on our corpus perform on other data?*

*In this chapter, we will start with a description of the algorithms that we used and the data preprocessing. We will then describe the experiments performed on the LIMSI corpora with several goals:*

- *number of classes to discriminate*
- *role of the context (Agent/Caller, Gender)*
- *relative importance of several cues*
- *combination of prosodic and linguistic information*

*We will then present the 'forced co-operation' CEICES (Combining Efforts for Improving Automatic Classification.of Emotional User States) in FP6-HUMAINE in which several sites compared and combined their expertise on a corpus of interactions between children and the sony dog AIBO.*

*Finally we will look into the performances of classifiers trained with our data on other data collected both in similar acoustic conditions (call center with a different task) and on acted speech (data collected by the university of Geneva).*

# 6. APPRENTISSAGE POUR LA DETECTION DES EMOTIONS

Dans les chapitres précédents, nous avons décrit en détail nos corpus, la manière dont ils ont été annotés, à l'aide d'un vecteur émotion pour chaque segment, et tous les indices que nous avons extraits par segment. Pour toutes les expériences qui vont suivre, nous sélectionnerons pour l'apprentissage et pour le test des segments « simples », c'est-à-dire pour lesquels le vecteur émotion n'a pour champ non nul que des étiquettes correspondant à une même classe. Au préalable, nous justifierons notre méthodologie et nos choix.

## 6.1. L'apprentissage automatique : cadre général pour nos travaux

L'apprentissage peut être défini comme « toute technique permettant d'améliorer les performances d'un système en cours d'utilisation »[Kodratoff et Barès 1991]

Nous nous intéresserons dans nos travaux à l'apprentissage supervisé : à partir d'un nombre limité d'observations (dans notre cas des segments avec une étiquette émotion), nous cherchons à estimer la classe de données[1] inconnues. Nous appellerons « classifieur » ou « modèle » (ou encore modèle computationnel pour bien le différencier d'un modèle théorique) l'objet permettant d'associer un nom de classe à une instance inconnue. On entraîne tout d'abord le classifieur sur un ensemble de données (appelé ensemble d'apprentissage ou « training set » en anglais). Le classifieur est évalué sur un ensemble de données étiquetées qui n'ont pas été utilisées pour l'apprentissage et le résultat de cette évaluation peut être représenté par une matrice de confusion.

Les algorithmes que nous avons utilisés sont principalement les arbres de décision et les Support Vector Machine] (ou SVM) [Vapnik 1998].

Pour toutes nos expériences, nous avons utilisé le logiciel libre Weka [Witten et Franck 2005] qui est un ensemble d'outils de fouille de données permettant le traitement et la sélection des paramètres et proposant différents algorithmes d'apprentissage. Ce logiciel est actuellement de plus en plus utilisé dans la communauté de reconnaissance des formes. Il englobe de nombreux

[1] Donnée : « Ensemble de valeurs prises par un ou plusieurs descripteurs d'un objet ou d'un évènement »[Kodratoff et Barès 1991]

algorithmes connus comme par exemple les SVM, les arbres de décision (J48), ainsi que Méta algorithmes[1].

Au cours de la thèse, le nombre de paramètre extraits a régulièrement évolué et la plupart des expériences décrites dans ce chapitre ont été réalisées avec l'ensemble de paramètres le plus récent décrits en détail dans le chapitre 5.
Avant de présenter nos résultats, nous allons tout d'abord rapidement décrire les principaux algorithmes qui ont été utilisés.

## 6.1.1. Algorithmes

### 6.1.1.1. Les arbres de décision

Les arbres de décision sont des méthodes de classification pour des instances représentées dans un formalisme attributs/valeur. Un arbre est « la représentation graphique d'une structure dans laquelle un nœud appelé le père est relié à un ou plusieurs autres nœuds, les fils. […] Formellement, on le définit comme un graphe connexe sans cycle. […] Un nœud sans père est appelé la racine de l'arbre. Le nœud sans fils est appelé une feuille. »[Kodratoff et Barès 1991]
Un arbre de décision est un « arbre dont chaque nœud correspond à un choix (une décision) et dont les fils sont les conséquences de ce choix »
Il existe différents algorithmes d'apprentissage des arbres de décisions comme par exemple les « Logistic Model Trees » [Landwehr et al. 2003], qui sont des arbres de classification avec des fonctions de régression linéaire aux feuilles ou l'« Alternative decision tree » ADTree [Freund et Shapire 1996] qui combine par vote pondéré les résultats de plusieurs arbres.

[1] Approche de plus en plus populaire qui consiste à combiner les sorties de différents modèles par un vote ou en moyennant des différentes prédictions dans le cas de prédictions numériques.

### 6.1.1.2. Les Séparateurs à Vaste Marge (SVM : Support Vector Machine)

L'idée des SVM (Support Vector Machine ou Séparateurs à Vaste Marge [Vapnik 1998]) est de trouver le meilleur hyperplan séparateur permettant de séparer deux ensembles de points, c'est-à-dire celui pour lequel la distance minimale aux exemples d'apprentissage est maximale. Cette distance est appelée « marge ». (*cf.* Figure 6-1).

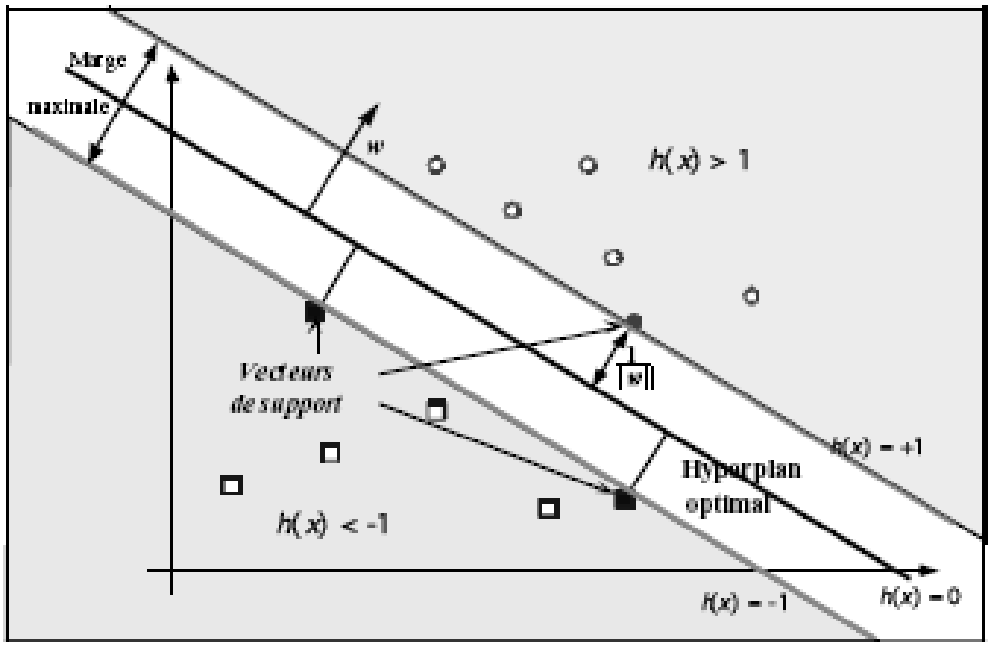



**Figure 6-1**. *Hyperplan optimal de marge 1/||w|| (schéma tiré de l'article de Cornuéjols [Cornuéjols 2002]).*

S est l'échantillon d'apprentissage S= {(x1, u1), (x2, u2), …, (xm, um)}

On cherche $h(x)= w_0 + w^T x$ tq $h(x) >0 => u_i=1$

$h(x) < 0 => u_i= -1$

soit encore $u_i(w_0 + w^T xi ) > 0$

La recherche de l'hyperplan optimal revient à minimiser ||w||. Le problème se résout mathématiquement [Cornuéjols et Miclet 2002] et la solution ne requiert que le calcul de produits scalaires. La contrainte des marges peut être relâchée en introduisant une variable ressort permettant de tolérer un certain nombre d'erreur. Une constante C définie par l'utilisateur va alors borner le nombre d'erreurs tolérées.

Pour des échantillons non linéairement séparables, la solution est de projeter les données dans un espace de dimension supérieur (potentiellement infini) dans lequel il existe un hyperplan

permettant de séparer linéairement les données. Cependant quand l'espace est grand, le calcul des produits scalaires devient impraticable. Une solution est alors d'utiliser des fonctions bilinéaires symétriques positives appelées fonctions noyaux, faciles à calculer et qui correspondent à un produit scalaire dans un espace de grande dimension.

En pratique, on choisit une fonction noyau que l'on sait correspondre à un produit scalaire dans un espace alors virtuel et on regarde si elle permet d'obtenir de bonnes fonctions de décision. (Il est nécessaire d'opérer alors par essai erreur). Les fonctions noyaux les plus utilisées sont indiquées dans le Tableau 6-1 ci-dessous.

| | |
|---|---|
| Linéaire | $K(x_i, x_j)=x_i^T x_j$ |
| Polynomiale | $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0.$ |
| RBF (a base radiale) | $K(x_i, x_j)= \exp(-\gamma\|\|x_i-x_j\|\|^2), \gamma > 0$ |
| Sigmoïdes | $K(x_i, x_j)=\tanh= (\gamma x_i^T x_j + r)$ |

**Tableau 6-1**. *Fonctions noyaux les plus utilisées. r, d et γ sont des paramètres des fonctions noyaux.*

Au final, pour utiliser les SVM, l'utilisateur doit simplement choisir le **coefficient C**, qui règle le compromis entre la marge possible entre les exemples et le nombre d'erreurs admissibles, **la fonction noyau et ses paramètres**. Il est conseillé de commencer avec les RBF car en pratique, ils donnent de bons résultats [Chih et al. 2003].

## 6.1.2. Méthodologie : Préparer et évaluer les données

Que le but soit d'évaluer une tâche sur les données dont nous disposons ou de pouvoir faire des prédictions sur des données inconnues, nous avons besoin d'un ensemble de données étiquetées n'ayant pas été utilisées pour l'apprentissage afin d'évaluer le classifieur et éviter le sur-apprentissage (le classifieur est très performant sur les données ayant servies pour le construire, mais il a un pouvoir de généralisation faible sur des données inconnues) [Cornuéjols et Miclet 2002].

Comme pour de nombreuses applications, nous sommes limités par la quantité de données dont nous disposons. Nous voulons créer des classifieurs le plus génériques possible, ce qui suppose d'avoir un nombre suffisant de données, mais il faut également suffisamment d'instances de test pour pouvoir généraliser sur les performances du classifieur sur des données inconnues. Enfin, il ne faut pas oublier que les performances des systèmes automatiques de détection des émotions seront comparées à la perception humaine.

### 6.1.2.1. Apprentissage et test / Validation croisée

Pour Cabena [Witten et Franck 2005 p60], 60% du travail pour la fouille de données est dans la préparation des données. Idéalement, il faut séparer les données en 3 ensembles : un ensemble d'apprentissage, un ensemble de calibration et un ensemble de test. Une solution lorsqu'on dispose de peu de données est la **validation croisée**. On divise les données en N ensembles, un est utilisé pour le test et les autres pour l'apprentissage et ceci pour les N sous ensembles. Le score de bonne détection est alors la moyenne des N scores avec un intervalle de confiance donné par la déviation standard des N scores. Cependant, la validation croisée sert principalement à évaluer les données et le logiciel Weka ne permet pas d'obtenir un classifieur afin de faire des tests sur des données nouvelles[1]. Nous aurions pu faire manuellement la validation croisée en séparant les données en N ensembles en prenant soin d'avoir des locuteurs différents dans chaque sous ensemble et une distribution à peu près identique des émotions ; puis entraînant un classifieur pour chaque sous ensemble. Finalement, pour les expériences les plus récentes, nous avons divisé les données en un ensemble d'apprentissage et un ensemble de test avec des locuteurs différents. Le nombre de

[1] Le logiciel weka, pour la validation croisée avec N ensembles, ne donne pas les performances respectives sur chaque sous-ensemble, mais seulement un score moyen. Si on lui demande de tester le « model » sur un ensemble de test distinct, il recréera un classifieur à partir de l'ensemble complet d'apprentissage. De plus, l'utilisateur a seulement accès aux résultats globaux et ne peut pas contrôler la distribution des émotions ou celle des locuteurs par sous ensemble.

segments utilisés pour l'apprentissage et le test varient selon les émotions que l'on cherche à discriminer. Toutes les expériences ont cependant été effectuées avec plus de 250 segments par émotion pour l'apprentissage.

## *Données non équilibrées*

Parce que les données sont majoritairement neutres et que la fréquence des émotions n'est pas la même pour toutes les classes, on dispose souvent de données non équilibrées avant de commencer l'apprentissage. Afin de ne privilégier aucune classe, il faut prendre en compte ce déséquilibre.
Pour certains algorithmes, des poids peuvent être fixés à l'apprentissage afin de pénaliser les classes les plus fréquentes, mais en pratique, on préfèrera sélectionner un ensemble de classes équilibrées pour l'apprentissage (Cette stratégie a été utilisée par tous les partenaires de CEICES). La rareté de certaines classes fait que si on sélectionne pour l'apprentissage le nombre d'instances de la classe la moins représentée fois le nombre de classes, on perd des informations. Une solution qui est souvent utilisée est de dupliquer ou tripler certaines instances comme dans l'exemple de Figure 6-2. Est-ce que cela biaise l'apprentissage ? Comment choisir le nombre optimal d'instances par classe pour l'apprentissage ? Il n'y a pas vraiment de règles. Cela va dépendre entre autres du nombre de classes que l'on cherche à discriminer, de la distance entre ces classes et du nombre d'indices calculés pour chaque instance.

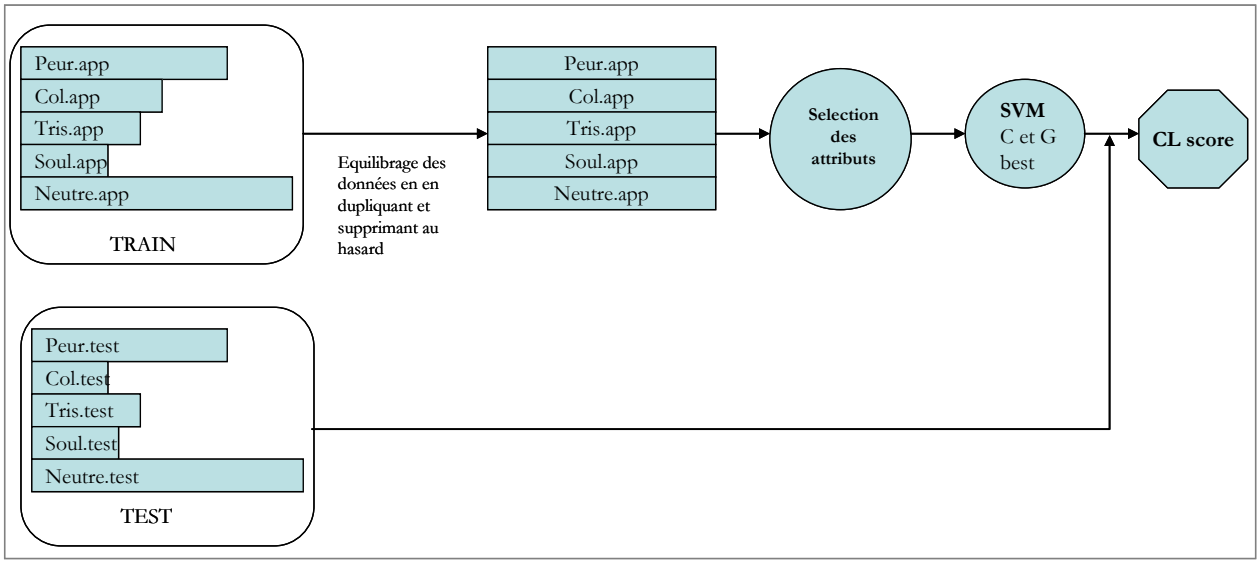



**Figure 6-2.** *Obtenir des données équilibrées pour l'apprentissage : un exemple pour une classification Peur/Colère/Tristesse/Soulagement/Neutre avec des données non équilibrées pour l'apprentissage et en utilisant des SVM*

.

### 6.1.2.2. Comment représenter et évaluer les résultats ?

N'ayant pas d'application précise en vue, nous cherchons à obtenir la meilleure détection possible sans favoriser d'émotions et en pénalisant les « falses negatives ».

Une représentation complète des résultats est la matrice de confusion, mais la plupart des études présentent également leurs résultats sous la forme d'un taux de détection, ce qui permet de comparer plus facilement les différentes expériences et de donner un résultat délié des données qui ont servi à l'obtenir.

Dans certaines des études, le score qui est donné est le score de bonne détection ou **RR rate** (Nombre de bonne détection/ Nombre total de segments). Dans le cas de classes non équilibrées, ce score n'est pas nécessairement très significatif[1], surtout si certaines émotions sont mieux reconnues que d'autres et le score RR risque de varier selon la distribution de l'ensemble de test.

Nous avons donc choisi d'évaluer nos résultats en utilisant le **CL score** (Class_wise : moyenne de la diagonale de la matrice, *cf.* Figure 6-3). Ainsi, les scores de détection ne dépendront pas de la distribution de l'ensemble de test et pour les meilleurs modèles, le taux de reconnaissance par émotion est à peu près celui donné par le CL score.

Une autre mesure intéressante est la précision par émotion, le nombre de fois où une émotion est correctement identifiée divisé par le nombre de fois où elle est identifiée (bien ou mal).

| Score | formule |
|---|---|
| RR (Recognition Rate) | $\frac{\sum m_{ii}}{Total}$ |
| CL (Class wise) | $\frac{1}{n}\sum \frac{m_{ii}}{m_i}$ |
| FF score | $\frac{RR + CL}{2}$ |
| Précision (par émotion i) | $\frac{m_{ii}}{m_i + m_{i.}}$ |
| Recall (par émotion i) | $\frac{m_{ii}}{m_i}$ |

Matrice de confusion :

| | | Reconnue comme | | | | | |
|---|---|---|---|---|---|---|---|
| | | $E_1$ | … | $E_i$ | … | $E_n$ | Total |
| Classe | $E_1$ | $m_{11}$ | … | $m_{1i}$ | … | $m_{1n}$ | $\mathbf{m_1}$ |
| | : | : | | : | | : | … |
| | $E_i$ | $m_{i1}$ | … | $m_{ii}$ | … | $m_{in}$ | $\mathbf{m_i}$ |
| | : | : | | : | | : | … |
| | $E_n$ | $m_{n1}$ | … | $m_{ni}$ | … | $m_{nn}$ | $\mathbf{m_n}$ |
| | Total | $\mathbf{m_{1.}}$ | … | $\mathbf{m_{i.}}$ | … | $\mathbf{m_{n.}}$ | |

**Figure 6-3.** *Différentes mesures de performances se déduisant de la matrice de confusion.*

[1] En prenant le cas extrême où la distribution des émotions correspond à celle des émotions dans des données réelles, avec en général 80% de données « neutre », un modèle qui classerait tout en Neutre aurait un très bon pourcentage de bonne détection, mais ne présenterait pas un grand intérêt.

## *Evaluer la fiabilité des résultats*

Certaines de nos expériences comparent les performances de différents classifieurs sur un même ensemble de test, mais d'autres comparent les performances de différentes tâches (Colère/Neutre vs. Peur/Colère par exemple). Pour ce type d'expérience, il est particulièrement nécessaire de vérifier que les différences de performances sont bien significatives. Comme indiqué page 127, le logiciel Weka permet de faire de la validation croisée et donne la moyenne et la déviation standard du RRscore des N expériences. La déviation standard informe sur la variabilité des performances[1] pour une tâche et permet de voir si les différences de performances entre 2 tâches différentes sont significatives. C'est ainsi que nous avons procédé pour les premières expériences en utilisant Weka [Vidrascu et Devillers 2005a]. Comme nous nous intéressons au score CL et que nous voulons contrôler l'équilibre des données de l'apprentissage, nous n'avons plus utilisé Weka pour les dernières expériences, mais nous avons procédé de manière similaire en regroupant les données utilisées pour l'apprentissage et le test et en répétant N fois l'expérience de choisir aléatoirement 75% des données pour l'apprentissage et le reste pour le test ; puis de regarder la moyenne et la déviation standard des scores RR et CL.

