Analyse et reconnaissance des émotions lors de conversations de centres d'appels

Christophe Vaudable

▶ **To cite this version:**

Christophe Vaudable. Analyse et reconnaissance des émotions lors de conversations de centres d'appels. Autre [cs.OH]. Université Paris Sud - Paris XI, 2012. Français. NNT : 2012PA112128 . tel-00758650

HAL Id: tel-00758650

https://theses.hal.science/tel-00758650

Submitted on 29 Nov 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Thèse présentée pour obtenir le grade de Docteur de l'université Paris 11

Spécialité : Informatique

Analyse et reconnaissance des émotions lors de conversations de centres d'appels

Christophe Vaudable

LIMSI-CNRS

Ecole doctorale informatique Paris Sud

11 juillet 2012

Membres du jury

- Pierre Zweigenbaum	Président
- Daniel Luzzati	Rapporteur
- Björn Schuller	Rapporteur
- Martine Garnier-Rizet	Examinateur
- Martine Adda-Decker	Examinateur
- Laurence Devillers	Directeur de thèse

Remerciements

Je tiens à remercier mon directeur de thèse, Laurence Devillers, de sa confiance, ses conseils et son enthousiasme pour une collaboration qui à débutée lors de mon stage de Master.

Ma pensée va également à mes rapporteurs, Björn Schuller et Daniel Luzzati qui ont acceptés de relire ce manuscrit malgré des emplois du temps, j'en suis sur, déjà très chargé.

C'est très naturellement que j'envoie également mes remerciements aux trois membres de mon jury Martine Garnier-Rizet, Pierre Zweigenbaum et Martine Adda-Decker

D'une manière plus large je souhaiterai témoigner ma reconnaissance à tous mes collègues du thème « émotion » (je les cite par ordre d'arrivée) Agnès, Marie, Clément, Caroline, Virginie, Julieta et Mariette pour leur contribution à l'excellente ambiance de travail qui règne au sein du groupe.

Je n'oublie évidement pas mes collègues situés dans mon second lieu de travail, EPITECH, pour leurs discussions, leurs idées et leur dynamisme. Pour cela je souhaiterai remercier tout particulièrement Nicolas et Thibault.

Au delà de mon entourage professionnel je tiens bien évidement à saluer très chaleureusement mes deux amis « Mr Poulet » et Thierry Lagrosse pour leur présence et leur soutien indéfectible depuis près de 15 ans.

Cette thèse n'aurait bien évidement pas pu être accomplie sans le soutien de ma famille : sœur, parents et grands-parents qui m'ont apporté encouragements et réconfort lors des moments de doute.

Enfin une mention toute particulière pour Anne-Claire, qui m'a quotidiennement accompagné, soutenu et encouragé notamment sur la dernière partie, sans doute la plus délicate, de cette thèse.

Table des matières

Table des matières ***3***

Introduction générale ***7***

1.1 Les émotions et l'ère des sciences « affectives » 8

1.2 Contexte de l'étude 8

1.3 L'évaluation automatique des communications en centre d'appels 9

1.4 Objectifs et apports de l'étude 10

1.5 Organisation du document 12

Partie 1 ..... ***15***

Chapitre 1 Modèles théoriques ***17***

1.1 Introduction 18

1.2 La représentation des émotions 18

1.3 Représentation des sentiments/opinions 25

1.4 Conclusion 27

Chapitre 2 Les corpus émotionnels : base de données et collecte d'information ***29***

2.1 Introduction 30

2.2 Les types de corpus 30

2.3 Les centres d'appels et les bases de données émotionnelles 34

2.4 Conclusion 35

Chapitre 3 Les indices mis en jeux lors de l'expression d'un état affectif dans un contexte de conversation ***37***

3.1 Introduction 38

3.2 La production vocale : 39

3.3 Indices Prosodiques 39

3.4 Indices linguistiques 41

3.5 Linguistique et parole 43

3.6 La synthèse de la parole émotionnelle : aperçu des méthodes employées 49

3.7 Fusion d'indices : méthodes et stratégies 51

3.8 La détection automatique des émotions dans les centres d'appels 57

3.9 Conclusion 61

Partie 2 ..... ***63***

Chapitre 4 Données utilisées et schéma d'annotation proposé ***65***

4.1 Collecte de données 66

4.2 Le corpus emoVox : description, segmentation des données et stratégies d'annotations 67
4.3 Validation du schéma d'annotation : choix des éléments descriptifs 74
4.4 Présentation du corpus retenu pour l'entrainement de modèles computationnels 86
4.5 Conclusion 90
Chapitre 5 Indices et méthodes de classification pour la détection d'émotions 93
5.1 Les indices 94
5.2 Apprentissage artificiel : algorithmes et méthodes 102
5.3 Quelles méthodes pour tester et évaluer des modèles ? Techniques et mesures employées. 106
5.4 Choix de l'ensemble de descripteurs acoustiques pour la reconnaissance des émotions 109
5.5 Choix du vocabulaire pour la reconnaissance des émotions 114
5.6 Conclusion 117
Chapitre 6 Détection des émotions real-life en centre d'appels : Fusion d'indices linguistiques et paralinguistiques 119
6.1 Vers plus de réalisme pour les systèmes de détection des émotions 120
6.2 Détection automatique des émotions naturelles en centre d'appels : Quels sont les challenges à surmonter ? 120
6.3 Fusion d'indices pour la détection d'émotions 127
6.4 Conclusion 137
Chapitre 7 Enrichissement automatique des modèles paralinguistiques 139
7.1 Introduction 140
7.2 Apprentissage semi-supervisé et sélection active de classes (active class selection) pour l'entrainement de données sur de grands corpus 140
7.3 Enrichissement de corpus : la solution proposée 142
7.4 La sélection de données pour l'enrichissement de modèle 143
7.5 Sélection fine des instances destinées à enrichir le modèle 151
7.6 Quel jeu de descripteurs ? 152
7.7 Enrichissement d'un modèle lexical 153
7.8 Enrichissement avec 100 heures de dialogues choisis aléatoirement 154
7.9 Conclusion 155
Chapitre 8 Evaluation de la qualité d'une conversation : vers une signature émotionnelle d'un dialogue 157
8.1 Introduction 158
8.2 Données utilisées 158
8.3 Test sur les débuts et fins de dialogues : résultats d'après une annotation manuelle des données 162

8.4 Test sur les débuts et fins de dialogues : résultats d'après une annotation automatique 164
8.5 Ajouts d'informations concernant le rôle du locuteur pour la classification au niveau global. 165
8.6 Prise en compte d'une séquence de prédictions émotionnelles 166
8.7 Validation des indices structurels avec un modèle automatique 167
8.8 Indices supplémentaires pour l'évaluation de la modalité négative ou positive d'un dialogue169
8.9 Combinaisons des indices, structurels et conversationnels pour la classification globale d'un dialogue 177
8.10 Conclusion 178
Conclusions et perspectives ***179***
8.11 Conclusions 179
8.12 Perspectives 180
Table des Figures ***182***
Listes des tableaux ***184***
Bibliographie ***187***

Introduction générale

Les fondements même de la société dans laquelle nous vivons sont fondés sur la communication. Cette communication, peut être vocale, gestuelle, textuelle répond à un ensemble de règles précises permettant à chacun de comprendre et de se faire comprendre. D'une manière générale ces formes de communication ont toutes pour but de traduire la pensée, de la représenter à l'aide d'un ensemble de mots choisis dans un lexique, de gestes répondants à une culture, de sons articulés dans le but de former des syllabes, des mots, des phrases.

Bien avant le langage écrit, la parole est l'un des plus anciens modes de communications, nombre de sociétés se sont fondées sur une tradition orale uniquement. Afin d'être intelligible pour autrui la parole est organisée selon les règles syntaxiques et sémantiques de la langue considérée. Ces règles, aussi importantes soient elles pour la communication, ne sont cependant pas suffisantes pour définir ce qu'est la parole. Au delà des mots et du contenu sémantique du message c'est le corps tout entier qui s'exprime à travers la vocalisation. La communication orale ne nous permet-elle pas de percevoir la tristesse ou la joie dans le message qui nous est adressé ? Ne nous est-il pas possible de percevoir une souffrance physique sans que cette dernière ne soit explicitement verbalisée ? C'est ainsi que la parole à travers la modulation de la voix permet la transmission d'un spectre très large d'émotions, de sentiments, de désirs et d'intentions.

Lors d'une communication orale, la verbalisation et le contenu sémantique et lexical permettent de préciser le message émis, de structurer la conversation et d'établir un dialogue. Le message écrit nécessite un degré d'abstraction supplémentaire par rapport à la seule conversation sonore car la communication d'un état mental ou affectif se doit d'être explicitée afin d'être comprise. Ce mode de communication ajoute ainsi une étape supplémentaire entre un ressenti et son expression.

Les règles sociales en vigueur imposent d'une manière générale une certaine retenue dans l'expression de nos états affectifs. Dans une conversation, nous voyons de ce fait se mettre en place une forme d'expression émotionnelle complexe mêlant expression verbale et vocale. Une communication qui n'a plus seulement pour but de transmettre un message mais également de montrer une forme de respect des normes sociales.

« *Les aspects paralinguistiques de la communication sont les éléments de signification (conventionnels ou non) produits au niveau de la face (et des membres, dans la modalité orale) parallèlement à la production de la parole ou du signe, ainsi que la prosodie, les attitudes et les postures corporelles, dans la mesure où tous ces éléments permettent de nuancer, de préciser, voire de contredire les messages transmis simultanément*. » (Losson 2000)

La communication orale ouvre ainsi la voie à un très large champ d'étude, proposant de comprendre, d'étudier, de formaliser voire de synthétiser la communication paralinguistique des états expressifs tels que les émotions.

1.1Les émotions et l'ère des sciences « affectives »

Dès l'antiquité, des philosophes tels qu'Aristote dans son ouvrage l'*Ethique à Nicomaque* se sont intéressés à l'étude des phénomènes affectifs en distinguant les affects (*pathè*) et le raisonnement logique (*prohairèsis*). Aristote note qu'émotion et raison contribuent ensemble à la définition de l'être. Au XVIIème siècle d'autres auteurs tels que Descartes ou Spinoza s'intéressent également aux émotions en décorrélant complètement les états affectifs de la raison pour le premier, le second plaidant quant à lui pour un rôle des affects dans le processus de raisonnement. L'étude des états affectifs et plus particulièrement des émotions a connu un nouvel essor il y a 150 ans avec des auteurs comme (Darwin 1872) qui s'est intéressé aux formes d'expressions émotionnelles des hommes mais également des animaux. Darwin note que les émotions primaires sont universelles au niveau de l'expression faciale. Un siècle plus tard des chercheurs comme Scherer proposent un nouveau thème d'étude : les sciences affectives avec une série d'études portant sur l'expression émotionnelle vocale et/ou l'universalité d'expression et de perception des émotions chez des sujets de cultures différentes (Scherer et al. 1972; Ekman and Friesen 1975; Scherer 1981; Scherer 1984). Cependant, comme nous allons le voir et comme le mentionne encore récemment Scherer, *« Defining ''emotion'' is a notorious problem »,* tant les définitions utilisées pour définir ce qu'est une émotion varient suivant les disciplines *(Scherer 2005)* . Depuis le début des années 2000 et l'ouvrage « *Affective Computing »* (Picard 1997) le nombre de publications et conférences scientifiques dédiées aux états affectifs n'a cessé de croître. Le nombre de papiers acceptés dans les conférences autour du thème de la détection des émotions dans la voix a été multiplié par 100 par rapport à l'année 2000[1]. Des réseaux d'excellence (ReX) comme les réseaux HUMAINE et SSPNet principalement dédiés à l'étude des états affectifs et des interactions sociales ont permis l'organisation de plusieurs conférences et workshop (Affective Computing and Intelligent Interaction par exemple) consacrés à l'affective computing ainsi que l'édition de plusieurs ouvrages permettant de synthétiser les avancées du domaine (Sander and Scherer 2009; Scherer et al. 2010; Douglas-Cowie et al. 2011; Petta et al. 2011). La mise en place de challenges réunissant plusieurs laboratoires comme CEICES (*Combining Efforts for Improving automatic. Classification of Emotional user States*) mettant en jeu une collaboration entre plusieurs équipes du réseau HUMAINE (Batliner et al. 2006) a été l'occasion de constater les avancées réalisées au cours des dernières années. De nombreuses thèses sont soutenues chaque année sur ce thème (Abrilian 2007; Clavel 2007; Vidrascu 2007; Steidl 2009; Mahdhaoui 2010; Mower 2010), traduisant un enthousiasme grandissant de la part de la communauté scientifique.

1.2 Contexte de l'étude

L'étude présentée dans cette thèse se situe dans le contexte de la détection d'émotions lors d'une interaction téléphonique entre deux humains dans des centres d'appels. Dans ce type d'interactions, le téléphone est le seul canal de communication. Cela rend les conversations téléphoniques particulièrement intéressantes à étudier pour nous puisque la voix devient le seul canal d'expression possible. Nous disposons dès lors d'une expression émotionnelle vocale d'une grande richesse ainsi

[1] Source Scopus : http://www.info.sciverse.com/scopus

que de stratégies de communication élaborées (exagération de l'expression émotionnelle pour convaincre, ironie).

Le corpus collecté ainsi que les études menées ont été réalisés dans le cadre du projet national FUI voxFactory soutenu par le pôle de compétitivité Cap Digital[2]. Ce projet propose d'analyser la qualité et l'efficacité de l'interaction client/télé-conseiller en s'appuyant sur les technologies de la reconnaissance de la parole, de la détection des émotions (audio mining) et de la fouille de données textuelles (text mining). L'approche est mixte fondée d'une part sur l'analyse textuelle de l'appel, à partir de sa transcription automatique, et d'autre part sur la prise en compte de paramètres paralinguistiques extraits du signal de parole pour la détection des émotions notamment.

Au cours du projet voxfactory les données ont été collectées par les sociétés EDF et Vecsys. Le corpus final à été traité à différents niveaux par plusieurs partenaires :

- Les transcriptions automatiques ont été fournies par la société Vocapia ainsi que l'équipe TLP du LIMSI
- Les annotations automatiques des thèmes des appels par les sociétés EDF, Sinequa et Temis
- Les annotations automatiques des opinions par les sociétés Sinequa et Temis
- Les transcriptions manuelles des conversations et des disfluences par la société Vecsys

1.3 L'évaluation automatique des communications en centre d'appels

L'analyse de conversations orales, « speech analytics », est un nouveau domaine qui ouvre de nombreuses applications pour une meilleure gestion des clients des centres d'appels. En utilisant ce type d'analyse, les centres d'appels peuvent suivre des conversations entières et identifier des sujets problématiques, repérer les interactions mettant en jeu des clients non satisfaits, identifier les agents fatigués, ceux qui auraient besoin d'une formation supplémentaire, etc.

Le but de cette thèse est la détermination de la satisfaction client pour qualifier les conversations téléphoniques entre agent et client à l'aide de la détection automatique d'émotions, considérées comme signes de contentement ou de mécontentement de la part du client. Dans le domaine de la relation clientèle deux objectifs primordiaux se distinguent. Le premier est bien évidement la conquête de nouveaux clients grâce aux diverses stratégies marketing et publicitaires. Le deuxième est la fidélisation des clients existants pour laquelle la connaissance de la satisfaction client est primordiale. Le téléphone, via les centres d'appels, est le moyen privilégié pour la détermination de cette satisfaction clientèle. Les récentes avancées de l'audio et du text-mining permettent la proposition d'une mesure automatique de la satisfaction client. Deux types d'utilisation de cette technologie sont possible i) Lors d'une interaction homme-machine ii) dans un cadre de fouille de données. C'est dans ce deuxième type d'utilisation que se situent les travaux de cette thèse.

L'analyse de la voix et de l'interaction, le traitement du langage, sont autant de moyens pour reconnaître et détecter le degré de satisfaction du client lors de son interaction avec le centre d'appels de l'entreprise. Le projet de détection des émotions mis en œuvre dans cette thèse s'appuie sur ces différentes analyses afin de fournir un modèle innovant de mesure de la satisfaction.

[2] http://www.capdigital.com/

1.3.1 Applications et retombées industrielles

D'un point de vue applicatif il n'existe pas, à notre connaissance, de produits commerciaux mettant en œuvre des systèmes de détection d'émotions pour le français. Quelques produits proposent des méthodes pour faciliter la recherche de conversations problématiques dans des bases de données contenant des enregistrements d'appels à l'aide de mots clés (en se basant sur les transcriptions des dialogues) ou du nombre de pauses (silences) contenues dans les conversations. Les outils les plus élaborés proposent des recherches par thématique d'appel (coupure, demande d'informations, etc) grâce aux techniques de « text mining ». Cependant, aucun n'effectue une mise en relation des indices acoustiques et linguistiques permettant d'obtenir une qualification beaucoup plus fine de l'appel. Il faut de plus noter que la plupart de ces produits sont basés sur la langue anglaise. Aux Etats-Unis le marché est florissant et 14 acteurs ont été recensés lors d'une étude Gartner en 2007[3]. Les trois principales sociétés étant Autonomy etalk[4], Nice Systems[5] et Verint[6] . D'après les sondages réalisés par les sociétés spécialisées dans les solutions d'hébergement/gestion de centre d'appels, 70% des clients se déclarent intéressés par l'acquisition d'outils permettant l'évaluation automatique de la qualité de la relation client-téléconseiller. Ces sociétés y voient un moyen d'améliorer la qualité de leur service mais également leur productivité. Deux modes de traitement principaux peuvent être envisagés : un suivi en temps réel permettant le suivi des conversations à problèmes par un superviseur et une approche en différé prenant en compte l'ensemble des évènements survenus au cours de la conversation. Pour cette seconde approche, qui est celle retenue pour cette thèse, les applications envisagées sont nombreuses et variées. Nous citons quelques exemples ci-dessous :

- Evaluation automatique de la qualité d'un appel d'après les affects exprimés au cours de la conversation,
- Conception de nouveaux plannings de travail (renforcement des équipes par exemple) en fonction du pourcentage d'appels se déroulant bien ou mal au cours de la journée,
- Formation des agents sur des cas d'écoles repérés automatiquement dans la liste d'appels problématiques.

Ce marché en pleine expansion est estimé à plusieurs dizaines de millions d'euros dans les années à venir et représente donc un enjeu stratégique pour les éditeurs de produits informatiques spécialisés dans les traitements vocaux.

1.4 Objectifs et apports de l'étude

A travers les travaux menés pendant cette thèse notre objectif était la mise en place d'un système permettant la mesure de la satisfaction client à travers les émotions exprimées par ce dernier à partir de corpus de données conversationnelles. Les expériences ont été menées sur un corpus de centre d'appel contenant un très grand volume de données. Ce projet offre plusieurs challenges d'un point de vue recherche qui sont les points les plus originaux de la thèse :

[3] MarketScope for Contact Center Quality Management, 2007, Gartner RAS Core Research Note G00151329
[4] Autonomy etalk : http://www.autonomy.com/
[5] Nice Systems :
[6] Verint : http://verint.com/corporate/

- L'utilisation de données naturelles implique des difficultés de traitements supplémentaires liés aux erreurs de segmentation et de reconnaissance. Ces erreurs sont notamment dues aux problèmes de parole superposée comme cela est couramment le cas dans les dialogues spontanés. Le fait d'utiliser des conversations enregistrées sur un seul canal audio (et non deux comme ce devrait être le cas dans les futurs enregistrements dans des centres d'appels) ajoute à la difficulté de la tâche. La combinaison des indices sur les deux modes paralinguistiques et linguistiques (indices acoustiques, sémantiques, structurels, dialogiques) dans les systèmes de détection des émotions est nécessaire afin d'essayer d'avoir des modèles plus robustes à des données extrêmement variables (locuteurs, bruits, accents, …) et avec des unités segmentales pouvant être erronées (par exemple un segment peut comprendre du signal venant du client et de l'agent). Dans cette thèse nous comparons les taux de détection de la valence (émotion positive/négative ou neutre) sur des segments de parole obtenus automatiquement ou manuellement et proposons des scores basés sur des tours de parole mais également sur le dialogue global. Nous nous basons pour cela sur des indices linguistiques et paralinguistiques mais également sur une combinaison, à différents niveaux, de ces derniers.
- L'important volume de données (1620 heures de conversation) dans notre projet est un des aspects les plus importants de ces challenges. D'un côté, il est particulièrement intéressant d'avoir accès à un grand nombre de données pour construire un modèle computationnel permettant de détecter les émotions. D'un autre côté il est compliqué de sélectionner, segmenter et annoter des données émotionnelles naturelles, les prétraitements manuels étant chers en temps humain. Des études précédentes sur les centres d'appels (Devillers et al., 2005) montrent que de 10% à 30% des tours de parole des dialogues sont émotionnellement marqués suivant les tâches (centre de renseignements, de services médicaux, d'appels d'urgence,…). La construction de modèles nécessite en effet de choisir des données appropriées, si possible en grand nombre et de regrouper ces dernières au sein de classes homogènes et équilibrées en terme de nombre d'instances. La sélection de ces données oblige à la mise en place de stratégies intervenant à plusieurs niveaux de recherche (acoustiques, linguistiques) dans le but d'extraire de manière automatique les données émotionnelles. Dans cette thèse, nous proposons des méthodes automatiques pour enrichir nos modèles avec de nouvelles instances.
- Les travaux réalisés en analyse conversationnelle montrent qu'une conversation ne met pas en œuvre des tours de parole isolés mais une construction ayant pour unité dialogique de base la paire adjacente. La co-construction de la conversation est donc un concept important qui implique de ne plus traiter isolément les tours de parole mais de considérer le dialogue comme un tout. Dans ce cas la notion d'erreurs de segmentation liées au backchannel ou à la parole superposée n'est plus aussi importante. Nous considérerons dès lors le couple client/agent comme un seul système co-évoluant en fonction des activités menées. La prise en compte d'indices dialogiques permet d'obtenir des modèles de détection plus robustes du fait de la combinaison d'indices.

1.5 Organisation du document

Cette thèse est organisée suivant deux grandes parties. La première présente le contexte dans lequel intervient cette thèse et met en avant les particularités i) des données utilisées (données naturelles) et ii) les raisons théoriques des approches présentées.

Trois chapitres composent cette première 1ere partie (chapitres 1, 2 et 3). Le chapitre 1 s'attache à présenter les théories existantes dans le domaine de la représentation des émotions ainsi que les différents indices acoustiques et linguistiques mis en œuvre lors d'une expression affective. Il démontre qu'une approche pluridisciplinaire est indispensable pour l'élaboration de modèles de détection se basant sur des données naturelles. Nous ferons dans le chapitre 2 un récapitulatif des différents types de corpus utilisés dans la communauté ainsi que leurs avantages et inconvénients pour la construction de modèles automatiques. Ce chapitre permet également de présenter un état de l'art des différentes méthodes d'annotation. Le chapitre 3 a pour objet de présenter les modèles computationnels mis en œuvre lors d'études sur les émotions. Nous nous attarderons particulièrement sur l'existant en matière de modèles de détection appliqués aux données recueillies en centre d'appels.

La partie 2 présente les travaux réalisés et met en avant i) la combinaison des indices pour améliorer les systèmes de détection des émotions, ii) l'enrichissement des modèles par ajout automatique de données et enfin iii) l'exploitation du dialogue entier pour trouver une mesure de valence (positive, négative, neutre) appelée signature émotionnelle du dialogue.

Cinq chapitres sont compris dans la deuxième partie de cette étude (chapitres 4, 5, 6, 7, et 8) et portent sur les travaux que nous avons menés sur les corpus de centres d'appels sélectionnés. Ainsi, dans le chapitre 4, nous présentons le schéma d'annotation, les taux d'accord inter-annotateurs ainsi que les résultats obtenus sur les expériences relatives aux annotations que nous avons effectuées. Nous détaillerons également les données qui composeront les modèles d'apprentissage utilisés pour la suite de l'étude. Le chapitre 5 fait quant à lui un récapitulatif des principaux algorithmes d'apprentissage utilisés et présente les méthodes que nous avons employées pour l'extraction des indices acoustiques et linguistiques. Ce chapitre nous permet de valider chacun des ensembles d'indices retenus. Dans le chapitre 6, deux études portent sur l'impact du mode de segmentation sur les performances des modèles de détection ainsi que les résultats que nous obtenons lors de la combinaison des différents types d'indices (acoustiques, lexicaux, sémantiques) que nous avons retenus pour l'étude. Dans le chapitre 7, nous avons expérimenté des méthodes pour l'enrichissement automatique de nos modèles à l'aide de données sélectionnées automatiquement pour améliorer les performances des modèles. L'approche finalement retenue se base sur une sélection de nouvelles instances destinées à être ajoutées au corpus d'entrainement d'après une sélection automatique. Cette sélection est effectuée dans une base de données grâce à deux types d'indice linguistique/acoustique. Le chapitre 8 propose quant à lui de ne pas considérer le dialogue comme une succession de tours de parole indépendants mais comme un objet co-construit. Nous incluons dans ce chapitre des indices relatifs à la structure d'une conversation (durées, temps de parole des locuteurs, etc.). Nous voyons ainsi dans quelle mesure il est possible de considérer un dialogue complet pour une tâche de classification en vue d'une application métier. Nous avons tiré

de ces différentes expérimentations un certain nombre de résultats, de bonnes pratiques et de réflexions qui seront présentés dans la conclusion de cette thèse.

Partie 1

La première partie de cette thèse, composée des chapitres 1, 2 et 3 à pour but d'exposer le contexte de l'étude.

Le chapitre 1 passe ainsi en revue les différents modes de représentations de concepts tels que les émotions, les sentiments et les opinions mais aussi de la théorie de l'énonciation. Les modélisations à base d'étiquettes, de dimensions continues ou discrètes ainsi que les théories d'évaluation cognitive sont présentées

Le chapitre 2 s'intéresse quant à lui aux corpus disponibles au sein de la communauté. Nous commençons ainsi par récapituler les différents types de corpus existants (actés, induits, naturels). Nous abordons ensuite le problème de la disponibilité et des conditions d'accès des données. Dans une dernière partie, nous présentons les corpus de centres d'appels existants.

Dans le chapitre 3, nous mettons en avant la complexité de la parole en montrant le lien inextricable entre expression verbale et prosodique. Nous insistons sur la complémentarité de ces deux approches lorsque nous nous plaçons dans la situation d'une détection automatique. Afin de pouvoir construire des modèles hybrides (acoustiques et linguistiques) nous présentons également les différentes méthodes de fusions d'indices utilisés (précoce et tardive).Nous finissons la partie 1 en récapitulant les différents types d'indices qui ont été combinés lors d'études portant sur les centres d'appels ainsi que les gains obtenus suite à cette fusion.

Chapitre 1 Modèles théoriques

Résumé

Nous présentons dans ce chapitre les différents modes de représentations utilisés dans la littérature pour modéliser des phénomènes aussi divers que les émotions, les sentiments ou les opinions. Nous verrons que la définition même de ces trois concepts pose encore problème et que des recouvrements dans la représentation de ces derniers existent.

Au long de ce chapitre nous nous attacherons dans un premier temps aux modes de représentation des émotions en présentant les différents types de théories utilisées pour les représenter (étiquettes, dimensions, etc.). Nous mentionnons également dans ce chapitre le langage de description emotionML proposé par le W3C pour essayer de normaliser les éléments traditionnellement mis en jeu pour la représentation d'une émotion.

Dans une seconde partie nous présentons les théories liées à la modélisation de phénomènes comme les sentiments ou les opinions à travers les représentations utilisées dans la théorie de l'évaluation ou encore la théorie de l'énonciation de Charaudeau.

1.1 Introduction

Dans cette thèse nous nous basons sur deux groupes de concepts différents pour réaliser nos analyses. Le premier est celui se référant aux émotions qui seront modélisées à partir de la voix de participants lors d'une conversation téléphonique. La différenciation entre des états affectifs tels que les émotions et les humeurs par exemple n'est pas une tâche triviale. Le second groupe est celui des opinions/sentiments. Ces deux concepts sont regroupés car, comme nous le verrons dans les paragraphes consacrés, les termes *d'opinion mining* et de *sentiment analysis* sont en pratique utilisés comme synonymes dans la littérature. Les recherches décrites sur les sentiments et les opinions ont été effectuées à partir des transcriptions de conversations et sont issues des études fournies par (Cailliau et al. 2010). Le support de ce deuxième groupe de concept est le langage écrit.

1.2 La représentation des émotions

Scherer souligne trois points principaux qui font des émotions des phénomènes de communication par excellence entre un individu et le monde extérieur (Scherer 1984) :

- L'évaluation de la situation par l'individu selon ses besoins et objectifs,
- La préparation à l'action selon différentes modalités (physiologique, psychologique),
- La communication des états internes et/ou des intentions à l'environnement extérieur.

Il n'existe pas réellement de définition d'une émotion ou d'un état émotionnel faisant consensus au sein de la communauté. La multiplication des domaines impliqués dans l'étude des émotions (psychologues, linguistes, acousticiens, informaticiens, etc.) en est probablement une des explications.

Depuis maintenant plus d'un siècle, des auteurs comme (James 1884) ou plus récemment (Scherer 2005) ont proposé différentes définitions d'une émotion. Scherer met en avant les nombreuses définitions entretenues par les auteurs autour de ce thème. Ainsi dans (Kleinginna and Kleinginna 1981), les auteurs relèvent plus de 100 définitions différentes du terme émotion. Pour tenter de clarifier les divergences entre les différents états affectifs, Scherer dans (Scherer and Peper 2001) propose de les classifier selon différents critères (Intensité, durée, etc.). Le Tableau 1-1 ci-dessous récapitule cette classification en cosidérant les modalités d'intensité, de durée ou de vitesse d'adaptation pour différents états affectifs (émotions, humeurs). Les systèmes actuels de détection des émotions considèrent sous le terme « émotion » le concept plus large d'état affectif incluant émotion, attitude, humeur, etc.

Phénomènes affectifs	Intensité	Durée	Synchronisation	Causalité	Evaluation	Vitesse d'adaptation	Impact sur le comportement
Emotions	+++++	+	+++	+++	+++	+++	+++
Humeurs	+++	++	+	+	+	++	+
Positions relationnelles	+++	+++	+	++	+	+++	++
Attitudes (préférences)	++	+++++	0	0	+	+	+
Dispositions affectives (traits de personnalité)	+	+++	0	0	0	0	+

Tableau 1-1 Classification des différents états affectifs (Scherer and Peper 2001)

Pour apprendre à une machine à détecter des états affectifs, il est nécessaire d'adopter un mode de représentation qui soit adapté à la tâche que nous souhaitons réaliser. Un grand nombre de théories existe dans ce domaine ; nous proposons ainsi une revue de ces théories dans les sections 1.2 et 1.3.

1.2.1 La théorie des émotions discrètes

Elle est fondée sur l'idée d'un nombre limité d'émotions primaires. En combinant ces émotions primaires, on obtient une infinité d'émotions. Il y a beaucoup de réponses à la question « Qu'est-ce qu'une émotion ? » et peu de consensus ; mais les recherches s'accordent toutefois sur l'existence d'un ensemble minimal d'émotions primaires dont le nombre varie très largement selon les chercheurs (Tableau 1-2).

Malgré le caractère somme toute relatif de la classification et du nombre des émotions, il existe une sorte de hiérarchie entre elles, dans la mesure par exemple où l'on est capable de chasser les moins instinctives d'entre elles par un acte de volonté. La distinction entre émotions primaires (ou élémentaires) et secondaires (ou sociales) est largement utilisée. Celles qui sont dites primaires pourraient être celles survenant de manière instinctive et dont il ne serait que très difficilement possible de faire abstraction lorsqu'elles surviennent.

Chercheurs	Emotions primaires
Plutchik	Acceptation, colère, anticipation, dégoût, joie, peur, tristesse, surprise
Arnold	colère, aversion, courage, découragement, désir, désespoir, peur, haine, espoir, amour, tristesse
Ekman	colère, peur, joie, tristesse, dégoût, surprise
Frijda	désir, intérêt, bonheur, surprise,
Gray	rage, terreur, anxiété, joie
Izard	colère, mépris, dégoût, détresse, peur, culpabilité, intérêt, joie, honte, surprise
James	peur, douleur/chagrin, amour, rage
Darwin	colère, peur, joie, tristesse, dégoût
Mowrer	souffrance, plaisir
Oatley	colère, dégoût, inquiétude, bonheur, tristesse

Tableau 1-2 : Les émotions primaires (élémentaires)

D'après P. Ekman le groupe des 6 émotions primaires (« big six ») qu'il utilise serait universel et donc présent aussi bien chez l'homme que chez l'animal (Ekman 1999). R. Plutchick (Plutchick 1984) qui postule huit émotions primaires (la joie, l'acceptation, la peur, la surprise, la tristesse, le dégoût, la colère, et l'anticipation) a inscrit dans une roue ces émotions élémentaires qui, en se combinant, produisent d'autres émotions : l'amour serait la résultante des émotions élémentaires de joie et d'acceptation ; la soumission résulterait de l'acceptation et de la peur. Il présente les émotions secondaires comme des mélanges d'émotions primaires. Cette approche assez simplificatrice n'est cependant qu'une première ébauche pour mélanger les émotions.

1.2.2La théorie dimensionnelle (Wundt 1913)

La théorie dimensionnelle décrit l'état émotionnel suivant trois dimensions : envie-aversion (positif/négatif), excitation-apaisement (actif/passif), contrôle-relâchement (de soi même). Cette troisième dimension est considérée comme étant moins stable. C. Osgood (Osgood et al. 1975) reprend ces trois dimensions de la façon suivante : l'Evaluation, l'Activation et le Pouvoir. Ce sont des dimensions continues, souvent discrétisées sur des échelles de 3 à 7 graduations. La dimension

d'Evaluation (ou Valence) a une graduation entre négative et positive ; la dimension d'Activation entre active et passive ; et la dimension de Pouvoir (potentiel de critique) entre fort et faible. Cette troisième dimension est nommée posture (S, stance, pour (Breazeal 2003)) ou encore contrôle, ouverture, ou dominance (Mehrabian and Russel 1974), et permet de distinguer les émotions donnant lieu à des réactions d'approche et de combat (comme la colère) de celles engendrant des comportements d'évitement ou de fuite (comme la peur). Ainsi on peut définir les émotions à partir de dimensions continues abstraites plutôt que de les nommer explicitement avec des catégories discrètes.

Les deux dimensions les plus largement utilisées tout supports d'études confondus (langage écrit et parlé, indices multimodaux, indices physiologiques) pour les études sont la valence et l'activation (Jones and Troen 2007; Schuller et al. 2009; Stickel et al. 2009). Ces dimensions peuvent être employées à travers des outils comme Feeltrace (Cowie et al. 2000) (Figure 1-1). Cependant d'autres dimensions sont nécessaires, par exemple pour distinguer la peur de la colère. De nouvelles dimensions abstraites ont été suggérées par les théories de l'évaluation, par exemple un axe par rapport au but favorable ou défavorable de l'évènement déclencheur d'émotions, i.e. le potentiel de chaque individu à gérer les conséquences de l'évènement émotionnel. Il convient de souligner que les étiquettes verbales et les états émotionnels peuvent être appréhendés facilement au travers d'un petit nombre de dimensions, telles la valence ou l'activité. Ces dimensions représentent des dispositifs descriptifs qui sont intéressants en eux-mêmes mais qui ne nous fournissent pas d'éclaircissement sur la nature de l'émotion.

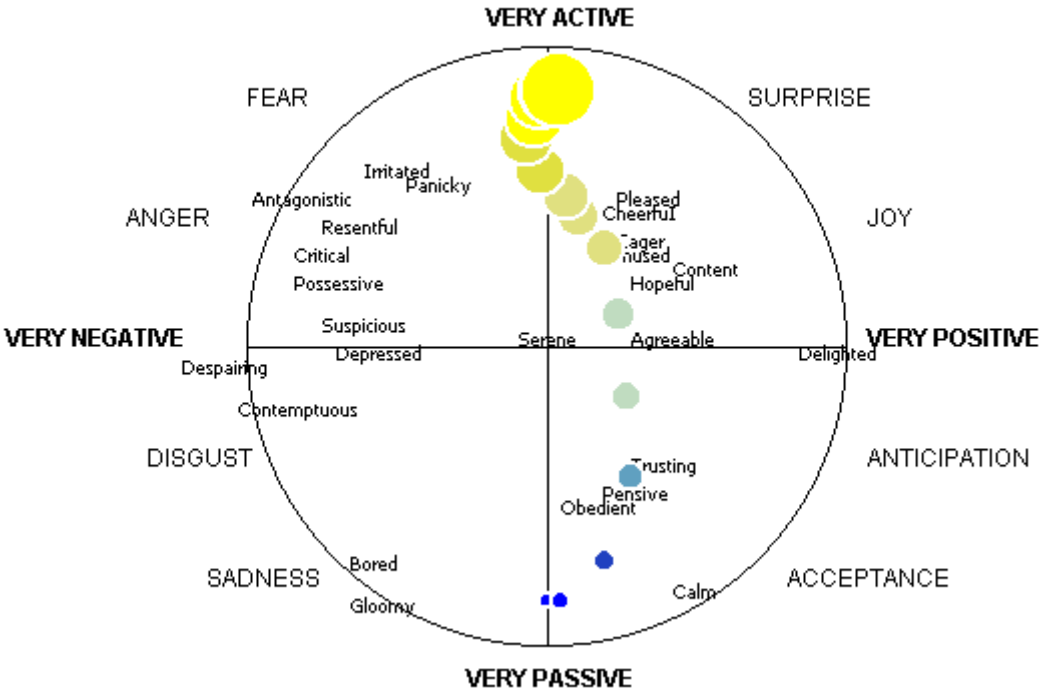


Figure 1-1 : *Feeltrace (Cowie et al. 2000)*

De plus en plus d'études comme (Eyben et al. 2008) utilisent des dimensions continues comme mode de représentation des émotions, mais cette approche est plus récemment utilisée dans les systèmes de détection ou de génération des émotions.

Une approche mixte mélangeant représentation discrète des émotion ainsi que théories dimensionnelle est présentée dans (Scherer 2005) à travers la « Geneva Emotion Wheel » telle que présentée dans la Figure 1-2 ci-dessous :

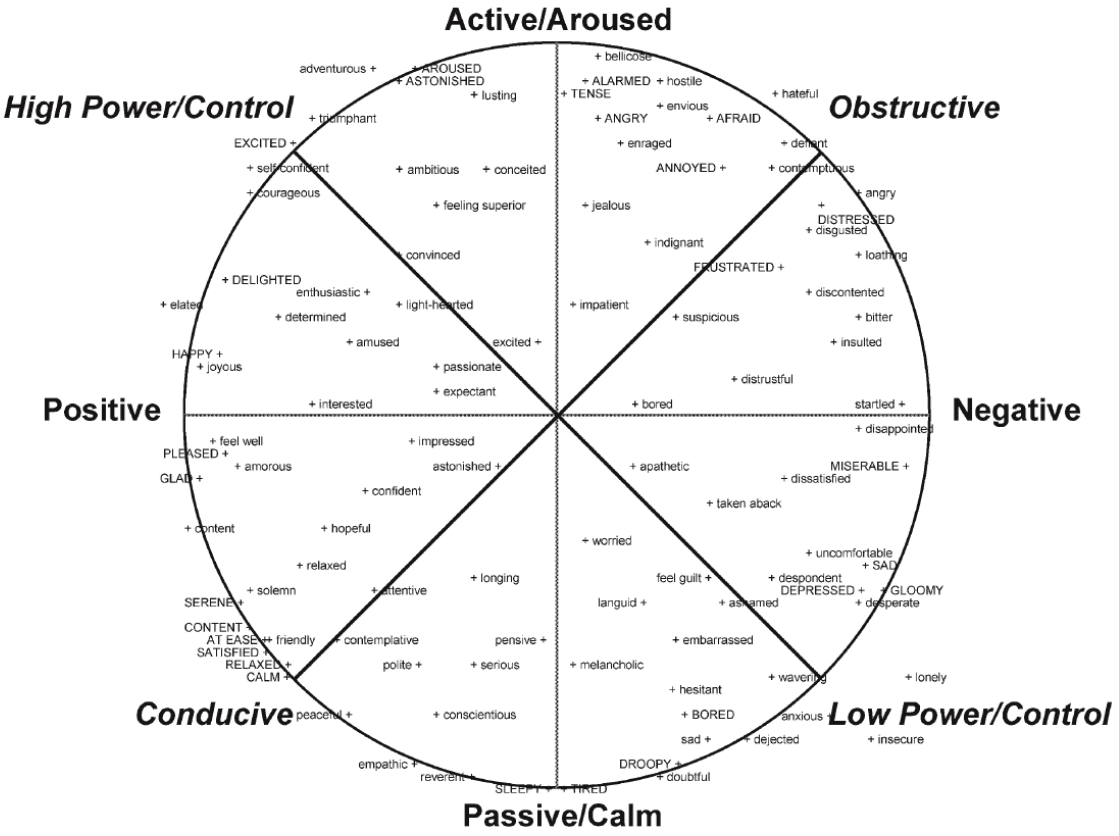


Figure 1-2 La roue des émotions (Scherer 2005) utilisant une approche mixte dimensions/étiquette

Dans cette représentation quatre dimensions sont utilisées, permettant de délimiter d'après leurs intersections graphiques 8 zones émotionnelles distinctes (exemple : positif avec une activation élevée et un fort contrôle). Dans chacune de ces zones une série d'étiquettes, retenues d'après une sélection, sont projetées dans le repère. Cette méthode a donc l'avantage de pouvoir définir un état émotionnel en deux temps, i) suivant une série de critères s'inscrivant sur des échelles (les trois dimensions) et ii) définir de manière plus précise à l'aide d'une étiquette l'état émotionnel auquel il est fait référence.

1.2.3 Les théories cognitives de l'évaluation

Ces théories considèrent que la nature de l'émotion est déterminée par une évaluation cognitive. Le modèle conçu par K. Scherer pour représenter cette évaluation a été nommé « *component process model* » (Scherer and Peper 2001). Cette modélisation offre une spécification détaillée des dimensions d'évaluation qui sont supposées être utilisées dans l'évaluation des évènements antécédents à l'émotion. C'est un modèle à plusieurs composantes. Un épisode émotionnel implique dans ce cadre théorique d'évaluer la pertinence d'un évènement par rapport aux buts à atteindre, de créer un état de préparation à l'action chez l'individu et s'accompagne de l'expérience subjective d'états mentaux distincts de nos états habituels.

La principale différence avec les représentations précédentes est de considérer que **l'émotion est un phénomène dynamique et non un état statique** défini par une catégorie statique. Pour K. Scherer, une émotion est une séquence de changements internes liés et synchronisés avec l'état de différents systèmes de l'organisme en réponse à l'évaluation d'un stimulus externe ou interne. « L'émotion est donc un processus cumulatif, une unité temporelle au cours de laquelle les différents composants de l'émotion se synchronisent et se désynchronisent ».

Ces différents composants sont la composante d'évaluation, de tendance à l'action, d'activité physiologique, d'expression motrice et enfin la composante de sentiment subjectif (feeling) reflétant les changements survenus dans les autres composants.

K. Scherer introduit la notion de « Stimulus Evaluation Check » (SEC) et 5 types de SEC (total de 18 variables) allant des plus universels aux plus spécifiques:

- **Nouveauté** (Soudaineté, Familiarité, Prévisibilité) : caractère inattendu ou non de l'évènement,
- **Agrément** (Intrinsèque, Global (désirabilité)) : expérience plaisante ou déplaisante,
- **Rapports aux causes et buts** (Causalité interne, Causalité externe, Pertinence, Degré de certitude dans la prédiction des conséquences, Attentes, Opportunité, Urgence),
- **Potentiel de maîtrise** (Contrôle de l'évènement, Contrôle des conséquences, Puissance, Ajustement) : possibilité de s'adapter,
- **Accord avec les standards** (Externes, Internes) : accords aux normes sociales et concepts de soi.

Après plus de 15 ans de recherches, (Martin and White 2005) ont publié la théorie de l'évaluation, connue en anglais sous le nom d'*appraisal theory*, celle-ci étant une extension des théories appelées *Systemic Functional Linguistics* (Halliday 1994).

Parmi les travaux d'analyse automatique basés sur ce cadre théorique, on retrouve (Whitelaw et al. 2005). L'analyse se base sur un lexique d'adjectifs classés selon les quatre types d'attributs qu'on trouve dans la Figure 1-3. Ces quatre types ont de nombreuses options, dont une est également illustrée avec les adjectifs associés dans la même figure.

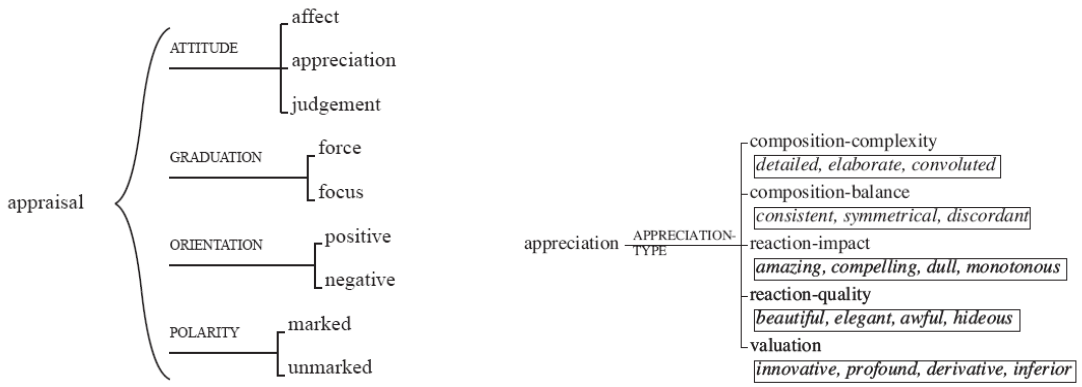


Figure 1-3 : Types d'attributs de l'Appraisal Theory et options du type appréciation avec exemples d'adjectifs

La ci-dessus illustre comment l'analyse d'un groupe d'évaluations est effectuée. En général, le groupe est constitué uniquement de l'adjectif.

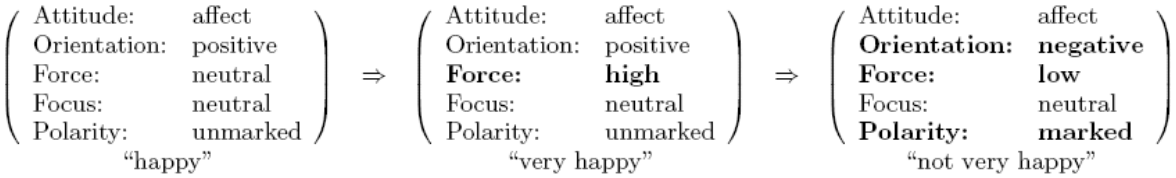


Figure 1-4 : Analyse du groupe d'évaluations *not very happy*

L'identification de l'origine et de la cible sont également des éléments centraux dans cette approche. (Bloom et al. 2007) et (Bloom et al. 2007) se situent dans le même cadre et continuent leurs travaux dans la même lignée.

Ces travaux ont donnés lieu à la mise à disposition du lexique anglais appelé *Appraisal Lexicon* qui est disponible en libre téléchargement.

1.2.4 Langages de représentations des émotions : emotionML

En 2005 un groupe de travail crée par le W3C et composé par des experts du réseau d'excellence HUMAINE (Schröder et al. 2007) a entrepris la création d'un langage de représentation des émotions dans un contexte d'utilisation technologique. Ce langage nommé emotionML se présente sous la forme d'un plug-in prévu pour être utilisable dans les situations suivantes : 1) annotation manuelle des émotions, 2) reconnaissance automatique des émotions ou des états affectifs, 3) systèmes de génération de comportements affectifs (avatars par exemple).

Le schéma proposé permet la représentation d'une émotion ou d'un état affectif à différents niveaux :

- Descripteur global sous la forme d'étiquettes. Du fait du manque de consensus au sein de la communauté sur les étiquettes de référence à utiliser, il est possible de sélectionner

parmi des listes existantes les libellés à prendre en compte parmi des listes telles que celles de (Ortony et al. 1988; Ekman 1999),

- Dimensions affective. Il s'agit d'un ensemble de dimensions ayant pour but la description d'un état émotionnel. Le jeu de dimension Valence, Activation, Contrôle est notamment proposé. Il est toutefois possible de choisir un ensemble de dimensions différent issu notamment de celles proposées dans (Ortony et al. 1988; Merhabian 1996),
- Dimensions de la théorie de l'évaluation cognitive (apparaisal theory). En pratique, ces éléments correspondent à des dimensions telles que la nouveauté, le caractère plaisant/déplaisant ou le rapport aux buts à atteindre. Comme pour les éléments précédents les dimensions peuvent être choisies individuellement en fonction de la tâche. Les dimensions proposées sont notamment issues des études de (Ortony et al. 1988; Scherer 1999; Gratch and Marsella 2004),
- Etiquettes relatives à la tendance à l'action. Ces libellés partent du constat que les états émotionnels ont un impact sur la motivation du sujet à effectuer une action. Des étiquettes telles que « l'évitement » ou le « rejet » composent cet ensemble d'éléments. Les éléments de tendance à l'action peuvent être vus comme le lien entre le résultat d'une action et les éléments issus de la théorie de l'évaluation. Les éléments de cet ensembles sont issus des travaux de (Fridja 1987),
- L'intensité. Elle représente la force avec laquelle est exprimée une émotion.

D'une façon générale tous ces éléments peuvent être inclus ou exclus d'une représentation en fonction des besoins d'une étude. De plus, étant donné la diversité des éléments pouvant être utilisés ce mode de représentation est adapté à tous les types de données (acoustiques, audio-visuelles, multimodales).

1.3Représentation des sentiments/opinions

Si la terminologie pour désigner les travaux du domaine n'est pas encore complètement stabilisée, nous employons le terme d'*analyse des sentiments*, traduction de *sentiment analysis*, qui semble s'imposer par rapport à *fouille d'opinion* (*opinion mining*). Nous suivons en cela (Pang and Lee 2008) qui mentionne également l'usage des termes *subjectivity analysis*, *review mining*, *appraisal extraction*, et *affective computing*. Rappelons qu'à l'origine l'analyse de sentiments désigne plutôt l'analyse en polarité positif-négatif du sentiment général régnant, alors que le terme d'opinion est plutôt utilisé concernant l'analyse des propriétés de produits en « bien – mixte – pas bien ».

Cette partie descriptive est tirée du livrable WP4 du projet Voxfactory (Cailliau et al. 2010).

1.3.1 Cadres théoriques

1.3.1.1 Private States Theory

La notion de « private state » est un terme générique pouvant englober des états mentaux tels que les croyances, les pensées, les sentiments, les buts et les évaluations. (Quirk et al. 1985) qui ont été les premiers à proposer ce concept, définissent les private state comme « des états mentaux ne pouvant être objectivement observés ou vérifiés ». C'est le cas par exemple d'une croyance telle

que : « je crois en l'action de notre gouvernement ». Cette déclaration ne pouvant garantir à elle seule que l'émetteur approuve réellement les mesures prises.

Nous retrouvons traditionnellement trois types d'expression relative aux « private states »

- a) Des mentions explicites d'états mentaux. Exemple : « *L'Europe **a peur** de l'extrême droite* », Jean-Arnault Dérens, le rédacteur en chef du Courrier des Balkans.
- b) Des évènements langagiers exprimant des états mentaux par l'intermédiaire de verbes comme « dire » par exemple Exemple : "C'**est une utopie louable**, mais totalement irréaliste", **selon** Florence Magne, directrice associée de CTPartners
- c) Des éléments expressifs subjectifs. Exemple : **est une utopie louable** dans l'exemple ci-dessus. Dans cet exemple, deux cadres sont donc créés, le premier pour l'événement langagier indiquant que l'état mental est exprimé dans ce qui s'est dit, et l'autre pour l'élément expressif subjectif même.

Dans (Breck et al. 2007) les auteurs ont utilisé le corpus journalistique présenté dans (Wiebe and Riloff 2005) pour entraîner des modèles de détection automatique avec des CRF (*Conditional Random Fields*). Le schéma d'annotation utilisé se base sur les états mentaux privés (private state) en utilisant néanmoins seulement deux des trois types d'expressions présentées dans les lignes ci-dessus. Les types d'expressions conservées pour l'annotation sont les mentions explicites d'états mentaux (a) et les éléments expressifs subjectifs (c).

(Somasundaran et al. 2006) a repris le cadre de (Wiebe et al. 2005) afin de permettre la prise en compte de plusieurs participants à une discussion. Dans cette étude deux catégories de phénomènes affectifs et langagiers sont retenus. Les sentiments et l'argumentation, pouvant chacun être exprimé de manière positive ou négative. Pour chaque catégorie les auteurs annotent un degré d'intensité et de certitude.

1.3.1.2 Théorie de l'énonciation

La théorie de l'énonciation se base sur la méthode définie dans (Charaudeau 1992). Sa théorie se situe dans la lignée des théories de l'énonciation, se référant notamment au principe de la subjectivité de (Benveniste 1970), qui est la capacité d'un locuteur à se poser comme sujet (emploi du je par exemple). Dans la théorie de l'énonciation telle que la propose (Charaudeau 1992) cinq modalités en particulier retiennent notre attention, car permettant de représenter la polarité d'un discours lors d'une interaction :

- Opinion : le locuteur évalue la vérité de son propos et révèle son point de vue
- Appréciation : présuppose un fait sur lequel le locuteur donne son sentiment, donc une valeur affective.
- Accord-désaccord : présuppose un message adressé au locuteur qui demande son adhésion, que le locuteur confirme ou non.
- Acceptation-refus : présuppose une demande d'accomplissement d'un acte auquel le locuteur répond favorablement ou non.

- Jugement : le locuteur pose une action réalisée dont l'interlocuteur est responsable et à propos de laquelle il déclare son approbation ou sa désapprobation.

Une application de cette théorie a été mise en œuvre dans le projet Blogoscopie ayant pour but la détection d'opinion dans un corpus de billets de blogs annoté avec ces modalités (Dubreil et al. 2008). Un lexique-grammaire a été constitué (Stern 2008), contenant 971 entrées dont essentiellement des adjectifs (493), des verbes (192) et des noms (166). Ce lexique est le point de départ des extractions des opinions dans les blogs par catégorisation automatique (Vernier et al. 2009).

1.4Conclusion

Nous avons vu dans ce chapitre les principales théories permettant de modéliser des phénomènes aussi variés que les émotions, les sentiments et les opinions. Des théories et définitions nombreuses et variées proposent de représenter ces derniers que ce soit lorsqu'il s'agit de désigner une émotion ou un sentiment. Si, à l'heure actuelle, les théories les plus mises en œuvre sont celles modélisant les états affectifs grâce à une série d'étiquettes. Il faut noter que des études utilisant des dimensions continues au lieu des étiquettes commencent à voir le jour (Eyben et al. 2008). D'une manière générale chaque méthode comporte des avantages : la méthode à base d'étiquette permet par exemple une compréhension aisée des classes à utiliser mais peut se heurter à des problèmes d'interprétation sémantique de la part des sujets qui auront à les utiliser (la définition d'une émotion peu varier d'une personne à une autre). La méthode à base de dimensions continues propose quant à elle de décomposer un état affectif en un certain nombre d'éléments atomiques mais peu souffrir d'une difficulté de mise en œuvre supérieure à celle des méthodes à base d'étiquette. Dans tous les cas, une analyse du problème est nécessaire afin de choisir la méthode de représentation la plus adaptée au problème traité.

Chapitre 2 Les corpus émotionnels : base de données et collecte d'information

Résumé

Ce chapitre se focalise sur la collecte, l'annotation et l'accessibilité des corpus de données. La construction de corpus a souvent été le parent pauvre du domaine, de par le coût considérable en temps et en argent que nécessitent la collecte et l'annotation de données.

Nous présenterons dans un premier temps les différents types de corpus existants (actés, induits et naturels) ainsi que leurs avantages et inconvénients respectifs. Il faut savoir qu'à l'heure actuelle, la grande majorité des corpus sont basés sur des données actées et prototypiques, ce qui peut poser problème lorsque le modèle construit sur ces données est testé sur des données naturelles beaucoup plus riches. Afin de nous positionner plus précisément nous passerons dans un dernier temps en revue les corpus de centres d'appels disponibles.

2.1Introduction

Toute étude scientifique en apprentissage numérique est extrêmement dépendante des données qui sont utilisées pour décrire le phénomène à modéliser. Dès lors, la collecte d'informations adaptées à la tâche que nous souhaitons modéliser devient un enjeu majeur pour l'obtention d'un modèle de détection avec un pouvoir de généralisation suffisamment fort. Nous allons voir dans ce chapitre les différents types de corpus existants ainsi que les avantages et inconvénients de chacun. Le problème de la disponibilité des données pour la communauté sera également abordé. Nous présenterons les corpus mis à disposition par le réseau d'excellence « HUMAINE » au sein de la communauté.

2.2 Les types de corpus

Lorsque nous nous intéressons aux corpus émotionnels, nous pouvons dégager trois grandes familles:

Les corpus actés : Ce type de corpus représente à l'heure actuelle la majorité des données sur lesquelles les études sont menées. Le cas le plus courant pour avoir des données émotionnelles est le recours à des acteurs qui vont jouer l'émotion demandée que ce soit hors ou en contexte. En 2003 (Juslin and Laukka 2003) montrent que sur 104 études réalisées, 84% d'entre elles étaient basées sur des données actées. Ce constat est également fait plus récemment dans (Batliner et al. 2011) où les auteurs calculent une large proportion de corpus actés dans les études de la communauté et plaident pour une utilisation plus importante de corpus de données naturelles. Dans des travaux réalisés sur les trois types de corpus actés, induits et spontanés, (Batliner et al. 2003) montrent par ailleurs que les modèles obtenus à partir de données actées ou induites ne sont pas utilisables pour des applications réelles ; les scores de détection correctes évoluant de façon inversement proportionnelle à la quantité de données naturelles présentes dans le corpus. D'autres études telles que (Scherer et al. 1991) arrivent également à la conclusion qu'il ne serait pas possible de généraliser des modèles obtenus à partir de données actées sur des données réelles.

- Les corpus avec induction d'émotions : Les données obtenues à partir d'émotions induites se situent à mi chemin entre des données artificielles et des données naturelles. Ce type d'enregistrement permet d'obtenir des comportements plus contrôlés de la part des sujets. Il est en effet proposé à ces derniers un contexte dans lequel se placer. Il est cependant possible d'obtenir, avec ce type de méthodes, des réactions émotionnelles dépendantes du contexte pouvant être relativement naturelles. Plusieurs techniques peuvent être employées dans ce cadre. Nous pouvons par exemple citer l'hypnose comme dans (Grossberg and Wilson 1968), la méthode dite de « rappel », l'utilisation de films, d'images ou de jeux permettant d'induire des émotions (Zara et al. 2007; Devillers and Martin 2008). Les méthodes utilisant l'induction offrent l'avantage de pouvoir recueillir des informations lors de tâches complexes et d'obtenir des données émotionnellement riches.
- Les corpus real-life : Les émotions dites « real-life » sont obtenues à partir d'enregistrements non contrôlés, collectés à partir de situations écologiques pouvant être extrêmement diverses. Même si la tendance actuelle est à l'utilisation, de plus en plus fréquente, de bases de données naturelles comme par exemple dans (Burkhardt et al. 2009; Polzehl et al. 2011), il faut noter que

ce type de corpus ne peut que très rarement être distribué et ce pour des raisons évidentes de vie privée. Les contextes dans lesquels sont recueillis ce type de données sont très variés (émission de TV, centre d'appels, interview de consommateurs).

2.2.1Disponibilité et réutilisabilité des corpus

Nous pouvons dater l'apparition des premiers corpus oraux au moment de la démocratisation du magnétophone, soit aux alentours des années 1950. Dans une étude qui retrace l'évolution des corpus depuis leur apparition Luzzati (Luzzati 2005) distingue différents types de corpus, chacun ayant sa spécificité en terme d'accessibilité pour les chercheurs d'une communauté donnée voir pour le grand public :

- Les données librement accessibles : Ce sont des données pouvant être utilisées, modifiées et exploitées librement.
- Les données pouvant être obtenues moyennant finance : En France l'organisme ELDA propose notamment la création de corpus (collecte et annotation) d'après un cahier des charge.
- Les données consultables : Ce sont des données généralement disponibles en ligne, permettant une lecture ou le lancement de requêtes de la part de l'utilisateur.
- Les données solicitables : Les données contenues dans ces corpus sont en généralement accessibles tout ou en partie sur demande au prêt de l'organisme propriétaire des données.
- Les données propriétaires : qui sont la propriété d'un laboratoire ou organisme et qui ne sont pas accessible par un tiers. Ces données sont souvent protégées par un accord de confidentialité ou de non divulgation

Les méthodes et les résultats employés pour la détection d'émotions varient considérablement selon les données sur lesquelles les travaux sont menés. Dès lors, la collecte d'informations et la constitution de corpus adaptés deviennent des éléments cruciaux pour la bonne marche des systèmes de détection, comme le font remarquer (Douglas-Cowie et al. 2003) et (Ververidis and Kotropoulos 2003). Ainsi, pour l'heure, la majorité des corpus contient des émotions actées ou induites comme (Engbert and Hansen 2007) ou (Burkhardt et al. 2005). Les corpus de données naturelles sont rares et souvent difficiles à distribuer pour des raisons de droits sur les données. Indépendamment de la nature des données considérées le volume de celles-ci peut également être problématique. Le traitement des données naturelles (collecte, segmentation, annotation) est souvent une tâche longue et coûteuse ce qui peux fortement limiter le volume de données réellement exploitées lors de la construction des modèles. A noter que ce problème concernant le volume de données disponibles pour un corpus se retrouve également pour les données actées. Ainsi, sur neuf corpus accessibles dans la communauté (Tableau 2-1), un seul propose plus de deux heures de paroles émotionnelles naturelle.

Corpus	Content	hh:mm	Speaker	Rec	kHz
ABC	German fixed	01:15	8 4 f	acted stud	16
AVIC	English variable	01:47	21 10 f	spont norm	44.1
DES	Danish fixed	00:28	4 2 f	acted norm	20
EMO-DB	German fixed	00:22	10 5 f	acted stud	16
eNTER	English fixed	01:00	42 8 f	acted norm	16
SAL	English variable	01:41	4 2 f	spont norm	16
SmartKom	German variable	07:08	79 47 f	spont noisy	16
SUSAS	English fixed	01:01	7 3 f	mixed noisy	8
VAM	German variable	00:47	47 32 f	spont norm	16

Tableau 2-1 Exemple de corpus émotionnels disponibles (Schuller et al. 2009)

Quelques études comme (Devillers et al. 2005), (Devillers et al. 2006) ou (Vaudable et al. 2009) utilisent des corpus proposant un volume de données plus important. Dans (Devillers et al. 2010) vingt heures de données « Real-Life » sont utilisées, mais ce type d'étude reste relativement rare au sein de la communauté. De plus ces corpus ont souvent la particularité de mettre en œuvre des émotions mélangées ce qui complexifie la tâche de modélisation. En effet, les sujets de ces études ont pour des raisons de désirabilité sociale tendance à masquer leurs émotions (Goffman 1982). A ce sujet, le modèle push/pull de (Scherer 1985) reprend l'hypothèse que les échanges émotionnels sont modifiés, d'un côté par l'effet « push », qui concerne l'effet des changements physiologiques sur l'expression vocale et multimodale, et d'un autre coté par l'effet « pull », qui pour des raisons sociales ou stratégiques, pousse le locuteur à masquer l'émotion qu'il ressent. Ces effets push/pull sont plus caractéristiques des corpus « real-life » où il y a un vrai ressenti de la part des sujets et un contexte d'interaction réelle.

Récemment de grands efforts ont été entrepris pour permettre à la communauté de disposer de corpus librement accessibles et d'un panel suffisamment représentatif de situations pouvant se produire dans un contexte réel. C'est dans cette optique que la base de donnée « HUMAINE » a été conçue (Douglas-Cowie et al. 2011). Ce regroupement comprend 11 corpus basés principalement sur

des données réelles ou induites. Un sous ensemble de 48 clips d'une durée comprise entre 3 secondes et 2 minutes a été annoté :

Raw data records	Data type	Number of clips selected
Belfast Naturalistic Database	Naturalistic	10
Castaway Reality Television Data Set	Naturalistic	10
Sensitive Artificial Listener (Belfast recordings in English)	Induced	12
Sensitive Artificial Listener (Tel Aviv recordings in Hebrew)	Induced	1
Activity Data/Spaghetti Data	Induced	7
Green Persuasive Data Set	Induced	4
EmoTABOO	Induced	2
DRIVAWORK (driving under varying workload) Corpus in German	Induced	1
GEMEP Corpus	Acted	1

Tableau 2-2 Les types de données rassemblées et annotées pour la base de données HUMAINE (Douglas-Cowie et al. 2011)

Le sous-ensemble sélectionné comprend des situations se déroulant dans plusieurs contextes différents (scènes en intérieur, extérieur, monologue, dialogue, …). Les auteurs ont également veillés à conserver des expressions émotionnelles très variées (émotions positives et négatives pouvant être actives ou passives ; combinaison d'émotions et transition d'une émotion à une autre au cours du temps, …). Des types d'indice très variés pour l'étude des émotions sont finalement disponibles dans cette base de données (gestes, visage, voix) et ce pour des langues et cultures différentes.

L'une des grandes forces de ce corpus réside dans sa capacité à fournir des données clés en mains, annotées de manière fine. Au total, ce sont plus de 20 éléments différents qui ont été annotés (étiquettes émotionnelles, dimensions continues, annotations sur le lexique utilisé, …). Un autre élément important concernant ce-sous ensemble est son accessibilité pour l'ensemble de la communauté, facilitant ainsi une comparaison des résultats obtenus par les différents chercheurs du domaine.

Le point négatif de cette base de données est le faible volume de données disponibles, ces données ne peuvent pas servir à la construction de modèles. En effet, en prenant une valeur moyenne par segment d'une minute environ, c'est moins d'une heure de données qui a été produite.

2.3 Les centres d'appels et les bases de données émotionnelles

2.3.1Corpus de centres d'appels existants

Nous assistons, au sein de la communauté, à un intérêt croissant pour les données naturelles issues de centres d'appels, qui sont étudiées sous différents aspects, comme la reconnaissance de la parole, l'analyse conversationnelle ou le traitement automatique des langues (Garnier-Rizet et al. 2008). Ce type de base de données est très intéressant du point de vue des émotions car il est possible de collecter un grand volume d'informations contenant une expression émotionnelle extrêmement riche. Un certain nombre de corpus composés de paroles collectées en centres d'appels existe ; comme (Devillers et al. 2010) qui est un corpus de 20 heures collectées dans un centre d'appels français dont le but est d'étudier les émotions lors de conversation collectées dans un contexte d'urgences médicales. Nous pouvons également citer les corpus présentés dans l'étude (Erden and Arslan 2011) qui utilise six heures de données collectées dans trois centres d'appels turc différents (assurance, finance et commercial). Pour leur part, les auteurs de (Polzehl et al. 2011) proposent une détection des émotions sur un corpus multilingue anglais/allemand composé de dialogues en provenance de centres d'appels d'un fournisseur internet et de téléphonie mobile. Il faut cependant noter que ces corpus ont tous comme dénominateur commun le fait qu'ils ne soient pas distribuables. Nous récapitulons dans le tableau ci-dessous les corpus issus de centres d'appels et ayant été utilisés pour des études ayant pour but la détection d'émotion. Il faut noter que nous ne nous sommes intéressés ici qu'aux corpus se basant sur des données naturelles. Certains comme (Petrushin 1999; Yacoub et al. 2003) ayant pour but une implémentation dans un service de centre d'appel mais se basant sur des données actées n'ont pas été retenus. Le Tableau 2-3 ci-dessous détaille le contenu de ces corpus :

Type de donnée : Dialogue Humain-Humain(DHH)/ Dialogue Homme Machine (DHM)	Nombre d'appels	Nombre d'étiquettes émotions	Auteurs
DHH	20	3	(Gupta and Rajput 2007)
DHM (2 corpus différents)	4682 et 1911	2	(Metze et al. 2010)
DHM	4683	2	(Burkhardt et al. 2009)
DHH (2 corpus différents)	100 et 688	6	(Vidrascu 2007)
DHM	1187	2	(Lee and Narayanan 2005)
DHH	385	2	(Erden and Arslan 2011)
DHM	5690	7	(Liscombe et al. 2005)

Tableau 2-3 Récapitulatif de corpus « call center » utilisés pour la construction de modèles de détection d'émotions

Dans l'optique d'une utilisation à grande échelle des modèles, ce qui est le cas pour notre étude, le choix de départ des dialogues ou des extraits de dialogues à utiliser est une tâche cruciale. Il est en effet indispensable de pouvoir sélectionner un nombre de locuteurs suffisamment important pour la phase de modélisation. Les émotions exprimées lors de ces dialogues doivent être assez marquées acoustiquement parlant pour obtenir un score d'inter-annotation suffisamment élevé, mais également ne pas être trop prototypiques pour obtenir un modèle assez générique et ne pas retomber dans les problèmes évoqués plus haut concernant les corpus actés. La sélection automatique d'instances au sein de corpus de grande taille dans le but d'enrichir des modèles composés d'annotation manuelle est par ailleurs l'un des challenges de cette thèse.