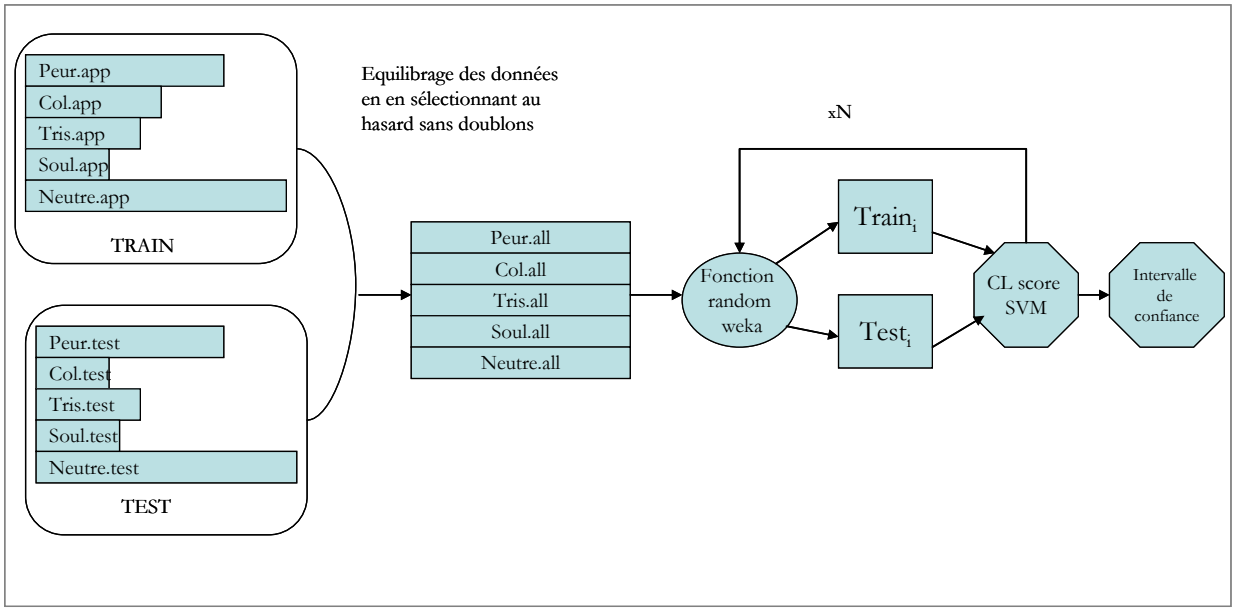



**Figure 6-4.** *Création de N classifieurs en faisant varier les ensembles d'apprentissage et de test afin d'avoir un aperçu de la variabilité des résultats.*

[1] Pour une même tâche, les performances varient suivant les données utilisées pour l'apprentissage (pour celles du test également, si on dispose de trop peu de données). Pour les tâches simples pour lesquelles on dispose de beaucoup de segments, cette variation est inférieure à 1%, mais elle peut être plus importante lorsqu'on travaille sur beaucoup de classes et peu de données.

### 6.1.3. La sélection des attributs

Le but est de diminuer le nombre de descripteurs (complexité) sans nuire à la qualité des résultats (performances) [Cornuéjols et Miclet 2002]. La diminution de la dimension de l'espace des attributs permet d'améliorer la vitesse et les performances de modèles de détection des émotions [Dellaert et al. 1996] car avec certains algorithmes, des attributs non pertinents peuvent faire baisser les performances [Lee et al. 2001]. La sélection des attributs peut également faciliter l'intelligibilité des données.

Deux types d'approches existent :

- l'**élimination** des attributs les moins pertinents
- l'**extraction d'attributs** qui réduit la dimension de l'espace d'entrée en appliquant des transformations des attributs (analyse en composantes principales par exemple)

Pour évaluer les performances des attributs, on peut soit utiliser des « wrappers », qui s'appuient sur un algorithme d'apprentissage ou des « filters » qui calculent des mesures d'entropies indépendantes de la méthode de classification.

Deux types de méthodes de sélection des attributs existent :

- des méthodes qui évaluent des sous ensemble d'attributs, comme par exemple la « Correlation based feature selection ». Hall a montré dans sa thèse qu'elle permettait d'avoir des performances souvent supérieures à celles obtenues en gardant tout les attributs, à l'exception de certains cas lorsque des attributs étaient éliminés qui étaient très bons prédicteurs d'une toute petite partie de l'ensemble des données [Hall 1999].
- d'autres qui évaluent la pertinence d'un attribut individuellement [Hall et Holmes 2003], soit par calcul d'un gain d'information (GainRatio, InfoGain, Relief), soit à l'aide d'un « wrapper » (SVM predictif) et classent les attributs par ordre de mérite.

Pour la sélection des attributs, nous avons utilisé les 4 méthodes (InfoGain, GainRatio, SVM, Relief) implémentées dans Weka. Nous avons comparé pour différentes tâches les performances de classification avec chaque méthode de sélection d'attribut et avons attribué un rang à chaque attribut en moyennant les rangs obtenus pour chaque méthode. Le fait de moyenner les rangs obtenus avec chaque méthode donnait des performances identiques ou supérieures à celles obtenues en n'utilisant qu'une seule méthode et c'est dons ainsi que nous avons procédé par la suite.

Les listes des meilleurs attributs sont différentes selon les données et les émotions que l'on cherche à différencier comme l'illustre le Tableau 6-2.

| Colère/Neutre | Peur/Neutre | Peur/Colère | Peur/Colère/Neutre |
|---|---|---|---|
| nbmots | nbmots | pSlopeWithoutOctave | maxhes |
| debitm | nbVoyelles | moyenneF1 | nbmots |
| sdF1 | Duree | rangeEn | meanhes |
| moyenneEn | Nbhes | quartileEn | debitm |
| debVoy | nbPhones | medianF21 | tquartileF0 |
| tquartileF0 | Debitm | HNR | moyenneF21 |
| voisNonvois | bb | tquartileF0 | pSlopeWithoutOctave |
| quartileEn | nbTrunc | jitterLocal | PI |
| slopeMaxF0 | minF1 | jitterPpq5 | nbhes |
| coeffRegMinSegEn | lengthmaxPhone | minF32 | moyenneF32 |
| moyenneF21 | slopeMaxF0 | MSEMaxSegEn | minF32 |
| maxF21 | sdBW3 | minF3 | debVoy |
| coeffRegMaxSegEnn | debVoy | MSEMaxF0 | medianF1 |
| MSEMeanF0 | debPhones | Maxhes | bb |
| medianF21 | medianF0 | Nbmots | MSEMeanF0 |
| Bb | sdBW2 | moyenneEn | HNR |
| nbVoyelles | rangeBW3 | sdEn | sdF1 |
| HNR | lengthmax | nbVoyelles | sdEn |
| spkgRate | spkgRate | lVoyelles | rangeF32 |
| tquartileEn | moyenneF32 | maxF0 | moyenneEn |
| sdEn | rangeBW2 | interF0 | coeffRegMinSegEn |
| lengthmax | PI | sdF1 | maxEn |
| moyenneF1 | Punvoiced | MSEEn | sdF0 |
| MSEEn | maxhes | nbhes | MSEMaxF0 |

**Tableau 6-2**. *24 meilleurs paramètres (sur 129) pour 4 tâches différentes. Peur/Neutre, Colère/Neutre, Peur/Colère et Peur/Colère/Surprise.*

Nous avons également comparé la sélection globale des paramètres (un algorithme de sélection des paramètres est appliqué à l'ensemble des paramètres) et la selection en faisant une sélection séparée pour chaque classe[1] : prosodie, énergie, formants…, mais n'avons pas observé de différence significative dans les performances, bien que celles avec la sélection globale semblent légèrement meilleures.

[1] ce qui permet les meilleurs paramètres de chaque classe.

## 6.2. Quel algorithme utiliser ? Premiers résultats : Transaction boursières / CEMO

Dans un premier temps [Vidrascu et Devillers 2005a], nous avons comparé différents algorithmes fréquemment utilisés dans la fouille de données sur les 2 corpus de centre d'appels CEMO et les données boursières (Emotions très contrôlées dans les données boursières et beaucoup moins dans CEMO) afin de vérifier que les performances étaient bien comparables et de choisir l'algorithme à utiliser pour la suite de nos recherches.

### 6.2.1. Comparaison de différents algorithmes sur les données boursières et CEMO pour la classification de 2 classes.

*Données boursières*

Nos premières expériences (voir Tableau 6-3) comparaient : des arbres de décision (C4.5 et ADTree), des SVM et un « voting » algorithme (Adaboost) [Freund et Shapire 1996] pour une tâche de détection Neutre/Négatif en procédant par validation croisée avec 50 paramètres.

| | **C4.5** | **AdaBoost** | **ADTree** | **SVM** |
|---|---|---|---|---|
| **5best** | 72,8 ( 5,2) | 71,2 (4,5) | 72,3(4,6) | 67,2(6,3 ) |
| **10best** | 73,0 ( 5,3) | 71,5( 4,8) | 73,0( 5,7) | 69,5( 5,6) |
| **15best** | 71,7 ( 6,4) | 71,1( 4,7) | 71,6( 4,9) | 70,8( 4,9) |
| **20best** | 71,8 ( 5,3) | 71,3( 4,3) | 71,8( 5,1) | 71,0( 4,9) |
| **Allatt** | 69,4 (5,6) | 71,7 (4,3) | 71,6 (4,8) | 69,6 (3,5) |

**Tableau 6-3.** *Algorithmes et sélection des attributs : comparaison des performances Neutre/Négatif (Peur et Colère); RR score avec les meilleurs attributs1 et Allatt : tous les attributs. Le tableau montre la moyenne de segments bien classifiés pour 30 executions. Le nombre entre parenthèses est la déviation standard.*

Ce type d'expérience a été répété pour différentes tâches et nous n'avons pas constaté de différences significatives entre les différents algorithmes, ni de détérioration en ne sélectionnant que peu de paramètres.

[1] Pour cette tâche, les paramètres les plus pertinents étaient principalement des paramètres liés à la F0 (plage F0, maximum F0, pente F0, F0 minimum, coefficient régression F0 et son erreur quadratique moyenne, déviation standard de la F0, Inter-segment F,0 Intra-segment F0), à l'énergie (plage de l'énergie, moyenne de l'énergie, énergie maximum) aux formants et largeur de bande (moyenne F1 , moyenne F2, moyenne BW1, plage F2), à des paramètres de disfluences(# hésitations ("euh"), #pauses) et à des marqueurs affectifs (bruits de bouche, nombre de rires).

## ***Données CEMO***

Les paramètres extraits étaient les mêmes que ceux extraits pour les données boursières avec en plus des disfluences. Sur ces données, nous avons comparé un SVM et un LMT (Logistic Model Tree [Smith et Abel 1999], qui est un arbre de classification avec des fonctions de régression au niveau des feuilles (voir Tableau 6-4 ci-dessous).

| | **SVM** | **LMT** |
|---|---|---|
| **5best** | 80,28 (3,71) | 80,69 (3,14) |
| **10best** | 82,68 (3,17) | 82,65 (3,28) |
| **15best** | 83,17 (2,94) | 83,49 (3,03) |
| **20best** | 83,36 (3,02) | 83,42 (3,35) |
| **Allatt** | 83,16 (2,74) | 82,85 (3,36) |

**Tableau 6-4.** *Algorithmes et sélection des attributs : comparaison des performances de détection Positif/Négatif avec les meilleurs paramètres ; Allatt: tous les paramètres. Le tableau montre la moyenne de segments bien classifiés pour 100 executions. Le nombre entre parenthèses est la déviation standard.*

Les mêmes tendances s'observent pour les deux corpus : il n'y **a pas de différences significatives entre les différents algorithmes et la sélection des paramètres n'a pas une incidence négative sur les performances**. En accord avec d'autres études dans plusieurs domaines incluant celui des émotions[Lee et al. 2002] [Schuller et al. 2005], les SVM, et en particulier ceux à noyaux RBF s'avèrent donner de bons résultats quelle que soit la tâche et sont assez simples à entraîner. La seule difficulté est de trouver les coefficients C et Gamma, qui dépendent du type de données et de la tâche. Comme il n'y a pas de règles ou de méthodes pour les choisir de manière optimale, nous les faisons varier afin de trouver les plus adaptés.

### 6.2.2. Intérêt de ne pas utiliser les mélanges : exemple Peur/Colère sur CEMO et données boursières.

L'annotation des données CEMO et l'utilisation d'un vecteur émotion comme étiquette permettent de distinguer les segments « simples » et les émotions complexes. Sur les données boursières, nous avions constaté (voir chapitre Annotation) que les scores de détection pour la classification Peur/Colère étaient assez faibles (60% de bonne détection environ) et que cela pouvait peut être s'expliquer par le nombre élevé de mélanges Peur/Colère parmi les segments utilisés pour l'apprentissage[Vidrascu et Devillers 2005a].

Nous avons effectué des expériences similaires de classification Peur/Colère pour les données CEMO avec des SVM en utilisant ou non des mélanges pour l'apprentissage. Les performances sont significativement meilleures[1] lorsque les mélanges sont retirés de l'apprentissage avec un score CL de 82% de bonne détection sans mélanges, contre 78% avec mélanges. Parallèlement, un même modèle aura des meilleures performances sur un ensemble de test sans mélanges. Toutefois, les performances, même avec les mélanges sont bien meilleures que pour les données boursières. Plusieurs raisons sont possibles. Tout d'abord plus de segments sont utilisés pour l'apprentissage (il y a plus de 800 segments par émotion dans CEMO contre 192 *Peur* et 243 *Colère* dans les données boursières). De plus, les émotions sont beaucoup plus contrôlées et moins intenses dans les données boursières.

Le fait de ne pas utiliser de mélanges d'émotions dans l'apprentissage des modèles permet ainsi d'avoir de meilleures performances et **pour toutes les expériences décrites dans ce chapitre, les segments correspondant à des mélanges d'émotions ne seront pas utilisés**.

### 6.2.3. Combien de données pour l'apprentissage ?

Pour des tâches simples, peu de données suffisent pour pouvoir construire de bons classifieurs. Par exemple pour la détection Peur/Neutre, on peut obtenir des scores CL de plus de 80% de bonne détection avec une cinquantaine d'instances par classe pour l'apprentissage. Toutefois, pour des tâches plus complexes, il devient important d'avoir assez de données pour l'apprentissage, comme l'illustre la Figure 6-5.

---

[1] Chaque expérience a été répétée 250 fois avec les mêmes données en test et en faisant varier l'ensemble d'apprentissage (avec et sans mélanges). Un t-test entre les 2 ensembles de résultats donnait $p<0.0001$ (différence très significative).

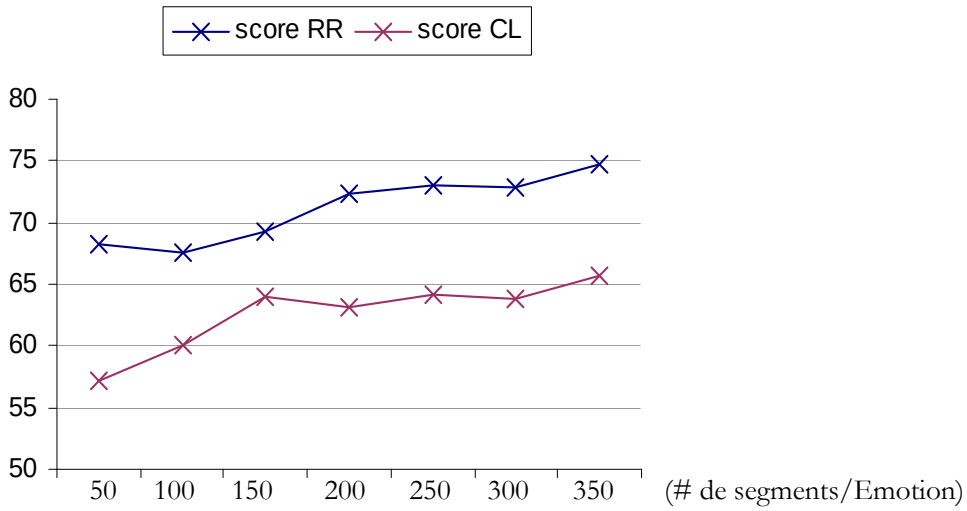



**Figure 6-5.** *Evolution des scores CL et RR sur un même ensemble de test pour la classification Peur/Colère/Neutre en faisant varier le nombre de segments par émotion pour l'apprentissage. (Il n'y a que 180 segments distincts pour la colère qui sont aléatoirement dupliqués au dessus de 250 segments par émotion. Les données de test ne sont pas équilibrées (moins de « Colère » qui est la classe la moins bien reconnue).*

Dans beaucoup de cas, le fait de dupliquer les segments de la classe la moins représentée permet d'améliorer les performances. C'est ce qui est illustré par Figure 6-5, qui donne les scores RR et CL pour une classification Peur/Colère/Neutre en fonction du nombre de segments par classe d'émotion. Le nombre de segments distincts pour la classe Colère est de 180, mais le fait d'en dupliquer afin d'avoir plus de variétés de Peur et de Neutre pour l'entraînement permet d'améliorer le modèle

### 6.2.4. Quelle normalisation ?

Différents tests ont été faits pour comparer les normalisations (cf. p112) sur des tâches de classification entre 2, 3 ou 4 émotions en comparant les performances sans normalisation, avec la Z-normalisation et avec la normalisation de Nearey. Il n'y avait pas de différences significatives dans les scores CL.

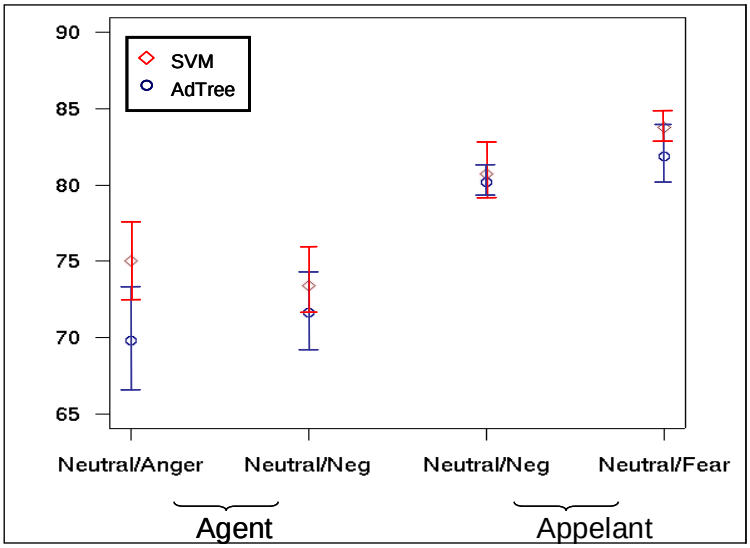
## 6.3. Sur les données CEMO

### 6.3.1. Informations contextuelles : Différences Agents/Appelants, Hommes/Femmes

Nous avons vu dans le chapitre 4 que les informations contextuelles pouvaient influer sur les émotions exprimées. Bien qu'il soit intéressant de regarder si elles influent sur la manière dont une même émotion va s'exprimer, elles sont souvent difficiles voire impossibles à extraire automatiquement d'un énoncé, à quelques exceptions près. En particulier, il est facile de distinguer entre agent et appelant et entre homme et femme.

#### *Agent/Appelant*

Nous avons tout d'abord regardé les différences suivant le « rôle » du locuteur (Agent/Appelant) en comparant les performances de systèmes de détection entraînés et testés uniquement sur des agents ou uniquement sur des appelants. Les émotions exprimées par les agents et clients étant différentes, il est difficile de comparer les classifications avec beaucoup de classes.




| *Rôle* | *Tâche* | *# Segments par émotion* | *ADTree* | *SVM* |
|---|---|---|---|---|
| *Agent* | Neutre/Colère | 450 | 70 (3.5) | 75 (2.5) |
| | Neutre/Négatif | 500 | 72 (2.5) | 73 (3) |
| *Appelant* | Neutre/Négatif | 2500 | 83 (1) | 83 (1) |
| | Neutre/Peur | 3000 | 82 (2) | 84 (1) |

**Figure 6-6.** *Comparaison des performances(RR score avec des ensembles équilibrés) de la détection Neutre/Négatif entre les agents et les appelants. Le nombre entre parenthèses est la déviation standard. Procédure de validation croisée avec N = 10 sous-ensembles et 10 exécutions.*

La Figure 6-6 ci-dessus illustre le cas de modèles à 2 classes. Les expériences ont été réalisées en début de thèse avec 50 paramètres. Elles avaient été effectuées avec une procédure de validation croisée N = 10 sous-ensembles. Les scores donnés dans le Tableau 6-7 sont des RR scores. Pour ces expériences, la répartition des données était équilibrée par classe. Le « rôle » des locuteurs semble avoir un impact sur les performances. Les appelants expriment plus clairement leurs émotions négatives que les agents (80 % vs 73 % de bonne détection), ce qui est tout à fait logique dans ce type de tâche.

Ces expériences effectuées sur 20h confirmaient des premières expériences réalisées sur un sous corpus de 10 heures [Devillers et Vidrascu 2006b] alors que toutes les données n'étaient pas transcrites

## *Hommes/Femmes*

Encore une fois à cause du nombre insuffisant de segments pour certaines émotions, les expériences ont été faites avec peu de classes et *Peur*, *Colère* et *Soulagement* pour la Figure 6-7 ci-dessous. Les expériences ont été faites en utilisant soit uniquement des hommes, soit uniquement des femmes pour l'apprentissage et en testant également séparément sur des hommes et des femmes.

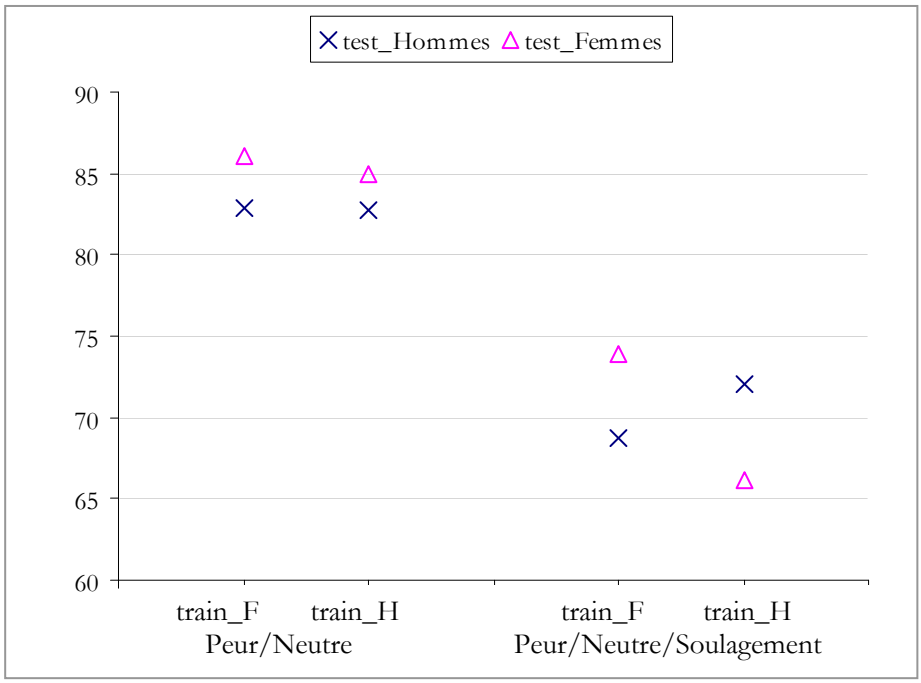


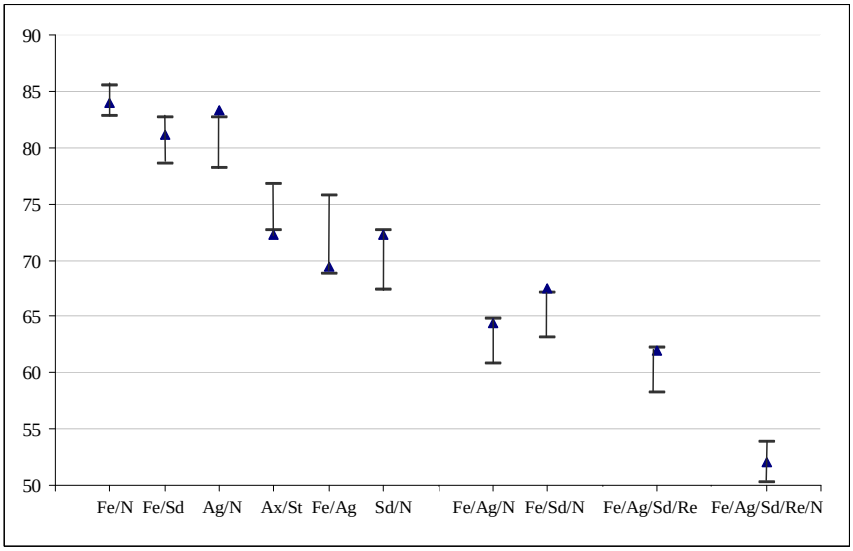

**Figure 6-7.** *Comparaison des performances pour des classifieurs entrainés seulement soit sur des hommes (train_H), soit sur des femmes (train_F).*

Dans le cas Peur/Neutre, les performances sont relativement comparables quelles que soient les données d'entraînement et de test. Ce n'est pas le cas pour la classification Peur/Neutre/Soulagement où les scores de bonne détection sont meilleurs lorsque le système a été

entraîné sur des locuteurs du même sexe. Il faudrait faire d'autres expériences pour vois si ces différences se manifestent pour des tâches de détection complexes (beaucoup de classes ou des classes très proches) ou pour certaines classes d'émotions spécifiques. Des études ont d'ailleurs montré des différences hommes/femmes dont l'expression des émotions, la colère étant par exemple plus passive chez les femmes [Fischer 1993], celles-ci pleurant plus que les hommes.

### 6.3.2. Variation du nombre de classes

Nous avons étudié la variation des performances lorsqu'on passe de 2 à 5 classes d'émotions à discriminer. Les résultats sont indiqués Figure 6-8. Sans surprise, les performances sont inversement proportionnelles au nombre de classes avec plus de 80% de bonne détection Peur/Neutre par exemple pour 2 classes et entre 50 et 55% de bonne détection avec les 5 classes Peur/Colère/Tristesse/Neutre/Surprise. De plus, elles sont meilleures pour des classes plus disjointes (Peur/Neutre mieux que Peur/Colère ou Anxiété/Stress), ce que nous avions déjà constaté en comparant les performances Peur/Surprise, Peur/Colère et Peur/Colère/Surprise [Devillers et Vidrascu 2006b].




| Classe | Apprentissage #Segments/#Locuteurs | Test #Segments /#Locuteurs |
|---|---|---|
| Neutre | 2000 (551 locuteurs) | 1448 (169 locuteurs) |
| Soulagement | 189 (122 locuteurs) | 108 (80 locuteurs) |
| Négatif | 500 (316 locuteurs) | 993 (164 locuteurs) |
| Tristesse | 250 (102 locuteurs) | 101 (40 locuteurs) |
| Colère | 180 (56 locuteurs) | 50 (24 locuteurs) |
| Peur | 2000 (555 locuteurs) | 808 (151 locuteurs) |
| Stress | 243 (138 locuteurs) | 83 (35 locuteurs) |
| Anxiété | 244 (180 locuteurs) | 243 (93 locuteurs) |
| Total | 5850 (678 locuteurs) | 4505 (209 locuteurs) |

**Figure 6-8.** *Résultats de classification en passant de 2 à 5 classes d'émotions ; Fe : Peur, N : Neutre, Sd : Tristesse, Ag : Colère, Ax : Anxiété, St : Stress, Re : Soulagement. Le nombre de segments distincts utilisés par émotion pour l'apprentissage et le test est indiqué dans le tableau. Les barres verticales indiquent la déviation standard des performances lorsque l'expérience est répétée 200 fois.*

### 6.3.3. Le poids des différents types d'attributs paralinguistiques : le cas de la détection dans le cas des 5 classes Peur/Colère/Tristesse/Soulagement/Neutre

Pour des tâches simples du type classification Négatif/Neutre, on obtient facilement des scores de détection de l'ordre de 80% en n'utilisant que très peu d'indices et en se limitant à une catégorie, voire sous catégorie de paramètres (par exemple seulement des indices déduits de la F0). Cependant, pour des tâches plus complexes (émotions moins distinctes, ou plus grand nombre de classes Emotion), il devient utile de mélanger des indices les plus variés possibles afin de tenir compte de la grande variabilité des expressions vocales dans le discours spontané.

Nous nous sommes intéressés à différentes catégories de paramètres pertinents pour la détection des émotions en nous intéressant particulièrement au cas de la classification en 5 classes émotions[Vidrascu et Devillers 2007]. Les paramètres ont été divisés en plusieurs types, similaires à ceux utilisés dans les études CEICES (voir p150) avec une distinction entre ceux qui peuvent être extraits automatiquement sans intervention humaine (« blind » : paramètres prosodiques, spectraux, microprosodie) et les autres (durées obtenues après alignement phonémique, paramètres extraits de la transcription) La liste des paramètres est résumée dans le Tableau 6-5.

| Type de paramètres | Description | # params |
|---|---|---|
| Paramètres prosodiques | F0 (normalisée par locuteurs)<br>position sur l'axe temporaire où F0 est maximum (resp. minimum)<br>Energie (normalisée)<br>Durées : débit, silences<br>ratio du nombre de segments voisés et non voisés | 45 |
| Paramètres spectraux | 3 premiers formants et leurs bandes passantes, F3-F2, F2-F1 (normalisés) | 48 |
| Microprosodie | jitter, shimmer, NHR, HNR | 14 |
| Indices et disfluences (transcription) | inspiration, expiration, bruits bouche, rires, pleurs, mots tronqués.<br>Disfluences : "euh" | 11 |
| Durées (alignement phonémique) | Longueur des phonemes<br>debit phonémique<br>taille des hésitations | 11 |

**Tableau 6-5.** *Les différents types d'indices extraits et leur nombre.*

## *Sélection des attributs*

Pour sélectionner les meilleurs attributs, un classifieur SVM a été utilisé et les résultats ont été comparés en utilisant les 15, 25, 40, 50, 70 et 80 meilleurs paramètres. Les meilleures performances étaient obtenues avec 25 paramètres. Cet ensemble peut encore contenir des attributs redondants et pourrait être encore optimisé.

| Type de paramètre | # parmi les 25 meilleurs |
|---|---|
| F0 | 4 |
| Energie | 5 |
| Microprosodie | 4 |
| Formants | 2 |
| Paramètres déduits de la transcription | 6 |
| Durées déduites de l'alignement phonémique | 4 |

**Tableau 6-6.** *Nombre de paramètres sélectionnés pour chaque classe de paramètres.*

Le Tableau 6-6 ci-dessus indique le nombre d'attributs sélectionnés par classe d'attributs. L'importance de combiner différentes classes est reflétée par le fait que des attributs de chaque classe ont été sélectionnés. Parmi les plus utiles, on trouve des marqueurs affectifs (pleurs, voix inintelligible), des disfluences (nombre et longueur des hésitations), des durées (débit (#mots/durée du signal ; #phonèmes/durée du signal, 1/durée moyenne des segments voisés)) et des paramètres de microprosodie (jitter, shimmer, HNR). Ceci s'explique aussi par le fait que ces paramètres sont peu présents dans le corpus bien qu'ils soient des marqueurs extrêmement utiles lorsqu'ils sont présents. Le traitement d'un grand nombre de paramètres est nécessaire même si pour certains ils sont plus exceptionnels (tout le monde ne pleure pas quand il est triste, mais si ce paramètre est présent on a une grande chance que la personne soit triste). Peu de paramètres liés aux formants ont été retenu.

## *Résultats avec les paramètres en mode "blind", c'est-à-dire sans aucune connaissance du contenu*

La Figure 6-9 indique les différentes performances en utilisant seulement les paramètres « blind » Les résultats sont au dessus du taux de hasard (20% avec 5 classes également distribuées) et les performances sont comparables pour F0, Energie et Formants. Le fait de les combiner améliore de manière significative les performances.

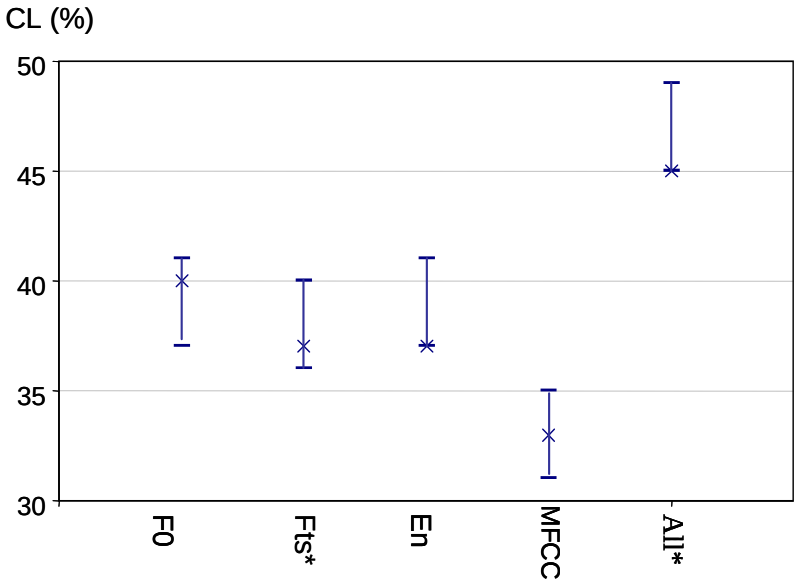



**Figure 6-9.** *Score CL (5 classes) avec F0 : seulement des paramètres relies à la F0, Fts: Formants et leur bande passante, En: énergie, MFCC, All* (107 paramètres) : tous les paramètres « blind ». Les barres verticales indiquent la déviation standard des résultats.*

Contrairement aux résultats obtenus avec des données de magicien d'Oz par [Vogt et Andre 2005], pour nos données téléphoniques, les MFCCs, même s'ils donnent des résultats au dessus du hasard ne sont pas aussi performants que les paramètres prosodiques ou les formants.

## *Indices « Blinds » vs indices semi-automatiques*

La Figure 6-10 montre l'impact des paramètres déduits de la transcription et de l'alignement phonémiques. Avec seulement 11 paramètres chacun, ils permettent d'obtenir d'assez bonnes performances avec à peu près 45% de bonne détection. Le mélange des indices (129 au total) augmente de manière significative le score CL. Il n'y a pas de différence significative entre les performances avec tous les attributs et les 25 meilleurs.

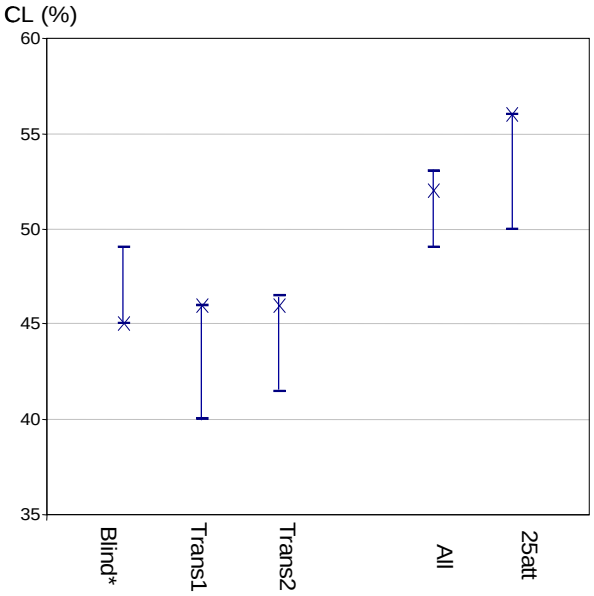



**Figure 6-10.** *CL score pour 5 classes Peur Colère Tristesse Soulagement et Neutre avec différents ensembles d'indices. Blind : extraits automatiquement (F0, formants, énergie, prosodie), correspond au All* de la figure précédente, trans1 : indices extraits de la transcription manuelle ; trans2: durées phonémiques, 25 best: 25 meilleurs paramètres.*

## *Résultats par émotion*

Intéressons nous maintenant aux performances par émotion, toujours pour le cas de la classification des 5 classes Peur, Colère, Tristesse, Neutre, Soulagement. La plupart des confusions ont eu lieu entre Tristesse/Neutre, Soulagement/Neutre et Peur/Colère. Si on regarde pour chaque type d'indice (« blind », paramètres extraits de la transcription et durées dérivées de l'alignement phonémique) les taux CL de reconnaissance pour chaque émotion (Figure 6-11), on remarque qu'un type de paramètre sera meilleur pour une émotion spécifique. Dans l'exemple de la Figure 6-11, le score de détection de la peur est de 40% environ avec juste des indices de durées, mais de plus de 50% avec les indices « blinds ». Pour le soulagement, c'est l'inverse avec moins de 50% de reconnaissance avec juste les indices « blinds », et plus de 60% avec ceux issus de la transcription. Les performances avec les 25 meilleurs paramètres sont encore globalement supérieures à celles par type d'indice. C'est particulièrement le cas pour la tristesse avec moins de 40% de bonne détection pour chaque type d'indice et près de 60% lorsque les indices sont combinés.

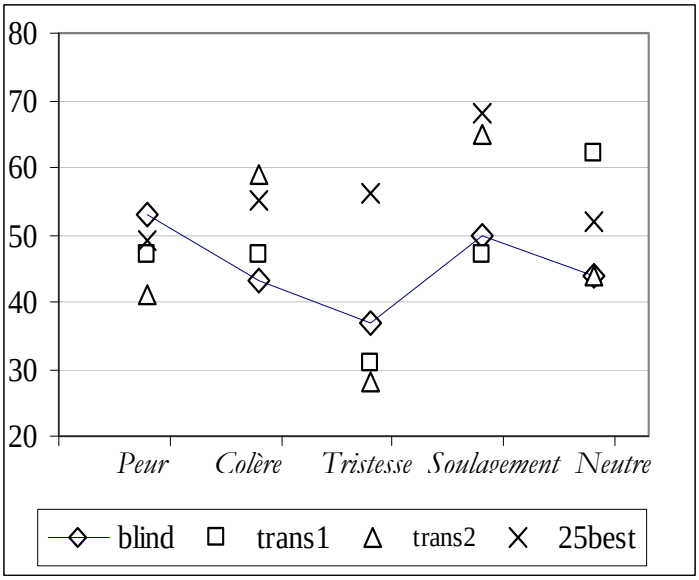



**Figure 6-11.** *CL score par émotion (Peur, Colère, Tristesse, Soulagement + état Neutre) pour les paramètres « blind » vs. paramètres déduits de la transcription (trans1) vs. paramètres déduits de l'alignement phonémique (trans2) vs. 25meilleurs paramètres.*

La bonne reconnaissance du Soulagement avec les indices déduits de l'alignement phonémique peut être expliquée par le fait qu'il y a moins d'hésitations et d'allongement phonémiques que pour les autres émotions. Pour l'état Neutre, il y a peu de marqueurs affectifs et d'hésitations par rapport aux émotions et le débit (nombre de mots/durée du signal) est plus lent que pour la parole

émotionnelle, ce qui pourrait expliquer les bonnes performances en utilisant uniquement les indices déduits de la transcription manuelle.

La Figure 6-12 est donnée à titre illustratif afin de montrer le poids des différents types d'indices.

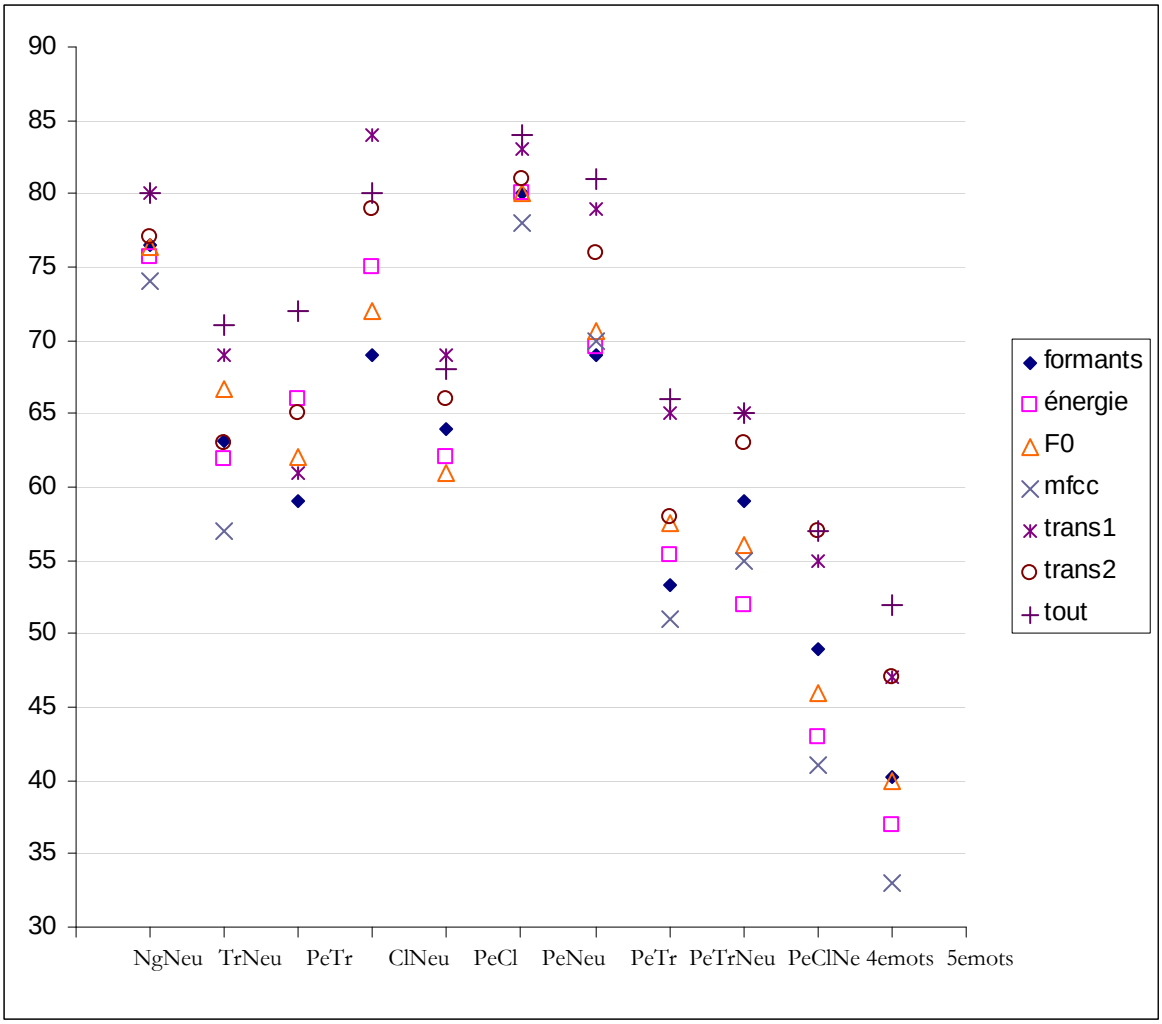



**Figure 6-12.** *Performances pour différentes tâches de classification en n'utilisant qu'un seul type d'indice (formant : formants et leur bande passante, F0 :F0 et durées, trans1 : indices extraits de la transcription, trans2 : indices extraits de l'alignement phonémique). Ng : Négatif ; Neu : Neutre ;Pe :Peur ; Cl :colère ;Tr :Tristesse ; 4emots :Peur/Colère/Tristesse/Neutre ; 5emots :Peur/Colère/Tristesse/Soulagement/Neutre.*

### 6.3.4. Combinaison indices lexicaux et prosodiques

Dans le paragraphe précédent, nous nous sommes intéressés aux apports des différents types d'indices paralinguistiques. Les informations obtenues grâce à ces différents indices peuvent être enrichies par des informations lexicales. Deux expériences ont été faites avec un système de détection des émotions basé sur un modèle unigramme développé au LIMSI. La première, sur le corpus de transactions boursières combinait les prédictions lexicales et acoustiques pour les 2 émotions Neutre/Négatif [Devillers et al. 2005b]. La deuxième sur le corpus CEMO comparait les performances pour 4 classes des 2 modèles [Devillers et Vidrascu 2006a].