2.4Conclusion

Ce chapitre nous a permis de passer en revue les différents types de corpus existants (artificiels, naturels). Nous avons pu voir que la grande majorité des études utilisaient des données actées dans le but de pouvoir contrôler plus facilement les situations et par conséquent pouvoir travailler sur des volumes de données plus importants. Nous avons également mis en avant le fait que les corpus n'étaient que rarement mis à disposition de la communauté pour des raisons de droits d'auteurs, de vie privée ou simplement de coûts. Nous nous sommes dans un dernier temps attardé sur les corpus basés sur des données naturelles et plus particulièrement les corpus de centre d'appels qui présentent l'avantage d'offrir des situations écologiques et donc particulièrement riches en mettant en œuvre une gamme d'émotions plus fines et plus complexes que celles trouvées dans les données actées souvent prototypiques.

Chapitre 3 Les indices mis en jeux lors de l'expression d'un état affectif dans un contexte de conversation

Résumé

Le but de ce troisième chapitre est de mettre en avant les différents types d'indices pouvant intervenir pour une tâche de détection des émotions. Les chapitres précédents nous ont permis de situer le contexte dans lequel se plaçait notre étude en définissant les théories de représentation propres aux émotions/sentiments/opinions ainsi que les données sur lesquelles se basaient classiquement ces théories.

Nous allons dans les paragraphes qui vont suivre, continuer notre tour d'horizon du domaine en présentant plus en détail les descripteurs utilisés lors de l'extraction des informations acoustiques pour une tâche de détection des émotions. Nous passerons par la suite en revue les indices de type linguistique utilisés dans la communauté de l'affective computing. Les méthodes et descripteurs linguistiques utilisés dans les domaines de la détection de sentiments/opinions sont également présentés car pris en compte dans la suite de l'étude.

Nous souhaitons également mettre en lumière dans ce chapitre les différents niveaux langagiers impliqués dans l'expression affective. Cette thèse portant sur la détection des émotions lors de conversations téléphoniques, seulement deux canaux de communication nous intéressent: le canal paralinguistique et le canal linguistique. Nous passerons en revue, selon des points de vue différents (linguistique, acoustique), la façon dont les deux canaux de communication se complètent et l'intérêt que cela représente en termes de détection automatique des émotions.

Enfin, pour conclure ce chapitre, nous présenterons un état de l'art des différentes approches sur ce sujet ainsi que les gains pouvant en pratique être obtenus lorsque la fusion d'indices acoustiques et linguistiques est appliquée sur des données extraites de centres d'appels.

3.1Introduction

Nous allons aborder dans ce chapitre l'ensemble des éléments permettant l'expression, et la compréhension, d'un état affectif par l'intermédiaire de la voix, au cours d'une conversation entre deux humains. Nous nous rendons compte, intuitivement, que cette expression émotionnelle suit un ensemble de processus complexes, intervenant à de nombreux niveaux. Le premier niveau mis en jeu est physiologique. Il nécessite de la part du sujet la production d'un signal sonore audible par un interlocuteur. Ce signal sera par la suite modulé suivant un ensemble de variables contextuelles, culturelles et langagières qui ont toute pour but de permettre la communication (émotionnelle ou non). Si nous nous plaçons dans le cas particulier d'une communication au cours de laquelle un état affectif (émotion, sentiment) devait être exprimé nous pouvons voir que ce dernier implique deux niveaux de communication principaux dans son expression. Le premier est le niveau acoustique qui englobe les différentes formes prosodiques et intonatives de la voix. En vue d'une analyse automatique, différents indices peuvent être calculés dans le but de déterminer, au moins partiellement, la nature de l'expression affective (positive, négative, neutre). Le second niveau est l'expression linguistique qui accompagne la production vocale. Le niveau linguistique transmet également une partie de l'expression affective à différents niveaux (lexicaux, syntaxiques, sémantiques, pragmatiques). La combinaison de ces deux niveaux répond au principe du double encodage décrit par (Fonagy 1983). Ce dernier défend la thèse qu'une expression (affective ou non) doit être décodée suivant les deux canaux de communication acoustique et linguistique pour être appréciée dans son ensemble. La Figure 3-1 ci-dessous illustre ce principe du double encodage :

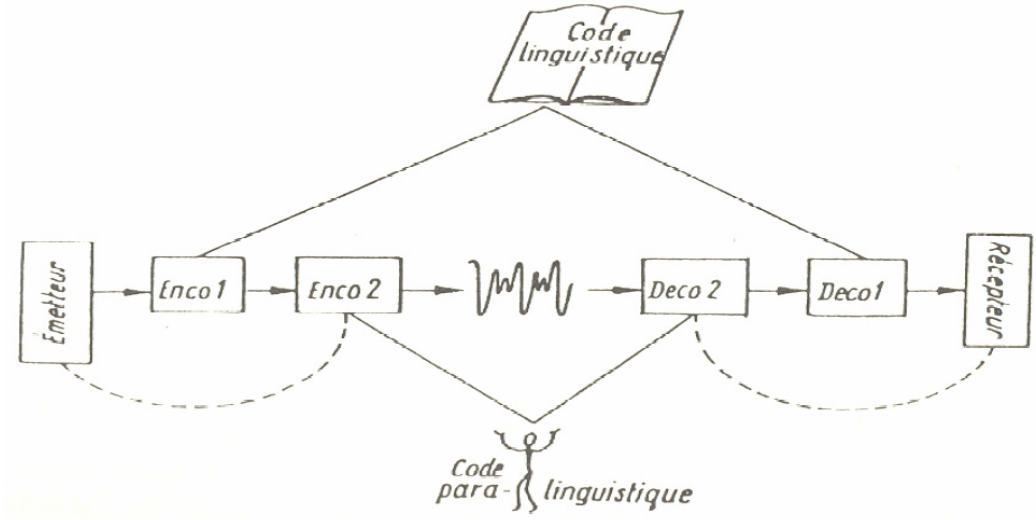


Figure 3-1 Le double encodage (Fonagy 1983) ou l'utilisation conjointe des canaux de communication linguistique et acoustique

Les différents paragraphes que nous allons présenter dans ce chapitre reprennent et détaillent les étapes de la production émotionnelle vocale ainsi que le lien qu'entretiennent ces deux éléments entre eux. Nous commencerons donc avec un récapitulatif de la production du signal vocal par l'humain. Nous verrons par la suite les cadres et études théoriques mettant en avant l'importance du rapport entre linguistique et acoustique. Notre but étant la conception de modèles automatiques, nous présenterons les types d'indices acoustiques et linguistiques généralement exploités dans le

cadre d'études sur la détection automatique des émotions et enfin nous observerons les gains que les différents auteurs arrivent à obtenir dans le cadre d'une détection automatique des émotions mêlant acoustique et linguistique par rapport aux études n'utilisant qu'un seul des deux canaux de manière indépendante.

3.2La production vocale :

La production vocale humaine se déroule en trois étapes successives :

- Inspiration : L'inspiration, par l'intermédiaire de la trachée permet de stocker l'air dans les poumons. Ces derniers jouent un rôle de réservoir. Au moment de l'expiration l'air suit le chemin inverse et, arrivant au niveau du larynx, permet la vibration des cordes vocales.

- Vibration des cordes vocales : Afin de passer de la trachée au gosier, le souffle traverse les cordes vocales. Si la glotte est fermée, le souffle force l'expiration et permet au passage la vibration des cordes vocales à différents niveaux de fréquence. La hauteur d'un son dépend de cette variation de fréquences. Nous appellerons ainsi « contenu voisé » le signal de parole produit grâce à la vibration des cordes vocales et « contenu non voisé » le signal produit sans vibration des cordes vocales (chuchotement par exemple).

- Génération et amplification du son des cordes vocales par l'intermédiaire de résonateurs : Qu'il permette l'émission ou non d'un son, le souffle traverse le « conduit vocal ». Ce dernier s'étend ainsi de la glotte jusqu'aux lèvres et narines. Le souffle peut emprunter trois passages différents durant son parcours : le canal guttural (gosier), le canal buccal et le canal nasal. Ainsi, si le voile du palais est abaissé, empêchant ainsi l'air d'accéder au canal nasal, le son est qualifié de gutturo-bucal (gosier bouche). A l'inverse, si le voile du palais est levé, le son est de nature gutturo-nasal (gosier-nez). Le son émis est ainsi affecté par le trajet emprunté par le souffle. Ces différentes trajectoires permettent l'enrichissement du son grâce à des harmoniques (multiples de la fréquence fondamentale). Certaines consonnes (comme les plosives et fricatives) sont produites sans vibration des cordes vocales mais sans envoi d'air dans le conduit vocal. Les phénomènes d'occlusion, de friction dans le conduit vocal permettent de réaliser ces sons.

3.3Indices Prosodiques

Les paramètres expressifs les plus largement étudiés sont ceux de la prosodie. La prosodie est généralement définie suivant 3 paramètres principaux :

- **la fréquence fondamentale (timbre de la voix)** est définie comme la fréquence de vibration des plis vocaux à un instant donné. On appelle voisement la présence d'une fréquence fondamentale, c'est-à-dire la vibration périodique des plis vocaux.
- **La durée (le rythme) :** Mesure d'un intervalle de temps nécessaire à l'émission d'un segment sonore
- **L'intensité :** qui représente l'énergie contenue dans le signal

Les différents paramètres acoustiques seront présentés plus en détail dans le chapitre 5.

3.3.1 Synthèse des résultats obtenus pour la caractérisation des émotions (Juslin and Laukka 2003)

Nous avons retenu pour cette partie, la synthèse réalisée dans (Juslin and Laukka 2003) qui a été obtenue d'après 102 études dont 12 réalisées sur des données naturelles. Le Tableau 3-1 ci-dessous récapitule les descripteurs mis en œuvre pour la caractérisation d'une émotion parmi 5 proposées :

Paramètres	Colère	Peur	Joie	Tristesse	Tendresse
Intensité moyenne (forte-moyenne-faible)	Forte (30/32)	Forte (11/22)	Forte (20/26)	Faible (29/32)	Faible (4/4)
Intensité variabilité (forte-moyenne-faible)	Forte (9/12)	Forte (7/12)	Forte (8/13)	Faible (8/11)	/
F0 moyenne (haute-moyenne-basse)	Haute (33/43)	Haute (28/39)	Haute (34/38)	Basse (40/45)	Basse (4/5)
F0 variabilité (forte-moyenne-faible)	Forte (27/35)	Faible (17/32)	Forte (33/36)	Faible (31/34)	Faible (5/5)
F0 contours (montants-descendants)	Montant (6/8)	Montant (6/6)	Montant (7/7)	Descendant (11/11)	Descendant (3/4)
Energie haute fréquence (forte-moyenne-faible)	Forte (22/22)	Forte (8/16)	Forte (13/17)	Faible (19/19)	Faible (3/3)
Débit de parole (rapide-moyen-lent)	Rapide (28/35)	Rapide (24/29)	Rapide (22/33)	Lent (30/36)	Lent (3/4)
Proportion de pauses (forte-moyenne-faible)	Faible (8/8)	Faible (4/9)	Faible (3/6)	Forte (11/12)	Forte (1/1)
Précision articulation (haute-moyenne-basse)	Haute (7/7)	NC (2-2-2)	Haute (3/5)	Basse (6/6)	Basse (1/1)
Formant 1 (hauteur) (haut-moyen-bas)	Haut (6/6)	Bas (3/4)	Haut (5/6)	Bas (5/6)	/
Formant 1 (largeur) (étroit-large)	Etroit (4/4)	Large (2/2)	Etroit (2/3)	Large (3/3)	/
Jitter (fort-faible)	Fort (6/4)	NC (4-4)	Fort (5-8)	Faible (5/6)	/

Tableau 3-1 Synthèse de la revue de résultats obtenus par (Juslin and Laukka 2003) pour la caractérisation des émotions. Les indices sur fond gris sont les plus couramment utilisés pour la détection d'émotions

Les résultats présentés sont obtenus d'après différentes études réalisées par les auteurs. Les chiffres entre parenthèses représentent le nombre d'études à avoir obtenu cette valeur pour une émotion donnée. Les cases notées « NC » (non concluant) indiquent qu'aucune tendance ne se dégage pour ce paramètre et cette émotion. Nous pouvons voir que les indices couramment utilisés pour la détection automatique des émotions, en gris dans le tableau, ne permettent pas de caractériser une émotion comme étant positive ou négative (colère et joie par exemple). A noter également que la différence des résultats trouvés entre chaque étude peut également être relativement importante (par exemple pour le cas de la peur). Ces résultats sont concordants avec ceux proposés par (Scherer et al. 2003).

3.4Indices linguistiques

3.4.1 Approches linguistiques pour la détection des émotions

Comme le mentionnent de nombreux chercheurs comme (Arunachalam et al. 2001) des informations émotionnelles peuvent être transmises par le texte. Cela se traduit en général par l'utilisation de mots appartenant à un champ lexical en particulier et d'altérations grammaticales ou syntaxiques (ex : « j'suis pas content »). Cette constatation est également valable même si les transcriptions sont fournies par un système de transcription de la parole contenant des données bruitées (Georgiou et al. 2011).

Dans le domaine de la détection d'émotions, beaucoup de méthodes différentes ont été expérimentées. Nous pouvons citer notamment le « word spotting » mis en œuvre par exemple dans (Cowie et al. 1999) et qui se propose de rechercher des séquences prédéterminées de mots dans un ensemble de données. Des études utilisant la sémantique permettent d'obtenir des résultats intéressants comme c'est le cas dans (Lu et al. 2006; Fang et al. 2011). Dans la première étude les auteurs combinent annotation sémantique (i.e. mise en relation de sujets avec un verbe d'action par exemple) et désambigüisation des mots inconnus grâce à des requêtes via des moteurs de recherche publics dans le but d'augmenter le vocabulaire. La méthode est par la suite implémentée dans un système de discussion. La seconde étude propose de combiner ontologie et logique floue dans le but de déterminer le degré de satisfaction client à propos d'un produit. Les auteurs parviennent grâce à l'utilisation des ontologies à augmenter leurs scores de 10% par rapport aux études précédentes.

En matière de détection des émotions au niveau linguistique, nous pouvons noter deux méthodes prédominantes.

La première se base sur les n-grammes ; elle a été employée notamment dans (Lee et al. 2002; Devillers et al. 2003). Dans cette méthode à partir d'une séquence de mots donnés (par exemple "*par exemple*"), il est possible d'obtenir la fonction de vraisemblance de l'apparition du mot suivant. Ainsi appliquée à la détection des émotions, l'idée est de construire à partir d'un corpus d'apprentissage l'ensemble des séquences de n-mots pour les classes du corpus afin de pouvoir prédire la probabilité a posteriori d'une instance. En pratique, du fait de la grande disparité des données utilisées et de la petite taille des corpus, c'est le modèle unigramme qui est le plus utilisé (N=1) comme c'est le cas dans (Lee et al. 2002; Devillers et al. 2003).

La seconde est la méthode dite des « sacs de mots » (Bag of Word) décrite par (Joachims 1998) et qui a été utilisée à de nombreuses occasions comme dans (Schuller et al. 2005) . Cette méthode propose de représenter les mots présents dans un segment par un vecteur. L'inconvénient de cette méthode est que chaque mot va représenter une dimension supplémentaire dans le vecteur. Il est donc possible de se retrouver avec des vecteurs de très grandes dimensions et donc des temps de calcul très élevés. Une extension naturelle de la méthode « bag of word » est la méthode « bag of n-gram » où une séquence de *n* mots est utilisée pour représenter une unité textuelle. Cette méthode offre l'avantage de pouvoir, dans une certaine mesure, pallier le manque de vocabulaire spécifique pour un corpus mais augmente rapidement, de manière artificielle, le nombre de dimensions d'un problème. En règle générale, *N* représentant le nombre de mots à retenir pour une séquence, ne

dépasse pas *N*=3 (trigramme), et ce principalement pour des raisons de temps de calcul et de taille de mémoire disponible évoquées ci-dessus. Il faut également noter que le nombre de n-gramme de fréquence nulle augmente avec *N* ce qui à pour conséquence de bruiter artificiellement les modèles.

Pour pallier ce problème, il est possible de se baser sur l'analyse de (Turney 2002) qui indique que toutes les classes de mots ne sont pas utiles pour une tâche de détection des sentiments ou de l'émotion dans le texte. En accord avec l'auteur il ne faudrait donc conserver que les classes de mots sémantiquement pleins de type adjectif, adverbe, interjection, nom. Les mots vides (connecteurs, pronoms, articles, auxiliaire…) et les entités nommées sont également éliminés. L'utilisation de « Part Of Speech tagger » permet d'étiqueter morpho-syntaxiquement un texte. Beaucoup d'outils existent et ont l'avantage de permettre un étiquetage correct avec une précision supérieure à 90% comme c'est le cas de (Brill 1995) ou (Schmid 1995). Il faut toutefois préciser que ces outils sont prévus pour traiter des données textuelles et non de la parole transcrite. De plus, l'utilisation d'un système de reconnaissance de la parole (ASR) pouvant donner lieu à des erreurs de reconnaissance comme c'est le cas dans notre étude, soulève des problèmes supplémentaires. Toujours dans le but de réduire la dimensionnalité d'un problème donné, l'utilisation de méthodes de lemmatisation (racinisation) permet de ne conserver que la forme canonique d'un mot (suppression des formes conjuguées pour retenir que le verbe à l'infinitif, utilisation du singulier seulement). Des algorithmes de lemmatisation très couramment utilisés en traitement du langage écrit, comme par exemple (Porter 1980; Savoy 2006), fournissent des performances avec un degré de fiabilité élevé (>90% de lemmatisation correcte).

Des domaines voisins à la détection des émotions comme l'analyse des sentiments ou des opinions peuvent proposer des pistes intéressantes et complémentaires à celles classiquement utilisées. Les principales techniques utilisées dans ces domaines sont présentées dans les paragraphes ci-dessous[7].

3.4.2 Approches linguistiques pour la détection des sentiments et opinions (sentiment analysis and Opinion mining)

Dans (Maurel and Curtoni 2007) les auteurs définissent des stratégies basiques, consistant à affecter à un document la classe qui possède le plus grand nombre d'occurrences appartenant à ladite classe d'après un dictionnaire préalablement constitué. D'autres auteurs, tels que (Vernier et al. 2007) ont des approches plus subtiles : ils commencent d'abord par donner un score pour toutes les expressions (qui combine le score de l'expression avec l'intensité qui est donnée par des adverbes). Ils additionnent le nombre d'annotations mais accordent un poids tout particulier aux expressions présentes dans l'introduction ou la conclusion (ce poids prend le pas sur tous les autres).

Certains auteurs, tels que (Bloom et al. 2007) exploitent une petite ontologie du domaine, pour typer la cible du sentiment, mais également, pour désambiguïser et typer l'expression du sentiment elle-

[7] Le texte du paragraphe 3.4.2 est tiré du livrable WP4 du projet voxfactory Cailliau F., Bouffier A., Kucharski M., Guillemin-Lanne S., Devillers L. and Vaudable C. (2010). "Etats de l'art sur la détection des émotions, des opinions et des sentiments dans des interactions clients agents des centres d'appels." Livrable du projet FUI voxFactory..

même. Par exemple, connaître certains traits de la cible (humain ou non) permet de distinguer entre un jugement et une appréciation.

Bien qu'une abondante littérature exploite uniquement des connaissances lexicales (adjectifs, adverbes + ressources), on ne peut que constater, à une échelle fine, que des connaissances lexicales isolées sont très ambigües et insuffisantes. D'où la nécessité de recourir à des relations entre unités lexicales. Dès lors, deux approches s'affrontent : une approche de type « fenêtre de mots » où la relation entre les unités lexicales est approchée par une mesure de proximité, et une approche plus précise exploitant des relations syntaxiques. Par exemple, (Dini and Mazinni 2002; Bloom et al. 2007) exploitent les résultats d'un « chunking » pour mettre en relation différentes unités lexicales. Les structures discursives sont peu exploitées bien qu'elles jouent un rôle manifeste pour l'analyse de l'opinion ou du sentiment. Par exemple, l'expression d'une opinion n'a pas le même poids selon sa position dans la structure rhétorique d'un texte : si elle est insérée dans une conclusion, elle est plus importante (Pang and Lee 2002) que si elle se situe dans le corps du texte ou une incise. Sans entrer dans une modélisation très fine de la structure réthorique des textes (en général coûteuse), certaines connaissances surfaciques simples à mettre en œuvre peuvent être mises en place avec profit : par exemple, la position de l'expression d'opinion dans le texte (début, milieu, fin) peut être un bon indice, ou encore la présence de marqueurs tels que «*en conclusion*», «*en résumé*».

D'autres phénomènes, fins, jouent également un rôle, tels que le caractère intégré ou périphérique de certains syntagmes (Jackiewicz et al. 2009). En effet, selon que le jugement évaluatif fait l'objet d'une prédication ou que celui-ci fait partie d'un constituant périphérique (« *hypocondriaque, le clarinettiste amateur est un hédoniste contrarié* »), la valeur de la prise en charge doit être nuancée. Si l'expression de l'opinion est insérée dans un constituant périphérique, l'information tend à être présentée comme communément partagée, comme une sorte de présupposé tandis que si elle fait l'objet d'une prédication, c'est réellement une information nouvelle, prise en charge par l'auteur. Cette nuance peut être tout à fait pertinente dans un but de hiérarchisation de l'information. Néanmoins, ces travaux restent, pour la plupart, du domaine de l'analyse linguistique fine et sont difficiles à implémenter aujourd'hui pour des études à grande échelle.

Il faut noter que dans les domaines de la détection des opinions/sentiments/émotions dans le texte, l'immense majorité des travaux se base sur du langage écrit. Les données orales sont très peu représentées. Nous pouvons cependant citer les travaux de (Hollard et al. 2005) et (Tomokiyo et al. 2005) qui ont pu travailler sur un corpus d'enregistrements de messages laissés sur le répondeur de l'assistance informatique d'un hôpital public. Le corpus correspond à 5 heures 30 minutes de parole. Les auteurs ont étudié les marqueurs lexicaux et phonologiques (ton, débit) et, d'après leurs observations, les deux types n'ont pas d'incidences systématiques l'une sur l'autre lorsque les deux modes d'expressions sont observés en parallèle.

3.5 Linguistique et parole

Selon la théorie du double encodage proposée par Fonagy (Fonagy 1983), une conversation entre deux humains implique un double canal de communication (linguistique et paralinguistique). Dans le premier niveau d'encodage (niveau linguistique), le message est découpé en unités de sens (syllabes,

mots, groupes de mots, phrases) en tenant compte du marquage prosodique qui va servir à faciliter l'accès pour l'interlocuteur aux différentes unités de sens précédemment mentionnées. Le deuxième niveau se situe sur le canal paralinguistique et va définir la façon dont va être exprimée un énoncé. Des modulations prosodiques, comme les proéminences, sont généralement émises pour insister sur un point particulier du discours ; dans ce cas, la prosodie est utilisée comme un instrument intervenant au niveau sémantique-pragmatique (voir paragraphe 3.5.1). D'autres informations relatives à l'affect peuvent également être mises en œuvre permettant de colorer émotionnellement un énoncé lexicalement neutre ou ambigu (voir paragraphe 3.5.2). Les paragraphes suivants s'attachent à présenter les liens entre les canaux linguistique et paralinguistique et mettent en avant l'utilité que peuvent avoir les méthodes décrites, dans le cadre de la conception de modèles de détection des émotions.

3.5.1 La prosodie et l'expression verbale non émotionnelle

Nous nous intéresserons notamment dans cette partie au lien segmental et syntaxique qui existe entre acoustique et linguistique. Il a en effet été noté qu'en matière de détection des émotions la manière de segmenter les données d'un corpus représente un enjeu crucial mais est un champ de recherche largement sous exploré (Batliner et al. 2010).

De manière générale, la prosodie reste un phénomène à définir de manière très contextuelle, c'est-à-dire en fonction du sujet d'étude et d'expérimentation. Aucune définition ne fait réellement consensus de part les divers aspects (cognitifs, acoustiques, linguistiques) qu'elle met en œuvre. Dans (Aubergé 2002) l'auteur note par exemple qu'en 1975 plus de 4000 références bibliographiques traitaient du sujet. Considérant les données qui ont été utilisées pour nos travaux (conversations audio et transcriptions de ces conversations dans des centres d'appels), nous nous intéresserons dans cette partie au lien qui unit la prosodie à l'expression verbale pour le français uniquement. Chaque langue ayant ses particularités (Hirst et al. 2001), un champ d'étude plus large ne serait bénéfique ni à la lisibilité de l'étude, ni à l'efficacité des méthodes proposées. Les études linguistiques prenant en compte la dimension prosodique d'un énoncé, ont débuté il y a maintenant plusieurs décennies (Fonagy 1983) ou (Boulakia 1979) mettent en évidence les liens existant entre syntaxe et prosodie.. Dans son étude (Lacheret-Dujour 2011) met en avant la double utilité de la prosodie pour l'articulation langagière. La prosodie interviendrait donc à un niveau phonétique (segmental) et sémantico-pragmatique aidant à définir le contexte d'une interaction. La prosodie dans une conversation permet d'effectuer des regroupements de mots (intonation) et de rendre saillant les concepts importants d'un énoncé (accentuation). Ils permettent ainsi aux locuteurs de segmenter une conversation et d'avoir accès aux éléments linguistiques nécessaires à la compréhension du discours (Lacheret-Dujour 2011). A noter que si le phénomène d'intonation/accentuation est présent dans toutes les langues, la façon dont les informations sont regroupées et structurées est quant à elle, propre à une langue donnée.

Dans son étude du lien entre prosodie et syntaxe, Lacheret (Lacheret-Dujour et al. 2011) mentionne la forte corrélation qu'il est possible d'observer entre « frontière prosodique et syntaxique ». L'étude met en avant une corrélation forte entre deux types d'unités prosodiques et linguistiques : les paquets intonatifs et les unités illocutoires :

- *Les paquets intonatifs peuvent être vus comme des unités segmentales basées sur des « groupes accentuels » qui sont des unités phonétiques ayant la particularité de présenter des ruptures prosodiques entre elles. Les paquets intonatifs s'inscrivent à leur tour au sein d'unités plus larges nommées « périodes ». Ces périodes correspondent aux différentes phases pouvant survenir lors d'un discours entre deux ou plusieurs interlocuteurs (période ou séquence d'ouverte, illustration, argumentation, clôture, etc.). A l'échelle du discours les périodes sont classiquement séparées entre elles par des pauses ayant une durée dépassant un seuil donné. L'amplitude d'un contour potentiellement terminal pour une période est représentée par un « saut mélodique » considéré comme étant marqueur d'une « réinitialisation » mélodique débutant dans la période suivante.*

- *Les unités illocutoires sont présentées comme des éléments autonomes du discours c'est-à-dire pouvant être reconnus comme des éléments syntaxiquement complets et bien formés. Le sens de ce type d'unité ne sera pas affecté en cas de mise en relation avec d'autres unités illocutoire. (Martin 2009).*

(Lacheret-Dujour et al. 2011) montre ainsi que 65% des paquets intonatifs (UI) qui peuvent être vus sont également des frontières d'unités illocutoires. Inversement, 87% des frontières d'unités illocutoires sont également des frontières de paquets intonatifs ; ce qui revient à pouvoir dégager de manière plus ou moins précise en fonction de l'unité servant d'étalon, des segments ayant sens à la fois linguistiquement et acoustiquement. L'enchevêtrement d'unités de type linguistique et acoustique observés met ainsi en avant les liens qu'entretiennent ces deux types d'unité ; l'un pouvant servir à structurer le second et inversement. Il faut toutefois noter que cette constatation n'est pas nécessairement applicable à tous les cas de figure du français. Certaines formes linguistiques, comme par exemple celle des « mots-phrases » échappent au processus d'empaquetage intonatif et sont donc plus difficilement analysables.

La seconde grande spécificité à laquelle nous pouvons nous intéresser concernant le lien entre prosodie et expression verbale dans la langue française a trait aux proéminences. Une proéminence est en général définie comme une entité syllabique qui se détache de son environnement comme une figure sur un fond, en vertu d'un certain nombre de paramètres acoustiques et perceptifs (Avanzi et al. 2011) . La mise en œuvre des proéminences consiste en l'accentuation d'une ou plusieurs syllabes d'un mot, ce qui se traduit par une élévation du niveau du pitch, de l'intensité de la voix ainsi que par un allongement syllabique (Hirst et al. 2001). Traditionnellement, une proéminence permet de mettre en avant un concept particulier dans une construction verbale. Elle a pour objet d'attirer l'attention du locuteur sur un point précis du discours (Di Cristo 2007). Nous pourrions donc considérer la proéminence comme un point d'ancrage sémantique et pragmatique mis en avant par la prosodie. Ces phénomènes langagiers étant envisagés comme des points d'ancrage dans le discours peuvent être particulièrement intéressants à étudier en vue d'une analyse automatique, car fournissant un indice précis sur la partie du discours à étudier. Des outils automatiques proposant un niveau de fiabilité correct commencent à être disponibles. Ainsi, dans (Avanzi et al. 2011), les auteurs proposent un algorithme de détection automatique des proéminences se basant sur un fenêtrage temporel de 3 syllabes précédant et suivant la syllabe en cours d'analyse. Se basant sur un alignement manuel signal/phonème, chaque syllabe se voit attribuer un score de proéminence. Les calculs sont basés suivant 4 paramètres :

(a) La durée de la syllabe courante par rapport aux 3 précédentes et aux 3 suivantes
(b) Calcul de la hauteur moyenne de la syllabe courante par rapport aux 3 précédentes et aux 3 suivantes
(c) Estimation de l'amplitude de la voyelle sur le ton montant de la syllabe
(d) Présence d'une pause après la syllabe courante (hors pause longue dénotant la fin d'un paquet intonatif tel que décrit dans les paragraphes ci-dessus)

Dans cette étude une syllabe est considérée comme proéminente si au moins un des quatre critères cités précédemment atteint un certain seuil défini d'après une annotation manuelle sur une moyenne d'environ 70 minutes de dialogues présents dans le corpus. Concrètement, le score total est une somme des scores attribués indépendamment à chacun des 4 critères de décision. Au final les auteurs obtiennent un taux de détection correct des proéminences égal à 91% dans le meilleur des cas.

Malgré le fait que ces résultats soient très intéressants d'un point de vue théorique, la méthode semble difficilement automatisable ou portable à grande échelle. L'alignement manuel en phonème de la parole sur le texte est en effet une opération extrêmement coûteuse en temps et en argent. Le découpage et l'annotation des 13h de parole de notre corpus semble inenvisageable de ce point de vue. La seconde problématique se posant pour le portage de ce genre de méthode est la transcription fournie par des systèmes de reconnaissance de la parole. Le taux d'erreurs fourni par les ASR (Automatic Speech Recognition) pouvant être assez élevé, l'alignement serait dès lors difficile, étant donné que la transcription ne correspondrait pas nécessairement à ce qui a été dit lors de la conversation. Enfin, dans le même ordre d'idée, en supposant un alignement automatique dans un contexte de corpus « real life », la présence de nombreux éléments perturbateurs (segments contenant de la parole superposée, bruit ambiant) et la probabilité de disposer de données suffisamment fiables semble assez faible mais pourrait être une piste à analyser.

3.5.2 Le lien entre l'expression verbale et l'expression émotionnelle

Nous allons voir dans cette section le lien entre l'expression verbale et l'expression émotionnelle. Si nous avons vu dans la section précédente qu'unités prosodiques et verbales étaient intimement liées, il nous importe maintenant d'observer l'impact que peut avoir un état émotionnel donné sur l'expression linguistique. Nous verrons donc dans quelle mesure l'expression émotionnelle peut être porteuse de sens dans le message verbal adressé à un interlocuteur au cours d'une conversation. Nous nous placerons cette fois-ci dans un cadre plus large que celui de la seule étude du français en présentant des travaux réalisés sur des langues aussi diverses que l'espagnol, l'hébreu, l'anglais ou le français en partant du principe que l'expression émotionnelle est universelle, du moins pour un petit nombre d'émotions «basiques» (Ekman 1999). Il faut noter que des études telles que celles réalisées par Scherer montrent un taux de reconnaissance émotionnelle d'environ 66% pour des sujets de 9 nationalités différentes. Dans cette série d'études il était demandé de juger l'expression affective d'un énoncé exprimé dans une langue différente de celle du sujet. L'étude montre que les erreurs commises par les volontaires sont semblables malgré la différence de nationalité des sujets (Scherer et al. 2001).