#### *Description du modèle lexical*

Le système de détection des émotions basé sur un modèle uni-gramme est détaillé dans [Devillers et al. 2003a]. Le modèle lexical est un uni-gramme, où la similarité d'une phrase et d'une émotion est le log du ratio de la probabilité entre un modèle spécifique à une émotion et un modèle général spécifique à la tâche (équation. 1).

(Equation. 1)
$$\log P(u/E) = \frac{1}{L_u} \sum_{w \in u} tf(w,u) \log \frac{\lambda P(w/E) + (1-\lambda) P(w)}{P(w)}$$

L'émotion d'une phrase inconnue est déterminée par le modèle obtenant le score le plus haut pour la phrase u étant donné le modèle d'émotions E basé sur les N émotions étiquetées dans le corpus ; où P(w/E) est la probabilité maximale estimée de la probabilité d'un mot w étant donné le modèle d'émotions, P(w) est la probabilité générale dépendant de la tâche du mot w dans le corpus d'entraînement, tf(w,u) sont les fréquences des termes dans la phrase inconnue u, et Lu est la longueur de la phrase en mots. Le modèle général a été estimé sur tout le corpus d'entraînement. Les scores de détection augmentent de manière significative lorsqu'on considère deux classes principales d'émotions, Positives vs Négatives.

#### *Combinaison linéaire entre les modèles lexicaux et prosodiques pour les données boursières*

Les données ont été divisées en 10 sous-ensembles de 50 tours de parole. Neuf étaient utilisés pour l'apprentissage et 1 pour le test ; l'expérience était répétée pour chaque sous ensemble [Vidrascu et Devillers 2005a].

Un score de prédiction par émotion était obtenu avec le modèle lexical uni-gramme et un autre avec un arbre de décision (AdTree, 39 indices extraits). Le pourcentage de bonne détection est de 71% à peu près avec le modèle lexical et avec le modèle prosodique.

Pour chaque ensemble de test, les prédictions par émotions avec les 2 modèles ont été combinées linéairement et les résultats sont donnés Figure 6-13. Le score moyen de reconnaissance après mélange est de 76,6%, soit un gain de plus de 5%. Ce résultat, même s'il n'est pas généralisable car obtenu sur peu de données, va dans le sens d'autres expériences [Forbes-Riley et Litman 2004] [Narayanan 2002], et montre que le lexical apporte de nouvelles informations utiles pour la détection des émotions.

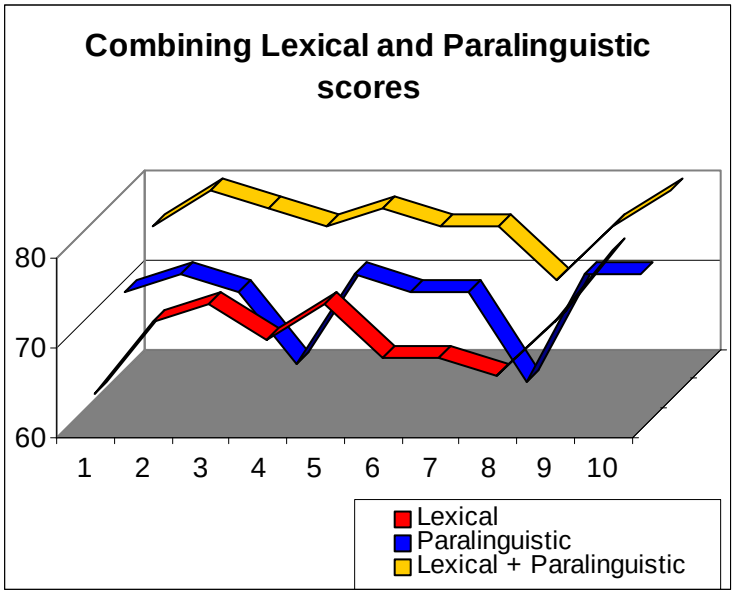



**Figure 6-13.** *Combinaison des scores lexicaux et prosodiques.*

## *Expériences sur le corpus CEMO*

Une expérience à également été menée sur le corpus CEMO afin de comparer les performances entre un modèle lexical et un modèle paralinguistique pour les 4 classes Peur, Colère, Tristesse, Soulagement. Le corpus utilisé pour cette expérience est décrit ci-dessus dans le Tableau 6-7. Les locuteurs du test sont différents de ceux du corpus d'apprentissage.

| Corpus | Apprentissage | Test |
|---|---|---|
| #Segments | 1618 | 640 |
| #Locuteurs | 501(182 H, 319F) | 179(60H, 119F) |
| Colère | 179 | 49 |
| Peur | 1084 | 384 |
| Soulagement | 160 | 107 |
| Tristesse | 195 | 100 |

**Tableau 6-7.** *Sous-corpus utilisé pour les tests avec un modèle lexical et paralinguistique.*

Les scores obtenus avec le modèle lexical et avec le modèle paralinguistique ont été comparés. Avec le modèle lexical après normalisation, les quatre émotions sont détectées avec environ 67,2% de bonne détection. Le Tableau 6-8 compare les scores de détection obtenus par classe avec le modèle lexical et le modèle paralinguistique.

| | Total | Colère | Peur | Soulagement | Tristesse |
|---|---|---|---|---|---|
| #Segments | *640* | *49* | *384* | *107* | *100* |
| % rec. modèle lexical | 78 | 59 | 90 | 86 | 34 |
| % rec. modèle acoustique | 61 | 43 | 58 | 71 | 68 |

**Tableau 6-8.** *Répartition pour les 4 classes avec les modèles lexicaux et prosodiques.*

Avec le modèle lexical, le meilleur score est obtenu pour la classe Peur et le pire pour la classe Tristesse. Le score élevé obtenu pour Soulagement est lié aux marqueurs lexicaux spécifiques de cette classe (tels que merci, d'accord). A l'inverse, la tristesse serait ici plus liée à des marqueurs syntaxiques et prosodiques que lexicaux. Les principales confusions ont lieu entre Peur et Tristesse d'une part et Peur et Colère d'autre part.

Comme pour le modèle lexical, la classe la mieux reconnue avec le modèle paralinguistique est la peur (64%) et la pire est la colère (39%), mais le score reste au dessus de la chance. Cela peut être du au fait que la Peur (Inquiétude/Stress) est souvent en arrière plan de tous les appels et que la colère est souvent mélangée. L'étape suivante serait de combiner plus astucieusement qu'avec une combinaison linéaire le résultat des deux modélisations.

## 6.4. Utilisation de nos méthodes sur des données différentes : CEICES (Combining Efforts for Improving Classification of Emotional user State)

### 6.4.1. Coopération dans le cadre du réseau d'excellence humaine

CEICES est une collaboration entre plusieurs équipes impliquées dans le réseau Humaine étudiant la classification des états émotionnels transmis par la voix : UKA-US, Université d'Erlangen-Allemagne, ITC-Italie, TAU-Israël, Université d'Augsburg-Allemagne et LIMSI-CNRS-France. Cette collaboration est née d'une volonté d'améliorer les performances de classification des états émotionnels pour des données naturelles et de répondre à plusieurs problématiques en partageant les compétences des différents sites.

- Les performances de classification pour des données naturelles sont beaucoup plus faibles que pour des données actées parce que la tâche est plus difficile et à cause de la difficulté à annoter de manière fiable, à extraire les paramètres appropriés et.
- De plus, comme indiqué dans l'état de l'art, il est souvent difficile de comparer les performances de différentes expériences et les paramètres extraits peuvent être assez obscurs[1].

Le site d'Erlangen, à l'initiative du projet, a fourni les fichiers audio et leur transcription manuelle, ainsi qu'une annotation par mot, par tout de parole, et par « cluster » émotionnel et une correction manuelle de la F0. Les différents sites ont dans un premier temps comparé les paramètres extraits et les différentes méthodes de classification en utilisant les mêmes ensembles d'apprentissage et de test. Ils se sont également réunis pour réfléchir à une dénomination plus explicite des paramètres. Des expériences ont également été menées en particulier en comparant les performances avec la F0 manuellement corrigée ou non et les impacts respectifs des différents types de paramètres. Cette coopération a conduit à plusieurs publications.

### 6.4.2. Le corpus AIBO

Le corpus est composé d'interactions en allemand entre des enfants de 11-12 ans et le chien robot AIBO de Sony (51 interactions, 9,2 heures de parole, 51393 mots). Il était demandé aux enfants de faire accomplir un parcours au robot en lui parlant comme ils parleraient à un ami. Ils pensaient

[1] Par exemple, on peut s'attendre à avoir de meilleurs résultats avec des paramètres calculés manuellement que automatiquement.

que le robot leur répondait, alors qu'il était en fait contrôlé par un opérateur humain. AIBO pouvait ainsi désobéir et provoquer des réactions émotionnelles du type colère.

Le corpus a été annoté au niveau du mot par 5 étudiants en linguistique avec un choix de 11 étiquettes « émotion ». A cause de l'insuffisance des données pour certaines classes, un sous corpus a été conservé avec les étiquettes *Motherese* (valence positive, ton maternel), Neutral (neutre : classe par défaut), Emphatic (insistance, situation « pré-négative ») Angry (colère) (Tableau 6-11).

Le corpus peut être étudié à plusieurs niveaux et en particulier le mot, la phrase et le « chunk » (les règles syntactiques et prosodiques pour le découpage sont détaillées dans [Batliner et al. 2007])

| | |
|---|---|
| Motherese | 586 |
| Neutral | 1998 |
| Emphatic | 1045 |
| Angry | 914 |

**Tableau 6-9.** *Fréquence des « émotions » dans le corpus AIBO pour le découpage en chunks.*

### 6.4.3. Schéma d'encodage des paramètres.

Un workshop a été organisé à Erlangen en décembre 2006 afin de se mettre d'accord sur des descripteurs de paramètres les plus complets possibles et a abouti à un schéma avec N champs codés sur des dizaines de bits et donnant des informations diverses comme le type de micro utilisé, les unités sur lesquels un paramètre est calculé, son type (linguistique, prosodique, …), si le paramètre est extrait de manière automatique ou manuelle et les différentes fonctions appliqués à un paramètre.

Par exemple pour le coefficient de régression de la F0, le champ F0 est mis à 1 et différents codes indiquent que le coefficient de régression est calculé pour chaque partie voisée puis que au niveau du chunk, on garde le minimum.

Un exemple de codage de paramètres LIMSI est donné Figure 6-14.

```
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F000200.N00.X1000000000.T0000000000PMaxF0
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F000500.N00.X1000000000.T0000000000PRangeF0
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F002100.N00.X1000000000.T0000000000PMedianF0
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F001000.N00.X1000000000.T0000000000PMeanF0
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F003000.N00.X1000000000.T0000000000PSdF0
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F002000.N00.X1000000000.T0000000000PFirstQuartileF0
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F002200.N00.X1000000000.T0000000000PThirdQuartileF0
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C1000010000.F016301.N00.X1000000000.T0000000000PCoeffF0min
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F016302.N00.X1000000000.T0000000000PCoeffF0max
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F016314.N00.X1000000000.T0000000000PCoeffF0mean
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F016101.N00.X1000000000.T0000000000PMinSlopeF0
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F016102.N00.X1000000000.T0000000000PMaxSlopeF0
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F016114.N00.X1000000000.T0000000000PMeanSlopeF0
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F016002.N00.X1000000000.T0000000000PF0MseRegMax
S2.I.M1.D3.R5111.L000000.A00.00.10.00.00.00.00.00.00.00.C0000010000.F016014.N00.X1000000000.T0000000000PF0MseRegMean
S2.I.M1.D3.R5111.L000000.A10.00.00.00.00.00.00.00.00.00.C1000010000.F003410.N00.X1000000000.T0000000000PSpeakingRate
S2.I.M1.D3.R5111.L000000.A10.00.00.00.00.00.00.00.00.00.C1100010000.F003447.N00.X1000000000.T0000000000PPercentVoicedUnvoiced
```

**Figure 6-14.** *Exemple de codage de paramètres LIMSI à l'issu du workshop à Erlangen.*

Malgré l'aspect complexe de ce protocole, il permet d'isoler facilement des types de paramètres afin de pouvoir les comparer.

### 6.4.4. Comparaison des performances par site

La première expérience a été de comparer les performances de chaque site en utilisant les mêmes ensembles d'apprentissage et de test. Le type et le nombre des paramètres utilisés ainsi que les techniques d'apprentissage étaient libres. Les types de paramètre/classifieur/performances par site sont donnés Tableau 6-10. Les résultats obtenus par les différents sites sont très proches (entre 54 et 57% de bonne détection).

| Site | *# paramètres* | | *# par type de paramètre* | | | | | | *domaine* | | *Classification* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | original (40204) | selection (381) | prosodic | spectral | MFCC | POS | lexical | genetic | Tour | mot | Classifier | RR | CL | ROVER |
| FAU | 303 | 87 | 19 | - | - | 6 | 62 | - | √ | √ | NN | 55.8 | 55.3 | √ |
| TUM | 980 | 103 | 9 | 17 | 22 | 2 | 50 | 3 | √ | - | SVM | 59.3 | 56.4 | √ |
| ITC | 32 | 32 | 26 | - | - | 6 | - | - | √ | √ | Random Forest (RF) | 57.6 | 55.8 | √ |
| UKA | 1320 | 25 | 6 | - | 5 | - | 14 | - | √ | - | Régression linéaire | 59.1 | 54.8 | - |
| UA | 1289 | 84 | 10 | 1 | 73 | - | - | - | √ | - | Naive Bayes | 50.9 | 52.3 | √ |
| **LIMSI** | **76** | **26** | **9** | **9** | **-** | **5** | **3** | **-** | √ | **-** | **SVM** | **54.9** | **56.6** | √ |
| TAU | 24 | 24 | 24 | - | - | - | - | - | √ | - | Rule-based | 48.9 | 46.6 | – |

**Tableau 6-10.** *Paramètres et classifieurs : par site, # de paramètres avant/après la sélection des attributs ; # par type de paramètres, et par domaine; classifieur utilisé, RR et CL scores, utilisé ou non pour le ROVER ; de [Batliner et al. 2006]*

En mélangeant les meilleurs paramètres de chaque site et en re-sélectionnant les meilleurs d'entre eux (Tableau 6-11), les performances ont été améliorées ; chaque site contribuant à cette amélioration. Lorsque les classifications sont combinées par ROVER, les scores CL et RR atteignent les 62% (voir [Schuller et al. 2007a] pour plus de détails)

| Classifieur | RR | CL |
|---|---|---|
| LDA | 58.8 | 56.3 |
| SVM | 61.8 | 57.9 |
| RF | 60.8 | 58.7 |

**Tableau 6-11.** *Classification en combinant les meilleurs paramètres parmi les 381 de tous les sites avec 3 classifieurs.*

### 6.4.5. Impact des erreurs d'extraction du pitch

La F0 a été manuellement corrigée pour les données AIBO. Des expériences ont été faite avec et sans F0 corrigée [Batliner et al. 2007] et bien que les résultats soient meilleurs avec la F0 manuellement corrigée, les différences sont peu prononcées.

## 6.4.6. Impact de différents types de paramètres

Les paramètres ont été séparés en différents types, tous sites confondus (*cf.*[Schuller et al. 2007a]). Les performances respectives des différents types de paramètres ont été évaluées et sont données Tableau 6-12**.**

| **feature set** | | **full** | | **reduced** | |
|---|---|---|---|---|---|
| type | # | $F_{SVM}$ | $F_{RF}$ | $F_{SVM}$ | $F_{RF}$ |
| voice quality | 153 | 51.5 | 51.1 | 51.6 | 50.8 |
| F0 | 333 | 56.1 | 56.6 | 55.1 | 55.1 |
| spectral/formants | 656 | 54.4 | 57.1 | 56.0 | 56.6 |
| cepstral | 1699 | 52.7 | 55.7 | 57.1 | 56.3 |
| wavelets | 216 | 56.0 | 56.5 | 56.3 | 56.7 |
| energy | 265 | 58.5 | 59.3 | 60.0 | 60.0 |
| duration | 391 | 55.1 | 60.1 | 60.0 | 59.8 |
| all acoustic | 3713 | 57.7 | 62.5 | 61.2 | 60.9 |
| disfluencies | 4 | 26.8 | 25.2 | – | – |
| non-verbals | 8 | 24.8 | 24.2 | – | – |
| part of speech[77] | 31 | 54.7 | 54.1 | – | – |
| higher semantics | 12 | 57.6 | 57.7 | – | – |
| bag of words | 476 | 62.6 | 60.2 | 62.6 | 58.6 |
| all linguistic | 531 | 62.6 | 60.2 | 62.4 | 59.0 |
| all | 4244 | 61.0 | 64.0 | 63.1 | 61.7 |

**Tableau 6-12.** *Résultats de la classification, # : nombre de paramètre par type d'attributs ; F-scores pour tous les paramètres (full) ou un ensemble avec un nombre réduit de paramètres ( reduced) en utilisant SVM ou random forrest ( RF)[Schuller et al. 2007a].*

Dans cette étude, le paramètre acoustique le plus important s'est avéré être l'énergie et le moins important la qualité vocale. Les paramètres lexicaux avaient un impact très important et en particulier les « Part of Speech » (pour les données AIBO, 6 classes comprenant: nom, verbe, auxiliaire … et décrites en détail dans [Batliner et al. 1998] ).

[77]Part of speech » : catégorie lexicale

### 6.4.7. Conclusions générales sur les données AIBO

Les données émotionnelles semblent moins sensibles au bruit que d'autres tâches liées au traitement de la parole. Les erreurs de détection des algorithmes de détection du pitch font peu baisser les performances. L'énergie est pour cette tâche le paramètre acoustique le plus important et la qualité vocale le moins, ce qui ne sera pas forcément vrai pour d'autres données. Les paramètres linguistiques ont un impact très important, mais il faudrait encore les comparer à la sortie de la reconnaissance automatique de la parole, ce qui devrait amener de nombreuses autres erreurs. Les performances sont meilleures lorsqu'on combine plusieurs types d'attributs. Elles sont également meilleures pour le découpage en « chunks ». Les meilleures performances obtenues pour la détection de 4 classes sont supérieures à 60%, ce qui est cohérent avec l'état de l'art.

Cependant il reste à répéter ces expériences sur d'autres données. Par ailleurs, il faut se méfier des performances de classification qui peuvent être « tuned » par des réglages transparents pour un lecteur « normal » (découpage de l'unité émotionnel, ensemble de prototypes ou de données actées …)

## 6.5. Portabilité sur des données différentes

Nous avons vu que notre méthodologie d'extraction des paramètres et de classification était efficace pour traiter des données différentes, indépendamment de la manière dont les données ont été enregistrées et pour différentes langues : français et allemand.

Une deuxième question qui se pose est celle de la portabilité des « modèles ». Quelles sont les performances des modèles entraînés sur les données CEMO sur d'autres données ? Deux expériences ont été réalisées, l'une sur le corpus 1 de transaction boursière (call center, données téléphoniques avec situation et comportements différents, même langue) et une autre sur des données actées, en français, mais pas en interaction.

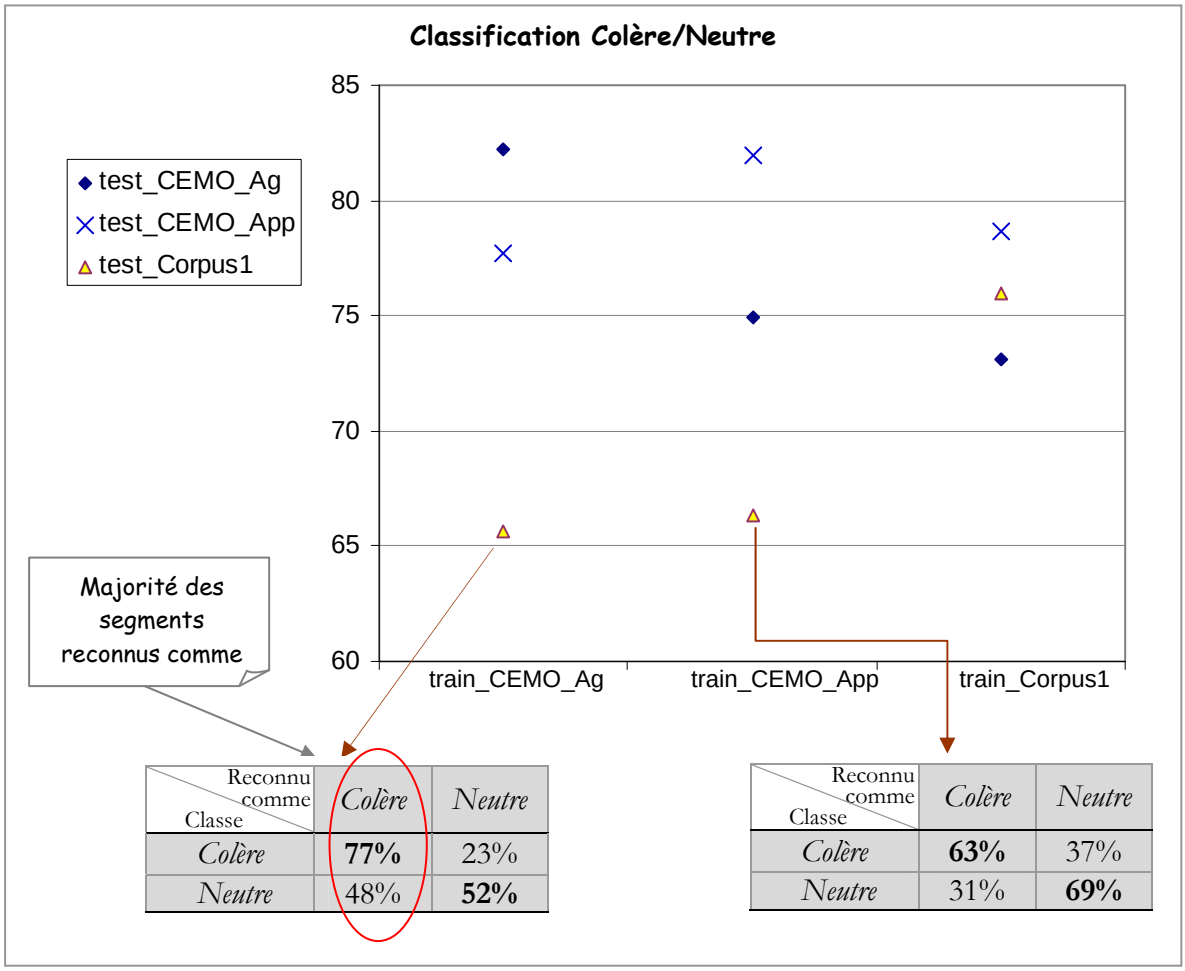
Parmi les paramètres extraits pour les données de CEMO, certains n'ont pas été calculés sur d'autres données (alignement phonémique….).