Nous mentionnions dans le paragraphe 3.5.1 l'utilité de paramètres tels que la fréquence fondamentale, l'intensité ou la durée pour l'analyse prosodique d'un discours. Dès lors que notre

focus se porte sur les états affectifs, une extension de ces paramètres s'avère nécessaire afin de pouvoir prendre en compte toute la richesse de la parole émotionnelle (Morel and Bäzinger 2004). Les descripteurs spectraux et de manière plus générale les descripteurs de qualité vocale, ont démontré leur efficacité dans le cadre d'une synthétisation prosodique des états affectifs (Bulut et al. 2002). Ces descripteurs ont été utilisés, pour l'aspect détection des émotions dans (Vidrascu 2007). Pour rappel, la qualité vocale est directement liée à la phonation, elle caractérise un type de voix. Elle est classiquement définie comme l'ensemble des paramètres pouvant différencier deux contenus lexicalement identiques. Laver (Laver 1991) propose dans son étude le terme « *setting »* qui à par la suite été largement repris pour définir la qualité vocale *: « There is an alternative, wider approach to the task of articulatory description, that concerns itself with both differences and similarities in vocal performance in speech and sees individual segments as momentary actions superimposed on a long term SETTING of the vocal apparatus. The setting accounts for the similarities and the segments for the differences, as it were. A setting gives a background, auditory colouring to sequences of short term segmental articulations* ». Cela peut se traduire, pour une même phrase, par une différence de hauteur, d'intensité, de nasalité ou de tension. La Figure 3-2 ci-dessous détaille à différents niveaux syntaxiques les paramètres classiquement mis en jeu lors de l'évaluation de la qualité vocale :

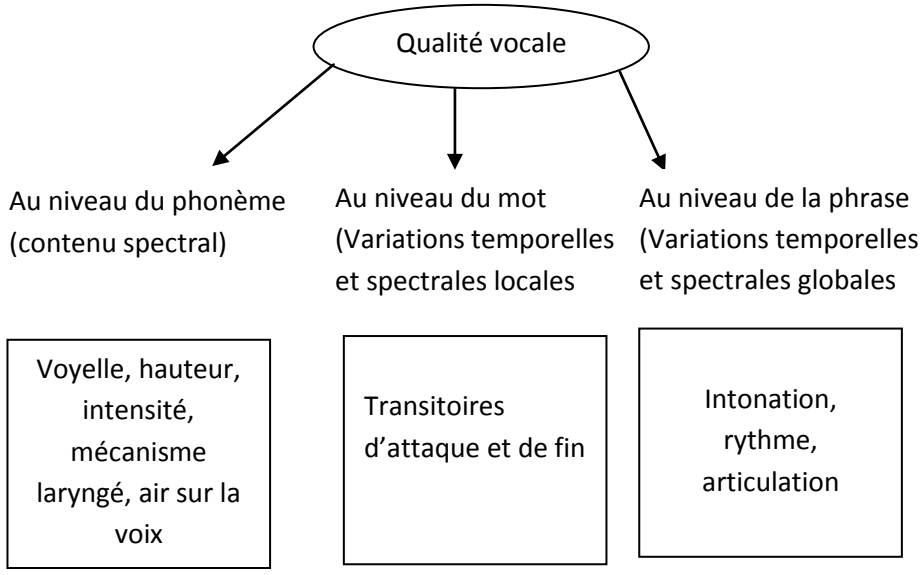


Figure 3-2 Description de la notion de qualité vocale (Garnier et al. 2005)

Dans une tentative de rapprochement entre linguistique et paralinguistique, Lacheret (Lacheret 2011) se base sur des travaux réalisés dans le cadre de la linguistique cognitive et notamment sur la définition apportée par Lambrecht. Ce dernier voit une construction grammaticale comme une forme globale et conceptualisée «*qui se définit comme un appariement entre forme (linguistique) et fonction (communicative, symbolique et représentationnelle) dont les propriétés formelles (syntaxiques, lexicales, prosodiques) et interprétatives ne se déduisent pas strictement des propriétés compositionnelles de la grammaire* » (Lambrecht 2008). Quoique prenant un champ d'observation plus large, l'auteur met ici en lumière un mécanisme qui est proche de celui de Fonagy et du double encodage (Fonagy 1983), puisque le discours est également défini selon deux sources d'informations

(linguistiques et prosodiques) à analyser de manière globale, l'une représentant la forme et l'autre le sens.

Pour illustrer son propos, Lacheret utilise un exemple qui est lexicalement porteur d'une valence positive telle que « ça va ». Il est ainsi expliqué que selon le contour mélodique mis en œuvre lors de l'expression de cette phrase, la valence peut passer du positif (contour montant/descendant avec un rythme ralenti) au négatif (contour descendant et plage de fréquence basse). Celle-ci met donc en lumière tout l'intérêt d'une analyse linguistique contextualisée par l'utilisation de la prosodie. Dans le cadre de notre étude sur la détection d'émotions issues de conversations humain-humain naturelles, de telles ambigüités se produisent fréquemment avec une utilisation de phénomènes langagiers comme l'ironie. Les canaux de communication linguistique et acoustique étant alors typiquement utilisés de façon contradictoire, comme par exemple dans la phrase suivante : « *C'est génial ! Vraiment génial* ». Dans ce cas, seule l'utilisation de l'ensemble des indices permet potentiellement de détecter une incohérence entre les deux canaux de communication. Cette approche peut, dès lors, donner des pistes pour la détection de ce type de pratiques verbales.

Pour aller plus loin dans la mise en relation de la parole et de l'expression verbale Argaman (Argaman 2010), étudie le lien entre lexique et intensité émotionnelle. L'expérience propose de mesurer les marqueurs linguistiques utilisés par 500 sujets pour exprimer leurs opinions à propos de films mettant en jeu des émotions positives (joie) et négatives (tristesse) plus ou moins intenses. La thèse défendue ici est une approche très lexicale de la catégorisation émotionnelle. L'argument avancé est qu'une personne peut facilement contrôler son récit lors de la description d'un événement, mais qu'il est beaucoup plus difficile de contrôler complètement le choix des mots employés (répétitions, expression à la première personne). Même en prenant comme sujet d'étude une personne ayant des capacités avancées à discourir, le contrôle de chaque mot énoncé est un exercice compliqué (Planalp 1999). L'analyse se situerait donc à un niveau descriptif se concentrant sur la façon dont est énoncé lexicalement un discours émotionnel.

Plus précisément, un focus est réalisé sur 10 points de l'expression lexicale. Les plus intéressants concernant notre étude, puisque pouvant potentiellement être traités de façon automatique, sont les suivants :

- l'emploi des « intensifieurs» (« très », « le plus », « extrêmement») et « d'atténuateurs » lexicaux (« assez », « plutôt »)
- Les répétitions : emploi d'un même terme à plusieurs reprises. Une distinction est faite entre les répétitions lexicales orientées vers une cible extérieure et les répétitions lexicales orientées vers le sujet lui-même.
- l'utilisation de la première personne du singulier (« je », « moi »)
- Expression exclamative ou de type affect burst (par exemple « wow », « merde »)
- Nombre de mots émotionnellement marqués
- Le « token ratio » qui est une notion prenant en compte le nombre de mots différents utilisés par rapport au nombre de mots total utilisés dans le texte

Les résultats montrent une différence très marquée en termes d'intensité pour les émotions de type joie forte et faible. A l'inverse, l'expression de la tristesse est quant à elle, assez peu marquée quelque soit l'intensité de l'émotion exprimée dans le film par le sujet observé. Si l'on analyse en détail les résultats obtenus pour les comparaisons des films exprimant la joie, nous pouvons voir que

seul le token ratio n'est pas significatif pour la différentiation de l'intensité. Dans l'ordre du plus discriminant au moins discriminant nous retrouvons : (1) Intensifieurs, (2) répétitions orientées vers soi-même, (3) nombre de mots émotionnels utilisés, (4) utilisation de mots atténuateurs, (5) utilisation de la première personne du singulier.

Même si ce type de méthode ne fournit pas d'indice sur la valence d'un texte donné, la possibilité de pouvoir détecter de manière automatique une intensité émotionnelle lexicale au cours d'une expression naturelle est importante pour le type de modèle que nous cherchons à définir. La frontière au niveau perceptif acoustique entre la colère froide et un discours neutre est extrêmement faible voire inexistante. Cette méthode nous fournirait donc une métrique afin d'observer si l'énoncé est lexicalement plus proche du neutre ou d'une classe émotionnelle. De notre point de vue, le grand avantage de ce type d'indices est qu'il intervient à un niveau lexical facilement manipulable : de simples méthodes de word spotting permettent d'extraire les indices voulus. Ces indices offrent donc potentiellement une résistance au bruit plus importante ; ce qui est un point crucial dans notre situation puisque nous nous plaçons dans le cadre d'un système complètement automatisé de détection des émotions issues de données naturelles. Il est de plus possible avec ce genre de méthode, d'analyser le dialogue suivant des fenêtres temporelles très variées (un ou plusieurs tours de parole, portions de dialogue, dialogue complet), ce qui permet de pouvoir catégoriser de manière plus globale une discussion pouvant s'étaler sur plusieurs dizaines de minutes. Cette approche fait partie de nos perspectives.

3.6La synthèse de la parole émotionnelle : aperçu des méthodes employées

La synthèse de la parole et de manière encore plus flagrante la synthèse de la parole émotionnelle est un exemple des plus représentatifs d'une application technologique mêlant expression verbale et expression prosodique. Le domaine de la synthèse s'intéresse à la représentativité de l'émotion exprimée pour que cette dernière soit simulée de la façon la plus réaliste possible compte tenu du contexte. Une technique classique de synthèse de la parole se base sur une phonétisation et une analyse morphosyntaxique d'un texte fourni en entrée. Des règles de transformations prosodiques (contours de la fréquence fondamentale par exemple) sont appliquées aux sorties morphosyntaxiques afin de réaliser la fonction linguistique segmentale et de hiérarchisation de la prosodie (voir paragraphe 3.5.1). Perceptivement parlant, les systèmes récents n'ont en général plus de problème d'intelligibilité et l'on arrive actuellement à obtenir une parole synthétisée relativement naturelle. Le nouveau challenge se porte à présent sur l'expressivité, définie comme étant la possibilité d'exprimer et de comprendre les émotions, intentions et attitudes à travers la modulation de la voix (Beller and Rodet 2007). Il faut préciser que le concept d'expressivité se propose de mesurer non pas une émotion ressentie mais une émotion exprimée, c'est-à-dire l'état émotionnel pouvant être réellement perçu par l'interlocuteur. Pour avoir des informations plus précises au sujet de l'état émotionnel réellement ressenti par un interlocuteur, plusieurs éléments supplémentaires doivent être pris en compte. Un récapitulatif est présenté dans la Figure 3-3 ci-dessous :

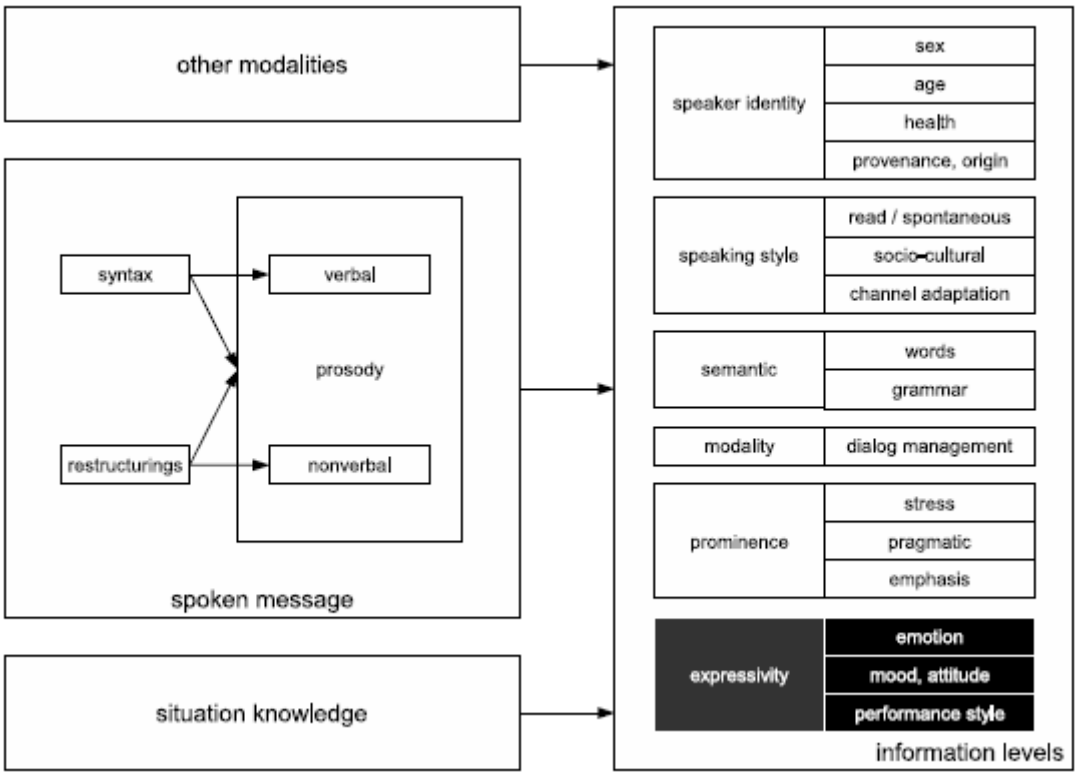


Figure 3-3 Présentation des différents niveaux d'information impliqués dans une communication verbale (partie droite). Le message parlé peut être décomposé par différents supports présentés dans la partie gauche. (Beller 2008)

L'étude présentée dans (Beller 2008) propose une revue détaillée des paramètres présentés sur la Figure 3-3. Il est ainsi possible d'après cette liste de construire un contexte dans lequel s'inscrit la prosodie émotionnelle générée. Ce contexte aura pour but de modifier les paramètres acoustiques employés au niveau de la syllabe. Ainsi chaque syllabe souhaitant être rendue expressive se verra modélisée par la stylisation de ses paramètres acoustiques (F0, intensité, débit de parole, degré d'articulation, etc.) grâce à un modèle fourni par le contexte. Cette méthode offre l'avantage de pouvoir définir par avance un cadre contextuel pertinent pour le type de voix que l'on cherche à synthétiser. Ainsi un contexte défini pour une application en vue d'un jeu vidéo (guerrier par exemple) ou d'une application médicale (infirmière) sera drastiquement différent au niveau contextuel mais avec une approche semblable du point de vue des paramètres utilisés et de la modélisation plus généralement. Cette méthode se base entre autres, sur les paramètres de qualité vocale nécessitant l'extraction des parties voisées et plus précisément des voyelles (Beller 2008).

Du point de vue de la détection des émotions il semble évident que l'expression d'un état affectif (et c'est encore plus vrai pour les données réelles que nous utilisons) est très clairement influencé par son contexte. L'utilisation de modèles permettant une prise en compte de l'expressivité et des paramètres contextuels permettrait très probablement une amélioration des méthodes utilisées. Il reste cependant que ces méthodes nécessitent un travail de prétraitement manuel important (définition du contexte, adaptation à un type de locuteur en fonction de critères sociaux culturels par exemple) qu'il n'est pas possible d'appliquer compte tenu du nombre de locuteurs très importants que nous avons dans les bases de données actuelles (plus de 2000 locuteurs différents).

3.7 Fusion d'indices : méthodes et stratégies

Nous avons vu dans les chapitres et paragraphes précédents que les émotions, voire les états affectifs de manière plus générale, peuvent mettre en jeu des indices très variés. Si les niveaux acoustiques et linguistiques sont les plus étudiés, de plus en plus d'études utilisent des indices multimodaux : audio-visuels, physiologiques (conductivité de la peau par exemple) ou même des signaux cérébraux. Il a été montré que la combinaison de ces indices apportait, dans tous les cas, des gains en termes de détection correcte. Nous allons voir dans ce paragraphe les deux principales méthodes employées pour combiner des indices provenant de sources de données différentes : la fusion précoce et la fusion tardive. La fusion précoce (*feature level fusion*) propose de combiner les descripteurs obtenus d'après les différentes sources de données entre eux pour obtenir un vecteur contenant l'ensemble des informations de chaque source de données. A l'inverse, la fusion tardive (*decision level fusion*) propose de combiner les résultats obtenus en sortie de classifieur. Les méthodes sont détaillées dans le paragraphe dédié à la fusion tardive.

La Figure 3-4 ci-dessous récapitule la différence entre fusion tardive et fusion précoce :

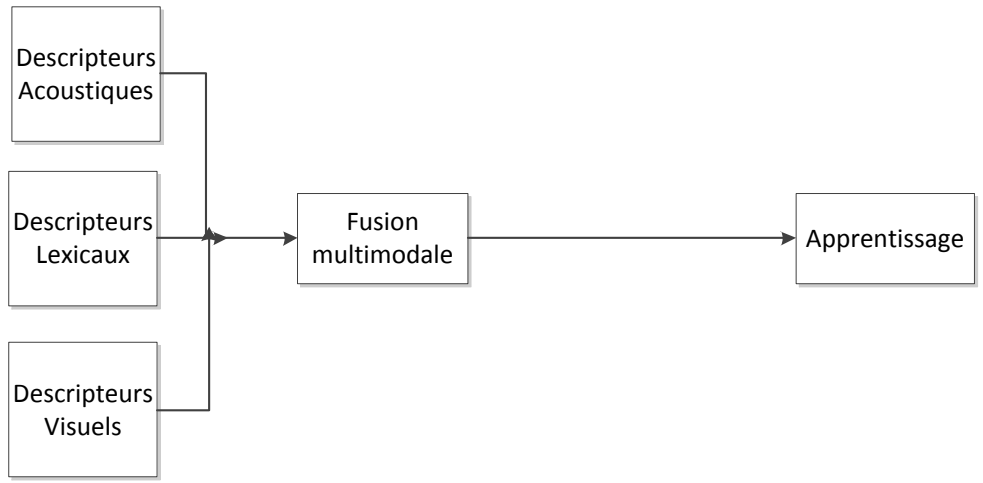


Figure 3-4 Illustration des concepts de fusion tardive (haut) et de fusion précoce (bas)

3.7.1 Fusion au niveau du descripteur (feature level fusion)

La fusion d'informations au niveau du descripteur (ou early fusion) propose de combiner les descripteurs obtenus à partir de sources différentes dans un même vecteur afin que ces derniers puissent être soumis au classifieur. C'est le cas dans (Busso et al. 2004) par exemple où les auteurs proposent une méthode de détection des émotions se basant sur des données issues d'expressions faciales et sur des données acoustiques. De manière générale, si l'on se place dans le cadre de fusions de données multimodales avec un nombre de sources de données supérieures ou égales à 3, la fusion au niveau du descripteur semble donner de meilleurs résultats que la fusion tardive. Dans (Kim 2007) l'auteur utilise plusieurs indices en provenance de signaux physiologiques et d'indices acoustiques pour déterminer la valence d'une émotion. Les résultats sont calculés pour une fusion tardive et une fusion précoce. C'est cette dernière qui donne les scores les plus élevés. Le constat est le même dans (Castellano et al. 2008) qui fusionne des indices venant de la parole, de la gestuelle et de l'analyse faciale. La fusion au niveau des descripteurs améliore la précision de la classification d'environ 4% par rapport à une fusion tardive.

Il faut cependant noter deux limites principales à cette méthode de fusion des données :

- La première est liée à la dimensionnalité du problème étudié. Nous avons énoncé dans les chapitres et paragraphes précédents que le nombre de descripteurs utilisés pour une détection acoustique pouvait être de plusieurs milliers (Schuller et al. 2009). De même nous mentionnions précédemment que les méthodes de représentation de données textuelles comme le sac de mots (Bag of Words) pouvaient donner lieu à des problèmes mettant en jeu un nombre très important de dimensions et que de ce fait, le nombre de mots pris en compte devait être réduit à l'aide de méthodes telles le « stemming ». L'utilisation de méthodes de fusion au niveau du descripteur amplifie encore donc un peu plus des inconvénients déjà existants. Des algorithmes de sélection de descripteurs peuvent être appliqués afin de réduire le nombre de dimensions du problème et éliminer d'éventuels descripteurs non significatifs pour la tâche en cours mais cette tâche est également très coûteuse en temps de calcul.
- La seconde provient du fait que pour obtenir des gains stables dans les performances en utilisant cette méthode les indices fusionnés doivent utiliser une métrique temporelle et segmentale proche. Dans (Busso et al. 2004) par exemple, les gains obtenus avec des méthodes de fusion au niveau du descripteurs sont légèrement plus importants (+1%) comparés à une méthode de fusion tardive. Cependant, seules deux émotions sur cinq (neutre et colère) voient leur taux de détection correcte augmenter. Certaines autres émotions telles que la colère, voient leur taux de reconnaissance baisser de manière importante (-9%). A contrario, la fusion tardive permet d'améliorer la détection de façon plus homogène sur l'ensemble des classes émotionnelles étudiées.

Principalement dans le domaine de la fusion multimodale, des stratégies sont mises en œuvre pour tenter de pallier le problème de cohérence des descripteurs réunis en un seul vecteur. Ainsi, dans (D'Mello and Graesser 2010), les auteurs proposent l'utilisation d'un même nombre de descripteurs de chaque canal d'information afin de ne pas privilégier artificiellement une source d'informations par rapport à une autre.

Les résultats sont nettement plus contrastés lorsque seulement deux sources d'informations sont utilisées ; particulièrement concernant des indices de type linguistique/acoustique. Dans (Schuller et al. 2005), les auteurs utilisent une fusion au niveau du descripteur (early fusion) pour combiner des informations linguistiques et paralinguistiques avec un gain, dans les scores de détection, inférieur à 4%. Les gains apportés par la fusion précoce dans (Steidl 2009) sont également modérés, laissant penser que la combinaison acoustique/linguistique se trouve de manière générale mieux valorisée par une fusion tardive. Le constat est également fait dans (Polzehl et al. 2011) qui délaissent la fusion précoce pour une tâche de combinaison tardive de ces deux types d'indices.

3.7.2 Fusion au niveau de la décision (decision level fusion)

La fusion tardive a pour but de combiner plusieurs sources de données obtenues en sortie de classifieur et pouvant être fortement hétérogènes. La plupart des études mettant en œuvre de la détection d'émotions basées sur une fusion acoustique/linguistique utilisent une fusion tardive, notamment pour les raisons évoquées dans le paragraphe précédent. La fusion tardive présente l'avantage de permettre à l'expérimentateur de jouer de manière fine sur les paramètres des algorithmes d'apprentissage afin d'obtenir un « tuning » précis du classifieur utilisé. Cette méthode est également considérée comme étant la plus robuste au bruit (Hussain et al. 2011). Le plus gros reproche généralement exprimé à l'encontre de la fusion tardive est la perte d'informations, du fait d'une analyse à plus haut niveau, pouvant résulter de la fusion. La combinaison des jeux de descripteurs en provenance des différentes sources de données étant alors impossible. Des méthodes de fusion hybride sont parfois employées, avec succès, pour palier ce problème comme dans (Hussain et al. 2011) où la méthode hybride se révèle être la plus performante parmi les trois méthodes de fusion testées. Celle-ci propose de combiner les sorties de classifieurs entrainées séparément (chaque sortie étant une source de données différente) comprenant un nombre limité et égal de descripteurs bas niveau en provenance des différentes sources de données. Récemment des études comme celles de (Schuller et al. 2011) proposent une fusion tardive en se basant non pas sur une combinaison de classes d'indices (linguistique acoustique par exemple) mais sur des modèles issus de l'entrainement en parallèle de données provenant de différents corpus.

Nous allons passer en revue dans les paragraphes ci-dessous les principales méthodes de fusion tardive existantes. Dans (Utthara et al. 2010) les auteurs décomposent ces méthodes en 3 grandes catégories : La combinaison de classes émotionnelles, le *Class Ranking* et la combinaison de probabilité en sortie des modèles. Ces trois approches sont détaillées dans les paragraphes ci-dessous.

3.7.2.1Combinaison de classes émotionnelles :

Le vote majoritaire

La méthode la plus simple dans ce domaine est celui du vote majoritaire (Kittler et al. 1998). Dans cette configuration, un choix est fait si, en sortie des classifieurs, une étiquette obtient un nombre d'occurrences pour une étiquette donnée plus élevé que les autres. Cette méthode de décision se traduit par l'équation suivante :

$$\sum_{i=1}^{L} d_{i,k} = \max_{j} \sum_{i=1}^{L} d_{i,j}$$

Équation 3-1 Vote majoritaire pour la fusion tardive de données

Où *L* représente les classifieurs utilisés, $d_{i,j}$ la décision du *i* ème classifieur pour la *j* ème classe

Pour que cette méthode de fusion soit acceptable il faut cependant que certaines contraintes soient respectées :

- Le nombre de classifieurs doit être impair afin d'être sûr de pouvoir prendre une décision suivant cette méthode,
- La probabilité d'obtenir une classification correcte pour une instance donnée doit être supérieure à celle d'obtenir une prédiction correcte par hasard pour chacun des classifieurs proposant une solution,
- Les classifieurs doivent être indépendants les uns des autres.

Il faut noter qu'il est courant de constater que les performances des classifieurs peuvent varier de manière importante suivant les sources de données. Dans ce cas, la fusion par vote majoritaire se retrouve alors biaisée, accordant une importance identique à une prédiction qui a une probabilité plus élevée d'être incorrecte qu'une autre provenant d'un classifieur différent. Le vote majoritaire pondéré corrige ce biais en appliquant un poids à chaque classifieur en fonction de sa capacité à reconnaître une classe d'émotions (Littlestone and Warmth 1994).

Combinaison de bayesienne

Pour que cette méthode soit valable il faut que les classifieurs soient considérés comme conditionnellement indépendants pour une classe émotionnelle donnée. Considérons $P(S_j)$ comme étant la probabilité que le classifieur d_j assigne une instance x à classe S_j. L'indépendance conditionnelle permet alors la représentation suivante (Kuncheva 2004):

$$P(\mathbf{s}|\omega_k) = P(s_1, s_2, \ldots, s_L|\omega_k) = \prod_{i=1}^{L} P(s_i|\omega_k)$$

Équation 3-2

1) La probabilité a posteriori nécessaire pour étiqueter une instance x est définie par :

$$p(d_i|S_1,\ldots,S_m) = \frac{p(S_1,\ldots,S_m|d_i)p(d_i)}{p(S_1,\ldots,S_m)},$$

Équation 3-3

Les différentes probabilités sont alors, dans ce cas, estimées indépendamment lors de l'entrainement de chacun des classifieurs.

2) Considérant l'hypothèse d'indépendance nous pouvons exprimer le problème de la manière suivante :

$$p(d_i|S_1,\ldots,S_m) = \frac{\prod_{j=1}^{m} p(S_j|d_i)p(d_i)}{\prod_{j=1}^{m} p(S_j)},$$

Équation 3-4

Au final la décision est généralement prise en attribuant à l'instance ***x*** une classe émotionnelle en fonction du maximum de la probabilité a posteriori qui est déterminée suite aux calculs précédents.

En pratique la combinaison bayésienne naïve a donné des résultats étonnamment bons du fait de la grande résistance au bruit de la méthode (Kuncheva 2004). Il faut en effet noter qu'en pratique les conditions d'indépendance statistique sont rarement respectées.

3.7.2.2Class ranking

La méthode des rangs est essentiellement applicable aux problèmes mettant en jeu un nombre important de classes de données. Dans cette méthode, des rangs sont attribués aux étiquettes de classes. Les rangs peuvent être attribués par un expert du domaine ou, comme c'est le plus souvent le cas, par un ensemble de classifieurs. L'idée de la méthode est de réduire progressivement le nombre de classes du problème tout en réordonnant les classes restantes à chaque itération. D'après cette méthode, la classe-solution du problème devrait être celle restante en tête du classement après *n* itérations.

La méthode de Borda count (Borda 1781) propose d'attribuer un rang à chaque alternative possible. A l'alternative rangée en première position, sont affectées *n* voix données par des électeurs. L'alternative suivante dans le classement obtient *n-1* voix et ainsi de suite. A la fin de la procédure, une somme est effectuée pour chaque alternative. L'alternative ayant la somme la plus élevée est

celle qui gagne l'élection. Lorsque cette méthode est appliquée, les classifieurs remplacent les électeurs et les classes, les alternatives (Van-Erp and Schomaker 2000).

Etant donné L classifieurs, portant sur M classes et ayant N instances dédiées à l'apprentissage réparties dans les M classes, la distance de l'instance inconnue par rapport aux instances appartenant à chaque classe donne une indication sur la classe d'appartenance de l'instance inconnue. On trie les prédictions des classifieurs (la distance attribuée à l'instance inconnue pour chaque classe) par ordre croissant. La prédiction de la classe correspondant le plus à l'une des M classes d'apprentissage (c'est-à-dire avec la distance la plus petite) se voit attribuer N votes ; la classe suivante N – 1 votes, et ainsi de suite. La classe avec la distance la plus lointaine se verra attribuer un nombre de votes égal à 1. Le nombre de votes est obtenu en additionnant les votes de chacun des classifieurs pour une classe et une instance donnée. Au final le choix se porte sur la classe ayant obtenu le score le plus élevé parmi les M classes.

Deux variantes existent pour ces méthodes (Van-Erp and Schomaker 2000). Dans la première, la valeur médiane est extraite à la place de la somme des votes relatifs aux indices des différentes classes. La méthode de Nanson exécute M – 1 fois la méthode de Borda originale et retire à chaque itération la classe ayant obtenu le nombre de votes le plus faible. L'arrêt se produit lorsqu'une seule classe reste présente.

3.7.2.3Combinaison de probabilités des sorties

Dans (Lam and Suen 1995) les auteurs décrivent une méthode permettant la conception de profils de décision en fonction des probabilités de sorties de chaque classifieur inclus dans le système. La méthode est décrite ci-dessous :

Dans un système comprenant plusieurs classifieurs, un vecteur x est classifié dans une des C classes utilisant L classifieurs. Chaque classifieur D_i fournit un ensemble de C valeurs comprises entre 0 et 1 pour chaque classe. Les valeurs fournies par le classifieur D_i pour la classe C_j est $d_{ij(x)}$. Traditionnellement, plus la valeur est grande, plus la confiance accordée à ce jugement est importante (ces valeurs peuvent être les probabilités pour une instance donnée d'appartenir à une classe donnée, les F-scores obtenus pour une classe de données, etc). La sortie des L classifieurs pour une instance X donnée est alors définie par un profil de décision DP(X) donné dans la Figure 3-5 :

$$\begin{pmatrix} d_{1,1}(x) & \dots & d_{1,j}(x) & \dots & d_{1,c}(x) \\ d_{i,1}(x) & \dots & d_{i,j}(x) & \dots & d_{i,c}(x) \\ d_{L,1}(x) & \dots & d_{L,j}(x) & \dots & d_{L,c}(x) \end{pmatrix}$$

Figure 3-5 Profil de décision établi d'après les valeurs de confiance obtenues en sortie des classifieurs

La fusion des indices peut être obtenue de deux façons différentes :

- La première propose de calculer un score par classe (en prenant en compte les différentes colonnes de la matrice séparément). Différentes opérations peuvent être appliquées (somme, produit, maximum, minimum). Suivant l'opération appliquée la décision peut être prise en fonction de la classe obtenant le score maximum, minimum etc

- La seconde approche prend en compte l'ensemble des données du profil de décision sans tenir compte des différentes classes. Deux méthodes principales ont été employées pour cette seconde approche. La première est nommée « decision template » par (Kuncheva et al. 2001). L'idée est, pour chaque classe, de retenir la combinaison la plus prototypique d'après le profil de décision. Des mesure de distance comme la distance euclidienne sont appliquées pour calculer la similarité entre un patron de décision et une instance **x** soumise pour la classification. La seconde méthode classiquement utilisée pour cette approche est la théorie de Dempster-Shafer qui propose de calculer un degré de croyance pour chaque sous ensemble d'un problème d'apprentissage (Shafer 1976).

3.8 La détection automatique des émotions dans les centres d'appels

Les chapitres précédents ont mis en évidence l'importance qu'a la mise en relation de la parole et de l'expression verbale. Il faut toutefois remarquer que dans la grande majorité des cas, les études précédemment mentionnées nécessitaient des prétraitements manuels importants qu'il est difficile d'avoir dans le cadre d'une détection des émotions à grande échelle comme nous proposons de le faire dans cette thèse. Hormis des contraintes de temps et d'argent, l'alignement par phonème ou l'annotation contextuelle d'un dialogue ne seraient plus disponibles dans le cadre d'un système automatique car non détectables dans le cadre d'une classification sur des dialogues non prétraités. Nous allons toutefois voir dans la section 3.8.1 ci-dessous, que l'alliance linguistique-acoustique est possible et bénéfique pour la conception de modèles automatiques sur des données réelles. Toutes les études que nous citons ci-dessous ont la particularité d'avoir été menées sur des données collectées dans des centres d'appels. Nous nous attacherons à montrer ici l'intérêt, en termes de taux de reconnaissance, d'allier différents indices lors de la détection d'émotions réelles en centre d'appel.

3.8.1 Détection automatique des émotions dans les centres d'appels : indices et méthodes utilisés

Les centres d'appels constituent un support privilégié pour l'étude des émotions dans la voix. Dans ce contexte très précis, la voix est le seul support permettant l'expression d'un état affectif de l'appelant vers l'appelé. Historiquement, il semble que la première étude réalisée sur la détection des émotions au téléphone fut proposée par (Petrushin 1999). Il y est proposé la détection de deux états affectifs :

- « calme » (comprenant principalement du neutre)
- et « agité » comprenant les émotions de type colère, joie et peur.

Les données utilisées sont des messages téléphoniques de 15 à 90 secondes contenant un seul locuteur. Après sélection, l'étude est réalisée à l'aide de 14 descripteurs (F0, énergie, débit de parole…). Les résultats montrent un taux de détection correct de l'ordre de 77% sur deux classes, l'émotion la mieux reconnue étant la colère.

Dans un contexte de dialogues homme-machine enregistrés lors de conversations téléphoniques réelles, Lee (Lee et al. 2001) propose la détection de deux états « négatif » vs. « Non

négatif». Les descripteurs utilisés sont relatifs à la fréquence fondamentale et à l'énergie. Dans cette étude les instances faisant intervenir des hommes et des femmes sont analysées séparément afin de palier à la variation importante de F0. Le taux d'erreur le plus bas obtenu dans cette étude est de 20% sur les instances où des femmes sont intervenues. L'une des premières études prenant conjointement en compte l'utilisation de descripteurs linguistiques et paralinguistiques fut présentée dans (Lee et al. 2002). La méthode de détection lexicale nommée *« emotional sailence »* propose d'attribuer une probabilité à un mot en fonction de sa fréquence d'apparition dans une classe émotionnelle donnée. Les descripteurs acoustiques sont, quant à eux, basés sur les paramètres relatifs à la fréquence fondamentale et à l'énergie. En combinant les deux types de descripteurs avec une approche bayesienne les gains en termes de détection vont dans le meilleur des cas jusqu'à 45% d'instances supplémentaires classifiées correctement.

Dans une autre étude portant sur la détection d'émotions composées lors d'interactions entre humains (Vidrascu and Devillers 2005a), les auteurs différencient le rôle de l'agent et du client dans le but d'obtenir des modèles de détection plus fiables. La prise en compte d'émotions composées (« blended emotions ») est également envisagée, afin de traduire la richesse des expressions émotionnelles observées dans un centre d'appels pour les urgences médicales. Dans le même ordre d'idée, les auteurs ont étudié des données collectées dans un centre d'appels s'occupant des échanges boursiers, un modèle est construit avec une prise en compte d'émotions complexes i.e pouvant être annotées à l'aide de plusieurs étiquettes. Les scores de détection correcte sont comparés entre une détection « négatif » vs « neutre » et « peur » vs « colère ». Il ressort que les émotions qui ont été annotées avec plusieurs étiquettes (comme peur/colère) obtiennent des scores de détection plus bas. L'étude met en lumière le besoin d'adapter les modèles aux données issues de corpus réels et peut-être de construire des classes représentant des données complexes (comme peur/colère) (Vidrascu and Devillers 2005b).

Dans le cadre d'un dialogue homme-machine et dans le but d'obtenir des modèles de détection plus robustes, il a été envisagé d'introduire d'autres types d'indices pour la classification des émotions. Ainsi dans (Lee and Narayanan 2005), des informations concernant le discours sont introduites. Cinq actes de dialogues sont donc définis et annotés manuellement : *rejection, repeat, rephrase, ask-start over* et *none of the above*. Ces informations sont combinées avec des indices acoustiques et lexicaux (emotionnal sailence) par le biais d'une fusion tardive. La combinaison de ces 3 niveaux d'information offre un gain de performance lors de la phase de détection de l'ordre de 40% par rapport à la configuration de base utilisant seulement des indices acoustiques.

D'autres indices non verbaux, comme les rires et les disfluences, ont été mis en œuvre pour la détection de 4 émotions (colère, peur, positif, neutre) (Devillers et al. 2005). Une fusion des indices linguistiques et paralinguistiques a été appliquée aux données en sortie des classifieurs, en utilisant une combinaison linéaire. Les transcriptions ont été réalisées manuellement et l'alignement par phonème automatiquement, dans le but d'annoter des descripteurs verbaux comme les disfluences. Les résultats montrent un gain d'environ 5% lorsque les indices sont fusionnés. Ce résultat peut sembler inférieur à ceux présentés dans les études citées ci-dessous mais la complexité des données justifie ce résultat. De nombreux événements tels que la parole superposée (10% des segments), l'expression d'émotions mélangées ou masquées, propres à la parole spontanée, sont en effet présents dans ce type de corpus ce qui rend plus difficile la détection automatique des émotions.

Récemment, des études cross-corpus permettant de tester le pouvoir de généralisation des modèles commencent à apparaître (Devillers et al. 2010). Dans cette étude, deux corpus basés sur des données issues de centres d'appels sont conjointement utilisés (services clients et urgences

médicales). Les résultats montrent que malgré un contexte identique (centres d'appels), les modèles sont difficilement généralisables, notamment du fait des différences des tâches effectuées, et par conséquent, des différences d'expressions pour une même émotion. L'expression d'émotions positives étant dans ce cas la classe la plus affectée par les changements de corpus passant de 34% de détection correcte à 10% lorsque le corpus d'entrainement était différent de celui de test. Ce résultat concernant la difficulté de généralisation des modèles a également été mis en avant lors d'une étude sur la détection de la colère utilisant 3 corpus dont 2 centres d'appels, l'un en anglais et l'autre en allemand, intervenant dans le domaine du service après vente (DHM) et 1 corpus où des enfants communiquent de façon spontanée avec un robot (Polzehl et al. 2011). Si les résultats restent stables concernant les tests effectués sur les corpus de centre d'appels, ceux obtenus sur le corpus de dialogue enfant-robot sont quant à eux nettement plus bas. Les auteurs mentionnent par ailleurs la difficulté de détecter correctement la colère de manière lexicale (notamment du fait du problème de la taille du vocabulaire disponible). La fusion tardive des informations lexicales et linguistiques offre dans cette étude un gain compris entre 15% et 18% de détection correcte pour les corpus de dialogue enfant-robot et de centres d'appels.

Le Tableau 3-2 récapitule l'ensemble des travaux concernant les centres d'appels cités dans les paragraphes ci-dessus :

Type de donnée : DHH/DHM	Nombre de corpus/Nombre d'instances	Nombre de locuteurs	Nombres de labels	Descripteurs	Auteurs
DHM	1/142	/	2 (Negative – non negative)	Paralinguistique	(Lee et al. 2001)
DHM	1/1179	/	2 (Negative – non negative)	Paralinguistique et lexicale	(Lee et al. 2002)
DHM	1/1351	/	2 (Negative – non negative)	Paralinguistique, lexicale, informations sur le discours	(Lee and Narayanan 2005)
DHH	1/800	/	2 (Negative-Positive)	Paralinguistique	(Vidrascu and Devillers 2005a)
DHH	1/5000	104	2 neutre-négatif Et Peur-colère	Paralinguistique	(Vidrascu and Devillers 2005b)
DHH	2/19045	514	4 (colère, peur neutre, positif)	Paralinguistique + linguistique + non verbal	(Devillers et al. 2005)
DHH	2/7452	/	4 (colère forte, colère faible, neutre, positif)	Paralinguistique	(Devillers et al. 2010)
DHM	3/17286	1126	2 (colère, autres)	Paralinguistique + linguistique	(Polzehl et al. 2011)

Tableau 3-2 Récapitulatif des corpus et indices utilisés lors d'études portant sur les centres d'appels

Nous avons pu voir à travers le récapitulatif des études existantes sur les centre d'appels que la fusion d'indices apporte systématiquement des gains lors de la phase de détection. De manière globale, la détection acoustique donne individuellement des gains supérieurs à celle de la détection lexicale. A noter également, que les corpus de très grande taille comme ceux utilisés dans (Devillers et al. 2005; Polzehl et al. 2011) mettent en œuvre un nombre de locuteurs important (>500), ce qui représente une difficulté supplémentaire pour la détection acoustique. La conception de modèles cross-corpus dans le but d'avoir des modèles prenant en compte plus d'instances semble également être une difficulté lorsque la tâche varie. Nous pouvons également voir que peu d'études mettent en œuvre des modèles intervenant sur plus de deux à quatre classes d'émotions, les scores chutant très rapidement lorsque plus de classes émotionnelles sont ajoutées (Vidrascu and Devillers 2005a).

3.9Conclusion

Nous avons présenté dans ce chapitre un état de l'art concernant les différents types d'indices acoustiques et linguistiques. Nous avons vu également les deux principales méthodes de fusion permettant de combiner ces indices. Lors du passage à une détection automatique et à l'utilisation conjointe de modèles acoustiques et linguistiques, la fusion des indices apporte systématiquement un gain par rapport aux deux méthodes prises séparément. Dans les différentes études que nous avons présentées dans ce chapitre, l'aspect linguistique est essentiellement abordé d'un point de vue lexical (calcul des probabilités, pour un mot ou une classe de mots). Il semblerait donc intéressant, étant donné les études linguistiques théoriques, de pouvoir enrichir les modèles avec des indices de niveaux sémantiques voir pragmatiques. Dans ce contexte les méthodes utilisées en détection des sentiments et opinions semblent pouvoir fournir un complément intéressant à celui des méthodes classiquement employées en détection des émotions. Nous nous sommes également penchés sur la façon dont ses états affectifs sont exprimés et sur le lien inextricable unissant prosodie et expression verbale. Nous avons pu observer à travers des études linguistiques que la prosodie intervient à tous les niveaux de l'expression verbale (Lacheret-Dujour 2011; Lacheret-Dujour et al. 2011) et qu'il est, dans une certaine mesure, possible de détecter de façon automatique les proéminences (Avanzi et al. 2011).

Partie 2

Comme nous l'avons vu dans les chapitres précédents la parole est un acte complexe mettant en jeu de nombreux éléments (indices acoustiques, syntaxiques, sémantiques, dialogiques). Chacun de ces types d'indices est porteur d'informations sur l'état affectif d'un locuteur si l'on se place au niveau du tour de parole ou à un niveau interactionnel. Les chapitres 4, 5, 6, 7 et 8 représentent les apports principaux de cette thèse. Du point de vue de la terminologie nous utiliserons dans les chapitres suivants le terme « modèle » pour désigner principalement des modèles computationnels. Le chapitre 4 présente les données et le schéma d'annotations. Le chapitre 5 décrit les indices et modèles utilisés ainsi que les performances obtenues sur des données « propres » (segmentation manuelle…). Les trois derniers chapitres 6, 7 et 8 utilisent au minimum deux des trois types d'indices présentés. Nous récapitulons dans la figure ci-dessous ces différents indices :

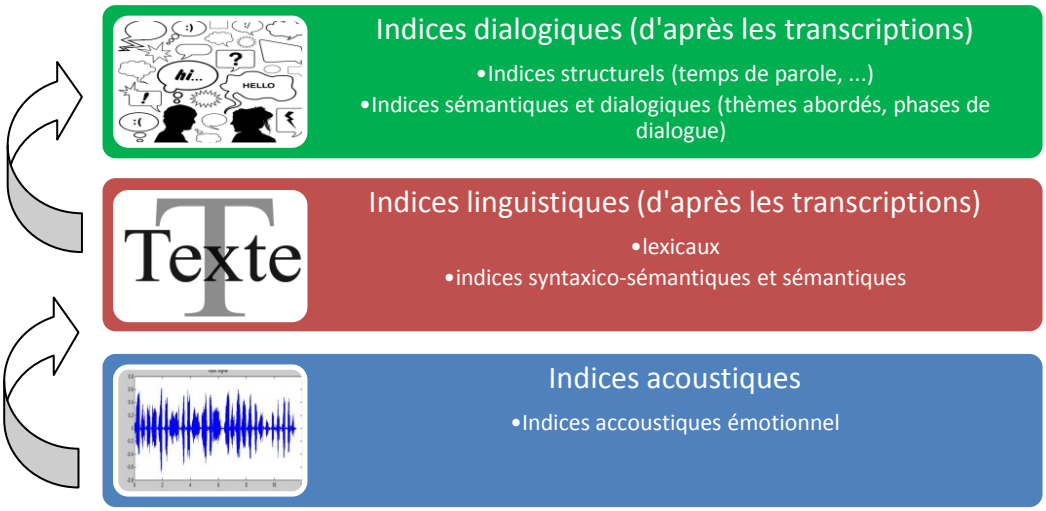


Figure 0-1 Les différents types d'indices utilisés pour la suite de l'étude

Le chapitre 6 se concentre sur la fusion des indices acoustiques et linguistiques (lexicaux, syntaxico-sémantiques et sémantiques). Il met également en avant la complexité des données de cette étude à travers la comparaison des performances obtenues en fonction du type de pré-traitement (automatique ou manuel). Le chapitre 7 se base sur une approche mixte acoustique-sémantique dans le but de sélectionner automatiquement des instances supplémentaires destinées à enrichir les modèles existants. Enfin le chapitre 8 se propose de considérer le dialogue comme unité de base. Les prédictions sur les tours de parole des modèles des précédents chapitres sont dans ce cas réutilisées et combinées avec les indices extraits au niveau du dialogue (indices structurels, …). Le but dans cette dernière partie est de présenter les éléments déterminants (émotionnels et interactionnels) pour la qualification de la satisfaction client.

Chapitre 4 Données utilisées et schéma d'annotation proposé

Résumé

Nous présentons dans ce chapitre les différentes étapes qui ont été nécessaires à la constitution de notre corpus d'apprentissage. Ces étapes commencent avec la sélection de dialogues contenant des expressions affectives. Cette sélection peut être réalisée de manière manuelle (comme c'est le cas pour ce chapitre) ou automatique (c'est le cas dans le chapitre 7). La première sélection effectuée à partir de critères lexicaux et acoustiques a permis d'obtenir 13h de parole à partir de dialogues. Notre but étant l'évaluation de la satisfaction client via l'expression émotionnelle de ce dernier nous avons choisi des dialogues avec une présence importante d'émotions négatives (réclamations) ou positive (satisfaction due à la résolution d'un problème).

Une fois les dialogues sélectionnés, une étape d'annotation manuelle est nécessaire. Nous présentons ainsi les étiquettes et dimensions retenues pour l'annotation de notre corpus. Nous présentons également les deux types de segmentations (automatiques et manuelles) utilisées pour le découpage du corpus en tour de parole.

Nos données se basent sur des conversations téléphonique qui sont, par définition, un enchainement de tours de parole organisés et établissant des liens sémantiques et pragmatiques entre eux. Nous avons souhaité annoter ce corpus en prenant en compte des informations sur les interactions dans le dialogue. Nous avons également expérimenté deux types d'annotations: discrète et continue. Nous concluons que malgré des performances similaires (scores d'accords, temps passé pour l'annotation), la première méthode basée sur une approche discrète semble préférable car jugée plus facile d'accès du de point de vue des annotateurs.

Etant donné que nous travaillons sur des conversations nous avons souhaité modéliser la dimension interactionnelle de nos données. Nous introduisons pour cela deux dimensions que nous nommons dimensions « d'implication affective » et qui ont pour but de représenter la façon dont les participants réagissent aux énonciations de leurs interlocuteurs. Ces deux dimensions offrent un outil théorique pour l'analyse de conversations naturelles mais sont difficilement automatisables en l'état. Le chapitre 8 s'appuie cependant sur certains éléments de l'implication affective pour proposer une signature émotionnelle sur des dialogues complets.

4.1Collecte de données

Les données que nous avons utilisées pour cette étude sont issues de deux centres d'appels EDF différents, l'un pour les professionnels et l'autre pour les particuliers. Les centres d'appels d'où sont issues les conversations traitent principalement des réclamations clients. Les appels abordent des thèmes variés (coupure de courant pour l'usager, demande d'échéancier pour la facturation, demande d'intervention pour rétablir une installation défectueuse). Nous cherchons dans cette étude à détecter le niveau de satisfaction des locuteurs via leurs expressions émotionnelles. Dans ce but des émotions considérées comme négatives (colère, agacement, peur) ou positives (joie, satisfaction) ont été collectées.

Etant donné le grand volume de données disponibles (1620 heures au total), une pré-sélection des dialogues ayant du contenu émotionnel a été nécessaire. Pour cette sélection nous avons utilisé l'outil développé dans le projet CallSurf (Garnier-Rizet et al. 2008), permettant de sélectionner, par l'intermédiaire d'une interface développée pour le projet, des conversations par sujet d'appel ou par mot clé. Cette navigation est rendue possible grâce aux transcriptions automatiques des conversations téléphoniques. Le système de reconnaissance de la parole inclus dans cet outil permet d'obtenir les transcriptions des appels et ainsi d'effectuer des recherches par mots clés ou thématique. Les méthodes de traitement du langage utilisées dans cet outil ont été optimisées pour être applicables au langage parlé (Cailliau and Giraudel 2008; Garnier-Rizet et al. 2008).

La sélection des dialogues s'est déroulée en deux temps :

1) Sélection à partir de critères lexicaux : Nous avons dans un premier temps commencé par effectuer des recherches par mots clés correspondant à différents sujets pouvant renvoyer à des conversations émotionnelles (coupure, facture, résiliation, rejet, frais) mais les résultats renvoyaient à des dialogues pas forcément émotionnellement marqués acoustiquement parlant. Cette première approche à partir de sélections de thèmes a abouti à l'extraction d'un grand nombre de dialogues. Nous avons cherché à utiliser des mots évaluatifs ou portant un certain degré d'affectivité souvent utilisés à l'oral lors de pertes du contrôle de soi. Ces mots sont le plus souvent des noms connotés en terme de niveau de langue (langage familier, ordurier) tels que *gueule ou conne*, ou de dérivés de noms notamment avec les suffixes –ard ou –asse (*connard*), ou bien de noms orientés négativement (*voleur, bagarre*). Il peut ensuite s'agir d'adverbes (*horriblement*), d'interjections ou de jurons qui permettent d'exprimer, à l'état brut, un sentiment plus ou moins vif du locuteur (*putain*) et de verbes (*écœurer, exaspérer*) dont une bonne part appartient à la classe des verbes de dire comme (*crier*, *hurler*) et plus particulièrement des verbes qui indiquent en plus du sens la façon dont les paroles ont été prononcées.

2) Sélection à partir de critères acoustiques : Dans un second temps, nous avons écouté l'ensemble des premiers dialogues sélectionnés (environ 150) afin de constituer un sous-ensemble basé cette fois-ci sur des critères acoustiques. Etant donné que nous partons du principe que les émotions des dialogues vont nous servir à estimer le degré de satisfaction client, nous avons basé la sélection suivant 3 grandes classes émotionnelles (positif, négatif et neutre) en veillant à ce que ces dernières soient à peu près équivalentes en terme de durée de parole. Concernant la classe de dialogue contenant des émotions négatives nous observons deux tendances principales. La première est celle

mettant en œuvre de la colère «chaude » et présentant une intensité élevée. La seconde contient des éléments plus nuancés comme l'irritation, l'agacement ou la tristesse. Ce choix a été fait dans le but d'avoir un spectre suffisamment large d'expressions émotionnelles négatives. La classe positive a été sélectionnée suivant le même protocole. Une partie de la classe contient de la parole positive présentant une intensité élevée dans la voix (joie, surprise positive) ; la seconde partie de cette classe présente une émotion avec une activation beaucoup plus faible (satisfaction, amusement).

La sélection finale d'un dialogue a été validée après l'accord de deux personnes expérimentées de sexe féminin respectivement de 30 et 35, de formation linguistique, entraînées à la tâche d'annotation.

Nous soulignons par ailleurs que les deux critères qui ont été utilisés pour la phase de sélection nous ont permis de nous assurer que les données étaient les plus représentatives possibles des appels téléphoniques pouvant être reçus dans les centres EDF. Il nous semblait en effet important de ne pas sélectionner des émotions trop prototypiques afin de pouvoir construire des modèles plus robustes face à la grande variabilité de l'expression émotionnelle.

4.2Le corpus emoVox : description, segmentation des données et stratégies d'annotations

Définition 1. *On appelle « extrait » une portion d'un dialogue constitué de un ou plusieurs tours de parole. Un extrait est choisi au sein d'un dialogue pour son contenu émotionnel*.

Définition 2 *Un segment manuel est constitué d'un ou de moins d'un tour de parole. Le découpage en segments a été réalisé dans le but d'avoir un contenu le plus homogène possible (en termes d'émotion exprimée) pour la durée du segment.*

Définition 3 *Un segment automatique est extrait du flux de parole. Les principaux critères permettant de délimiter un segment automatiquement sont les silences* (pause ou respiration*) ou le changement de locuteur.*

Le corpus issu de la sélection, que nous avons nommé emoVox, est constitué d'après les critères mentionnés dans la section 4.1 et comprend un total de 107 dialogues différents pour un temps de parole d'environ 13h. La durée d'un dialogue est comprise entre une et trente minutes. Il faut noter que pour 72 dialogues du corpus emoVox nous avons sélectionné seulement des extraits (voir définition 1) pour éviter d'avoir des classes émotionnelles trop disproportionnées en nombre de segments (les segments annotés comme étant « neutre » sont en effet très largement majoritaires (90%) même dans des dialogues dit « émotionnels »). Notre corpus comprend un total de 139 locuteurs différents répartis en 115 clients et 24 agents. Lors de la sélection des dialogues pour emoVox (i.e. après les deux étapes de présélection) 55 dialogues ont été choisis pour leur contenu plutôt positif et 52 pour leur contenu jugé plutôt négatif. Nous n'avons pas sélectionné spécifiquement de dialogues pour la classe neutre, cette dernière classe d'émotion étant toujours largement majoritaire même dans des dialogues qualifiés de manière globale de positifs ou négatifs.

Deux méthodes de segmentation, manuelle (voir définition 2) et automatique (voir définition 3), ont été utilisées pour le découpage des dialogues sélectionnés en segments. Le segmenteur automatique

inclus dans le système de reconnaissance se base principalement sur des informations acoustiques pour déterminer les frontières de chaque segment. Les critères acoustiques utilisés sont principalement le silence (pause dans le discours) ou la détection du changement de locuteur (agent ou client). Il faut de plus noter que la stratégie utilisée dans le système de reconnaissance de la parole privilégie le taux de reconnaissance des mots de la conversation à une optimisation de la segmentation. Cela explique la différence, en termes de durée, entre les segments automatiques et manuels comme nous le verrons dans les paragraphes suivants. D'un point de vue théorique, contrairement au segment manuel, aucune information pragmatiquo-sémantique n'est utilisée pour segmenter le dialogue, ce qui a pour conséquence de fournir un nombre des segments longs beaucoup plus important que la méthode manuelle (11798 segments pour la segmentation manuelle vs 7094 segments pour la segmentation automatique). Une analyse détaillée de l'impact du type de segmentation est donnée en chapitre 6.

4.2.1Stratégies d'annotation de données émotionnelles

Dans l'étude des phénomènes émotionnels, la fiabilité des données sélectionnées, leurs représentations ainsi que leurs annotations se révèlent être un problème particulièrement épineux. Le réseau d'excellence (ReX) « HUMAINE » a déjà, par le passé, proposé de nombreuses synthèses sur ce thème et contribué à l'élaboration d'un langage de description des émotions (Schröder et al. 2006; Douglas-Cowie et al. 2007; Douglas-Cowie et al. 2011). Ce schéma se base sur l'utilisation conjointe de représentations catégorielles à l'aide d'étiquettes, de dimensions ainsi que d'informations contextuelles dont un état émotionnel serait le résultat. Dans (Devillers et al. 2005) les auteurs proposent la possibilité de sélectionner plusieurs étiquettes émotionnelles (majeur pour la ou les émotions dominantes et mineur pour les émotions pouvant être perçues moins distinctement) pour un même segment dans le but de pouvoir représenter toute la richesse des données réelles. Ces dernières peuvent être couplées à une valence (positive, négative et neutre). Dans une étude sur la détection de la peur (Clavel 2007) se base sur une annotation utilisant des étiquettes seulement, l'approche dimensionnelle n'ayant pas donné de résultats probants en terme de taux d'accord entre les différents annotateurs. Une étude effectuée dans le cadre du projet européen SEMAINE propose une annotation avec une approche dimensionnelle couplée à une annotation continue des dimensions de valence et d'activation (Eyben et al. 2008). Pour une autre expérience sur des vidéos (Vaudable et al. 2009), nous avons souhaité vérifier la corrélation entre les dimensions et les étiquettes verbales. Nous avons créé un repère composé de dimensions valence et activation et avons demandé à 30 sujets de placer des étiquettes relatives à des états émotionnels au sein de ce repère. Chaque sujet devait ensuite annoter une série de vidéos en utilisant l'espace émotionnel préalablement construit. L'image ci-dessous donne un exemple d'espace émotionnel construit par un sujet :

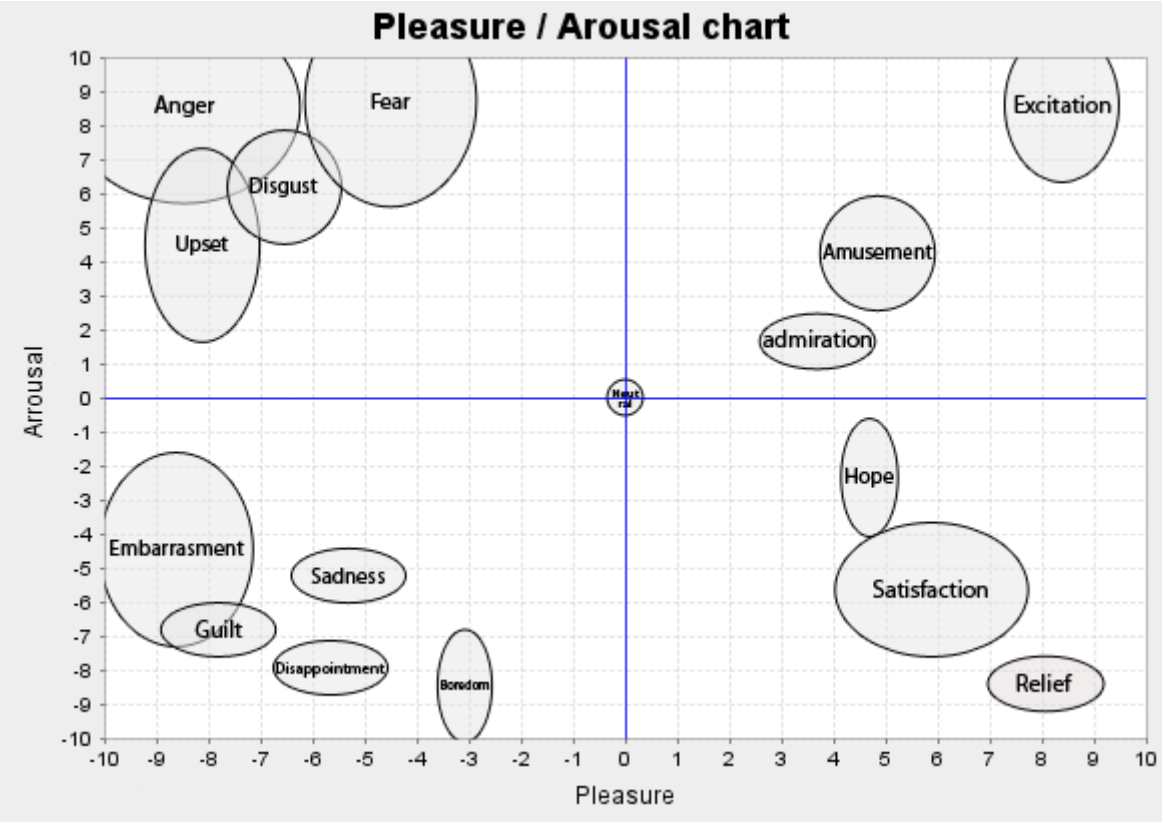


Figure 4-1 Espace émotionnel construit par un utilisateur se basant sur une série d'étiquette et sur les dimensions activation et valence (emoTool)

Les résultats obtenus ont montré des taux d'accord proche de 100% en ce qui concerne le placement positif ou négatif des étiquettes émotionnelles dans le repère (dans les 4 sous-parties) et d'environ 74% pour le placement de ces même étiquettes si l'on considérait les deux axes valence et activation. L'étude met également en avant la différence des placements de ces étiquettes au sein du repère en fonction de la personnalité des sujets (introverti/extraverti, neurotiscisme, etc.).

4.2.2Schéma d'annotation : labels et dimensions

Nous avons, pour l'étude sur les données de centres d'appels, choisi une approche à base d'étiquettes (voir paragraphe 4.2.2.2) et de dimensions abstraites (paragraphe 4.2.2.1) permettant l'annotation en interaction d'une conversation téléphonique et qui sont détaillés dans les paragraphes ci-dessous. Notre stratégie d'annotation a été mise en œuvre à travers la réalisation d'un outil d'annotation dédié que nous avons nommé « *emoTool »(Fig. 4.1).* Cet outil à l'image de Feeltrace (Cowie et al. 2000) par exemple permet l'annotation d'un dialogue avec une approche mixte (continue et discrète). La possibilité d'annoter les dialogues en contexte (les segments sont présentés dans leur ordre d'apparition au cours de la conversation) ou hors contexte (segments dans le désordre) est également une spécificité de cet outil.

4.2.2.1Les dimensions émotionnelles abstraites :

Les dimensions émotionnelles utilisées se basent sur le modèle *Pleasure Arousal Dominance* défini par (Mehrabian 1998) :

La valence : elle permet de déterminer si une émotion est agréable/désagréable et ce à différents niveaux d'intensité (de -2 pour une émotion très négative comme la colère par exemple à +2 pour une émotion très positive comme la joie).

L'*activation* : elle mesure la force avec laquelle le locuteur marque l'émotion qu'il exprime. Nous nous attachons à annoter l'intensité acoustique perceptible et non pas le vocabulaire (le lexique est également chargé de sens portant une orientation d'une part de valence, et d'autre part une information d'intensité : *être en colère* versus *être furieux)*. Enfin, certaines émotions sont plus aisément extériorisables : c'est le cas de la joie et de la colère contrairement à la satisfaction par exemple.

La dominance : elle s'attache à mesurer le contrôle opéré par les interlocuteurs sur eux-mêmes lors d'une interaction. L'émotion peut submerger l'interlocuteur qui perd le contrôle sur cette émotion ressentie. Cette dimension est fortement corrélée avec l'intensité : une émotion exprimée avec intensité aura tendance à renvoyer un contrôle faible.

4.2.2.2Les étiquettes émotionnelles :

Les 16 étiquettes utilisées sont présentées dans le Tableau 4-1 ci-dessous et ont été regroupées sous 6 « macro classes ». Les macros classes apparaissent en gras dans le tableau :

Macro-classe de l'émotion	Valeur d'annotation	Macro-classe de l'émotion	Valeur d'annotation
1) POSITIF	Satisfaction	**3) PEUR**	Doute
	Soulagement		Inquiétude
	Joie		Stress
	Excitation		
	Surprise positive	**4) NEUTRE**	Neutre
	Amusement	**5) TRISTESSE**	Déception
	Espoir		
2) COLERE	Irritation		Désespoir
	Surprise négative		Impuissance

Tableau 4-1 Récapitulatif des macro-classes / micro-classes utilisées pour l'annotation de emoVox

Combinaison d'étiquettes pour une annotation plus précise : Il va de soi qu'au fur et à mesure de l'annotation, un savoir pratique sur la façon la plus adéquate de décrire un segment se stabilise. Il se trouve que c'est davantage dans la combinaison des diverses étiquettes et dimensions qu'on peut être en mesure d'annoter un segment complexe (i.e. qui n'est pas une "colère franche" par exemple). Ce point nous amène à évoquer le cas des segments jugés d'emblée comme "neutres". En effet, les possibilités de combinaison de l'outil permettent de traiter des marqueurs affectifs de type

“mh hm”, pas seulement et unilatéralement comme des continuateurs neutres mais comme des marqueurs d’hésitation voir de désapprobation. Ainsi, il apparaît que la multitude d’occurrences de ces items rend compte de la grande diversité d’expressions affectives qui y sont logées : d’un désintérêt à peine masqué à l’exhibition d’un acquiescement, ou de l’empathie tempérante. On pourra ainsi annoter un segment Neutre en Majeur (émotion la plus représentative pour un segment donné), avec une activation à 0, et enrichir ensuite avec le label Mineur avec une émotion de type irritation par exemple (émotion de second plan), ainsi qu’avec les degrés d’intérêt, de rétroaction et d’induction.

Dans la mesure où le corpus est constitué de conversations téléphoniques, il est apparu qu’une description ne se limitant pas à annoter les segments “individuellement”, mais proposant une caractérisation “inter-segments”, permettait de rendre compte de la dimension séquentielle et pragmatique d’une expression affective (Vaudable et al. 2010).

4.2.2.3 Implication affective : Dimensions abstraites et fondements théoriques

Les travaux de (Sacks 1992) en Analyse Conversationnelle ont démontrés :

(1) que les participants s’orientaient vers les dimensions prospectives et rétrospectives des conduites verbales pour organiser une succession cohérente de cours d’actions, intelligibles et justifiables, “faisant conversation” ;

(2) que l’expression affective, bien que pouvant être décrite en soi, si elle émerge en interaction, constitue une ressource pour s’y engager, i.e., déployer un niveau d’engagement.

L’expression affective peut ainsi être décrite en tant que pratique située dans une position séquentielle spécifique pendant une interaction (Goodwin and Goodwin 1992; Goodwin and Goodwin 2000). De plus, la manière dont une expression affective est construite est particulièrement liée à l’activité dans laquelle elle émerge, ainsi qu’à une analyse par le participant de l’événement déclencheur. Dans le cas qui nous intéresse un refus de rétablissement de l’électricité suite à une coupure peut par exemple faire office de déclencheur.

Ainsi par exemple, le “cri en réponse” (*Response cry)* décrit par (Goffman 1981), n’est pas une expression langagière suite à un événement, mais une expression affective attachée à un événement, d’une organisation visible en tant qu’élément d’un processus plus large : *Triggering Event and Response Cry are fitted to each other as subcomponents of a larger activity system ; each implies the other (Goodwin and Goodwin 2000)*. Un rire, par exemple, peut constituer une conduite s’inscrivant dans un environnement séquentiel, où celle-ci exhibe un alignement sur un cours d’actions (un premier rire par exemple) : “*laughter may not always be a matter of flooding out, to be accounted for as something that happens to a speaker such that he “can’t help laugh”, but can be managed as an interactional resource, as a systematic activity that warrants and rewards more than a naming of its occurrence, but close attention to just how and where it occurs”* (Jefferson, 1985).

Le principe de l’annotation de l’implication affective (Vaudable et al. 2010) repose d’une part sur la dimension prospective et rétrospective d’une conduite verbale en conversation, considérant le fait que l’expression affective contribue à structurer (axes Induction et Rétroaction) ; et d’autre part sur

le niveau d'intérêt exprimé : intérêt simple (quasi machinal), sympathie (plus ou moins équivalent à la politesse), surenchère.

Pour résumer, la notion d'implication affective se décompose en deux dimensions, d'une part l'induction qui va représenter le fait de provoquer une réaction émotionnelle chez son interlocuteur et d'autre part, la rétroaction, qui va mesurer la réaction de l'interlocuteur face à l'évènement qui a eu lieu dans un des tours de parole précédents. Ces deux axes et leurs niveaux d'intensité respectifs sont représentés sur la Figure 4-2 ci-dessous :

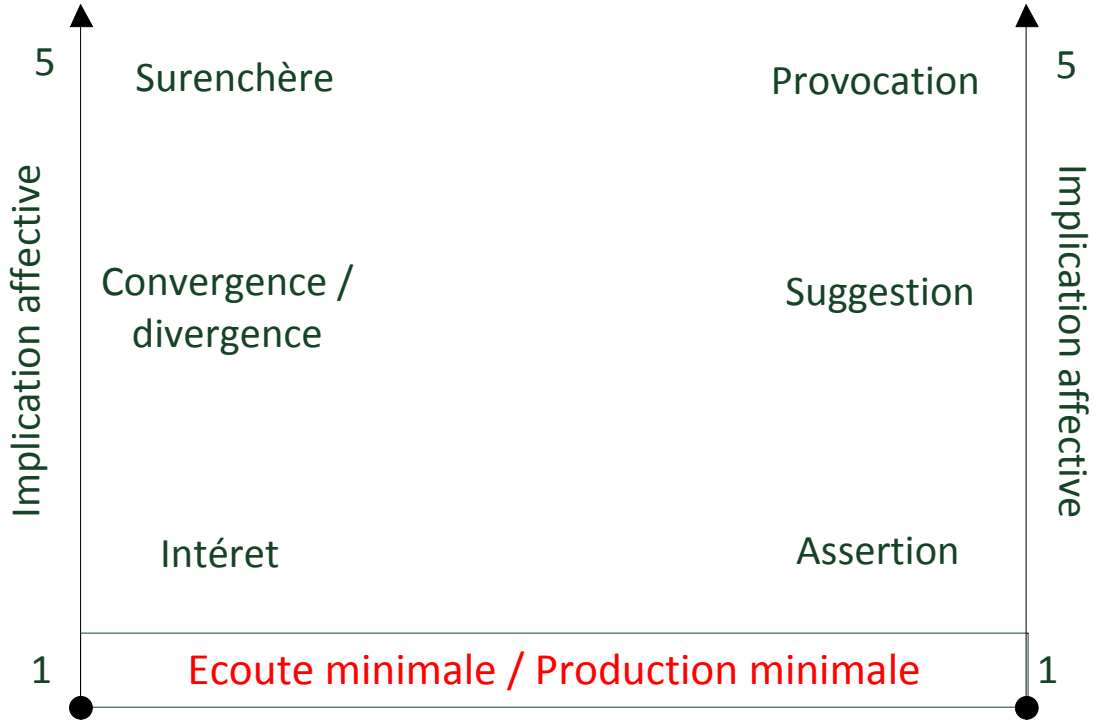


Figure 4-2 Les deux dimensions d'implication affective utilisées pour la mise en relation des segments d'une conversation

4.2.2.4 Implication affective : étiquettes associées

L'implication affective décrit la manière dont le locuteur s'implique dans sa relation de communication avec l'interlocuteur. Durant la phase d'annotation, on la représente par les étiquettes suivantes :

Axe rétroaction :

Ecoute minimale : Neutre. Le locuteur est présent dans la relation de communication, mais son attitude ne semble pas être affectée par le comportement de l'interlocuteur. Il ne tente pas non plus de modifier par sa conduite l'attitude de son interlocuteur.