### 6.5.1. Sur les données boursières

Les données boursières sont comparables aux données CEMO en ce qu'elles proviennent d'interactions téléphoniques. Les expériences ont été réalisées avec des SVMs et 116 attributs extraits par segments (F0, formants, énergie, microprosodie, marqueurs affectifs).

#### *Tâche "simple" Colère/Neutre*

Une expérience à d'abord été effectuée pour la classification Neutre/Colère. Un ensemble d'apprentissage et un ensemble de test avec des locuteurs distincts ont été constitués pour respectivement les agents du corpus CEMO, les appelants du corpus CEMO et les appelants du corpus de données boursières. Les classifieurs entraînés sur chaque ensemble d'apprentissage ont été testés sur chacun des ensembles de test. Les résultats sont représentés Figure 6-15.




| Classe \ Reconnu comme | *Colère* | *Neutre* |
|---|---|---|
| *Colère* | **77%** | 23% |
| *Neutre* | 48% | **52%** |

| Classe \ Reconnu comme | *Colère* | *Neutre* |
|---|---|---|
| *Colère* | **63%** | 37% |
| *Neutre* | 31% | **69%** |

**Figure 6-15.** *Score CL pour la classification Colère/Neutre avec différentes données en apprentissage et en test : train_CEMO_Ag : agents du corpus CEMO, train_CEMO_App : appelants du corpus CEMO, train_Corpus1 : appelants du corpus de données boursières. Les matrices de confusion sont données en grisé pour les cas où le corpus 1 est en test avec les agents ou appelants de CEMO en apprentissage.*

On voit sur la Figure 6-15 que lorsque l'apprentissage et le test sont effectués sur les mêmes données, plus de 80% de bonne reconnaissance est obtenue sur les données CEMO et environ 75% sur les données boursières. Comme nous l'avions vu avec nos premières expériences (p 137), la colère des agents et des appelants dans CEMO s'exprime différemment, non seulement en intensité, mais aussi en qualité (colère froide contre colère chaude). Ainsi la colère des agents est mieux reconnue par un classifieur entraîné sur des agents que par un classifueur entraîné sur des appelants et inversement pour les appelants. Il semblerait que la colère des appelants CEMO soit mieux reconnue avec un « modèle agent » que celle des agents avec un « modèle appelant ». Toutefois, les scores sont supérieurs à 75% de bonne reconnaissance dans touts les cas. Par contre, les scores de reconnaissance pour la colère des appelants des données boursières avec les « modèles CEMO » sont assez bas (de l'ordre de 65%), même s'ils sont supérieurs au hasard. On peut remarquer dans l'expérience representée Figure 6-15 qu'avec le modèle « Agent CEMO », la plupart des segments des données boursières sont reconnus comme de la colère alors que pour le modèle « Appelant CEMO », on est plus proche du point d'égale erreur, ce qui semblerait indiquer que le modèle « appelant CEMO » est meilleur que celui « agent CEMO » pour reconnaître la colère des « appelants données boursières ». D'ailleurs, si on fait l'expérience inverse en entraînant un système sur les données boursières et en testant sur les agents et appelants CEMO, la colère des appelants CEMO est bien mieux reconnue que celle des agents CEMO. Qui plus est, elle est même mieux reconnue que l'ensemble de test des données boursières. On peut noter également que dans l'expérience, la colère des appelants CEMO est mieux reconnue par un modèle « Appelant données boursières » que par un modèle « agent CEMO », malgré les différences certaines dans l'enregistrement des données.

Nous avons voulu vérifier que ces tendances observées sur les appelants CEMO et appelants de données boursières s'observaient également pour d'autres tâches (nombre d'émotions et classes d'émotions différentes). Les résultats sont données Figure 6-16 avec la détection Peur/Neutre et Peur/Colère/Neutre pour les appelants CEMO ou BOURSE. Là encore, pour les classifieurs entraînés sur les données CEMO, les scores sont bien meilleurs sur les données CEMO que sur les données boursières, bien que supérieures au taux du hasard et pour les classifieurs entraînés sur les données boursières, les scores sont à peu près identiques pour les données boursières et les données CEMO. Dans tous les cas, les performances sont meilleures quand l'apprentissage et le test sont réalisés sur les mêmes données.

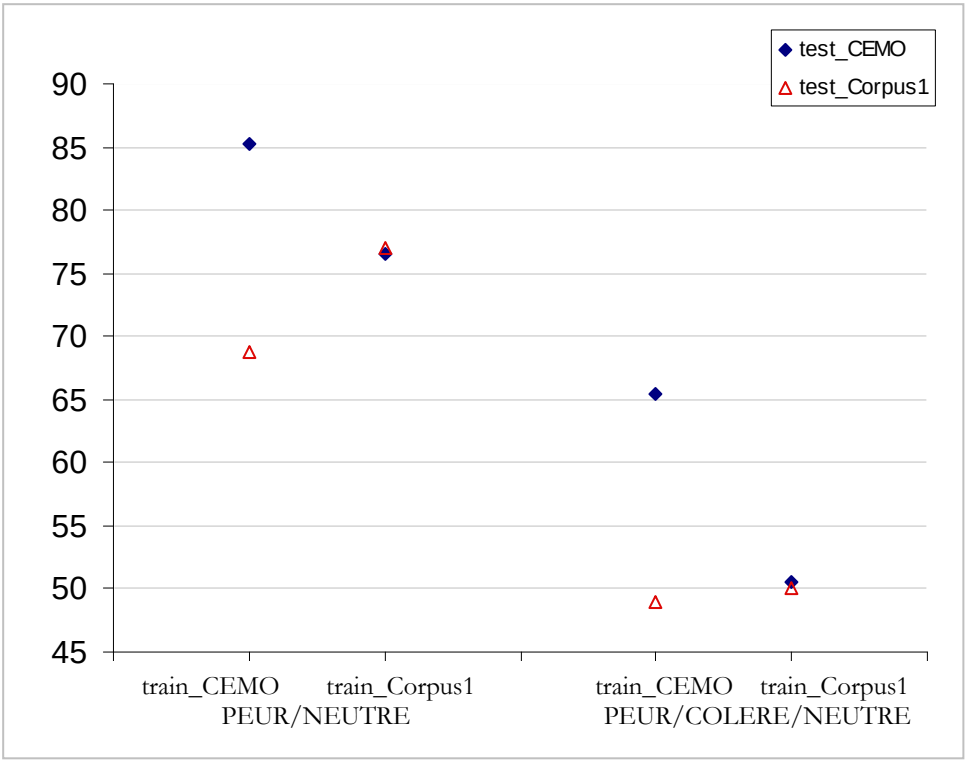



**Figure 6-16.** *Score CL pour la classification Peur/Neutre et Peur/colère/Neutre avec des classifieurs entraînés et testés sur les appelants CEMO ou sur les appelants du corpus de données boursières : train_CEMO : appelants du corpus CEMO, train_Corpus1 : appelants du corpus de données boursières.*

Ces expériences semblent indiquer qu'il est tout à fait envisageable d'utiliser un classifieur sur des données issues d'application différentes que les données ayant servies à l'entraîner. Il faut toutefois être alors très précis sur la définition des émotions que l'on cherche à reconnaître et prendre en compte également leur intensité. On aura de meilleures performances en utilisant un classifieur entraîné sur des données moins intenses que plus intenses. Cependant, même en ne tenant pas compte de toutes ces variations (intensité, définition des émotions, etc.), les scores restent supérieurs au niveau du hasard.

### 6.5.2. GEMEP (GEneva Multimodal Emotion Portrayals)

Nous nous intéressions à la portabilité des résultats entre données actées et naturelles. Avec cette expérience issue d'une collaboration entre le NCCR in Affective Sciences (UNIGE) et le LIMSI-CNRS au sein de HUMAINE, nous voulions tester nos « modèles » obtenus à partir de données naturelles sur des données actées.

#### *Description des données*

Les données sont décrites dans [Bänziger et Scherer. 2007]. Il a été demandé à 10 comédiens professionnels (5 hommes et 5 femmes) de jouer 18 émotions (admiration, amusement, attendrissement, colère chaude, dégoût, désespoir, fierté, honte, inquiétude, intérêt, irritation, joie exaltée, mépris, peur panique, plaisir, soulagement, surprise, tristesse) (Tableau 6-13) dans différentes conditions et avec différents degrés d'intensité.

| Emotion | Définition |
|---|---|
| admiration | émerveillement devant les qualités extraordinaires d'un personnage, d'un paysage ou d'une œuvre d'art |
| amusement | hilarité débordante face à quelque chose d'extrêmement drôle |
| attendrissement | être ému par un comportement touchant |
| colère chaude | mécontentement violent causé par l'action stupide ou malveillante de quelqu'un |
| dégoût | répulsion face à un objet ou un environnement répugnant |
| désespoir | Détresse face à un problème existentiel sans issue, couplé d'un refus d'accepter la situation |
| fierté | réaction triomphante suite à une réussite ou une performance personnelle (ou celle d'une personne proche) |
| honte | Amour-propre mis en cause par une maladresse ou une erreur dont on se sent responsable |
| inquiétude | crainte des conséquences (d'une situation) qui pourraient être potentiellement néfastes pour moi ou mes proches |
| intérêt | être attiré, fasciné, ou captivé par quelqu'un ou quelque chose |
| irritation | être confronté à quelque chose ou à quelqu'un qui me contrarie fortement, sans me faire perdre totalement mon sang-froid |
| joie exaltée | être transporté par une chose magnifique qui nous arrive de manière inattendue |
| mépris | Aversion pour le comportement moralement ou socialement répréhensible d'autrui |
| peur panique | Se sentir menacé par un danger imminent mettant en cause la survie ou l'intégrité physique |
| soulagement | Se sentir réconforté et rassuré suite à la résolution ou la fin d'une situation inconfortable, désagréable ou même dangereuse |
| surprise | être confronté, souvent de manière brusque, à un événement inattendu et insolite (sans connotation positive ou négative) |
| tristesse | Se sentir déprimé et découragé par la perte irrévocable d'un être proche ou d'un objet ou d'un environnement familier |

**Tableau 6-13.** *Définition des émotions exprimées dans GEMEP.*

Les différentes conditions incluaient :

- deux phrases pseudo-linguistiques (sans contenu linguistique) élaborées à l'aide d'un phonéticien, la première réalisée comme une affirmation, la seconde comme une question
- une expression prononcée sur une voyelle soutenue « A »
- de la parole improvisée avec un contenu lexical libre. Le projet GEMEP étant en cours, la qualité expressive des séquences n'avait pas été évaluée au moment de l'expérience (pas de test perceptif pour vérifier que les émotions jouées par les acteurs étaient bien reconnues).

Dans un premier temps, nous avons sélectionné les étiquettes comparables à celles du corpus CEMO. Elles correspondent aux grandes classes *Peur* (inquiétude peur-panique), *Colère* (irritation, colère chaude), *Tristesse* (tristesse/désespoir) et *Soulagement* (soulagement)[78]. Les caractéristiques des données que nous avons utilisées sont indiquées Tableau 6-14.

| Emotion | Mode | Contenu |
|---|---|---|
| Inquiétude<br>Peur-panique<br>Irritation<br>Colère chaude<br>Tristesse<br>Désespoir<br>Soulagement | Normal<br>Moins intense<br>Plus intense<br>Masqué | Phrase 1 : Né kal ibam soud molèn !<br>Phrase 2 : Koun sé mina lod bélam ?<br>Jouer de manière naturelle avec contenu verbal libre<br>Expression vocale basée sur une voyelle soutenue « AAA » |

**Tableau 6-14.** *Les données GEMEP (5 hommes/5 femmes).*

En utilisant les mêmes outils que pour les données CEMO, nous avons extraits avec Praat les indices acoustiques[79] (F0, formants, microprosodie) pour chaque segment GEMEP.

Nous avons sélectionnées les données correspondant aux deux phrases et au contenu verbal libre et les avons testées sur les classifieurs Peur/Colère/Tristesse/Soulagement et Peur/Colère entraînés sur le corpus CEMO avec uniquement des indices acoustiques. (Il n'a pas été demandé aux acteurs de jouer un état neutre, bien que l'intérêt dans GEMEP corresponde à un état affectif d'intensité relativement faible qui pourrait s'en rapprocher. Cependant, comme la valence est positive et que nous avons une étiquette Intérêt dans CEMO, nous ne pouvions pas assimiler la classe Intérêt de GEMEP au Neutre de CEMO et n'avons donc pas test avec des classifieurs Neutre/Négatif).

[78] Nous n'avons pas regardé la surprise et l'intérêt car ces émotions ont été peu étudiées dans CEMO.

[79] L'énergie dépendant des conditions d'enregistrement, nous ne l'avons pas extraite pour les expériences décrites ici. Les mêmes expériences avec extraction de l'énergie en plus donnaient à peu près les mêmes résultats.

## *Classification Peur/Colère/Tristesse/Soulagement*

**Apprentissage sur CEMO/Test sur GEMEP**

Sur les données CEMO (voix d'appelants) testées avec des locuteurs différents de ceux utilisés par l'apprentissage (*cf.* matrice de confusion Tableau 6-15), les performances sans indices lexicaux n'étaient pas très élevées bien que supérieures au niveau du hasard (taux de reconnaissance CL de l'ordre de 51% *cf.* Tableau 6-15)

| | Peur | Colère | Tristesse | Soulagement |
|---|---|---|---|---|
| Peur ( 808 sgts) | **58** | 21 | 12 | 9 |
| Colère (79 sgts) | 19 | **44** | 19 | 18 |
| Tristesse (105 sgts) | 20 | 17 | **49** | 14 |
| Soulagement (106 sgts) | 5 | 16 | 25 | **55** |

**Tableau 6-15.** *Matrice de confusion pour le classifieur Peur/Colère/Tristesse/Soulagement (avec uniquement des indices acoustiques) pour des segments du corpus CEMO(appelants) en apprentissage et en test avec des locuteurs différents de ceux utilisés pour l'apprentissage ; sgts indique le nombre de segments classifiés. Les résultats sont donnés en pourcentage par émotion. Par exemple, 21% des segments « Peur » ont été reconnus comme de la colère.*

Les pourcentages de reconnaissance par émotion pour les données GEMEP (en ne gardant que les phrases 1 et 2 et les phrases improvisées) avec le même classifieur sont données Tableau 6-16. Globalement, les classifieurs entraînés sur CEMO ne fonctionnent pas du tout sur les données GEMEP. Pour les modes 'peu intense' et 'masqué', toutes les émotions GEMEP sont principalement reconnues comme de la tristesse. C'est quasiment le cas également pour le mode 'normal' et 'intense', à l'exception de la colère qui est assez bien reconnue (presque mieux que la colère des données CEMO). La colère semble être encore mieux reconnue pour les données plus intenses (2/3 des segments colère reconnus contre 44% pour les données normales), mais il faudrait plus de segments pour pouvoir le vérifier. Par contre l'irritation n'est absolument pas reconnue comme de la colère. La peur-panique et l'inquiétude ne sont reconnues non plus comme de la peur, ce qui peut être du en partie aux grandes variations dans l'expression de la peur dans le corpus CEMO assez éloignées de son expression prototypique. De même, l'étiquette « soulagement » avait été utilisée pour décrire un type de réaction émotionnel très spécifique dans le corpus CEMO (état de l'appelant en fin de conversation lorsqu'il sait qu'une aide va lui être apportée), qui peut être assez éloignée de la manière dont les acteurs vont l'exprimer.

Nous avons finalement décidé de nous focaliser sur les données normales et intenses.

La question s'est posée de savoir si les mauvaises performances étaient dues aux différentes conditions d'enregistrement.

**normal: CL=20%**

| | Peu | Col | Tris | Soul |
|---|---|---|---|---|
| inq (518) | **3** | 8 | **80** | 9 |
| peu (276) | **8** | 34 | **45** | 14 |
| irr (370) | 6 | **6** | **79** | 9 |
| col (261) | 11 | **44** | 19 | 26 |
| tris (317) | 8 | 4 | **76** | 13 |
| des (269) | 11 | 31 | **51** | 7 |
| soul (391) | 14 | 9 | **71** | **7** |

**Peu intense CL=13%**

| | Peu | Col | Tris | Soul |
|---|---|---|---|---|
| Inq (59) | **5** | 0 | **95** | 0 |
| Peu (47) | **13** | 6 | **70** | 11 |
| irr (28) | 0 | **4** | **96** | 0 |
| col (52) | 8 | **23** | **52** | 17 |
| Tris (34) | 15 | 3 | **79** | 3 |
| Des (39) | 3 | 0 | **77** | 21 |
| Soul (41) | 12 | 2 | **66** | **20** |

**intense: CL=28%**

| | Peu | Col | Tris | Soul |
|---|---|---|---|---|
| inq (94) | **2** | 23 | **52** | 22 |
| peu (42) | **19** | 45 | **21** | 14 |
| irr (39) | 23 | **13** | **64** | 0 |
| col (29) | 17 | **66** | 7 | 10 |
| tris (49) | 4 | 4 | **92** | 0 |
| des (46) | 4 | 26 | **48** | 22 |
| soul (57) | 25 | 9 | **63** | **4** |

**masqué: CL=10%**

| | Peu | Col | Tris | Soul |
|---|---|---|---|---|
| inq (102) | **15** | 17 | **45** | 24 |
| Peu (82) | **20** | 9 | **55** | 17 |
| irr (69) | 32 | **6** | **55** | 7 |
| col (53) | 21 | **4** | **45** | 30 |
| Tris (73) | 3 | 0 | **90** | 7 |
| Des (47) | 11 | 0 | **72** | 17 |
| Soul (76) | 16 | 3 | **68** | **13** |

**Tableau 6-16.** *Matrices de confusion pour les segments du corpus GEMEP (inq : inquiétude ; peu : peur ; irr : irritation ; col : colère ; tris : tristesse ;des :désespoir ; soul : soulagement ; le nombre entre parenthèses donne le nombre de segments par émotion) avec le même classifieur que le Tableau 6-15 entrainé sur les données CEMO. Les résultats sont donnés en pourcentage par émotion pour chaque mode (normal, peu intense, intense, masqué). Par exemple en mode normal, 8% des segments inquiétude ont été reconnus comme de la Peur.*

### Transformation des signaux GEMEP

Les signaux audio GEMEP ont été transformés afin de pouvoir être comparé à des données téléphoniques :

- rééchantillonnage pour passer de 44kHz à 8Hz
- élimination des basses fréquences avec un filtre passe bande (bande téléphone 300Hz-3.4kHz)
- ajout d'un bruit de fond téléphonique (obtenu à partir d'un fichier CEMO)

Cette transformation n'a pas eu d'incidence sur les résultats.

**Elimination des « mauvais » acteurs**

Comme le remarquent d'autres chercheurs ayant travaillé sur les données GEMEP[80], les performances varient significativement suivant les acteurs. Nous avons regardé les performances par acteur toujours avec le classifieur entraîné sur les données CEMO, et retiré 3 acteurs pour qui aucune émotion n'était reconnue. Cela peut être du à de mauvaises performances ou à un prototype de la colère différent de la colère exprimée dans les données CEMO. Les résultats pour les données normales et intenses sont donnés Tableau 6-17.

**7 meilleurs locuteurs**

| | Peu | Col | Tris | Soul |
|---|---|---|---|---|
| Inq (394) | **4** | 11 | 71 | 13 |
| peu (135) | **10** | 37 | 38 | 15 |
| Irr (216) | 9 | **7** | 75 | 8 |
| col (127) | 13 | **61** | 9 | 17 |
| tris (201) | 6 | 2 | **81** | 10 |
| des (157) | 11 | 36 | **45** | 8 |
| soul (250) | 20 | 6 | 70 | **4** |

**Phrase 1**

| | Peu | Col | Tris | Soul |
|---|---|---|---|---|
| inq (130) | **5** | 13 | 74 | 8 |
| peu (51) | **18** | 43 | 31 | 8 |
| irr (66) | 15 | **9** | 70 | 6 |
| col (38) | 26 | **63** | 3 | 8 |
| tris (79) | 9 | 0 | **84** | 8 |
| des (74) | 14 | 42 | **34** | 11 |
| soul (114) | 28 | 3 | 66 | **4** |

**Phrase 2**

| | Peu | Col | Tris | Soul |
|---|---|---|---|---|
| inq (64) | **2** | 14 | 84 | 0 |
| peu (40) | **8** | 42 | 48 | 2 |
| irr (49) | 6 | **4** | 84 | 6 |
| col (36) | 6 | **72** | 11 | 11 |
| tris (41) | 2 | 0 | **90** | 7 |
| des (38) | 8 | 37 | **53** | 3 |
| soul (74) | 9 | 9 | 80 | **1** |

**Phrase libre**

| | Peu | Col | Tris | Soul |
|---|---|---|---|---|
| inq (200) | **4** | 10 | 66 | 20 |
| peu (44) | **5** | 25 | 36 | 34 |
| irr (101) | 7 | **8** | 74 | 11 |
| col (53) | 9 | **51** | 11 | 28 |
| tris (81) | 5 | 6 | **74** | 15 |
| des (45) | 11 | 24 | **58** | 7 |
| soul (62) | 18 | 6 | 66 | **10** |

**Tableau 6-17.** *Matrices de confusion pour les données (normales + intenses) du corpus GEMEP après avoir retiré 3 « mauvais » locuteurs (inq : inquiétude ; peu : peur irr : irritation ; col :colère ; tris :tristesse ; des :désespoir ; soul : soulagement ; le nombre entre parenthèses donne le nombre de segments par émotion) avec le même classifieur que le Tableau 6-15 entrainé sur les données CEMO. Les résultats sont donnés en pourcentage par émotion puis en détaillant par rapport au type de contenu. Par exemple pour la phrase 1 «Né kal ibam soud molèn ! », 5% des segments prononcés avec inquiétude ont été reconnus comme de la Peur.*

[80] 3ème école été humaine :www.emotion-research.net/ws/HPirker_featuring_GEMEP_1.pps

Seule la colère est reconnue[81] à 61%. Elle semble d'ailleurs être mieux reconnue avec les phrases sans contenu linguistique, peut-être parce que tout doit être codé dans la prosodie (plus de 60% de reconnaissance de la colère pour les 2 phrases, contre 51% pour l'improvisation).

**Apprentissage sur GEMEP/Test sur CEMO**

Nous avons tout d'abord regroupé {inquiétude et peur-panique} en une classe *Peur*, {irritation et colère chaude} en une classe *Colère* et {tristesse et désespoir} en une classe *Tristesse*. Nous avons gardé 7 acteurs pour l'apprentissage et 3 pour le test. Les résultats de classification Peur/Colère/Tristesse/Soulagement étaient très bas, peut être à cause des différences entre des étiquettes traditionnellement appartenant à la même catégorie (peur vs. inquiétude, colère vs. irritation). Cela confirme d'ailleurs les observations du paragraphe précédent sur les différences entre les taux de reconnaissance pour la colère et l'irritation de GEMEP avec un classifieur entraîné sur les données CEMO.

Finalement, nous avons conservé les étiquettes peur-panique, colère, tristesse[82] et soulagement et un SVM a été utilisé pour entraîner les données[83] en suivant exactement la même procédure que pour les données CEMO. La matrice de confusion sur les 3 locuteurs du test est donnée Tableau 6-18. Y figurent également les performances pour les émotions inquiétude, irritation et désespoir, qui ne sont pas utilisées dans l'apprentissage.

| | Peur | Colère | Tristesse | Soulagement |
|---|---|---|---|---|
| *inquiétude (171)* | *9* | *39* | *21* | *30* |
| **peur (75)** | **69** | 31 | 0 | 0 |
| *irritation (95)* | *4* | *35* | *17* | *44* |
| **colère (64)** | 19 | **73** | 0 | 8 |
| **tristesse (104)** | 2 | 4 | **66** | 28 |
| *désespoir (93)* | *59* | *38* | *2* | *1* |
| **soulagement (123)** | 6 | 11 | 14 | **69** |
| **CL=69%** | | | | |

**Tableau 6-18.** *Résultat en pourcentage par émotion pour la classification Peur/Colère/Tristesse/Soulagement sur les données GEMEP en apprentissage et en test. Les données ont été entrainées avec un SVM sur les émotions peur, colère, tristesse et soulagement de 7 locuteurs et testées sur les 3 locuteurs restants. Les nombres entre parenthèses correspondent au nombre de segments testés.*

On pourrait s'attendre à ce que l'irritation soit majoritairement reconnue comme de la *Colère* et l'inquiétude comme de la *Peur*, mais ce n'est pas du tout le cas.

[81] Il n'est pas possible de tirer de conclusion pour la tristesse à cause du nombre important de fausses détections.
[82] Nous avons également essayé avec désespoir à la place de tristesse en pensant que l'émotion serait mieux reconnue étant plus « forte », mais ce n'était pas le cas.
[83] Mêmes données que pour l'expérience précédente : phrases 1,2 et « libre » dans les modes normal et intense

Le score CL pour GEMEP, en ne comptabilisant que les émotions utilisées pour l'apprentissage, est de l'ordre de 70% (pas d'optimisation tel que sélection des meilleurs locuteurs, attributs…). L'expérience a été répétée en sélectionnant des locuteurs différents pour le test et l'apprentissage et le score CL restait autour de 70% de bonne détection. Nos outils et paramètres extraits semblent donc se transposer assez facilement pour des données actées. Les performances de détection sont d'ailleurs bien meilleures avec les données actées.
Lorsque les données CEMO sont utilisées en test pour ce même modèle (Tableau 6-19), rien n'émerge. La majorité des émotions semblent reconnus comme de la colère.

| | Peur | Colère | Tristesse | Soulagement |
|---|---|---|---|---|
| peur (1168) | 1 | 47 | 6 | 45 |
| colère (382) | 3 | 59 | 6 | 33 |
| tristesse (334) | 5 | 45 | 11 | 39 |
| soulagement (295) | 8 | 59 | 10 | 23 |

**Tableau 6-19.** *Matrice de confusion en pourcentage par émotion pour les données CEMO testées avec un modèle entraîné sur GEMEP.*

## *Classification Peur/Colère*

Aucun résultat concluant n'a été obtenu en testant les données GEMEP sur un modèle Peur/Colère CEMO. Nous avons essayé deux modèles, un entraîné sur des clients CEMO qui classifiait 80% des données comme de la colère, et un entraîné avec la colère des agents (colère froide), où cette fois-ci tout était classé comme de la peur. La peur exprimée dans les données GEMEP semble ne rien avoir en commun avec celle exprimée dans CEMO et la colère GEMEP pourrait correspondre à la colère chaude CEMO.

## *Conclusion pour les données GEMEP*

Les problèmes de portabilité d'une tâche à une autre peuvent avoir de nombreuses causes :

- le jeu des acteurs : des tests perceptifs ont été fait par l'équipe de Genève, mais n'ont pas encore été publiés et ne sont pas encore disponibles
- la définition derrière une étiquette émotion qui peut varier énormément selon les applications. De plus, les définitions des émotions dans CEMO sont plus homogènes car elles ont été classées par les mêmes annotateurs, alors que pour GEMEP, elles dépendent de l'acteur et de son interprétation.
- le caractère multimodal des données GEMEP, qui fait que l'émotion ne va pas nécessairement s'exprimer par la voix.

Il est intéressant de noter que la colère (colère chaude) est la seule émotion reconnue.

## 6.6. Vers une modélisation plus fine et temporelle

Une maquette de démonstration a été construite en java. Cette maquette est à la fois un outil de recherche et un démonstrateur. La première version a été réalisée sur des données actées (projet de TER sciences affectives, PXI). Elle a été ensuite adaptée aux données CEMO**.** Elle permet d'effectuer tout le traitement d'un ensemble de fichiers audio et éventuellement de leurs transcriptions soit en entraînant des modèles avec ou non sélection de paramètre, soit en les utilisant en test de modèles connus. Une capture d'écran pour l'extraction des paramètres est donnée Figure 6-17**.**

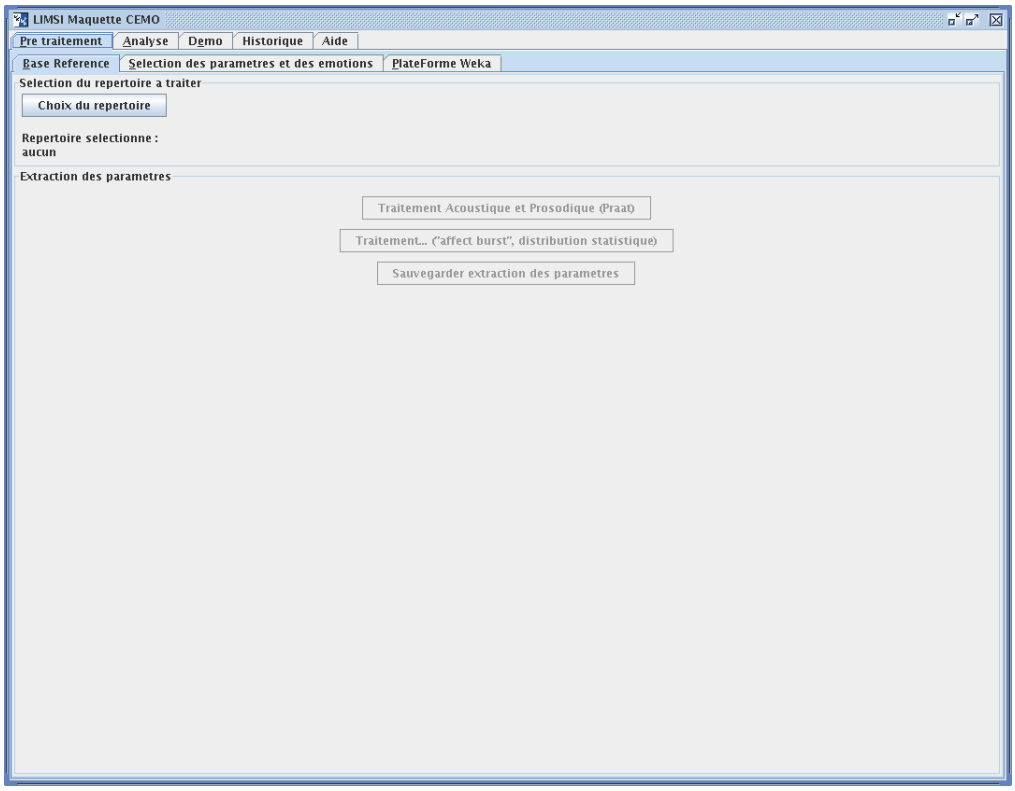



**Figure 6-17.** *Capture d'écran de l'onglet d'extraction des paramètres qui permet de calculer entre autre les paramètres acoustiques à l'aide de Praat et ceux déduits de la transcription s'ils sont fournis.*

Par exemple, pour traiter les 20 heures de données CEMO (30 000 fichiers : ~147Go, soit en moyenne 4,9 Mo pour chaque fichier), il faut 30 heures pour extraire tous les paramètres acoustiques avec Praat (F0, energie, formants toutes les 10ms et marqueurs affectifs), puis 5 heures pour associer à chaque fichier tous ses indices. Le temps d'apprentissage dépend de la taille de l'ensemble d'apprentissage et des algorithmes utilisés. Les résultats pour un ensemble de fichier sont présentés sous la forme d'une matrice de confusion. L'interface permet également de voir les probabilités avec un modèle donné de classe émotionnelle pour un segment émotion isolé, le

temps de décodage est alors 2 * TR et le résultat peut être présenté sous la forme d'un histogramme (*cf.* Figure 6-18).

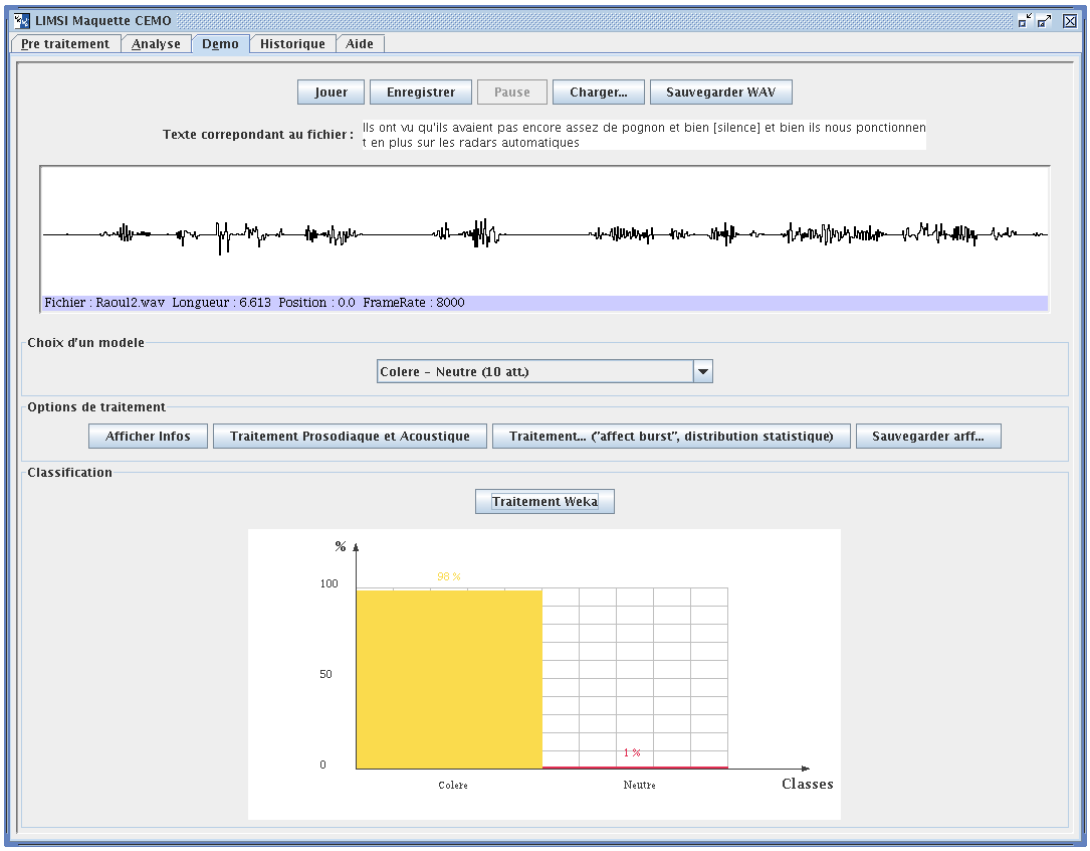



**Figure 6-18.** *Exemple de fichier en test d'un classifieur Colère/Neutre.*

La maquette nous a permis de faire des premiers tests des modèles entraînés sur CEMO sur des segments provenant de données réelles en anglais, ainsi que de données actées et réelles en Français et pour un classifieur Colère/Neutre, les résultats semblaient très prometteurs.

Les données réelles que nous voudrons tester ne seront pas toujours découpées en segments et dans le futur, il faudra pouvoir avoir une détection temporelle dynamique des émotions. Nous avons utilisé l'interface pour explorer deux découpages assez grossiers pour découper un flux audio, l'un en choisissant une fenêtre de taille fixe réglable qui se déplace avec un pas également réglable ; et l'autre en découpant le signal au niveau des « silences » (partie non voisée de durée supérieure à un seuil défini par l'utilisateur). Un exemple est donné Figure 6-19 en découpant le signal à l'aide des parties non voisées de plus de 30ms. Ces expériences ont été réalisées sur quelques signaux audio assez courts (quelques minutes) mais sont assez intéressantes et montrent l'aspect dynamique des émotions ainsi que les transitions rapides d'une émotion à l'autre.

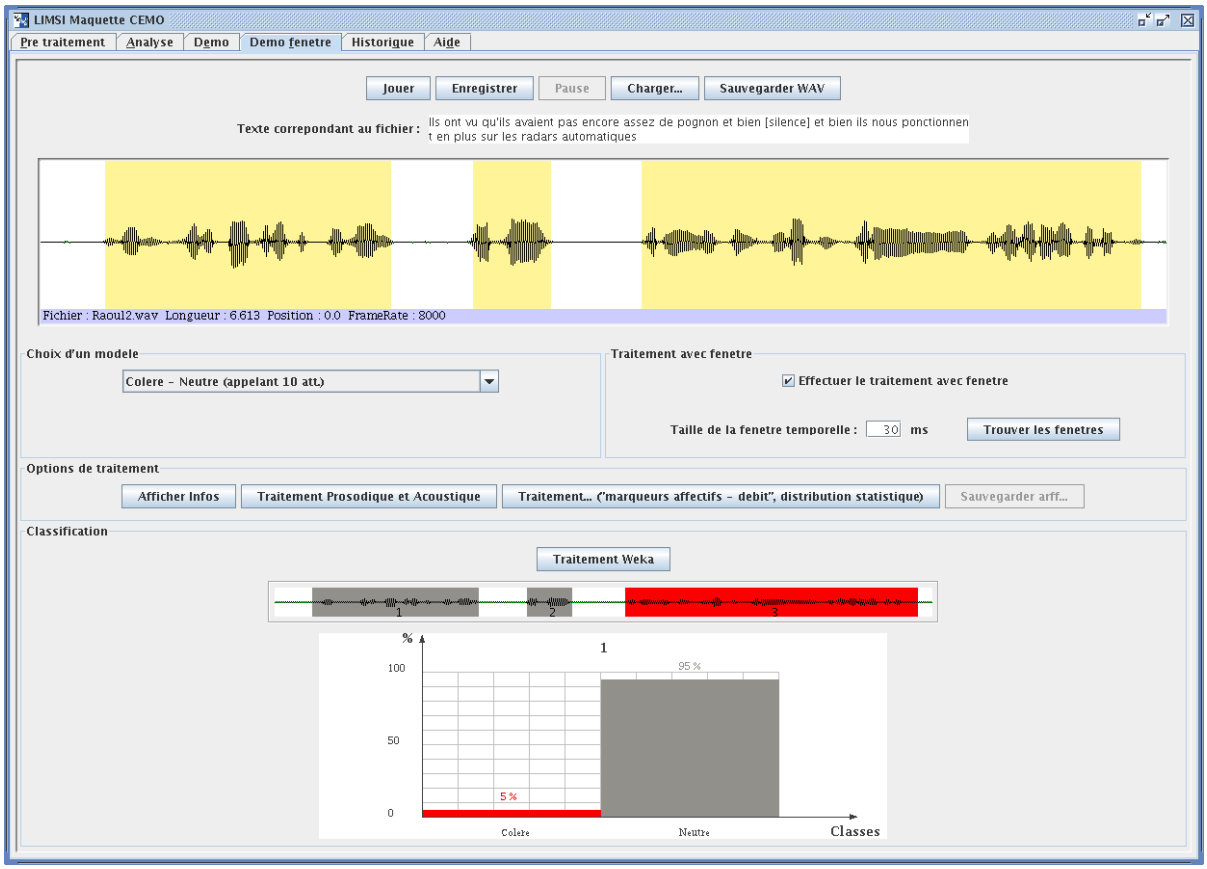



**Figure 6-19.** *Exemple de découpage en 3 segments émotions pour un classifieur Neutre/Colère, chaque émotion est représentée par une couleur et il est possible d'obtenir des précisions pour chaque segment.*

## 6.7. Conclusion

Des expériences préliminaires ont permis de vérifier qu'il n'y avait pas de différences significatives entre les différents algorithmes et que la sélection des paramètres semblait améliorer les performances. Nous avons donc décidé d'utiliser des SVM à noyau radial pour la suite de nos expériences et de combiner les résultats de différents algorithmes pour sélectionner les attributs. Nous avons également vérifié le bien-fondé du retrait des segments complexes de l'ensemble d'apprentissage et du test pour avoir de meilleurs systèmes et de doubler certains segments de l'ensemble d'apprentissage pour les classes peu représentées.

Une fois arrêté sur ces choix, nous avons dans un premier temps effectué une batterie d'expériences sur le corpus CEMO afin d'étudier les différences de contexte facilement observable (homme/femmes, agents/appelants), les différentes classes d'attributs et enfin les performances en faisant varier le nombre de classes discriminées (de 80% de bonnes détection avec 2 classes à 55% avec 5 classes). Nous avons également vérifié l'importance de l'extraction d'un grand nombre d'indices, à la fois lexicaux et paralinguistiques, les différents indices étant plus ou moins pertinents selon l'émotion à laquelle on s'intéresse.

Notre participation à CEICES nous a permis de vérifier que notre méthodologie fonctionnait bien sur des données différentes dans une autre langue et que les performances obtenues avec nos systèmes étaient comparables à celles d'autres sites s'intéressant à la détection des émotions. Cette collaboration nous a également amené à aborder d'autres sujets de réflexions et notamment le calcul des indices extraits et la manière de les nommer la plus explicite possible. Nous avons également pu comparer la F0 obtenue par Praat avec une correction manuelle de la F0[84].

Nous avons également constaté que les modèles entraînés sur nos données pouvaient donner de bonnes performances sur des données similaires, mais ne marchaient globalement pas sur des données actées. Certaines émotions comme la colère semblent plus robustes au changement de contexte.