Intérêt: L'intérêt tel qu'il est employé dans le schéma d'annotation consiste à manifester un degré « moyen » d'implication envers son interlocuteur sans pour autant comprendre l'état affectif ou mental dans lequel se trouve le locuteur.

Convergence/Divergence : Désigne à la fois une aptitude psychologique et les mécanismes qui permettent la compréhension des ressentis d'autrui. Dans l'étude des relations

interindividuelles on admet le fait que la réponse empathique aux états affectifs d'autrui se produit sans que l'on ressente soi-même la même émotion ou même une émotion quelle qu'elle soit.

Surenchère : Un même sentiment est ressenti par les deux individus et entraine une spirale émotionnelle inflationniste au fur et à mesure qu'ils entendent l'autre parler (ex : les deux locuteurs rigolent de plus en plus).

Axe induction :

Assertion : Proposition qu'on affirme et qu'on soutient comme vraie

Suggestion : Un des interlocuteurs tente d'inspirer une idée ou une action à l'autre. Dans tous les cas, cette étiquette désigne une action ayant pour but de faire réagir son interlocuteur.

Provocation : L'un des locuteurs tente de déclencher une réaction auprès de l'autre. Cela peut passer par la séduction ou l'un des intervenants essaie de charmer son interlocuteur. Tout comme l'étiquette précédante, il désigne une action ayant pour but de faire réagir son interlocuteur, mais de manière beaucoup plus explicite que précédemment.

Associations de segments entre eux

Dans la mesure où (1) les méthodes de segmentation ont conduit à un découpage la plupart du temps exogène par rapport à la sémantique des tours de paroles du dialogue, et (2) où les conduites verbales s'inscrivent dans une séquentialité que les participants exploitent pour ordonner leurs actions successives (Sacks et al. 1974), une fenêtre dans l'outil permet de désigner si le segment annoté est lié – en raison d'un de ces deux points – à des segments précédents ou suivants. La référence de ces segments est donnée par l'identifiant du segment courant.

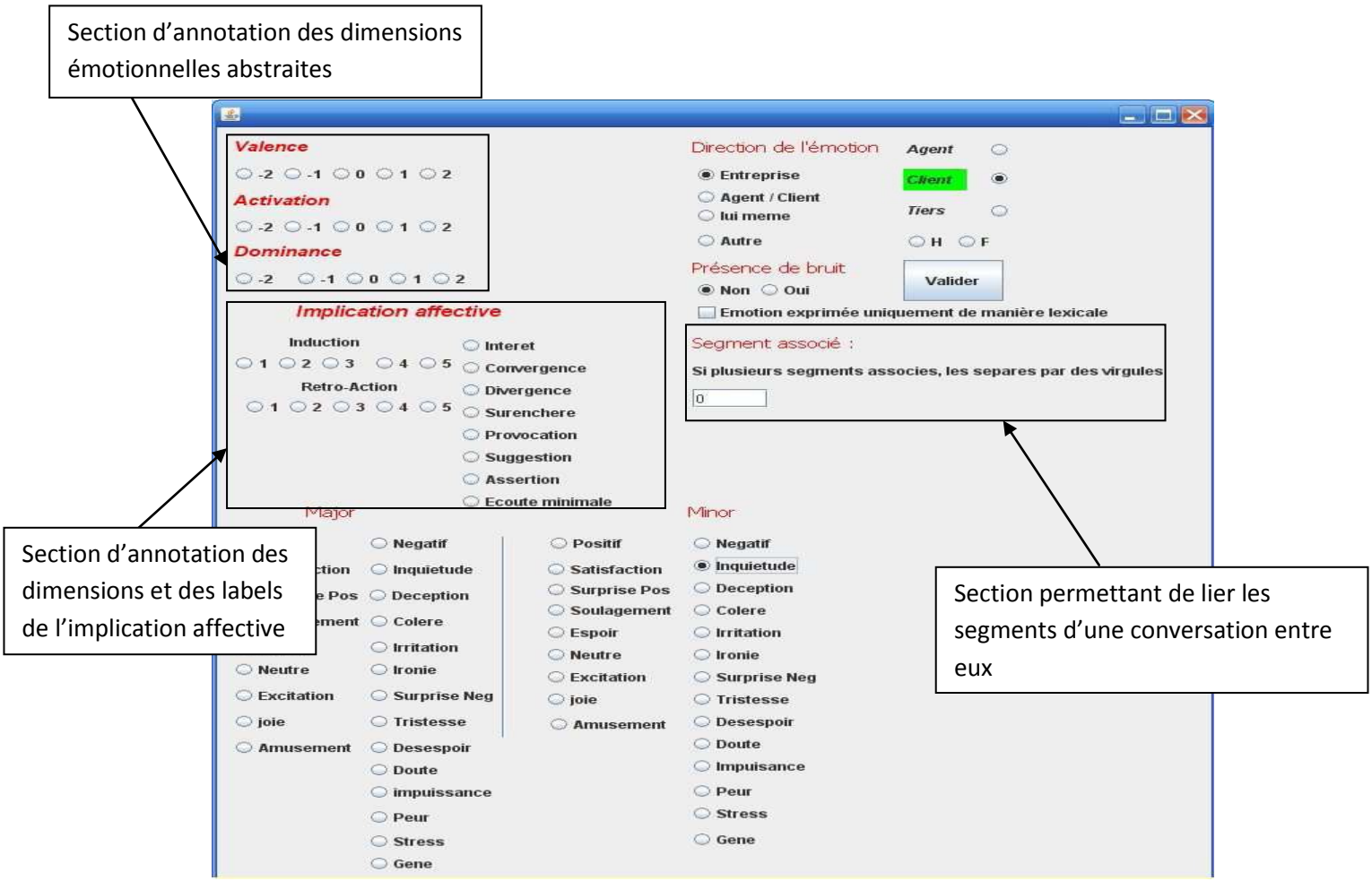


Figure 4-3 Interface d'annotation d'emoTool pour l'annotation des conversations du corpus emoVox

4.3 Validation du schéma d'annotation : choix des éléments descriptifs

Deux personnes (Ann1, Ann2) se sont occupées des tâches de pré-processing et d'annotation. La première personne est une femme de langue maternelle française âgée de 29 ans, linguiste de formation et ayant une expérience de plusieurs années en matière d'annotation. La seconde personne (Ann2) est une femme âgée de 34 ans de langue maternelle espagnole parlant couramment français, également linguiste de formation.

Compte tenu des contraintes en temps et en argent qu'implique une tâche d'annotation, le schéma contenant les dimensions émotionnelles et d'implication affective ainsi que leur label respectif n'a été utilisé dans son ensemble que sur une partie réduite du corpus. Le reste du corpus n'a été annoté qu'à l'aide des dimensions émotionnelles classiques et d'un nombre réduit de labels émotionnels. Les détails sont donnés dans le paragraphe ci dessous. Le Tableau 4-2 ci-dessous décrit les tâches d'annotations effectuées ainsi que leurs champs d'application respectifs dans le corpus emoVox :

Concept	Champ d'application	Nombre de segments pris en compte
Dimensions émotionelles (valence, activation, dominance)	Tout le corpus (107 dialogues)	11798 segments
Dimensions d'implication affective (induction, rétroaction)	Corpus réduit (38 dialogues)	450 segments
Etiquettes émotions	Tout le corpus (macro-classe) Corpus réduit (macro + émotion)	11798 segments 450 segments
Etiquettes d'implications affectives	Corpus réduit	450 segments

Tableau 4-2 Récapitulatif des champs d'applications des dimensions et labels utilisés pour le schéma d'annotation

Pour chacun des niveaux d'annotation que nous avons présentés dans le Tableau 4-2, nous avons calculé l'accord inter-annotateur à l'aide du Kappa de Cohen. Nous présentons les méthodes sélectionnées dans le paragraphe ci-dessous.

4.3.1 Mesure de l'accord entre juges : le Kappa de Cohen (Cohen 1960) :

L'accord observé entre des jugements qualitatifs ou non, résulte de la somme d'une composante «**aléatoire»** et d'une composante d'accord «**véritable**». Le coefficient Kappa ***K*** propose de chiffrer l'intensité ou la qualité de l'accord réel entre des jugements qualitatifs appariés. Il exprime une différence relative entre la proportion d'accord observée P_o et la proportion d'accord aléatoire P_e qui est la valeur espérée sous l'hypothèse nulle d'indépendance des jugements, divisée par la quantité disponible au-delà de l'accord aléatoire. En définitive, *K* est un pourcentage de l'accord maximum corrigé de ce qu'il serait sous le simple effet du hasard.

$$K = \frac{P_o - P_e}{1 - P_e}$$

avec

$$P_o = \sum_{i=1}^{r} p_{ii} = \frac{1}{n} \sum_{i=1}^{r} n_{ii}$$

et

$$P_e = \sum_{i=1}^{r} p_{i.}\, p_{.i} = \frac{1}{n^2} \sum_{i=1}^{r} n_{i.}\, n_{.i}$$

Dans la pratique, on considère généralement que l'homogénéité de l'instrument est satisfaisante lorsque la valeur du coefficient est au moins égale à 0.80. Elle est correcte entre 0 .6 et 0.8 et moyenne entre 0.4 et 0.6. Le Tableau 4-3 ci-dessous proposé dans (Landis and Koch 1977) récapitule les significations pouvant être données aux scores obtenus d'après une mesure d'accord inter juges :

Mesure *k*	Interprétation
< 0	Désaccord
0.0 – 0.20	Accord très faible
0.21 – 0.40	Accord faible
0.41 – 0.60	Accord modéré
0.61 – 0.80	Accord fort
0.81 – 1.00	Accord presque parfait

Tableau 4-3 Signification des taux d'accord obtenus après le calcul du Kappa de Cohen

Il faut cependant noter que ces ordres de grandeur ne font pas consensus au sein de la communauté scientifique notamment du fait que ces interprétations ne tiennent pas compte du nombre de catégories présentes pour les données observées. Un petit nombre de catégories se traduisant habituellement par un Kappa plus élevé.

Nous avons vu que la valeur maximale de Kappa est égale à 1 lorsque $P_o = 1$ et $P_e = 0{,}5$. Ceci n'est vrai que dans le cas où les marginales sont égales ($p_{i.} = p_{.i}$), puisqu'il suffit de prendre les effectifs diagonaux (ceux qui expriment l'accord dans le tableau de contingence) égaux aux marginales et les effectifs non diagonaux égaux à 0. Pour des marginales données, Cohen propose de déterminer la valeur maximale de Kappa (K_m)

$$K_m = \frac{P_m - P_e}{1 - P_e}$$

Avec P_m ci-dessous qui représente la proportion d'accord maximale :

$$P_m = \sum_{i=1}^{r} \inf\left(p_{i.}, p_{.i}\right)$$

Cette mesure nous permet de déterminer la proportion d'accord obtenue d'après la valeur maximale K_m et ainsi avoir une vision plus juste de la valeur du taux d'accord. Cette mesure est couramment utilisée et notamment dans des études comme (Craggs 2004; Abrilian et al. 2006; Vidrascu and Devillers 2006; Clavel 2007).

4.3.1.1Les mesures de Kappa pondérées

Des variantes de cette mesure existent avec pour objectif de minimiser certaines différences d'appréciations entre les juges. Pour pouvoir utiliser ces méthodes il faut que les catégories considérées entretiennent des relations entre elles (exemple dans notre cas, la catégorie « positive » est plus proche de la catégorie « neutre » que de la catégorie « négative », si l'on replace ces 3 étiquettes sur un axe de valence). Compte tenu du fait que nous travaillons avec des données naturelles, qui sont par définition plus difficiles à caractériser à cause d'un possible mélange

d'émotions, il peut être intéressant de ne pas prendre en compte que les concordances parfaites entre étiquettes mais également de possibles correspondances entre des étiquettes proches.

Ainsi pour une pondération linéaire (sur les lignes seulement de la matrice de confusion) cela revient à calculer un poids L défini comme suit :

$$L = 1 - \frac{|\text{Distance}|}{\text{Distance maximum possible}}$$

Si nous considérons *k* classes dans notre problème la distance maximum possible entre deux classes est *k – 1.*

Le tableau ci-dessous donne un exemple des distances pour un problème à 3 classes pour deux juges :

		A		
		Pos	**Neut**	**Neg**
B	**Pos**	**1**	0.5	0
	Neut	0.5	**1**	0.5
	Neg	0	0.5	**1**

Tableau 4-4 Exemples de coefficients linéaires calculés pour un problème à 3 classes

Pour obtenir le résultat définitif il faut multiplier chaque proportion d'accord observée représenté par (**Po**) par le coefficient correspondant dans le tableau ci-dessus. La suite de l'opération est effectuée comme pour le calcul de Kappa classique.

4.3.2 Choix de la méthode d'annotation : test sur un sous-ensemble

Etant donné la complexité de notre schéma d'annotation : au total 29 étiquettes (21 étiquettes émotionnelles (16 + 5 macro-classes (neutre étant une classe et une macro-classe) et 8 étiquette d'implication affective) et 5 dimensions (3 dimension émotionnelles et 2 dimensions d'implication affective), ainsi que le volume de données disponibles dans le corpus (13h de parole), nous avons sélectionné un sous-corpus composé de 38 extraits (450 segments soit environ 30 minutes de parole) qui seront annotés en utilisant l'ensemble des descripteurs fournis par le schéma. Un extrait comprend en moyenne 12 segments. Cette première étape nous a permis ainsi de faire des choix entre les éléments et dimensions les plus précis et les plus descriptifs d'un segment émotionnel.

4.3.2.1 Annotation à base de dimensions continues vs annotation à base de dimensions discrètes

Nous avons vu dans le paragraphe 4.2.2.3 que les conversations étaient structurées de façon à ce qu'un contenu verbal soit rendu intelligible et cohérent pour chacun des participants. Dans ce contexte chaque participant s'appuie sur les tours de paroles précédents pour construire un nouveau tour ayant du sens par rapport aux précédents. Ce cadre contextuel composé de l'ensemble des tours de paroles passés facilite la compréhension d'une discussion. Cependant, il faut noter que lors

d'une tâche de détection des émotions à l'aide d'un modèle automatique, les segments sont examinés de manière individuelle sans prise en compte du contexte lors de la tâche de classification. Nous avons voulu pour cette première phase d'annotation comparer les différences de perception que nous pouvions constater entre une annotation continue (i.e sur une portion de dialogue comprenant plusieurs tours de parole) et une annotation discrète (où chaque segment a été obtenu suite à un découpage manuel et correspond à la définition numéro 2 fournie dans le paragraphe 4.2)

a) L'annotation continue : protocole suivi

L'annotation a été effectuée sur des extraits de dialogues (d'après la définition numéro 1 présentée dans le paragraphe 4.2). Ces extraits ont des longueurs comprises entre 1 et 3 minutes. Il a été demandé aux annotateurs de ne considérer que le client pour la tâche d'annotation. L'annotation portait sur les trois dimensions émotionnelles « plaisir », « activation » et « contrôle » et a été réalisée en deux temps.

Dans un premier temps, il était demandé d'annoter le dialogue suivant les dimensions « plaisir » et « activation » et dans un seconde temps, « contrôle » et « activation ». La dimension « activation » servait donc de contrôle pour mettre en lumière d'éventuelles erreurs dues à une mauvaise manipulation de l'outil d'annotation ou une erreur flagrante d'annotation. La Figure 4-4 Interface d'annotation continue destinée à des extraits de dialogues :

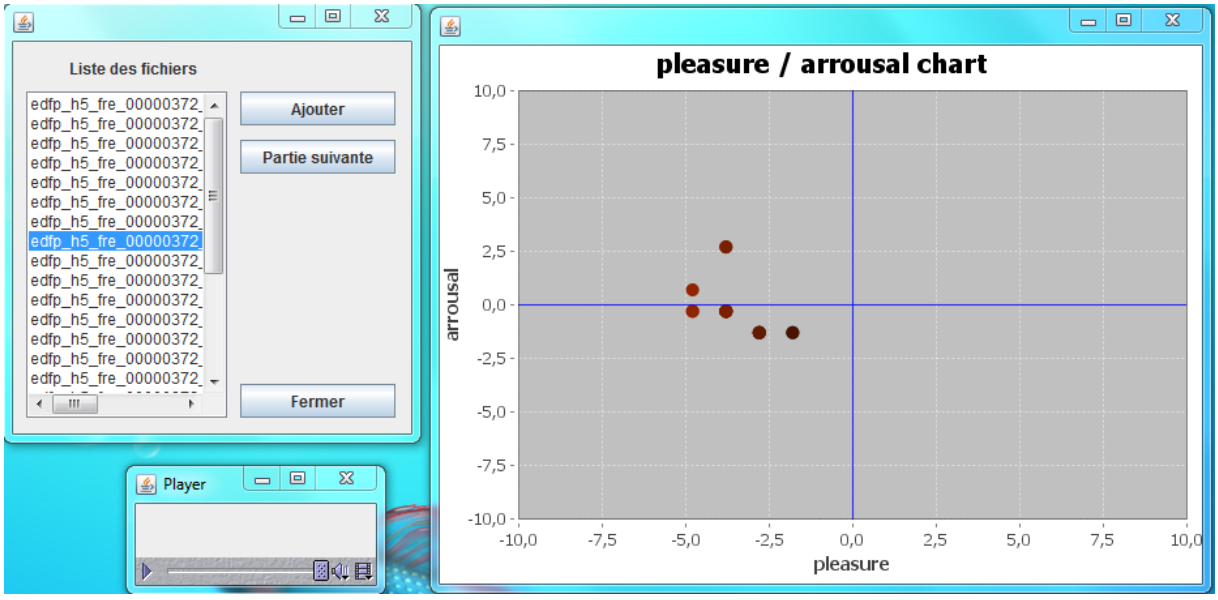


Figure 4-4 Interface d'annotation continue destinée à des extraits de dialogues

Lors de l'écoute du dialogue, l'annotateur devait déplacer sa souris dans le repère représenté sur la Figure 4-4. L'échelle allait de -10 (représentant un très faible niveau de plaisir/activation/contrôle) à 10 (représentant un haut niveau de plaisir/activation/contrôle). La région comprise entre -3 et 3 était considérée comme représentant un état neutre. Pour chaque extrait, le protocole prévoyait une première écoute sans annotation, suivi d'une seconde écoute ou l'annotateur effectuait la tâche d'annotation. Si l'annotateur considérait avoir fait une erreur il était possible de ré-annoter le même extrait.

b) L'annotation discrète

Concernant la phase d'annotation discrète, les mêmes dimensions et dialogues étaient proposées à l'annotateur. Cette fois ci, l'unité à annoter était le segment manuel. La durée d'un segment variait de 1 à 10 secondes. Les segments étaient ordonnés par ordre d'apparition dans la conversation. Chaque dimension devait être annotée suivant une échelle discrète allant de -2 à 2. Les annotateurs avaient la possibilité d'écouter plusieurs fois le segment avant de valider. La Figure 4-5 ci-dessous représente l'interface telle qu'elle était proposée aux annotateurs dans le cas de l'annotation discrète.

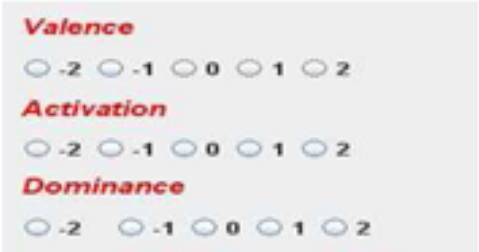


Figure 4-5 Interface pour l'annotation discrète des émotions

4.3.2.2 Méthodologie et résultats de la comparaison des annotations discrètes/continues

Afin de pouvoir comparer les deux méthodes d'annotation, nous avons discrétisé à posteriori l'unité la plus large (qui est l'extrait utilisé lors de la phase d'annotation continue). Les frontières sont basées sur les *timecodes* des segments utilisés pour l'annotation discrète. Cela nous permet donc de comparer exactement le même passage de dialogue et ainsi pouvoir constater les différences de perception entre les deux méthodes d'annotation.

Les Figure 1-5 et 1-6 ci-dessous montrent les étapes conduisant à la comparaison de chaque méthode.

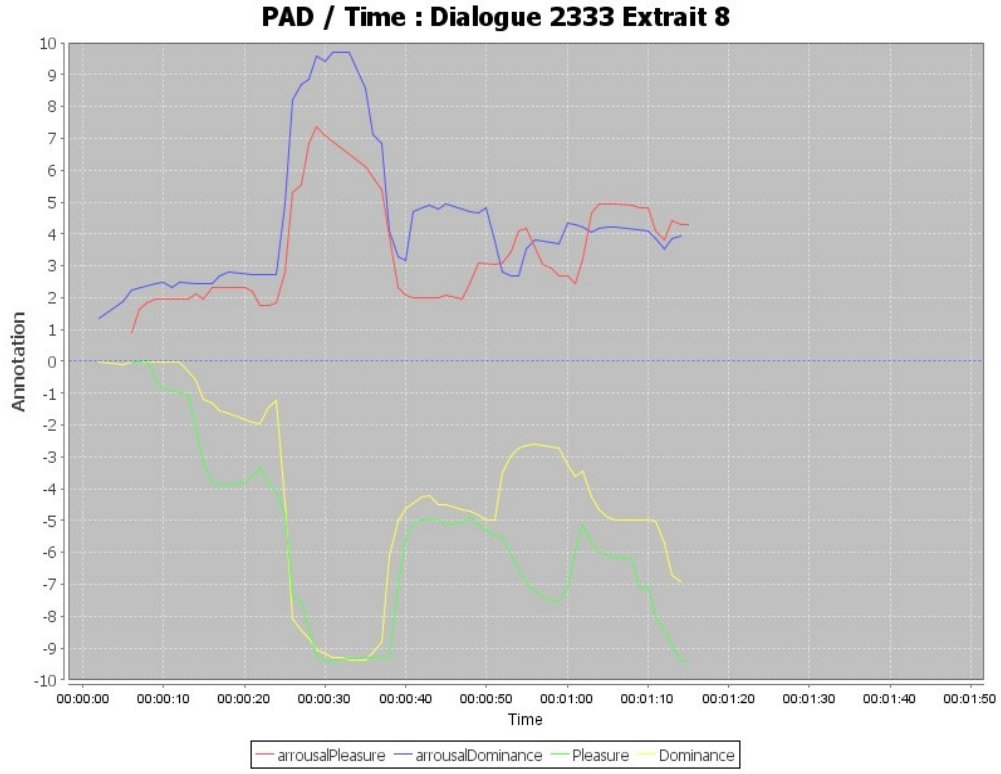


Figure 4-6 Représentation graphique des annotations continues (obtenu sur un extrait complet)

Afin de discrétiser les annotations continues nous avons découpé l'espace en cinq parties égales dans le but de représenter chaque valeur de l'annotation discrète. Ainsi les valeurs de -10 à -6 représentent la valeur discrète -2 ; celles de -5,9 à -2 la valeur -1, celle de -1.9 à 2 la valeur 0.

L'attribution d'un segment continu à une classe discrète à été calculée d'après la moyenne de 3 valeurs prélevées au hasard parmi les valeurs continues fournies. Exemple :

Valeurs de l'annotation continue choisie aléatoirement	Moyenne	Valeurs discrètes correspondantes
-1.26564	-0.8307	0
-0.48498		
-0.74156		

Tableau 4-5 Exemple de discrétisation d'une annotation continue

Le résultat peut être représenté graphiquement par la Figure 4-7 :

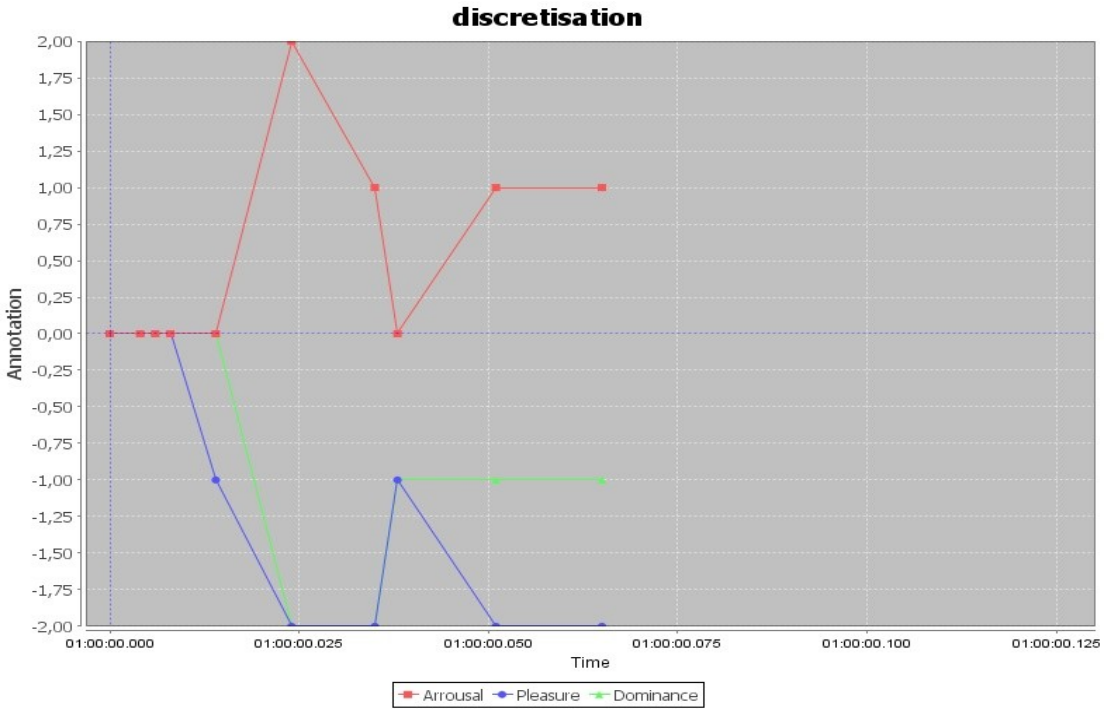


Figure 4-7 Discrétisation du flux continu obtenu d'après l'annotation continue en segment (d'après les times code des segments utilisés pour l'annotation discrète). Chaque point représente le début d'un segment

Afin de vérifier la cohérence des annotations continues nous avons dans un premier temps effectué une mesure de contrôle à l'aide de la dimension activation qui était présente lors des deux phases de l'annotation continue. Afin de valider ce mode d'annotation, un taux d'accord élevé devait être obtenu. Dans le cas contraire cela aurait mis en lumière une incohérence de la part de l'annotateur avec lui-même. Le score intra-annotateurs est exprimé selon le Kappa de Cohen à corriger en fonction de l'accord maximum pouvant être obtenu sur ces données et tel que défini dans le paragraphe 4.3.1. La deuxième mesure présentée est le pourcentage d'accord entre les deux annotations. Les scores obtenus sont présentés dans le tableau ci-dessous :

Annotateur 1 Kappa de Cohen (Pourcentage)	Annotateur 2 Kappa de Cohen (Pourcentage)
0.80 (85%)	0.92 (95%)

Tableau 4-6 Score d'accord intra annotateur pour la dimension activation obtenue sur les 2 phases d'annotations continues

Les taux d'accord obtenus pour la dimension de contrôle sont qualifiés de « presque parfait » suivant la typologie utilisée dans le paragraphe 4.3.1. Cela nous permet donc de valider la cohérence de l'annotation continue de chacun des annotateurs.

Par la suite, nous avons calculé le taux d'accord intra-annotateurs pour chaque dimension. Le but était de vérifier si les annotateurs indiquaient les mêmes annotations suivant les méthodes employées (discrètes ou continues).

Le Tableau 4-7 ci-dessous présente les résultats intra-annotateurs obtenus :

Dimensions	Accord intra-annotateur pour les annotations discrètes et continues Kappa (Pourcentage)	
	Ann1	Ann2
Valence	0.49 (58%)	0.61 (56%)
Activation	0.15 (53%)	0.29 (61%)
Contrôle	0.27 (41%)	0.23 (68%)

Tableau 4-7 Résultats des annotations discrètes/continues obtenus sur un sous ensemble du corpus emoVox

Nous pouvons observer que l'accord intra-annotateurs est relativement faible du point de vue du Kappa (entre 0.15 et 0.61). Le taux d'accord exprimé selon le pourcentage de segments annotés de façon identique est également relativement bas (de 41% à 68% dans le meilleur des cas). A titre de comparaison les annotateurs obtenaient un taux d'accord intra-annotateur de 85% entre 3 phases d'annotation (Vidrascu 2007). Il faut cependant considérer que la méthode d'annotation était la même d'une séance sur l'autre. Ces faibles scores d'accord intra-annotateurs peuvent être expliqués par plusieurs hypothèses :

- La première serait l'impact du contexte sur la perception de l'émotion. Un segment de dialogue considéré isolément ne porterait pas avec lui toutes les informations pragmatiques qui sont mises en œuvre lors d'un échange en contexte. Les choix exprimés dans la méthode discrète ne se baseraient donc pas sur les mêmes informations.
- La seconde serait la difficulté d'utilisation de l'outil d'annotation : Concernant l'annotation continue, des erreurs peuvent facilement être faites (faux mouvements, relâchement de l'attention durant une courte période, etc.). Cette méthode a été qualifiée de difficile d'utilisation par les annotateurs.

Considérant que pour un même segment audio et un même annotateur, des désaccords apparaissent lorsque des méthodes d'annotation différentes sont utilisées, nous avons voulu vérifier si les taux d'accord étaient plus élevés entre des annotateurs utilisant la même méthode. Le Kappa de Cohen a donc été calculé pour les annotateurs 1 et 2 successivement sur les dimensions discrètes et continues. Les résultats sont présentés dans le Tableau 4-8 ci-dessous :

	Valence Discret/continu	Activation Discret/Continu	Contrôle Discret/Continu
Taux d'accord entre Ann1 et Ann2	0.71/0.69	0.59/0.58	0.64/0.56

Tableau 4-8 Scores inter-annotateurs obtenus d'après une annotation continue et discrète

Nous pouvons voir que les scores obtenus sont très proches pour les dimensions valence et activation. La différence entre le discret et le continu pour la dimension contrôle est elle un peu plus marquée avec 8% d'écart. Le seul taux d'accord ne semble donc pas être un élément déterminant pour le choix de la méthode d'annotation.

Considérant que l'annotation continue a été déclarée comme étant difficile à prendre en main du fait de l'attention plus soutenue qu'elle demandait (arrêt du mouvement à chaque changement de locuteur, déplacement du curseur selon 2 dimensions en même temps) nous avons donc décidé de retenir l'annotation discrète comme méthode d'annotation des dimensions pour la suite de l'étude.

4.3.2.3 Annotation en interaction : l'implication affective

Nous avons présenté dans les paragraphes 4.2.2.3 et 4.2.2.4 deux nouvelles dimensions prenant en compte l'interaction lors d'une conversation que nous avons souhaité mettre en œuvre lors de la phase d'annotation. Nous avons dans un premier temps commencé par calculer le taux d'accord concernant les deux dimensions Induction et Rétroaction qui composent l'implication affective. L'échelle de valeur allait de 1 à 5 pour l'annotation de ces dimensions. Pour le calcul des scores d'accord nous avons considéré deux états possibles pour chaque dimension : « Forte » pour les valeurs supérieures à 3 et « faible » pour les valeurs inférieures ou égales à 3.

Les scores apparaissent dans le Tableau 4-9 ci-dessous :

	Score
Induction	0.77
Rétroaction	0.73

Tableau 4-9 Scores d'accord pour les dimensions d'implications affectives

Les bon taux d'accords obtenus témoignent d'une compréhension correcte des dimensions par les annotateurs. Pour rappel ces deux dimensions ont pour but la mise en relation des différents segments les uns avec les autres. Nous avons donc dans un premier temps essayé de mesurer l'impact qu'avait un degré élevé d'induction et de rétroaction sur la production d'émotion. En d'autres mots l'idée était de savoir comment la production émotionnelle d'un locuteur interagissait sur la production émotionnelle du second. Cette analyse semble pertinente dans notre étude au sens où elle permet de modéliser comment le comportement d'un locuteur (l'agent par exemple) va permettre de modifier l'état affectif observé chez le deuxième intervenant.

Dans un but de contrôle de la cohérence de nos dimensions interactionnelle avec celles des dimensions et étiquettes affectives utilisées nous avons dans un premier temps souhaité observer la relation entre les dimensions d'induction et de rétroaction avec les émotions neutres et colère. Le but était de confirmer que les expressions émotionnelles et les dimensions interactionnelles telles que nous les avons définies ont des impacts mesurables les unes sur les autres. Nous pouvons voir les chiffres obtenus dans le Tableau 4-10 ci-dessous :

Macro-Classe	Valence / Activation / Contrôle (échelle de -2 à 2)	Induction / Réaction (échelle de 1 à 5)
Neutre	0.0 / 0.0 / 0.1	1.8 / 1.9
Colère	-1.4 / 1.4 / -1.2	3.4 / 3.5

Tableau 4-10 Scores obtenus sur les axes d'induction et de réaction en fonction de l'état émotionnel exprimé

Nous pouvons voir que lorsque nous considérons un état affectif « neutre » (absence d'émotion exprimé) les valeurs obtenues sur les axes d'induction et de réaction sont très bas, ce qui semble cohérent avec ce que nous avons cherché à modéliser (à savoir l'interaction affective entre le client et l'agent). A l'inverse une expression émotionnelle exprimée plus nettement se voit attribuer des scores d'induction et de réaction plus élevés. En d'autres termes nous voyons que lorsque l'un des interlocuteurs exprime une émotion dans un but de communication (expression d'une émotion de type colère chez le client par exemple) cela provoque une réaction chez l'interlocuteur. A l'inverse lorsque qu'aucune émotion n'est exprimée les niveaux d'induction et de réaction restent faibles.