Enfin, les données réelles que nous voudrons tester ne seront pas toujours découpées en segment, aussi nous avons commencé à réfléchir à une détection des émotions sur des sous-segments temporels. Une interface de démonstration a permis d'explorer deux découpages assez grossiers pour découper un flux audio, l'un en choisissant une fenêtre de taille fixe réglable qui se déplace

[84] Cas particulier des voix d'enfants

avec un pas également réglable ; et l'autre en découpant le signal au niveau des « silences ». Une évaluation de la détection des émotions sur ces sous-unités est un de nos projets à court terme.

# Chapitre 7

# Conclusion et Perspectives

# 7.CONCLUSION ET PERSPECTIVES

## 7.1. Conclusions

Contrairement à une majorité des études sur les émotions réalisées au commencement de la thèse, nous disposions d'une importante quantité de données spontanées avec une très grande variabilité dans les locuteurs et dans les émotions exprimées. Nous avons présenté un protocole d'annotation original adapté à la complexité des données réelles avec une large palette d'étiquettes émotions et la possibilité d'annoter plusieurs émotions à la fois, et nous avons adopté un « vecteur émotion » pour représenter ces mélanges. Nous nous sommes également particulièrement interrogées sur les mesures de validation des annotations et sur la notion d'annotateur « expert », avec notamment des mesures de cohérence intra-annotateur.

Nous avons effectué plusieurs tests perceptifs sur les mélanges d'émotions, qui ont été peu étudiés empiriquement et avons constaté par exemple que le seul niveau paralinguistique pouvait dans certains cas permettre de les percevoir, bien que le contexte soit également important.

Plus d'une centaine de paramètres pertinents ont été extraits par segment, leur poids relatif a été étudié et des expériences ont été menées avec 2 à 5 classes d'émotion sur nos données, en utilisant principalement des SVM après avoir essayé d'autres techniques. Nous avons montré l'importance de mélanger les différents types d'indices et montré également que les indices les plus importants varient d'une tâche à l'autre et bien sûr en fonction du type d'émotions que l'on veut détecter.

D'autres expériences ont montré que nos méthodes étaient robustes au changement de langue (français, allemand) et de type de données (actées/spontanées, adultes/enfants), bien que ce ne soit pas forcément le cas des classifieurs et paramètres. En particulier, les expériences de détection sur des données actées avec des modèles entraînés sur des données réelles et inversement donnaient peu de résultats. Toutefois, certains classifieurs semblent pouvoir se généraliser à d'autres tâches. Les premiers essais avec un modèle Neutre/Colère sur des données en anglais (centre d'appel) et en français (spontané mais pas en interaction) semblent probants, mais nécessitent évidemment une phase rigoureuse de validation. Nous avons également commencé à explorer la détection de sous-segments émotionnels en décomposant des segments tests en une succession d'émotions. L'aspect dynamique des émotions est un des aspects importants très rarement pris en compte par les systèmes actuels.

## 7.2. Perspectives

Les travaux réalisés dans le cadre de cette thèse peuvent être prolongés dans plusieurs directions. Les schémas et expériences d'annotation des émotions sont réutilisables après adaptation aux différentes tâches et ont contribué à la définition des besoins en termes d'annotation des émotions au sein du groupe de travail W3C Emotion Incubator Group (http://www.w3.org/2005/Incubator/emotion/). Nous n'avons d'ailleurs pas pris en compte l'annotation des dimensions continues dans nos travaux et il serait intéressant de les exploiter, par exemple en les ajoutant à l'ensemble d'indices ou en étudiant les corrélations entre étiquettes et axes.

L'aspect multilingue est un des aspects qui nous intéressent. Cependant, bien que nous ayons abordé cet aspect multilingue à travers nos expériences sur le corpus AIBO (corpus en allemand) et sur quelques données en anglais, il est probable que nos modèles ne soient pas efficaces sur des données très différentes comme par exemple pour les langues asiatiques ou les différences sont à la fois culturelles et tonales. En plus, certains paramètres, comme les marqueurs affectifs peuvent avoir des significations complètement opposées d'une langue à l'autre. Par exemple « bah/boa » qui serait un bon indicateur de dégoût en français ou en anglais exprime l'admiration en allemand. Nous visons également un traitement complètement automatique de l'extraction des indices. Si on peut envisager une détection automatique de certains marqueurs comme le rire ou les hésitations, ce n'est pas le cas pour d'autres indices pourtant très performants (longueur des phonèmes, mots tronqués ou inintelligibles…). Par contre, il pourrait être intéressant de prendre en compte, comme le font déjà certaine études, le contexte dialogique et par exemple d'ajouter les actes de dialogue, qui eux commencent à pouvoir être détectés automatiquement [Rosset et al. 2007] et une perspective serait de comparer notre approche avec ces nouvelles tendances. Nous voudrions aussi poursuivre nos expériences afin de combiner au mieux un modèle lexical et un modèle paralinguistique.

Il faudrait également envisager une approche plus dynamique pour traiter les émotions afin d'avoir une analyse en temps réel. Actuellement les segments sont considérés comme des unités statiques desquels on extrait un ensemble de paramètre, mais on pourrait extraire des paramètres sur des fenêtres temporelles ainsi que nous avons commencé à l'étudier et prendre en compte les émotions des segments précédents, par exemple en utilisant des HMMs.
Nous n'avons pas remis en cause l'utilisation des SVM, mais des travaux récents semblent montrer que des arbres du type 'random forest' permettent d'obtenir de meilleurs modèles [Schuller et al.

2007] et la tendance est aux méta-algorithmes qui combinent les sorties de différents algorithmes d'apprentissage. Une autre tendance est d'annoter sur des axes abstraits, quitte à les projeter ensuite dans un espace émotionnel, ce qui a pour avantage de ne pas nécessiter de choisir un nombre d'émotions à détecter et de ne pas nécessiter de définitions[Grimm et Kroschel 2007].

Une autre problématique est de trouver la meilleure façon d'exploiter les mélanges d'émotions. Nous les avons analysé et le fait de les annoter nous permet de filtrer les segments utilisés pour l'apprentissage lors de la discrimination d'émotions « pures ». Est-ce que ces mélanges sont spécifiques à nos données. Peut on envisager de pouvoir les détecter ?

Enfin, pour ce qui est de savoir si un module de détection des émotions pourrait permettre d'améliorer les systèmes de reconnaissance de la parole, un premier travail est en cours pour évaluer les performances d'un système de reconnaissance de la parole sur des données émotionnelles. Ce travail montre, en premier résultat, que l'impact des émotions sur les performances du système de reconnaissance va dépendre du type d'émotion présente. Certaines comme le soulagement dans le cas du corpus CEMO seront exprimées assez souvent par le canal linguistique avec des phrases assez 'simples' (*merci beaucoup*) et sont susceptibles d'être bien reconnues, ce qui sera moins évident pour l'expression de la peur ou du stress.

A plus long terme, nous nous interrogeons également sur la détection de l'audio téléphonique comparée à celle de l'audio dans la multimodalité. Les indices sont-ils différents. Comment prendre en compte les informations données par les autres modalités ?

En conclusion, peu de travaux sur les émotions portent sur l'étude des corpus oraux spontanés, tout d'abord parce que la collecte de tels corpus est difficile pour des raisons de confidentialité liées aux données et également parce qu'elle est très coûteuse. Comme les différentes études se font rarement sur un même corpus, il n'existe pas encore de protocoles d'évaluation des systèmes de détection des émotions. L'expérience CEICES du réseau Humaine est pour l'instant unique. Les résultats obtenus dans cette thèse sont principalement des schémas d'annotation ainsi que des protocoles de validation, des tests perceptifs, des études sur les indices caractérisant certaines émotions en majorité négatives (comme peur, colère) dans des données spontanées et enfin la mise en œuvre de systèmes de détection des émotions pour différentes tâches ainsi que des premières évaluations sur leur robustesse. Nous avons également mis en avant la présence d'émotions complexes mélangées dans des données orales réelles.

De nombreuses études sur des données naturelles sont encore nécessaires pour détecter les comportements émotionnels complexes ou proches (par exemple différencier l'irritation du stress ou de la colère) même si sur peu de classes, par exemple 2 classes (colère, neutre), les scores de prédiction sont déjà intéressants pour imaginer dans un futur proche de premières applications notamment en fouille de données.

# IV Annexes

# ANNEXE1: QUELQUES DEFINITIONS DE L'EMOTION

---

http://www.alleydog.com/glossary/definition.cfm?term=Emotion:
"Emotion: Most people have little problem recognizing and identifying when we are having an emotion. However, emotion is one of the most difficult concepts in Psychology to define. In fact, emotion is such a difficult concept to define adequately that there are at least 90 different definitions of emotions in the scientific literature. A simple definition of emotion is that it is a response by a whole organism, involving (1) physical arousal, (2) expressive behaviors, and (3) conscious experience."

---

[Lang et al. 1997 p173]: "For the layman the basic datum of an emotion is a state of feeling, i.e.., a direct experience or internal apprehension, requiring no further definition."

---

[Averill 1996]: "…'emotion' is derived from the latin e + movere. It originally means to migrate or to transfer from one place to another. It also was used to refer to states of agitation and perturbation, both physical and psychological."

[Gellhorn et Loofbourrow 1963 p409]: "…emotion is a fact upon which all introspection agrees. Anxiety, depression, elation, indifference, anger, fear, pleasurable anticipation and dread, for example, are undeniable because there are states which we have experienced personnaly."

---

[Caffi et Janney 1994] : «… phénomène empiriquement investigable, généralement transitoire et d'une certaine intensité qui se manifeste au niveau linguistique de différentes manières par le choix des mots, l'intonation, les exclamations »

---

[Caffi et Janney 1994 p327]: « Western psychologists commonly distinguish between feelings, a broad, complex class of subjective personal sensations or states of inner physiological arousal; emotions, a restricted subset of empirically investigable phenomena within this general class that are relatively transitory, of a certain intensity, and are attached to, or triggered by, particular objects, ideas, or outer incentive events; moods, which are said to be of longer duration than emotions, and not necessarily attached to specific inner states or definite objects; and attitudes, or

transitory feeling states with partly uncontrollable subconscious psychobiological components and partly controllable expressive components, which are said to be instrumental in maintaining social and psychological equilibrium and adapting to different situations."

"The term 'affect' is usually reserved for feeling states that are ascribed to others on the basis of their observable behaviour in different situations. In cognitive psychology, notions of affect range from 'hot' to 'cold' extremes. At the hotter end, 'affect' is used almost synonymously with emotion as defined above. At the cooler end, it is used to refer simply to human preferences, attitudes, or likes and dislikes, and to adaptive choices related to these. […] In linguistics, on the other hand, the term 'affect' is often simply used as a broad synonym for 'feeling'."

---

(Scherer 1999, http://emotion-research.net/restricted/contract/technical_annex.pdf).
« We consider emotion in an inclusive sense rather than in the narrow sense of episodes where a strong rush of feeling briefly dominates a person's awareness – we have called those 'fullblown emotions'"

[Scherer 1993] : "Episode of temporary synchronisation of all major subsystems of organismic functioning represented by five components (cognition, physiological regulation, motivation, motor expression and monitoring/feeling) in response to the evaluation of an external or internal stimulus event as relevant to central concerns of the organism"

---

[Schachter et Singer 1962 p380] cité dans [Cornelius 1996]: "[A]n emotional state may be considered a function of a state of physiological arousal and of a cognition appropriate to this state of arousal"

---

Toates dans [Hamilton et al. 1988]:
p15: "Emotion is seen as an evolutionary development that accompanied the emergence of flexibility and learning skills in relatively advanced animals. It serves motivation and learning. In the present model, emotion is triggered in part by comparison between an expectation based upon a goal set by the motivation system ('Sollwert') and the actual state that prevails ('Istwert'). Emotion can be positive (outcome equal to or better than expected) or negative (outcome worse than expected, as assumed by Grey, 1971)

---

p16 : "I would suggest that what we call 'emotion' in everyday speech refers to subjective feelings arising from a compound of the stimuli that impinge upon us, their appraisal, the memories that they evoke and the course of goal directed activity that is investigated, or at least suggested, by their appraisal."

---

[Plutchik et Kellerman 1990 p4]: "A major element in both the implicit and explicit views of emotion is that an emotion is a subjective feeling of a certain kind -- the kind for which labels such as angry, disgusted, and afraid are appropriate. However, there is considerable evidence to suggest that this is too narrow a way to define emotions"

---

[Lazarus et al. 1980 p198]: "Emotions are complex, organized states … consisting of cognitive appraisals, action impulses, and patterned somatic reactions. Each emotion quality (e.g. anger, anxiety, joy) is distinguished by a different pattern of components, which is what urges the analogy to a syndrome. Moreover, the three components of emotion are subjectively experienced as a whole, that is, as a single phenomenon as opposed to separate and distinct responses. When one component is missing from the perception the experience is not a proper emotion although it may contain some of the appropriate elements"
action impulse : the action is set in motion internally (psychophysiologically ) need not be carried out, can be suppressed, denied, transformed.

---

[Averill 1980 p313] "An emotion is a transitory social role (a socially constituted syndrome) that includes an individual's appraisal of the situation and that is interpreted as a passion rather than as an action"

---

# TABLE DES FIGURES

# LISTE DES TABLEAUX

# BIBLIOGRAPHIE


Abrilian, S., L. Devillers and J.-C. Martin (2006). Annotation of Emotions in Real-Life video Interviews: Variability between Coders. *LREC*.

Adank, P. M.*Vowel normalization : a perceptual-acoustic study of Dutch vowels. These de doctorat*,2003-Radboud University Nijmegen

Adda-Decker, M., Lamel, L. (1999). Pronunciation variants across system configuration, language and speaking style. *Speech Communication*. **29**: p. 83-98.

Albrecht, I., M. Schröder, J. Haber and H.-P. Seidel (2005). Mixed feelings: Expression of non-basic emotions in a muscle-based talking head. *Virtual Reality*. **8**: p. 201-212.

Ang, J., R. Dhillon, A. Krupski, E. Shriberg and A. Stolcke (2002). Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog. *International Conference on Spoken Language Processing*, Denver. **3:** p. 2037-2040.

Arnold, M. B. (1960). *Emotion and personality*, New York: Columbia University Press.

Austin, J. L. (1962). *How to Do Things With Words*. Cambridge, Paperback: Harvard University Press.

Averill, J. R. (1975). A semantic atlas of emotional concepts. *JSAS: Catalog of Selected Documents in Psychology*. **5**: p. 330.

Averill, J. R. (1980). A Constructivist View of Emotion. *Emotion theory, research and experience vol1. Theories of Emotion*. R. Plutchik and H. Kellerman. New York, Academic Press: p: 849-855.

Averill, J. R. (1994). In the eyes of the beholder. *The nature of emotion*. P. Ekman.

Averill, J. R. (1996). An Analysis of Psychophysiological Symbolism and Its Influence on Theories of Emotion. *The Emotions. Social, Cultural and Biological Dimensions*. R. Harré and W. G. Parrott.

Averill, R. J. (1989). A constructivist view of emotion. *Emotion theory, research and experience vol1*. H. Kellerman: p: 305-339.

Ax, A. F. (1953). The physiological differentiation between fear and anger in humans. *Psychosomatic Medicine*. **15**: p. 433-442.

Banse, R. and K. R. Scherer (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*. **70(3)**: p. 614-636.

Bänziger, T., H. Pirker and K. S. Scherer (2006). GEMEP - GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions. *Proceedings of LREC'06 Workshop on Corpora for Research on Emotion and Affect*, Genoa**:** p. 15-19.

Bänziger, T. , K. R. Scherer (2007). Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus. *ACII* : p. 476-487

Barras, C., E. Geoffrois, Z. Wu and M. Liberman (2000). Transcriber : Development and Use of a Tool Assisting Speech Corpora Production. *Speech Communication*. **33 (1):** p. 5-22.

Batliner, A.; Warnke, V.; Nöth, E.; BUCKOW, J.; HUBER, R.; NUTT, M.( 1998) How to label accent position in spontaneous speech automatically with the help of syntactic-prosodic boundary labels**.** Technical Report.

Batliner, A., K. Fisher, R. Huber, J. Spilker and E. Noth (2003). How to Find Trouble in Communication. *Speech Communication*. **40**: p. 117-143.

Batliner, A., C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russell and M. Wong (2004). "You stupid ting box"- children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. *4th internacional Conference on Language Resources and Evaluation***:** p. 171-174.

Batliner, A., R. Kompe, A. Kießling, M. Mast, H. Niemann and E. Nöth (1998). M = Syntax + Prosody: A syntactic-prosodic labelling scheme for large spontaneous speech databases. *Speech Communication*. **25**: p. 193-222.

Batliner, A., S. Steidl, B. Schuller, D. Seppi, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous and V. Aharonason (2007). The Impact of F0 Extraction Errors on the Classification of

Prominence and Emotion. *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken**:** p. 2201-2204.
Bergeri, I., R. Michel and J. P. Boutin (2002). Pour tout savoir ou presque sur le coefficient kappa... *Médecine tropicale*. **62**: p. 634-636.
Boersma, P. and D. Weenink (2005). Praat: doing phonetics by computer (Version 4.3.14) [Computer program]. Retrieved May 26, 2005, from http://www.praat.org/.
Boite, R., H. Bourlard, T. Dutoit, J. Hancq and H. Leich (1999). *Traitement de la parole*. Lausannes.
Brenner, C. (1980). A psychoanalytic theory of affects. *Emotion theory, research and experience vol1*. R. Plutchik and H. Kellerman. New York, Academic Press: p: 341-348.
Cacioppo, J. T., G. G. Berntson, J. T. Larsen, K. M. Poehlmann and T. A. Ito (2000). The psychophysiology of emotion. *Handbook of emotions*. R. Lewis and J. M. Haviland-Jones. New York: Guilford: p: 173-191.
Caffi, C. and R. W. Janney (1994). Toward a pragmatics of emotive communication. *Journal of pragmatics*. **22**.
Campbell, N. and P. Mokhtari (2003). Voice Quality : the 4th Prosodic dimension. *15th ICPhS*(Barcelona).
Carver, C. S. (2001). Affect and the functional bases of behavior: On the dimensional structure of affective experience. *Personality and*
*Social Psychology Review*. **5**: p. 345-356.
Chih, W. H., C. C. Chi and J. L. Chih (2003). A practical Guide to Support Vector Classification. *http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf*.
Childers , D. G. (1978). *Modern spectrum analysis*, IEEE Press.
Clavel, C.*Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales*,2007-Doctorat Signal et Images, TSI Traitement du Signal et des Images, ENST, p.195
Clore, G. L. (1994). Why emotions are felt. *The nature of emotion: Fundamental questions*. P. Ekman and R. J. Davidson, New York: Oxford University Press: p: 103-111.
Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ Psychol Meas*. **20**: p. 27-46.
Cohen, J. (1968). Weighted kappa : nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*. **70**: p. 213-220.
Cornelius, R. R. (1996). *The science of emotion : research and tradition in the psychology of emotions*, Upper Saddle River, NJ: Prentice-Hall.
Cornuéjols, A. (2002). Une introduction aux SVM. *Bulletin n°51 de l'AFIA (Association Française d'Intelligence Artificielle)*.
Cornuéjols, A. and L. Miclet (2002). *Apprentissage artificiel*, Eyrolles.
Cowie, R. (2000). Emotional states expressed in speech. In describing the emotional states expressed in speech. *Proc ISCA ITRW on Speech and Emotion: Developing a Conceptual Framework for Research***:** p. 224-231.
Cowie, R. (2007). Emotion: concepts and definitions(and perhaps a declaration). *humaine Conceptualising emotion workshop*, Haifa**:**
http://emotion-research.net/ws/conceptualizingemotion/concepts%20and%20definitions%203.ppt/view.
Cowie, R., E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey and M. Schröder (2000). Feeltrace: an instrument of recording perceived emotion in real-time. *Proc ISCA ITRW on Speech and Emotion: Developing a Conceptual Framework for Research*: p. 224-231.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*. **16**: p. 297-334.
Damasio, A. (1994). *L'erreur de Descartes*, New York: Grosset/Putnam.
Darwin, C. (1872). *The expression of emotions in man and animals*. New York, Philosophical library.
Dellaert, F., T. Polzin and A. Waibel (1996). Recognizing Emotion In Speech. *ICSLP*.

Devillers L., (2006) Les émotions dans les interactions homme-machine : perception, détection et génération. Thèse d'Habilitation à diriger des Recherches, *Université Paris-Sud/LIMSI* .

Devillers, L., S. Abrilian and J.-C. Martin (2005a). Representing real life emotions in audiovisual data with non basic emotional patterns and context features. *ACII*.

Devillers, L., I. Vasilescu and L. Lamel (2002). Annotation and detection of emotion in a task oriented human-human dialog corpus. *International Standards for Language Engineering*, Edinburgh.

Devillers, L., I. Vasilescu and L. Lamel (2003a). Emotion detection in task-oriented dialogs corpus. *IEEE ICME*(Baltimore).

Devillers, L., I. Vasilescu and C. Mathon (2003b). Acoustic cues for perceptual emotion detection in task-oriented human-human corpus. *15th International Congress of Phonetic Sciences*.

Devillers, L., I. Vasilescu and L. Vidrascu (2004). Anger versus Fear detection in recorded conversations. *Speech Prosody*, Nara, Japon.

Devillers, L. and L. Vidrascu (2006a). Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs. *Interspeech*.

Devillers, L. and L. Vidrascu (2006b). Représentation et Détection des émotions dans des données issues de dialogues enregistrés dans des centres d'appels : des émotions mixtes dans des données réelles. *numéro spécial " Interaction Emotionnelle ", Revue Des Sciences et Technologies de l'Information, série Revue d'Intelligence Artificielle*. **20**(4-5) : p. 447-476.

Devillers, L. and L. Vidrascu (2007). Positive and Negative emotional states behind the laugh in spontaneous spoken dialogs. *workshop The phonetics of Laughter*, Saarbrucken.

Devillers, L., L. Vidrascu and L. Lamel (2005b). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*. **18**.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*. **6**: p. 169-200.

Ekman, P. and R. J. Davidson (1994). *The nature of Emotion : Fundamental questions*, New York: Oxford University.

Ekman, P. and A. J. Fridlung (1987). Assessment of facial behavior in affective disorders. *Depression and Expressive Behavior*. J. D. Maser. Hillsdale: p: 33-56.

Fehr, B. and J. A. Russell (1984). Concept of emotion viewed from a prototype perspective. *Journal of experimental psychology : General*. **113**: p. 464-486.

Fernandez, R. and R. Picard (2003). Modeling Drivers' Speech Under Stress. *Speech Communication*. **40**.

Fischer, A. (1993). Sex differences in emotionality: fact or stereotype. *Feminism and Psychology*. **3**: p. 303-318.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*. **76**: p. 378-382.

Fónagy (1983). *La vive voix. Essais de psycho-phonétique*.

Forbes-Riley, K. and D. Litman (2004). Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources. *Chapter of the Association for Computational Linguistics (HLT/NAACL)*.

Freund, Y. and R. E. Shapire (1996). Experiments with a new boosting algorithm. *19th International Conference on Machine Learning*: p. 148-156.

Frick, R. W. (1986). The prosodic expression of anger: Differentiating
threat and frustration. *Aggressive Behavior*. **12**: p. 121–128.

Fridlund, A. J. (1991). The sociality of solitary smiles: Effects of an implicit audience. *Journal of Personality and social psychology bulletin*. **60**: p. 229-240.

Gauvain, J. L. (2002). The LIMSI broadcast news transcription system. *Speech Communication*. **37 no. 1-2**: p. 89-108. .

Gellhorn, E. and G. N. Loofbourrow (1963). *Emotions and Emotional Disorders: A Neurophysiological Study*. New York.

Greasley, P., C. Sherrard and M. Waterman (2000). Emotion in Language and Speech: Methodological issues in Naturalistic Approaches. *Languaga and Speech*. **43**: p. 355-375.

Grimm, M. and K. Kroschel (2007). Emotion Estimation in Speech Using a 3D Emotion Space Concept. *Robust Speech Recognition and Understanding*. M. Grimm and K. Kroschel. Vienna, Austria, I-Tech Education and Publishing.

Grimm, M., K. Kroschel, E. Mower and S. Narayanan (2007). Primitives-Based Evaluation and Estimation of Emotions in Speech. *Speech Communication*. **49**(10-11).

Gross, J. J. and R. W. Levenson (1995). Emotion elicitation using films. *Cognition and Emotion*. **9**: p. 87–108.

Grossberg, J. M. and H. K. Wilson (1968). Physiological changes accompanying the visualization of fearful and neutral situations. *Journal of Personality and Social Psychology*. **10**: p. 124-133.

Hall, M. A.*Master Thesis, Correlation based feature selection for Machine Learning*,1999-Department of Computer Science, University of Waikato,

Hall, M. A. and G. Holmes (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge & Data Engineering*. **15**: p. 1437-1447.

Hamilton, V., G. Bower and N. Frijda (1988). *Cognitive perspectives on emotion and motivation*, Springer.

Hardy, H., K. Baker, L. Devillers, L. Lamel, S. Rosset, T. Strzalkowski, C. Ursu and N. Webb (2002). Multi-layer Dialogue Annotation for Automated Multilingual Customer Service. *International Standards for Language Engineering workshop*.

Hebb, D. O. (1972). *Textbook of psychology*, Philadelphia: Saunders.

Hess, U. (2006). Emotion ressentie et simulée. *Cognition et émotions*, Kirouak, G.: p: 115-127.

Hirst, D. and A. Di Cristo (1998). A survey of intonation systems. *Intonation systems A survey in twenty languages*. D. Hirst and A. Di Cristo. Cambridge, CambridgeUniversity Press: p: 1-45.

Hochschild, A. R. (1979). Emotion work, Feeling rules, Social structure. *American Journal of Sociology*. **85**: p. 551-575.

Howell, D. C. (1999). *Méthodes statistiques en sciences humaines*.

Izard (1972). *Patterns of emotions : a new analysys of anxiety and depression*. New York, Academic Press.

James, W. (1884). What is an Emotion? *Mind*. **9**: p. 188-205.

Juslin, P. N. and P. Laukka (2003). Communication of emotions in vocal expression and music performance: different channels same code? *Psychological Bulletin*. **129 (5)**: p. 770-814.

Juslin, P. N. and K. R. Scherer (2005). Vocal expression of affect. *The New Handbook of Methods in Nonverbal Behavior Research*. J. Harrigan, R. Rosenthal and K. R. Scherer. Oxford, UK, Oxford University Press: p: 65-135.

Kappas, A., U. Hess and K. R. Scherer (1991). Voice and emotion. *Fundamentals of nonverbal behavior*. R. S. Feldman and B. Rimé, Cambridge and New York: Cambridge University Press.: p: 200-238.

Kaufman, L. and P. J. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York.

Kleinginna, P. R. and A. M. Kleinginna (1981). A Categorized List of Emotion Definitions, with Suggestions for a Consensual Definition. *Motivation and Emotion*. **5 (4)**: p. 345-359.

Kodratoff, Y. and M. Barès (1991). *Base terminologique de l'intellingence artificielle*. Paris.

Landis, J. R. and G. G. Koch (1977). The measurement of observer agreement for categorical data. *Biometrics*. **33**: p. 159-174.

Landwehr, N., M. Hall and E. Frank (2003). Logistic Model Trees. *ECML*.

Lang, P. J., M. M. Bradley and B. N. Cuthbert (1997). Motivated attention: Affect, activation, and action. *Attention and orienting: Sensory and motivational processes*. N. Mahwah, Lawrence Erlbaum.: p: 97-135.

Larsen, R. J., E. I. Diener and p. (Ed.), 13. Newbury Park, CA: Sage. (1992). Promises and problems with the circumplex model of emotion. *Review of personality and social psychology*. M. S. Clark: p: 25-59.

Lazarus, R. S. (1991). *Emotion and Adaptation*, New York: Oxford University Press.

Lazarus, R. S. (1998). *Fifty years of the research and theory of R.S. Lazarus*.

Lazarus, R. S., A. D. Kanner and S. Folkman (1980). Emotions: A cognitive-phenomenological analysis. *Theories of emotion*. R. Plutchik and H. Kellerman. New York: Academic Press: p: 189-217.

Lee, C. M. and S. Narayanan (2004). Towards detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*.

Lee, C. M., S. Narayanan and R. Pieraccini (2001). Recognition of Negative Emotions from the Speech Signal. *Automatic Speech Recognition and Understanding ASRU*, Trento, Italy.

Lee, C. M., S. Narayanan and R. Pieraccini (2002). Classifying Emotions in Human-Machine Spoken Dialogs. *ICME*.

Lee, C. M., S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee and S. S. Narayanan (2004). Emotion recognition based on phoneme classes. *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea**:** p. 889-892.

Levenson, R. W., L. L. Carstensen, F. W. V. and P. Ekman (1991). Emotion, physiology,and expression in old age. *Psychology and Aging*. **6**(28-35).

Levenson, R. W., P. Ekman and W. V. Friesen (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*. **27**: p. 363-384.

Liberman, P. and S. B. Michaels (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *J.Acoustic. Soc. America*. **34**: p. 922-927.

Liénard, J. S. (1977). *Les processus de la communication parlée. Introduction à l'analyse et à la synthèse de la parole*. Paris, Masson.

Liscombe, J. (2006). Detecting Emotion in Speech: Experiments in Three Domains. *Proceedings of HLT/NAACL*, New York.

Liscombe, J., G. Riccardi and D. Hakkani-T ür (2005). Using Context to Improve Emotion Detection in Spoken Dialogue Systems. *Interspeech*, Lisbon, Portugal.

Marchal, A. (1980). *Les sons et la parole*. Montréal.

Martin, L. L. (1986). Set/Reset: Use and Disuse of Concepts in Impression Formation. *Journal of Personality and Social Psychology*. **51 (3)**: p. 493-504.

Narayanan, S. (2002). Towards modelling user behaviour in human-machine interactions: Effect of Errors and Emotions. *ISLE Workshop*(Edinburgh).

Nass, C., I.-M. Jonsson, H. Harris, B. Reaves, J. Endo, S. Brave and L. Takayama (2005). Increasing
safety in cars by matching driver emotion and car voice emotion. *CHI 2005*, Portland, Oregon, USA.

Ortony, A., G. L. Clore and A. Collins (1988). *The Cognitive Structure of Emotions*, New York: Cambridge University Press.

Ortony, A. and T. J. Turner (1990). What's basic about basic emotion? *Psychological Review*. **97**: p. 315-331.

Osgood, C., W. H. May and M. S. Miron (1975). *Cross-cultural Universals of Affective Meaning*. Urbana, University of Illinois Press.

Osgood, C. E., G. J. Suci and P. H. Tannenbaum (1957). *The measurement of meaning*, Urbana: University of Illinois Press.

Oudeyer, P. Y. (2003). The production and recognition of emotions in speech: features and algorithms. *Int. J. Hum. Comput. Stud.* **59**(1-2): p. 157-183.

Parrott, W. G. and R. Harré (1996). The social dimension of emotions. *The Emotions. Social, Cultural and Biological Dimensions*. London, Sage publications: p: 39-56.

Petrushin, V. (1999). Emotion in Speech: Recognition and Application to Call Centers. *Artificial Neural Net. In Engr. (ANNIE)*: p. 7-10.

Philippot, P. (1993). Inducing and assessing differentiated Emotion-Feelind states in the laboratory Louvain. *Cognition and emotion*: p. 171-193.

Picard, R. (1997). *Affective computing*. Cambridge, MIT Press.

Plutchik, R. (1962). *The Emotions: Facts, Theories and a New Mode*. New York, Random House.
Plutchik, R. (1984). Emotions: A General Psychoevolutionary Theory. *Approaches to Emotion*. K. R. Scherer and P. Ekman, Erlbaum Hillsdale NJ: p: 293-317.
Plutchik, R. and H. Kellerman (1990). *Emotion theory, research and experience vol1. Theories of Emotion.* New York, Academic Press.
Polzin, T. and A. Waibel (1998). Detecting Emotions in Speech. *Cooperative multimodal communication*, Tilburg Netherlands.
Roesch, E. B., J. R. Fontaine and K. R. Scherer (2006). The world of emotions is two-dimensional .. or is it? *Presentation at the 3rd HUMAINE Summer School*, Genova**:**
http://emotion-research.net/ws/summerschool3/.
Rosenberg, A. and E. I. Binkowski (2004). Augmenting the
kappa statistic to determine inter-annotator reliability
for multiply labeled data points. *Proceedings of the Human Language Technology Conference and Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL).*
Rosset, S., D. Tribout and L. Lamel (2007). Multi-level Information and Automatic dialog Act Detection in Human-Human Spoken Dialogs. *Speech Communication.*
Rossi, M., A. Di Cristo, D. Hirst, P. Martin and Y. Nishinuma (1981). *L'intonation De l'acoustique à la sémantique.*
Russell, J. A. and A. Mehrabian (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*. **11**: p. 273-294.
Sander, D., D. Grandjean and K. R. Scherer (2005). A systems approach to appraisal
mechanisms in emotion. *Neural Networks*(18): p. 317-352.
Schachter, S. and J. E. Singer (1962). Cognitive, social and physiological determinants of emotional states. *Psychological Bulletin*. **69**: p. 379-399.
Scherer, K. R. (1986). Vocal affect expression: A review of research paradigms. *Psychological Bulletin 99*: p. 143-165.
Scherer, K. R. (1989). Vocal correlates of emotion arousal and affective disturbance. *Handbook of Psychophysiology: Emotion and social behavior*. H. Wagner and A. Manstead. London: Wiley.: p: 165-197.
Scherer, K. R. (1993). Neuroscience projections to current debates in emotion psychology. *Cognition and Emotion*. **7**: p. 1-41. .
Scherer, K. R. (1998). Analysing Emotion Blends. *ISRE*.
Scherer, K. R. (2003). Vocal communication of emotions : A review of research paradigm. *Speech Communication*. **40**: p. 227-256.
Scherer, K. R., R. Banse, H. G. Wallbott and T. Goldbeck (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion*. **15**: p. 123-148.
Scherer, K. R. and G. Ceschi (2000). Criteria for emotion recognition from verbal and nonverbal expression: Studying baggage loss in the airport. *Personality and social psychology bulletin*. **26**: p. 327-339.
Scherer, K. R., T. Johnstone and G. Klasmeyer (2003). Vocal expression of emotion. *Handbook of the Affective Sciences*. R. J. Davidson, H. Goldsmith and K. R. Scherer. Oxford/New York, Oxford University Press: p: 433-456.
Scherer, K. R. and J. Sangsue (2006). Le système mental en tant que composant de l'émotion. *Cognition et émotions*, Kirouac, G.: p: 11-37.
Scherer, K. R., T. Wranik, J. Sangsue, V. Tran and U. Scherer (2004). Emotions in everyday life: Probability of occurrence, risk factors, appraisal and reaction pattern. *Social Science Information*. **43**: p. 499-570.
Schlosberg, H. (1941). A scale for the judgment of facial expressions. *Journal of Experimental Psychology and Aging*. **29**: p. 497-510.
Schröder, M. (2000). Experimental study of affect bursts. *ISCA workshop "Speech and Emotion"*: p. 132-137.

Schröder, M., E. Zovato, H. Pirker, C. Peter and F. Burkhardt (2007). W3C Emotion Incubator Group Final Report. Published online: http://www.w3.org/2005/Incubator/emotion/XGR-emotion-20070710.

Schuller, B., D. Arsic, F. Wallhoff and G. Rigoll (2006). Emotion recognition in the noise applying large acoustic feature sets. *Speech Prosody*, Dresden, Germany.

Schuller, B., A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous and V. Aharonson (2007a). The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. *Interspeech*.

Schuller, B., S. Reiter, R. Müller, M. Al-Hames, M. Lang and G. Rigoll (2005). Speaker Independent Speech Emotion Recognition by Ensemble Classification. *ICME , 6th International Conference on Multimedia and Expo, IEEE*, Amsterdam, The Netherlands.

Schuller, B., D. Seppi, A. Batliner, A. Maier and S. Steidl (2007b). Towards more Reality in the Recognition of Emotional Speech. *ICASSP, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii**:** p. 941_944.

Shafran, I., M. Riley and M. Mohri (2003). Voice Signature. *IEEE Automatic Speech Recognition and Understanding Workshop***:** p. 31-36

Shaver, P., J. Schwartz, D. Kirson and C. O'Connor (2001). Emotion knowledge: Further Exploration of a Prototype Approach. *Emotions in social psychology*. W. Parrott. Philadelphia, Psychology Press: p: 26-56.

Smith, C. A. and P. C. Ellsworth (1985). Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*. **48**: p. 813-838.

Smith, C. A. and S. Kleinman (1989). Managing emotions in medical school: Students' contacts with the living and the dead. *Social Psychology Quarterly*. **52**: p. 56-69.

Smith, C. A. and R. S. Lazarus (1993). Appraisal Components, core relational themes, and the emotions. *Cognition and Emotion*. **7**: p. 233-269.

Smith, J. O. and J. S. Abel (1999). Bark and ERB Bilinear Transforms. *IEEE Trans. Speech and Audio Proc.* **7**: p. 697-708.

Soren, B. and N. Zacharov (2006). *Perceptual Audio Evaluation - Theory, Method and Application*. Chichester, John Wiley & Sons.

Steidl, S., M. Levit, A. Batliner, E. Nöth and E. Niemann (2005). "Of All Things the Measure is Man" Automatic classification of emotions and inter-labeler consistency. *IEEE International Conference on Acoustics Speech and Signal Processing*.

Teukolsky, S. A., W. T. Vetterling and B. P. Flannery (1992). *Numerical Recipes in Fortran 77, 2nd ed.* Cambridge, U.K., Cambridge University Press.

Traum, D. and P. Heeman (1997). Utterance Units Spoken Dialogue Processing in Spoken Language Systems. *Lecture Notes in Artificial Intelligence*. E. Maier, M. Mast and S. LuperFoy, Springer-Verlag Heidelberg: p: 125-14.

Vapnik, V. N. (1998). The Nature of Statistical Learning Theory. **Springer**.

Vidrascu, L. and L. Devillers (2005a). Annotation and Detection of Blended Emotions in Real Human-Human Dialogs Recorded in a Call Center. *ICME*.

Vidrascu, L. and L. Devillers (2005b). Detection of Real-Life Emotions in Call Centers. *In InterSpeech*, Lisbon.

Vidrascu, L. and L. Devillers (2005c). Real-life Emotions Representation and Detection in Call Centers. *ACII*, Bejing: p. 739-746.

Vidrascu, L. and L. Devillers (2006). Real-life emotions in naturalistic data recorded in a medical call center. *LREC*, Genoa.

Vidrascu, L. and L. Devillers (2007). Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features. *Paraling'*, Saarbrucken, p. 11-16.

Vieru-Dimulescu, B. and P. Boula de Mareüil (2006). Perceptual identification and phonetic analysis of 6 foreign accents in French. *ICSLP*.

Vogt, T. and E. Andre (2005). Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition. *ICME*.

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal behavior*. **17**.

Wierzbicka, A. (1999). *Emotions across Languages and Cultures: Diversity and Universals*, Cambridge University Press.

Williams, C. E. and K. N. Stevens (1972). Emotions and speech : Some acoustical correlates. *Journal of the Acoustical Society of America*. **52**: p. 1238-1250.

Witten, I. H. and E. Franck (2005). *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. San Francisco.

Wrede, B. and E. Shriberg (2003). Spotting "Hot Spots" in Meetings : Human Judgements and Prosodic Cues. *Eurospeech*, Geneva.

Wundt, W. (1896). *Grundrisse der Psychologie [Outlines of psychology]*. Leipzig , Germany.

Yacoub, S., S. Simske, X. Lin and J. Burns (2003). Recognition of Emotions in Interactive Voice Response Systems. *Eurospeech*.

Zwick, R. (1988). Another Look ar Interrater Agreement. Psychological Bulletin. 103: p. 374-378.

# PUBLICATIONS

### Chapitre d'ouvrages

Laurence Devillers, Laurence Vidrascu, *Emotion recognition* in « Speaker classification II », Christian Müller Susanne Schötz (eds.), Springer,-Verlag, p. 34-42.

### Revues

Laurence Devillers, Laurence Vidrascu, *Représentation et Détection des émotions dans des données issues de dialogues enregistrés dans des centres d'appels : des émotions mixtes dans des données réelles*, numéro spécial « Interaction Emotionnelle », Revue Des Sciences et Technologies de l'Information, série Revue d'Intelligence Artificielle, Volume 20, n4-5/2006, p. 447-476.

Laurence Devillers, Laurence Vidrascu, and Lori Lamel. *Challenges in real-life emotion annotation and machine learning based detection.* Journal of Neural Networks, 18/4, 2005, p.407-422.

### Actes de colloques internationaux

Laurence Vidrascu, Laurence Devillers, *Five emotion classes detection in real-world call center data: the use of various* types of paralinguistic features, Paraling'07, 2007, p. 11-16.

Laurence Devillers, Laurence Vidrascu, *Positive and Negative emotional states behind the laugh in spontaneous spoken dialogs* The phonetics of Laughter, Saarland, 2007

Björn Schuller, Anton Batliner, Dino Seppi, Stefan Steidl, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, Vered Aharonson., *The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals*, Interspeech 2007, Anvers pp. 2253-2256

Batliner, Anton; Steidl, Stefan; Schuller, Björn; Seppi, Dino; Vogt, Thurid; Devillers, Laurence; Vidrascu, Laurence; Amir, Noam; Kessous, Loic; Aharonason, Vered. *The Impact of F0 Extraction Errors on the Classification of Prominence and Emotion* In: IPA (Eds.) Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007 Saarbrücken 2007, pp. 2201-2204

Anton Batliner, Stefan Steidl, Björn Schuller, Dino Seppi, Kornel Laskowski, Thurid Vogt, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, Vered Aharonson: *Combining Efforts for Improving Automatic Classification of Emotional User States,* IS-LTC,2006

Laurence Vidrascu, Laurence Devillers, *Real-life emotions in naturalistic data recorded in a medical call center*, LREC,Genoa, 2006

Laurence Devillers, Laurence Vidrascu, *Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs*, Interspeech 2006.

Laurence Vidrascu and Laurence Devillers. *Real-life Emotions Representation and Detection in Call Centers*. In ACII, Bejing, October 2005 : p. 739-746.

Laurence Vidrascu and Laurence Devillers. *Annotation and Detection of Blended Emotions in Real Human-Human Dialogs Recorded in a Call Center*. In ICME, June 2005.

Laurence Vidrascu and Laurence Devillers. *Detection of Real-Life Emotions in Call Centers*. In InterSpeech, Lisbon, September 2005.

Laurence Devillers, Iona Vasilescu, and Laurence Vidrascu. *Anger Versus Fear Detection in Recorded Conversations*. In Speech Prosody, Nara, March 2004.

### Actes de colloques français

Laurence Vidrascu, Laurence Devillers. *Détection de 2 à 5 classes d'émotions sur des données naturelles enregistrées dans un centre d'appel*, RJCP 2007.