Les différents types de dimensions interactionnelles et affectives semblent donc pouvoir fonctionner de manière cohérente entre elles.

Afin d'avoir un niveau de détail plus fin nous avons observé l'impact sur les 3 dimensions (valence, activation et contrôle) d'un faible et d'un fort niveau d'induction et de rétroaction. Pour rappel chaque émotion est représentée par les trois dimensions plaisir, activation, contrôle à l'aide d'une échelle discrète allant de -2 à 2.

Scores obtenus sur les dimensions émotionnelles VAD lorsque nous observons les segments ayant obtenu un haut score d'induction	Valence	Activation	Contrôle
Emotions Positives	1.56	1.44	1.00
Emotions Négatives	-1.60	0.00	-1.30

Tableau 4-11 Corrélation entre un haut score d'induction et les valeurs des dimensions plaisir, activation, contrôle

Nous pouvons voir qu'un haut degré d'induction (< 3/5) est fortement corrélé avec des valeurs dimensionnelles extrêmes. Cette constatation est plus marquée pour la dimension représentant la valence. Pour rappel, nous corrélons un haut degré d'induction avec un acte destiné à provoquer une réaction de la part de l'interlocuteur. Le constat est également le même lorsque nous corrélons des valeurs élevées concernant l'axe de rétroaction et les valeurs des dimensions émotionnelles. Les résultats sont présentés sur le Tableau 4-12 ci-dessous :

Scores obtenus sur les dimensions émotionnelles VAD lorsque nous observons les segments ayant obtenu un haut score de retroaction	Valence	Activation	Contrôle
Emotions Positives	1.30	1.40	0.00
Emotions Négatives	-1.65	1.00	-1.50

Tableau 4-12 Corrélation entre un haut score de rétroaction et les valeurs des dimensions plaisir, activation, contrôle

Dans 90% des cas, les segments ayant un fort degré de rétroaction sont produits par le client. La constatation est la même pour les segments mettant en jeu un haut degré d'induction (85% des cas). Ces résultats paraissent logiques dans la mesure où l'agent est tenu à une réserve dans les réactions que celui-ci peut avoir vis-à-vis des attitudes du client.

4.3.2.4 Exemples de situations interactionnelles prototypiques

Nous avons identifié trois situations prototypiques pouvant représenter des situations dialogiques précises (évolution de la conversation vers des situations de conflits, tentatives de prise de contrôle du cours de la conversation par l'agent)

Situation 1 : Alignement émotionnel d'un locuteur sur le second

Cette situation apparait principalement lorsque l'un des deux locuteurs se retrouve soudainement à exprimer un état affectif très marqué (joie ou colère), le second a alors tendance à s'aligner sur l'état du premier. Nous observons un exemple dans le Tableau 4-13 ci-dessous :

Numéro du tour de parole	locuteur	Valeurs sur les axes VAD	Induction	Retro Action	Annotations Majeur/Mineur
N	Agent	0/0/0	2	2	Neutre / Neutre
N + 1	Client	-2/2/-2	4	3	Colère / Colère
N + 2	Agent	-1/1/-1	1	3	Négatif / Négatif

Tableau 4-13 Exemple d'alignement émotionnel du client sur l'agent

Nous pouvons voir progressivement dans cet exemple les valeurs obtenues sur les axes valence/activation/dominance se rapprocher avec pour résultat un alignement de l'agent sur l'état du client. Nous donnons un exemple de conversation illustrant cette situation :

N (Agent) : On a un délai maxi, je dis bien maxi de 5 jours
N + 1 (Client) : 5 Jours ?!! Non mais... On est vraiment dans un pays de fous là, on est dans un pays de fous
N + 2 (Agent) : Ecoutez Monsieur, là la discussion est complètement stérile

Situation 2 : Tentative de maintien de contrôle sur la conversation

La deuxième situation prototypique retenue est celle où un client entre dans une situation de grande colère. La réaction de l'agent pour tenter de revenir à une situation de dialogue normale est de conserver son calme et répondre au client de manière posée

Numéro du tour de parole	locuteur	Valeurs sur les axes VAD	Induction	Retro Action	Annotations Majeur/Mineur
N	Agent	0/0/0	2	2	Neutre / Neutre
N + 1	Client	-2/2/-2	4	2	Colère / Colère
N + 2	Agent	0/0/0	1	2	Neutre / Neutre

Tableau 4-14 Tentative de maintien de contrôle de la part de l'agent en ne répondant pas à d'éventuelles provocations

Dans cette deuxième situation l'état affectif de l'agent n'évolue pas. Ce dernier répond et ne s'occupe que des éléments factuels en ignorant les parties agressives ou plaintives des propos du client. Ce dernier s'attache donc à garder le contrôle du dialogue en faisant progresser la résolution du problème.

N (Agent) : Et votre lettre, quand est elle partie ?
N + 1 (Client) : Je sais plus ! Il faut que je ressorte mon dossier… Alors la semaine dernière… Je vous dis ça commence à me me fatiguer là ! J'ai passé deux heures avec quelqu'un de chez vous la semaine dernière !
N + 2 (Agent) : hum hum… Alors moi dans le dossier je vois des frais de pénalité

Situation 3 : Tentative de maintien de contrôle sur la conversation

La 3ème situation met quand à elle en jeu un client énervé et un agent gardant son calme et répondant avec courtoisie dans le but de calmer ce dernier.

Numéro du tour de parole	locuteur	Valeurs sur les axes VAD	Induction	Retro Action	Annotations Majeur/Mineur
N	Agent	0/0/1	2	2	Neutre / Neutre
N + 1	Client	-2/2/-2	3	2	Colère / Colère
N + 2	Agent	1/0/1	2	3	Neutre / Positif

Tableau 4-15 Tentative de maintien de contrôle de l'agent en positivant la situation

N (Agent) : Oui, il y a un créneau de 2 heures
N + 1 (Client) : Et moi je peux pas être là, bien sûr, évidement ! Vous m'emmerdez avec ça !
N + 2 (Agent) : moi je suis là pour vous aider monsieur

4.4 Présentation du corpus retenu pour l'entrainement de modèles computationnels

4.4.1 Résultat de l'annotation

La tâche d'annotation sur un volume important de données nécessite un investissement en temps et en argent conséquent. Afin d'avoir à disposition un nombre d'instances d'entrainement suffisant en un temps raisonnable, nous avons réduit le nombre d'étiquettes et de dimensions à annoter. Côté étiquettes nous conservons les macro-classes « Neutre », « Négatif », « Positif ». Pour les dimensions la valence et l'activation sont retenues. La tâche d'annotation est réalisée de manière indépendante sur les données obtenues d'après la segmentation automatique et manuelle (voir définition paragraphe 4.2). Les taux d'accord (Kappa) sur chacune de ces deux configurations sont présentés dans le Tableau 4-16 ci-dessous :

	Segmentation manuelle	Segmentation Automatique
Positif / Neutre	0.47	0.32
Négatif / Neutre	0.77	0.42
Positif / Négatif / Neutre	0.58	0.40

Tableau 4-16 Taux d'accord en fonction du type de segmentation

Les taux d'accords obtenus sont très variables selon i) le type de segmentation considéré et ii) les états émotionnels que nous comparons. Nous remarquons dans un premier temps que le taux d'accord des juges sur la segmentation automatique est dans tous les cas plus bas que celui que nous obtenons sur la segmentation manuelle. Cette différence est la plus marquée lorsque nous cherchons à annoter les émotions négatives. Cette différence peut être expliquée par plusieurs facteurs. D'une part, les segments considérés dans la segmentation automatique sont nettement plus longs que ceux de la segmentation manuelle, ce qui peut amener les annotateurs à se focaliser sur des parties du discours différentes au moment de rendre leur jugement. Un deuxième point à prendre en compte lorsque nous examinons des segments issus du découpage automatique du dialogue est la présence de nombreux segments mettant en jeu plusieurs locuteurs, ce qui complique également la tâche d'annotation. D'une manière générale, dans ce corpus, les émotions négatives semblent être plus discriminantes que les émotions positives, les scores d'accord obtenus sur ces premières sont nettement plus élevés (0.47 vs 0.77) pour la segmentation manuelle et (0.32 vs 0.42) pour la segmentation automatique.

Nous présentons dans les figures ci-dessous la répartition des segments pour chacun des deux types de segmentation. Sur chacune de ces deux figures trois types de segments sont considérés : Les segments « non consensuels » qui ont obtenu une annotation avec une valence différente (positif pour un annotateur et neutre pour l'autre par exemple) ; les segments consensuels, ayant obtenu une annotation avec une valence identique pour chacun des deux annotateurs. Les segments « bruités », c'est-à-dire contenant de la parole superposée, un bruit d'arrière plan ou tout évènement sonore pouvant perturber la tâche d'annotation. La Figure 4-8 présente la répartition des segments pour un découpage automatique, la figure 4.9 présente la répartition pour un découpage manuel :

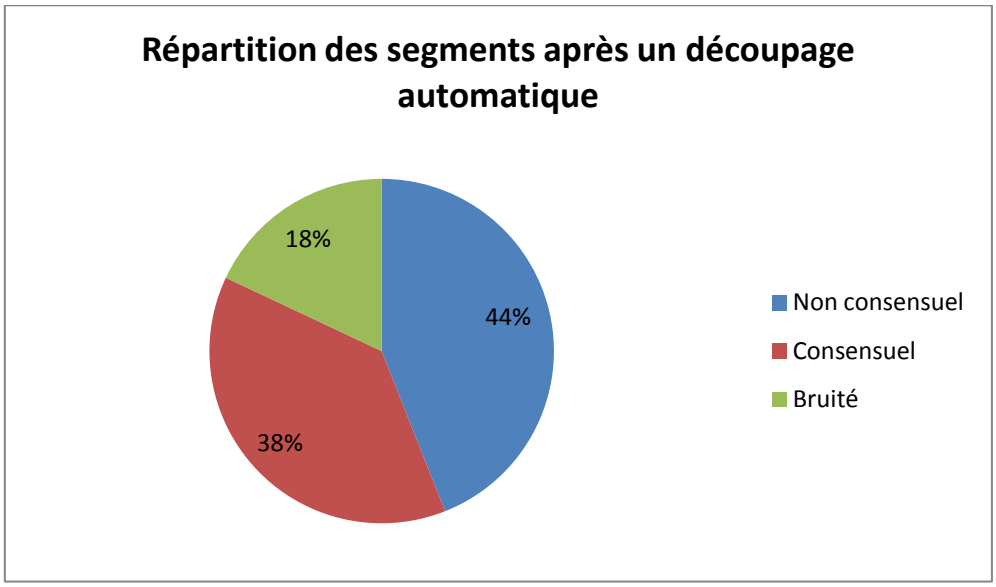


Figure 4-8 Répartition des segments après annotation de segments issus d'un découpage automatique des conversations

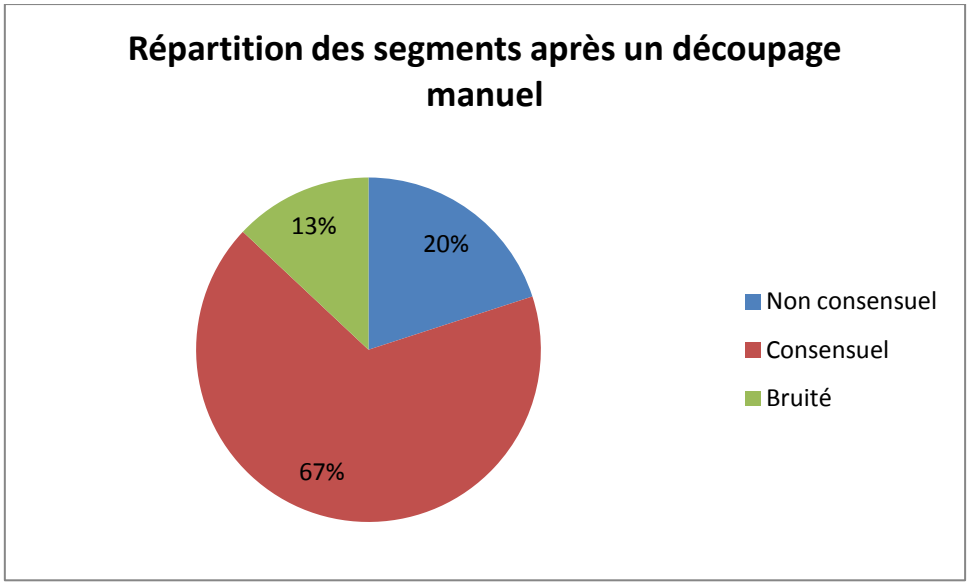


Figure 4-9 Répartition des segments après annotation de segments issus d'un découpage manuel des conversations

La répartition des données que nous pouvons observer sur les Figure 4-8 et Figure 4-9 sont assez différentes. Nous pouvons ainsi noter une sur-représentation des données non consensuelles lorsque l'on observe la segmentation automatique. Cette tendance est cohérente avec les taux d'accord présentés dans le Tableau 4-16. Les données bruitées sont également plus présentes. Cette différence est due à un nombre important de segments contenant plus de deux locuteurs du fait du découpage automatique.

Toujours dans le but de comparer les deux modes de segmentation nous avons mesuré la différence de durée entre des segments obtenus automatiquement et manuellement. Nous pouvons voir sur la Figure 4-10 ci-dessous la durée moyenne d'un segment selon sa catégorie émotionnelle et son mode de segmentation :

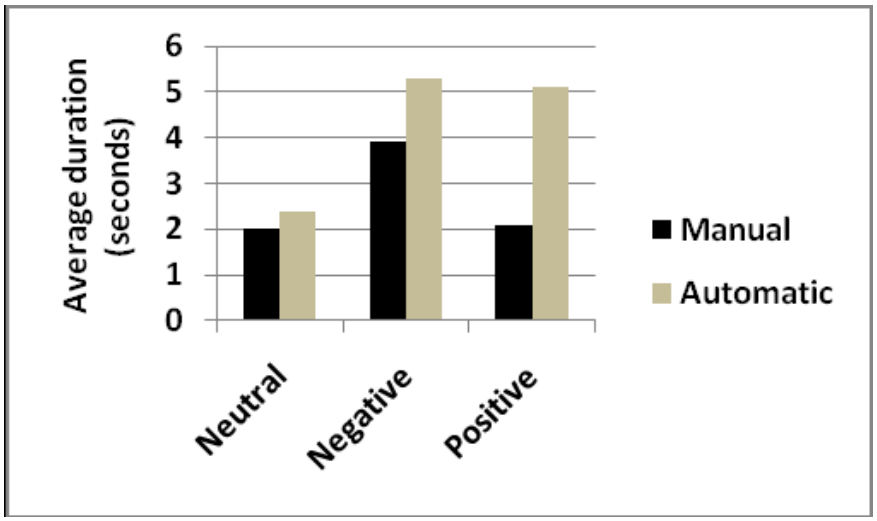


Figure 4-10 Durée moyenne d'un segment selon la catégorie émotionnelle et le type de segmentation

Nous pouvons voir que les durées entre segments manuels et automatiques peuvent avoir des longueurs relativement variables. Ceci est particulièrement vrai dans le cas des segments « émotionnels ». Nous observons ainsi, pour les segments négatifs, une durée moyenne de 3.9 secondes dans le cas d'une segmentation manuelle contre 5.3 secondes pour une segmentation automatique. La différence est encore plus marquée pour les segments positifs. Ces derniers durent en moyenne 2.1 secondes en manuel contre 5.1 seconde en automatique. Les segments neutres ont quant à eux des longueurs proches (2 secondes en moyenne pour les segments manuels, 2.3 secondes pour les segments automatiques).

4.4.2 Contenu du corpus emoVox

Le Tableau 4-17 ci-dessous récapitule le contenu de chacun des corpus. Les chiffres mentionnés ne prennent en compte que les segments ayant été annoté de manière consensuelle, donc potentiellement éligibles à une utilisation dans l'ensemble d'entrainement ou de test des modèles automatiques.

	Positif	Négatif	Neutre
Segmentation automatique	263	1173	2246
Segmentation manuelle	1083	2373	5778

Tableau 4-17 Nombres de segments retenus pour la construction des modèles automatiques

D'autres informations complémentaires concernant la nature des locuteurs intervenant dans le corpus peuvent être collectées (fonction de l'interlocuteur i.e. Agent/client, sexe).

Nous pouvons ainsi voir sur la Figure 4-11 la répartition des émotions exprimées en fonction du locuteur (agent ou client).

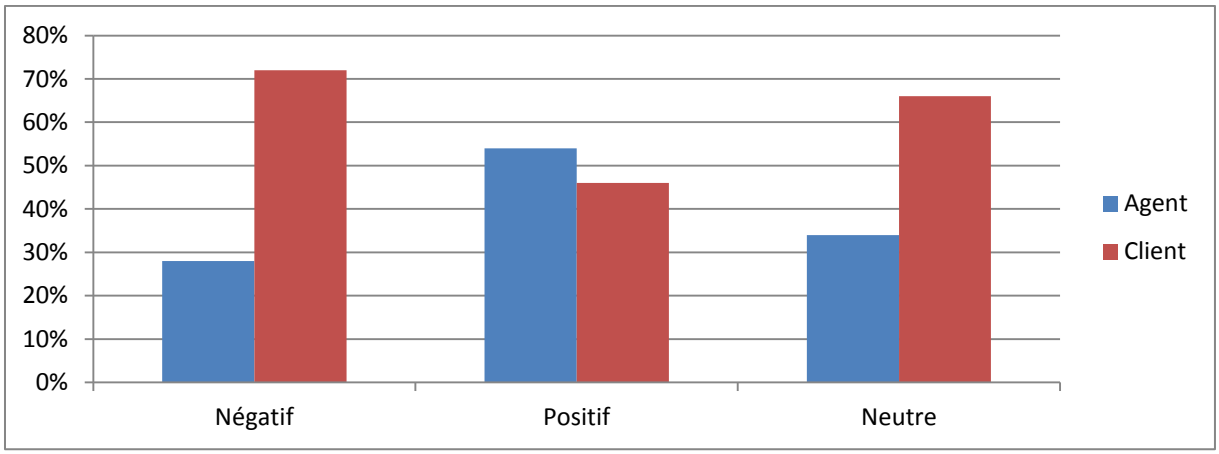


Figure 4-11 Répartition des émotions exprimées en fonction du rôle du locuteur

Nous voyons que la répartition de l'expression des émotions est différente suivant le locuteur. Comme nous pouvions l'envisager, étant donné le contexte commercial, les expressions négatives sont en très large majorité exprimées par le client (72% des cas). Les expressions positives sont en revanche plus souvent exprimées par l'agent, ce dernier ayant tendance à minimiser une situation négative ou à positiver un certains nombre d'évènements au cours d'une conversation. L'état neutre est quant à lui le plus souvent exprimé par le client mais cette répartition est probablement plus due au fait que nous avons effectué une sous-sélection de segments qu'à une réalité dialogique.

Durant la phase de segmentation manuelle des informations relatives à l'âge du locuteur ont été annotées. Cette information est justifiée par le fait que des différences au niveau acoustique existent entre une voix d'enfant et une voix de personnes plus âgées ou atteintes d'une pathologie particulière (Tahon et al. 2010). Il est de ce fait intéressant que différents types de voix soient représentés dans le corpus afin d'obtenir des modèles ayant un pouvoir de généralisation plus important. Les Figure 4-12 et Figure 4-13 ci-dessous présentent la répartition du type de voix perçue par les annotateurs après écoute des dialogues :

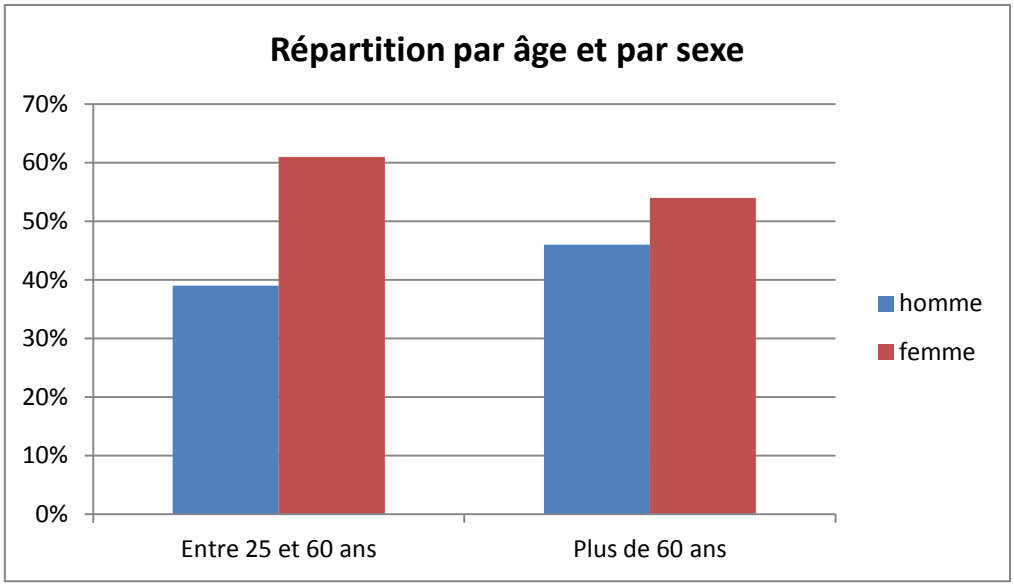


Figure 4-12 Répartition par âge et par sexe pour les locuteurs du corpus

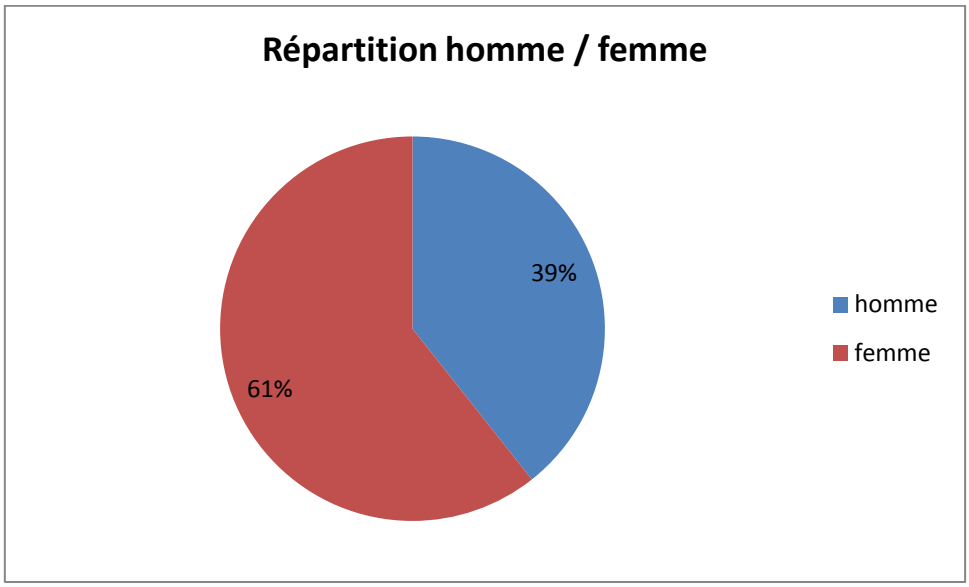


Figure 4-13 Répartition des tours de paroles de paroles consensuels par sexe

4.5Conclusion

Ce chapitre a été l'occasion d'expérimenter deux modes d'annotation différente, discrète et continue sur un sous-ensemble du corpus. Si en termes de taux d'accord les résultats restent assez proche les deux méthodes ne se valent pas en termes de facilité d'annotation. La méthode continue ayant été considérée comme difficile par les deux annotateurs, la méthode discrète a été retenue pour l'annotation du reste du corpus. Outre le fait d'être plus simple à mettre en œuvre elle permet une annotation hors contexte (les tours de parole sont alors présentés dans le désordre, ce qui permet aux annotateurs de ne pas être influencés par les éléments contextuels de la conversation tels que de potentiels tours émotionnels intervenant précédemment durant la conversation). En plus des dimensions émotionnelles classiques nous introduisons deux nouvelles dimensions d'implication

affective, qui ont pour but de mesurer l'interaction entre l'agent et le client. Nous avons vu que ces dimensions interactionnelles entretenaient des liens étroits avec l'expression affective (un fort taux sur la dimension d'induction/rétroaction par exemple est très fortement corrélé à une expression négative intense). D'après les tests effectués nous avons également relevé trois situations prototypiques mettant en lumière les stratégies conversationnelles adoptées par l'agent dans une situation conflictuelle. Bien que ce type d'annotation soit difficilement automatisable ces situations nous ont montré que le thème de la satisfaction client devait être abordé d'une manière générale, comme un enchainement de tours de paroles produisant du sens. De ce point de vue nous proposerons dans le chapitre 8 de nouveaux indices obtenus sur l'ensemble du dialogue. Nous émettons l'hypothèse que cette approche nous permettra de sélectionner des éléments pertinents pour l'évaluation des émotions, et par extension de la satisfaction client sur un dialogue complet.

Nous avons pour finir retenu un nombre limité d'informations pour l'annotation des 13h de dialogues composant notre corpus final suivant deux modes de segmentation. Le premier, une segmentation des conversations manuelles a donné des taux d'accord nettement plus élevés que le second qui utilise une segmentation automatique. Des différences en terme de taille des tours de parole obtenus ont également été identifiées entre la segmentation manuelle et automatique ce qui peut potentiellement poser problème pour une tâche de détection. Enfin une série de mesures statistiques nous a permis d'observer le contenu de notre corpus (répartition des données suivant le sexe, l'âge) qui semble être composé de données suffisamment variées pour couvrir un large éventail de situations.

Chapitre 5 Indices et méthodes de classification pour la détection d'émotions

Résumé

Ce chapitre a pour but de présenter les différentes techniques utilisées dans les chapitres suivants pour bâtir et évaluer les modèles de détection des émotions. Nous présentons dans un premier temps les indices acoustiques (basés sur le jeu de descripteurs utilisés pour le challenge Interspeech 2009) et lexicaux retenus (adjectifs, adverbes, interjections, noms communs).

Ayant eu recours à une étape de sélection de descripteurs, nous décrivons l'algorithme retenu pour notre étude qui repose sur le calcul d'un score individuel pour chacun des descripteurs.

Nous présentons également, d'un point de vue théorique, les algorithmes de classification que nous utiliserons dans les chapitres prochains. Notre choix se porte sur des approches de type SVM, largement connus pour leurs bonnes performances lors de tâches de classification mettant en jeu un nombre de dimensions élevées et très utilisés pour la classification à base de descripteurs acoustiques et linguistiques. Nous mentionnons également un second algorithme ayant donné des performances intéressantes lors de nos tests et qui utilise une approche mêlant un algorithme bayésien naïf et des tables de décision.

Les mesures d'évaluation utilisées pour nos modèles se basent d'une part sur le F-score (une combinaison des mesures de rappel et de précision) et d'autre sur un intervalle de confiance permettant d'établir la validité statistique des résultats obtenus.

A l'aide des méthodes citées nous avons effectué une première évaluation de différents ensembles de descripteurs acoustiques et lexicaux. Etant donné l'important volume de données à notre disposition nous avons pris soin de limiter le nombre de dimensions des modèles utilisés.

5.1Les indices

Ce chapitre s'attache à présenter les différents éléments techniques utilisés pour bâtir un système de détection des émotions. La Figure 5-1 ci-dessous en présente un résumé :

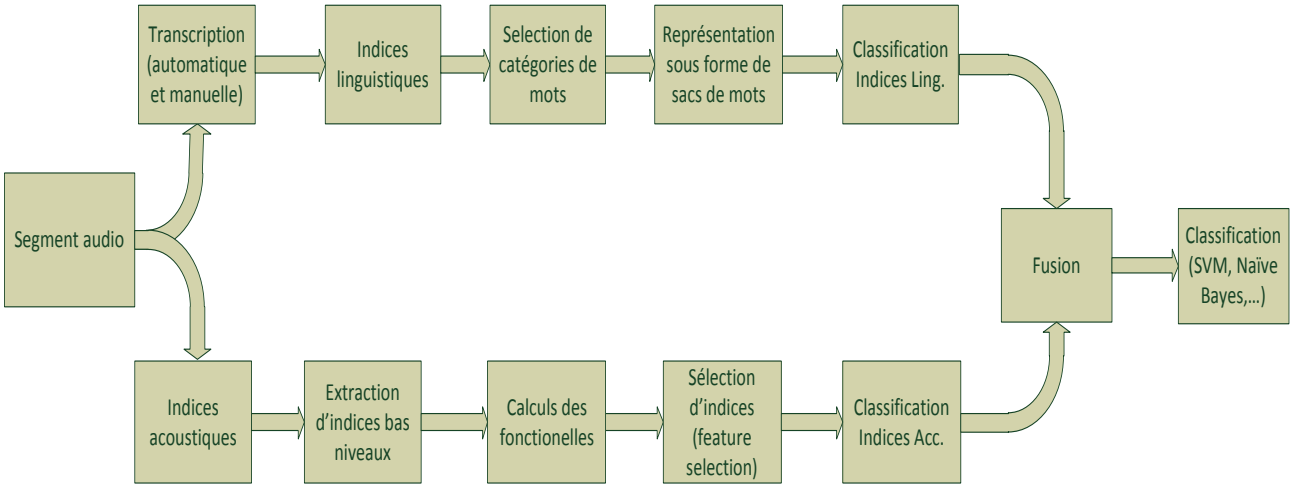


Figure 5-1 Briques procédurales impliquées dans un système de détection des émotions

Les paragraphes suivants décrivent l'ensemble des autres éléments : extraction des paramètres lexicaux et acoustiques, sélection des paramètres et algorithmes d'apprentissage retenus. Les méthodes et algorithmes présentés dans ce chapitre sont ceux qui ont été utilisés lors des tâches qui seront présentées dans le chapitre suivant.

5.1.1 Descripteurs acoustiques

A l'heure actuelle, il n'y a pas de consensus sur le meilleur ensemble de descripteurs à prendre en compte pour un système de détection automatique. La pratique la plus courante reste alors d'en sélectionner un grand nombre dans le but d'avoir une classification plus riche. Cependant, trop augmenter le nombre des descripteurs pour un corpus de taille réduite peut amener à une dégradation des performances et donc être contre productif. Deux types de descripteurs paralinguistiques (acoustiques) sont habituellement distingués : les descripteurs au niveau « frame level » qui sont calculés sur des fenêtres temporelles inférieures à 1 seconde (généralement 30ms) et au niveau « supra-segmental » intervenant sur une fenêtre temporelle plus large (1 seconde ou plus). Ces descripteurs peuvent être également calculés à partir de différents niveaux du signal sonore. Nous présentons ces différents niveaux dans les paragraphes ci-dessous

5.1.2 Descripteurs spectraux

Les descripteurs spectraux sont issus d'analyses à très court terme du signal et sont composés de mesures comme les *Mel-Frequency Cepstrum Coefficients* (MFCCs) qui sont basées sur une échelle perceptive non linéaire nommée « Mel ». Le principe des *MFCCs* repose sur le fait que, perceptivement, l'augmentation de l'intensité d'un signal n'est pas perçue de manière linéaire mais exponentielle. Ce type de répartition non linéaire permet en théorie de reproduire plus fidèlement la perception qu'un humain peut avoir d'un son. D'autres indices temporels perceptifs comme les

Linear Predictive Coding (LPC) (Rabiner and Juang 1993) ou les *Perceptual Linear Prediction* (PLP) (Hermansky and Morgan 1994) sont également couramment utilisés en reconnaissance de la parole notamment.

De manière générale, ces trois types de coefficients sont des caractéristiques des résonances du conduit vocal à un instant donné. Il est d'usage de les calculer sur une fenêtre temporelle donnée (Ex : 30 ms avec un décalage de 10ms pour la seconde fenêtre) et de calculer leurs dérivées premières et secondes dans le but de déterminer leur évolution dans le temps. D'autre type d'indices comme les formants et leur largeur de bande sont traditionnellement pris en compte. Ces derniers correspondent à des pics d'énergie dans des bandes de fréquence. Les trois premiers formants et surtout les différences entre les formants peuvent être des indices de comportement affectif.

5.1.3 Descripteurs prosodiques

La prosodie concerne le suprasegmental et englobe des phénomènes tels que l'accentuation, les variations de hauteur, de durée et d'intensité. Un trait suprasegmental, ou trait prosodique, est une caractéristique qui affecte un segment plus long que le phonème ; c'est ce qui donne un ton naturel et cohérent à la parole. La prosodie 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. Les trois classes principales de descripteurs prosodiques sont :

- la fréquence fondamentale (F0), qui représente la vitesse de vibration des cordes vocales. Cette dernière ne peut donc être calculée que sur les parties dites « voisées » du signal. La fréquence fondamentale est définie comme étant l'inverse de la période T : $F0 = \frac{1}{T}$. Du point de vue de l'extraction automatique, la méthode la plus couramment utilisée est celle implémentée dans le logiciel *Praat* (Boersma and Weenink 2009). Cette méthode est basée sur une fonction d'auto corrélation où le but est de rechercher des ressemblances entre des versions décalées du signal. La fréquence fondamentale est dans ce cas définie comme l'inverse de la période pour laquelle la fonction d'auto corrélation a une valeur maximale. Bien que la fréquence fondamentale ait prouvée son efficacité pour la reconnaissance des émotions, il est à noter que ce paramètre a l'inconvénient majeur d'être extrêmement variable d'un locuteur à un autre.

- L'énergie, qui représente la force avec laquelle est émis le signal sonore et qui se mesure en décibels (perception auditive). A l'inverse de la fréquence fondamentale, ce paramètre n'est pas lié à la physiologie du locuteur. Il varie également beaucoup et présente l'inconvénient d'être difficilement normalisable. Sa mesure peut en effet varier très fortement lors d'une conversation téléphonique par exemple, en fonction de la distance à laquelle l'interlocuteur se trouve du microphone. c'est aussi un paramètre très important : des émotions de type « colère » présentent habituellement une intensité forte contrairement aux états émotionnels de type « neutre » ou « tristesse ».

- La durée et le rythme, qui prennent en compte les phénomènes de successions de parole/silence mis en jeu lors d'une discussion à différents niveaux de granularité (syllabes/mots/parties voisées). Le nombre et la durée de ces silences peuvent varier en fonction de l'émotion exprimée. Lors d'une expression de type « gêne » ou « peur », les silences sont généralement plus longs et plus fréquents que lors d'un discours émotionnellement neutre.

Les descripteurs prosodiques permettent de modéliser les accents, le rythme, l'intonation, la mélodie de la phrase et sont très pertinents pour la modélisation de l'état émotionnel du locuteur. Ces derniers, tels que la fréquence fondamentale ou l'intensité, sont fréquemment utilisés dans le domaine de la parole émotionnelle. Ces paramètres sont généralement extraits des segments grâce à des librairies comme *Praat* (Boersma and Weenink 2009) ou *openEAR* (Eyben et al. 2009).

L'extraction des descripteurs prosodiques consiste d'abord en une analyse bas niveau (analyse à court terme sur le signal) pour extraire des indices tous les *x* ms (10, 25, 30, 50). Ensuite, une analyse de descripteurs « supra-segmentaux » est menée. Les descripteurs supra-segmentaux calculés sont des pentes, maxima, « range », etc. Il est à noter que des paramètres tels que la fréquence fondamentale sont très dépendants du locuteur et peuvent varier de manière importante (Clavel et al. 2007) . La fréquence fondamentale par exemple est plus élevée chez la femme (moyenne de 250 Hz) et l'enfant (moyenne de 350Hz) que chez l'homme (moyenne de 150 Hz). La solution envisageable peut être la mise en œuvre de modèles prenant en compte la différence de genres comme c'est le cas dans (Vidrascu and Devillers 2005a). D'autres paramètres, comme l'âge ou la présence d'accents régionaux, sont autant d'influences susceptibles de faire varier les scores de détection des systèmes. L'étude (Tahon et al. 2010) montre par exemple que la reconnaissance d'un locuteur est plus aisée avec des sujets âgés qu'avec des sujets jeunes, les premiers présentant une variabilité beaucoup moins importante des paramètres vocaux calculés.

Certains ajoutent dans la prosodie une quatrième dimension, la qualité de la voix (timbre, voix rauque, chuchotée, grinçante, voix de fausset...), due à des caractéristiques laryngales ou supra laryngales. L'extraction de paramètres de qualité vocale est un sujet intéressant mais qui nécessite une très bonne qualité d'enregistrement du signal ce qui peut souvent poser problème pour des données collectées au téléphone ou hors laboratoire. La qualité vocale correspond à un ensemble de paramètres perceptifs. C'est donc une notion subjective. Des indices tels que le timbre de l'émission vocale, l'articulation, la hauteur ou le rythme peuvent être relatifs à la qualité vocale. Des algorithmes prenant en compte la qualité vocale à travers la modification physique de la parole (voix soufflée, vibrante, etc.) ont été conçus et mis en œuvre dans des études comme celle de (Campbell and Mokhtari 2003) qui étudie notamment le rôle de la parole non voisée.

5.1.4 Descripteurs de micro-prosodie

Les descripteurs les plus connus pour qualifier la micro-prosodie sont le jitter, shimmer et HNR. Le coefficient Shimmer mesure les variations d'amplitude entre deux cycles d'oscillation. Le coefficient Jitter 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).

Le pourcentage de fenêtres non voisées dans le segment peut également être un indicateur pris en compte. Le signal sonore peut, dans ce cas, être divisé en plusieurs fenêtres (trames) voisées ou non voisées (pas de signal F0). Le pourcentage des fenêtres non voisées dans la phrase est alors révélateur de l'influence des 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 avec la colère ou la peur.

De manière générale les calculs effectués pour obtenir ces indices peuvent être relativement dépendant de la qualité du signal ou de l'interaction (parole superposée, bruit d'arrière plan) il convient donc souvent de procéder par essais (ou d'appliquer un algorithme de sélection de descripteur) afin de déterminer l'ensemble d'indices les plus appropriés à la tâche.

5.1.5 Descripteurs acoustiques retenus

Le choix des descripteurs retenus pour la conception des modèles de reconnaissance se base, d'une part, sur l'expérience tirée des challenges passés comme (Schuller et al. 2009; Schuller et al. 2010; Schuller et al. 2011) et d'autre part, sur le fait qu'un nombre limité de descripteurs devait être retenu si le système était mis en œuvre afin de réduire la complexité des calculs à réaliser. Au cours de cette thèse nous avons testé différents jeux de paramètres avec différents outils et méthodes d'extractions qui seront décrites ci-dessous :

5.1.5.1 openEAR/openSMILE (Eyben et al. 2009)

Lors d'une étude précédente ayant pour but de tester la robustesse de nos modèles de détection des émotions à travers l'utilisation de plusieurs corpus real life, nous avons montré que les scores obtenus à l'aide de deux jeux de descripteurs (respectivement 988 et 384) avaient des performances comparables (Devillers et al. 2010). Nous avons donc, pour la suite de l'étude, utilisé seulement le jeu de descripteurs de 384 paramètres afin de réduire la dimension du problème à étudier et, par conséquent, alléger le temps de calcul nécessaire aussi bien du point de vue de l'extraction des descripteurs que du temps d'entraînement des modèles. Les 384 descripteurs retenus sont ceux utilisés comme ensemble de référence lors du challenge INTERSPEECH 2009. Lors de la phase d'évaluation aucun participant n'avait été en mesure de proposer une amélioration significative des résultats en proposant d'autres jeux de descripteurs (Schuller et al. 2011).

Le Tableau 5-1 ci-dessous issu de (Schuller et al. 2009) récapitule les 16 descripteurs bas niveau et leurs 12 fonctionnelles associées :

Descripteurs bas niveau (16 * 2)	Fonctionnelles (12)
(Δ) ZCR **(Δ)RMS Energy** **(Δ)F0** **(Δ)HNR** **(Δ)MFCC 1-12**	mean Standard deviation Kurtosis, skewness Extremes : value, rel. Position, range Linear regression : offset, slope, MSE

Tableau 5-1 Récapitulatif des descripteurs bas niveau (LLD), de leur dérivée et des fonctionnelles associées (Schuller et al. 2009)

Nous disposons donc au final des 16 descripteurs bas niveau (Low-Level Descriptors) suivants : le « zero crossing rate » correspondant au tau de passages par 0 du signal pour une fenêtre temporelle donnée ; l'énergie moyenne quadratique par frame (RMS energy) permettant de prendre en compte la variation de l'énergie du signal ; la fréquence du pitch ; le coefficient HNR (harmonic to noise ratio), qui se base sur une fonction d'auto corrélation (recherche de ressemblances sur une unité temporelle donnée), permettant de mesurer le bruit dans le signal (nombre de parties bruitées remplaçant une harmonique) .

Ce jeu de descripteurs bas niveau est complété par les 12 premiers coefficients MFCC (mel-frequency cepstral coefficients). Pour chacun des 16 coefficients bas niveau, 12 fonctionnelles sont calculées (moyenne, déviation standard, coefficient d'aplatissement de Pearson (kurtosis), coefficient de dissymétrie (skewness), valeurs maximale et minimale d'un descripteur donné, position relative, plage de valeurs et régression linéaire (pente, erreurs quadratiques moyennes (MSE)).

5.1.5.2 Praat (Boersma and Weenink 2009) + bibliothèques LIMSI

La seconde méthode de calcul de descripteurs que nous avons utilisée est basée sur le logiciel Praat (Boersma and Weenink 2009) et complétée par une librairie développée au LIMSI (Brendel et al. 2010). Le détail des descripteurs retenus est présenté dans le Tableau 5-2 ci-dessous :

Descripteurs bas niveau (31 * 2)	Fonctionnelles (12)
(Δ)F0 **(Δ)BarkEnergy 1-12** **(Δ)MFCC 1-12** **(Δ)Loudness** **(Δ)Intensity** **(Δ)Formant 1-3**	Min, max, mean, range Standard deviation Median, Quartile, TQuartile Kurtosis, Skewness Linear Regression : slope, Rsquared

Tableau 5-2 Récapitulatif des descripteurs bas niveau (LLD) et des fonctionnelles associées

Nous retrouvons dans cet ensemble des descripteurs bas niveaux et des fonctionnelles communs avec l'ensemble d'openEAR vu dans le paragraphe précédent. Dans cet ensemble de test nous n'avons augmenté le nombre d'indices bas niveau en rajoutant notamment les descripteurs basés sur l'échelle perceptive Bark (Bark scale) ainsi que le calcul des 3 premiers formants et l'intensité. Comme nous le verrons plus tard ces descripteurs peuvent être calculés sur des bases différentes (sur tous les points voisées du segment considéré, sur tous les points voisés ou non, les maxima, les moyennes, etc). Une comparaison sera effectuée avec le jeu de descripteur présenté dans le paragraphe précédent que nous considérerons être le jeu de descripteur standard aux vues des résultats du challenge 2009 sur la détection d'émotions (Schuller et al. 2009)

5.1.6 Extraction et description des descripteurs lexicaux utilisés

Nous avons choisi de représenter les descripteurs lexicaux sous forme de sacs de mots (bag of word). Dans cette méthode, chaque mot d'un segment donné va être inclus dans un vecteur. Chaque segment a été étiqueté morpho-syntaxiquement à l'aide de l'étiqueteur proposé dans (Brill 1995) et entrainé à l'INALF[8] sur la base textuelle FRANTEXT. Considérant que dans ce mode de représentation chaque mot est représenté comme une dimension supplémentaire du vecteur, il est conseillé de ne conserver que les mots dits « pleins », tels que les adjectifs, les adverbes, les verbes et les noms qui ont été considérés comme pertinent pour décrire un état émotionnel (Turney 2002). Ces constatations ont toutefois été faites à partir de données purement textuelles et non de transcriptions de dialogues. Nous avons donc souhaité enrichir les catégories grammaticales retenues avec des marqueurs verbaux de type *affect bursts*, spécifiques au langage parlé, et ayant été définis comme pouvant être discriminants lors de la reconnaissance des états affectifs par un humain (Schröder 2001). Toujours dans le but de réduire le nombre de dimensions du problème de classification, une procédure de lemmatisation a été appliquée afin de ne conserver que la forme canonique de chaque mot (infinitif pour les verbes, mise au singulier des mots au pluriel, etc). Une « *stop list* » de 130 mots contenant principalement des entités nommées (nom de compagnie, nom propre) a été utilisée. Des occurrences d'informations telles que les chiffres ou les dates ont été remplacées sous forme d'étiquettes génériques de type *[chiffre], [date]*. D'autres balises textuelles (demande de prise de parole, identification de l'identité du client, souffle, rire, bruit de bouche) fournissent des informations conversationnelles sur la conversation.

Lors de l'utilisation de formes de représentation comme le sac de mots, il est courant de normaliser les composantes de chaque vecteur. Plusieurs méthodes peuvent être employées comme par exemple la représentation des occurrences de mots au format binaire. Dans cette méthode, chaque mot du dictionnaire est remplacé par un 0 ou un 1 marquant respectivement son absence ou sa présence dans le texte. D'autres approches existent comme la fréquence d'apparition du terme dans le document (term-frequency) ou la fréquence inverse d'apparition (inverse document frequency). Dans cette seconde approche un poids plus important est donné aux termes apparaissant dans le moins de documents. Ceux-ci étant dès lors considérés comme plus discriminants. La mesure *tf-idf* (*Term Frequency-Inverse Document Frequency)* prend quant à elle en compte la fréquence d'apparition d'une occurrence de mots ainsi que sa fréquence inverse en multipliant les deux valeurs entre elles.

5.1.6.1Analyse lexicométrique des conversations du corpus

Le vocabulaire présent au sein du corpus est au total composé de 106563 occurrences de mots et 5783 mots uniques. Nous retrouvons 2905 hapax au sein de ce vocabulaire (i.e. mot n'apparaissant qu'une seule fois). Le mot avec la fréquence d'apparition la plus élevée est le « vous ». Si nous observons les mots les plus fréquents en fonction du type d'émotion (positive, négative ou neutre) nous pouvons voir que le mot le plus fréquent pour les segments de type négatif est le mot « je » contrairement aux segments annotés comme positifs où le mot le plus courant est le « vous ». Une tournure de phrase employant la première personne du singulier serait donc dans ce corpus plus

[8] http://www.limsi.fr/TLP/grace/www/sys.id.html

susceptible d'appartenir à une classe négative. Cette constatation est également valable lorsque nous observons les fréquences de mots par locuteur (agent/client). Ainsi le mot le plus fréquent pour le client est également le mot « je ». Le mot « vous » est quant à lui celui qui est le plus exprimé par les agents.

Au delà de la simple fréquence d'apparition d'un mot dans le texte nous avons calculé la spécificité positive pour l'agent et le client de termes apparaissant dans le corpus. Nous nous sommes basé sur l'outil « Lexico 3 » pour réaliser ce calcul (Fleury et al. 2003). La spécificité permet de mesurer le caractère sur-représentatif d'un mot ou d'une expression dans un sous ensemble du corpus d'après sa fréquence d'apparition. Le calcul de spécificité repose sur un modèle hypergéométrique et défini par l'équation suivante :

$$P(X = x) = h(x; n, M, N) = \frac{\binom{\mathrm{M}}{\mathrm{x}} \binom{\mathrm{N} - \mathrm{M}}{\mathrm{n} - \mathrm{x}}}{\binom{\mathrm{N}}{\mathrm{n}}}$$

où :

x = succès_échantillon n = nombre_échantillon

M = succès_population N = nombre_population

Ainsi la forme sera spécifique si sa fréquence est supérieure à celle attendue théoriquement d'après sa distribution dans l'ensemble du corpus. Nous détaillons dans les deux tableaux ci-dessous les 14 termes ayant la spécificité positive la plus élevée par locuteur et par émotion dans le second tableau :

Agent	**Client**
Hum	J'ai
Vous	Heu
[nom_client]	Je
Avez	Me
Votre	Mon
Donc	Avais
Bleu (ciel)	Mes
Contrat	Dit
Alors	Ma
Hop	parce (que)
Allez	Moi
Accord	Et
[bruit de bouche]	Que
Frais (couts)	Puis

Tableau 5-3 Termes ayant le plus haut taux de spécificité positive en fonction du locuteur

Nous pouvons découper en 3 grandes catégories le vocabulaire de l'agent. La première catégorie est relative au travail de l'agent avec les termes « bleu ciel » (un nom de contrat EDF), « contrat », [nom_client] utilisé dans un contexte d'identification de ce dernier. La seconde catégorie, comme prévu, est relative à l'assistance avec des verbes et pronoms déclinés à la seconde personne du pluriel « vous », « avez », « allez » et des mots comme « donc » et « alors » relatifs à l'inférence et la

résolution de problèmes. Enfin la troisième catégorie peut être associée à l'écoute et/ou au maintien de la conversation avec des petits mots de types « hum », « hop » ou des bruits corporels de type [bruit de bouche] destinés soit à commenter l'action en cours soit à signifier au client que l'agent est à l'écoute de son problème ou est actif dans la résolution de celui-ci.

Comme cela était prévisible, le vocabulaire spécifique du client est quant à lui en très grande partie tourné vers sa propre personne. Nous comptons 7 mots sur 14 utilisés par le client qui sont destinés à se désigner lui-même (« j'ai, je, me, mon, mes, ma, moi). Une seconde catégorie plus énonciative peut également être vue dans les termes spécifiques du client avec des mots comme (parce (que), dit, et, que) utilisés principalement lors de l'explication de son problème auprès de l'agent.

Positif	Neutre	Négatif
Rire	hum	Pas
Voilà	[nom_clientAgent]	Mais
D'accord	personne	Je
Remercie	oui	Ne
Merci	bonjour	Que
(au) revoir	ouais	Me
Ok	alors	Moi
Oui	ok	Elle
(Bonne) journée	Bleu (Ciel)	Rien
Infos	[bruit de bouche]	Payer
Gentil	D'accord	Veux
Ah	Avez	paie
Très	(heures) Creuses	peux
Plaisir	Donc	puisque

Tableau 5-4 Termes ayant le plus haut taux de spécificité positive en fonction de l'émotion

Les résultats obtenus dans le Tableau 5-4 ci-dessus concernant le taux de spécificité positif relatif aux émotions négatives sont intéressants. D'une part ils permettent d'établir un lien potentiel entre émotions négatives et client de par la présence importante de mot relatifs à la personne (je, me, moi) que nous retrouvons également, pour certains, dans la liste des termes propres au client (voir Tableau 5-3). Sans surprise les termes de négations de type (« ne », « pas », « rien ») côtoient des termes marquant la volonté (ou plutôt ici l'absence de volonté) caractéristiques des conversations à problème (« ex : je ne paierai pas cette facture », « je vous dis que je ne peux pas »). Les termes relatifs aux émotions positives ont quant à eux la particularité d'appartenir aux phases de clôture du dialogue. Ils sont en majorité représentés par des remerciements et des formules de politesse (remercie, au revoir, bonne journée). Ce résultat est intéressant puisqu'il nous permettrait de situer de manière plus précise la partie du discours où il est possible de trouver les émotions positives les plus marquées. Les termes neutres quant à eux sont composés d'éléments divers permettant de structurer une conversation (« hum », « [bruit de bouche] »]), d'acquiescement ou de terme métier (« heures creuses », « [nom_clientAgent] »).

Les résultats présentés ci-dessus ont été effectués sur l'ensemble du vocabulaire du corpus. Comme nous le mentionnions dans le paragraphe précédent, nous n'avons dans certains cas retenu qu'une partie du vocabulaire (verbes, adjectifs, noms, adverbes, interjections), celui-ci étant jugé comme le plus discriminant pour la détection des émotions. Nous présentons dans le tableau ci-dessous les résultats par émotions obtenus sur ce vocabulaire réduit :

Positif	Neutre	Négatif
[rire]	Hum	pas
Remercie	[chiffre]	recevoir
Bon	Oui	Venir
Beaucoup	Mettre	Payer
Ah	Bleu	Savoir
Gentil	Alors	Jamais
Grand	Nouveau	Ne
Bah	Résilitation	Pourquoi
Vrai	[bruit de bouche]	Même
Très	Creux	[Turn request]
Prier	Donc	Là
Bien	Ancien	Pf
Oui	Combien	Ça
Super	Faillir	comprendre

Tableau 5-5 Termes ayant le plus haut taux de spécificité positive en fonction de l'émotion et d'une sélection du vocabulaire en fonction de la catégorie grammaticale

5.2 Apprentissage artificiel : algorithmes et méthodes

L'apprentissage artificiel s'intéresse à l'écriture de programmes d'ordinateur capables de s'améliorer automatiquement au fil du temps, soit sur la base de leur propre expérience, soit à partir de données antérieures fournies par d'autres programmes (Cornuéjols et al. 2002). Nous pouvons distinguer deux types d'apprentissages distincts :

- L'apprentissage non supervisé qui propose d'organiser un ensemble de données non étiquetées en partitions considérées comme homogènes. Les méthodes de clustering sont classiquement prises en exemple pour illustrer ce qu'est l'apprentissage non supervisé. L'un des inconvénients majeurs de cette méthode est la difficulté d'attribuer une sémantique (i.e une classe) aux groupes de données formées automatiquement.

- L'apprentissage supervisé nécessite quant à lui un étiquetage préalable des données à traiter. Dans cette situation, on cherche à dégager des éléments permettant de faire des discriminations entre les classes de données. Dans le domaine de la détection des émotions ces classes sont généralement des étiquettes émotionnelles (peur, joie, tristesse, etc).

Nous utiliserons essentiellement dans cette étude des algorithmes développés pour l'apprentissage supervisé. La première étape traditionnellement mise en œuvre avant une tâche d'apprentissage est la sélection des paramètres les plus pertinents pour la résolution du problème étudié. Dans ce cadre différents algorithmes d'apprentissage existent. Nous avons dans notre cas choisi l'algorithme de sélection défini dans (Chen and Lin 2005) que nous présentons dans le paragraphe suivant.

5.2.1 Algorithmes d'apprentissages employés au cours de l'étude

Nous présentons dans les paragraphes ci-dessous les algorithmes utilisés pour la construction de nos modèles de détection. Ces algorithmes sont issus des méthodes statistiques (SVM) et probabilistes (DTNB).

5.2.1.1 Les Séparateurs à Vaste Marge (Support Vector Machine)

L'idée des Support Vector Machine (SVM) (Vapnik 1998) est d'arriver à trouver le meilleur hyperplan séparateur permettant de séparer deux ensembles de points. Le meilleur hyperplan sera celui pour lequel la distance minimale pour chacun des deux ensembles de points sera maximale vis-à-vis de la frontière séparant les différents groupes de données. Cette distance porte le nom de « marge ». Les échantillons entre la marge et l'hyperplan le plus proche sont appelés « support vector ».

Concrètement ce principe peut être illustré par la Figure 5-2 suivante :

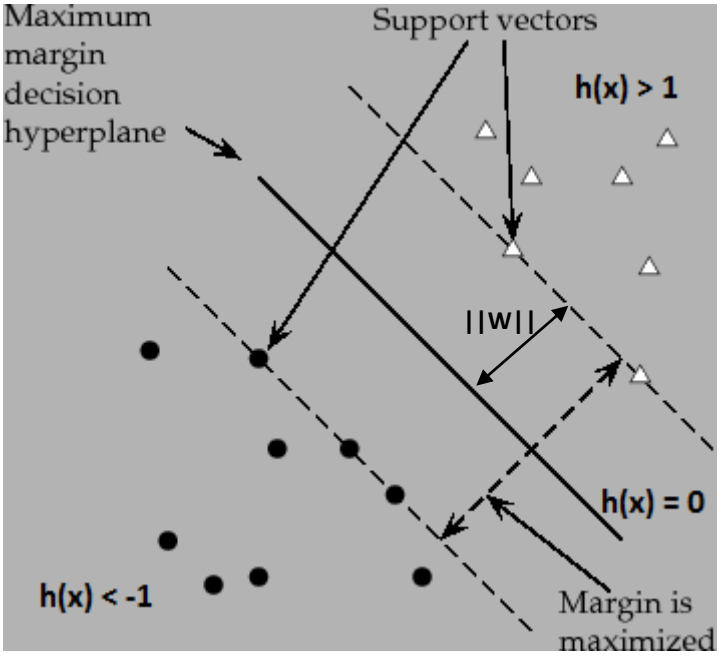


Figure 5-2 Exemple d'un hyperplan optimal(Manning et al. 2008)

Nous nommerons S le sous ensemble d'apprentissage $S = \{(X_1, E_1), (X_2, E_2), \ldots (X_n, E_n)\}$ et supposerons dans un premier temps que le problème de classification que nous posons est linéairement séparable.

Dans ce problème nous recherchons la fonction discriminante $h(x) = 0$ telle que $h(x) = w^T x + w_0 = 0$

Où

$h(x) > 0 \Rightarrow \quad E_i = 1$
$h(x) < 0 \Rightarrow \quad E_i = -1$

où

E représente une classe pouvant prendre les valeurs {-1 ; 1}. Dans l'exemple proposé sur la Figure 5-2, afin de déterminer l'hyperplan optimal il faut minimiser **|| w ||**. En pratique, il n'est pas toujours possible de trouver une fonction séparatrice fournissant la meilleure solution pour une configuration donnée. Dans ce cas, il est possible de relâcher la contrainte des marges par l'introduction d'une variable ressort. Ainsi, afin de tolérer une marge d'erreur dans la résolution du problème il est possible de définir une constante *C* représentant le nombre d'erreurs pouvant être admises lors de la construction du modèle.

Si nous nous plaçons dans le cas d'un problème non linéairement séparable, l'idée est alors de projeter l'instance dans un espace de dimension supérieure, appelé espace de re-description, potentiellement infini dans lequel il est possible de séparer linéairement les ensembles. Les fonctions permettant de passer d'un espace non linéairement séparable à un espace linéairement séparable sont appelées « fonction noyau ».

Les fonctions noyaux les plus couramment utilisées sont indiquées dans le Tableau 5-6 ci-dessous (Hsu et al. 2010) :

Linéaire	$k(x, x') = x.x'$
Polynomiale	$k(x, x') = (x.x')^p$
RBF	$k(x, x') = e^{-\gamma\left\|x-x'\right\|^2}$
Laplacien	$k(x, x') = e^{-\left\|x-x'\right\|1/\partial}$

Tableau 5-6 Fonctions noyaux utilisées pour des problèmes utilisant des classifieurs de type SVM

En pratique il est nécessaire de procéder par étape (essai-erreur) afin de déterminer la fonction noyau la plus adaptée à un problème donné.

L'optimisation des paramètres avec la méthode « grid search »

L'un des inconvénients des classifieurs à base de SVM est la très grande variabilité des performances pouvant être obtenues en fonction des valeurs des paramètres choisies. L'idée de l'algorithme grid search est d'explorer de manière exhaustive l'ensemble des plages de valeurs pouvant être attribuées afin d'obtenir un gain dans les scores des modèles de détection. Afin de pouvoir évaluer les différences de résultats suivant les différents couples de paramètres testés il est nécessaire de se baser sur une métrique (la précision de la classification par exemple).

Si nous prenons l'exemple d'un noyau polynomial (voir Tableau 5-6) il est possible de faire varier deux paramètres : C (erreurs de classification admises durant la construction du modèle) et P (puissance de la fonction polynomiale). Itérativement chaque couple de valeur (C, P) sera testé. Cette méthode peut cependant être coûteuse en temps de calcul. Pour palier à ce problème, (Hsu et al. 2010) recommandent de définir une « région » dans l'espace de recherche pour lesquels les résultats sont les meilleurs et d'itérativement affiner les valeurs retenues pour aboutir à un résultat optimal.

5.2.1.2 Algorithme bayésien naïf

Comme son nom l'indique ce classifieur est issu du théorème de Bayes. L'adjectif naïf a été octroyé à cet algorithme du fait que pour limiter la complexité des calculs nous supposons que les différents

descripteurs utilisés pour la représentation des classes de données sont indépendants entre eux. Cela signifie en pratique que l'absence ou la présence d'un attribut a_2 n'est pas dépendant de la présence ou de l'absence d'un attribut a_1. Si nous cherchons à résumer le fonctionnement de cet algorithme nous pouvons dire que le but est de calculer la probabilité d'un document d'appartenir à une classe de donnée en fonction de la valeur de ses attributs par rapport à la distribution observée de la valeur des attributs de l'ensemble d'entrainement.

La probabilité à estimer est donc : $P(c_j \mid a_1 , a_2 , a_3 , ..., a_n)$

où

- c_j est une catégorie
- a_i est un attribut

À l'aide du théorème de Bayes, on obtient : $\frac{P(a_1,a_2,a_3,\ldots,a_n|c_j)P(c_j)}{P(a_1,a_2,a_3,\ldots,a_n)}$

On peut omettre de calculer le dénominateur, qui reste le même pour chaque catégorie.

Etant donné l'indépendance conditionnelle (au moins supposée) des différents attributs on peut calculer : $P(a_1 , a_2 , a_3 , ..., a_n \mid c_j)$ que l'on peux écrire ainsi : $\prod_{i=1}^{n} P(a_i|c_j)$. En pratique, et malgré sa simplicité, l'algorithme bayésien naïf a donné des résultats étonnamment bons (Ilieva 2004). Il reste encore couramment utilisé du fait de sa facilité de mise en œuvre.

5.2.1.3 Combinaison de tables de décision et de classification bayésienne semi-naïve

Cet algorithme propose de combiner deux méthodes très largement utilisées : les tables de décision et l'algorithme de classification bayésienne semi-naïve (Hall and Frank 2008).

Dans une table de décision, chaque entrée est une série d'attributs à laquelle est associée une probabilité, la plupart du temps représentée par la fréquence de la distribution de la classe sur le corpus. Durant la phase d'apprentissage, une sélection du sous-ensemble de descripteurs les plus performants (i.e. les plus discriminants pour chacune des classes) est effectuée. La classification bayésienne semi-naïve reprend les principes énoncés pour la classification bayésienne naïve mais permet l'établissement de liens de dépendance entre certains descripteurs. De ce fait cet algorithme peut être vu comme un réseau bayésien simple ou chaque attribut devient à tour de rôle le parent d'un ou plusieurs autre attributs afin de déterminer la combinaison la plus adaptée au problème étudié. La combinaison reprend le principe utilisé pour l'entrainement des règles d'apprentissage mais sépare en deux sous-ensembles les attributs. Une partie sera utilisée pour les tables de décision, la seconde pour le classifieur bayésien.

Finalement, les probabilités fournies par chacune des deux méthodes sont combinées de la façon suivante où E^t est le sous ensemble sélectionné par les tables de décision et E^n est le sous ensemble sélectionné par l'algorithme bayésien semi-naïf :

$$\mathrm{P}(\mathrm{e_i}|\mathrm{E}) = \partial \times P_{dt}(e_i|E^t) \times \frac{P_{nb}(e_i|E^n)}{P(e_i)}$$

Avec

$P_{dt}(e_i|E^t)$ représentant les probabilités estimées par l'algorithme utilisant les tables de décision et $P_{nb}(e_i|E^n)$ les probabilités estimées par le modèle bayésien semi-naif. Dans ce cas, ∂ est estimée comme constante de normalisation et $P(e_i)$ comme la probabilité a priori de la classe considérée.

Cette combinaison a donné des résultats intéressants dans le cas de problèmes mettant en jeu un nombre de dimensions peu élevé comme cela peut être le cas lors d'une fusion tardive.

5.3 Quelles méthodes pour tester et évaluer des modèles ? Techniques et mesures employées

La construction de modèles automatiques nécessite idéalement trois ensembles indépendants servant respectivement à l'entrainement (train set), à la calibration (development set) et au test (test set). Dans ce cadre, l'utilisation de ces 3 ensembles de tests conduit à ne pas utiliser une partie non négligeable des données dans l'ensemble d'entrainement afin de prévenir les problèmes de sur-apprentissage (overfitting) (Cornuéjols et al. 2002). Ces trois ensembles doivent contenir des données préalablement étiquetées. Comme nous l'avions mentionné dans un précédent chapitre, la quantité de données disponibles est régulièrement un obstacle lorsque l'on souhaite construire un système de détection des émotions, empêchant les modèles d'être bâtis sur une variété de situations suffisamment importante et limitant de ce fait leur pouvoir de généralisation. Plusieurs solutions peuvent être envisagées pour, à la fois conserver des ensembles d'entrainements et de tests indépendants et utiliser malgré tout la totalité des données disponibles. La méthode la plus couramment utilisée est la méthode de test par validation croisée (cross validation) présentée dans le paragraphe ci-dessous. Nous présenterons également dans cette section les différentes mesures permettant d'évaluer la qualité des modèles construits et testés.

5.3.1.1 Test par validation croisée

L'idée du test par validation croisée est de sélectionner aléatoirement X% (généralement X=20 ou X=10) des données disponibles pour l'ensemble de tests et de construire le modèle à partir des Y% de données restantes. Cette méthode présente le grand avantage de ne jamais pouvoir retrouver simultanément une instance dans l'ensemble de test et dans l'ensemble d'entrainements. L'opération d'entrainement/test est renouvelée tant que toutes les données n'ont pas été utilisées dans l'ensemble d'entrainements et dans l'ensemble de tests. Le score final est obtenu d'après la moyenne des résultats donnés par le test réalisé lors de chacune des itérations. Des outils mettant en œuvre des algorithmes d'apprentissage tel que (Chih-Chung et al. 2001), fournissent des méthodes de test par validation croisée. Cette méthode appliquée à des données acoustiques, comme c'est le cas dans cette étude, présente cependant l'inconvénient de ne pas pouvoir sélectionner précisément les données à conserver dans l'ensemble d'entrainement et celles à placer dans l'ensemble de test. En pratique, des descripteurs comme la fréquence fondamentale de la voix qui est très spécifique à un locuteur donné peuvent potentiellement être appris durant la

phase d'entrainement et testés, sur des instances d'un même locuteur, lors de la phase de test (c'est le cas de la cross-validation utilisé dans la plateforme weka (Hall et al. 2009)). Cela a pour conséquence de fournir des scores de sortie artificiellement plus élevés. Des variantes de cette méthode comme la technique de test « Leave One Out Cross Validation » (LOOCV) proposent d'utiliser pour chaque pli de test un ensemble de données ayant toutes un dénominateur commun. Cela peut être un locuteur, un dialogue, etc. Cette méthode permet de construire des ensembles de tests complètement indépendants du locuteur, mais a l'inconvénient d'être très coûteuse en temps, chaque pli d'entrainement/test devant être réalisé manuellement ou par un algorithme non trivial. Nous avons, dans notre cas utilisé cette méthode dans le but de pouvoir conserver un équilibrage le plus précis possible entre les différentes catégories de locuteur (agent/client) et de type de voix (homme/femme).

5.3.1.2 Représentation des résultats

Il existe plusieurs façons de présenter les résultats obtenus. Le but de chacune de ces représentations est de rendre compte des points forts et des points faibles des modèles. La méthode la plus simple est la matrice de confusion dont un exemple est donné sur la Figure 5-3 ci dessous :

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

Figure 5-3 Exemple de matrice de confusion. Les instances correctement classifiées apparaissent sur la diagonale de la matrice

Dans ce mode de représentation les résultats corrects apparaissent sur la diagonale de la matrice. La matrice de confusion fournie ainsi des résultats bruts qui ne sont pas toujours pratiques à comparer surtout dans le cas ou plusieurs tests doivent être effectués. Dans ce cadre, plusieurs mesures comme la *recognition rate* (RR) qui représente le nombre d'instances bien reconnues par rapport au nombre total d'instances, sont régulièrement utilisées. Cependant, cette dernière mesure est biaisée si le nombre d'instances est inégalement réparti entre les différentes classes. Un cas d'école serait celui où la distribution des émotions correspond à celle des émotions dans des données réelles, avec en général, entre 80% et 90% de données de type « neutre ». Ainsi, un modèle qui classerait tout en Neutre aurait un très bon pourcentage de bonne détection, mais ne présenterait pas un grand intérêt. Nous donnerons donc nos résultats sur des ensembles équilibrés (UAR).

Nous avons choisi dans notre cas de présenter nos résultats sous la forme de f-score. Cette mesure est basée sur deux mesures intermédiaires : le rappel et la précision. Le rappel est le rapport entre le nombre d'instances correctement **attribuées** à une classe et le nombre d'instances **appartenant** à cette classe. La précision est quant à elle représentée par le rapport entre le nombre d'instances correctement **attribuées** à une classe émotionnelle *i* et le nombre de documents total **attribués** à la classe émotionnelle *i*.

En combinant ces deux mesures, nous obtenons un taux moyen permettant d'évaluer, d'une part la capacité de rappel du modèle et d'autre part sa capacité à sélectionner les instances intéressantes pour notre requête. Cette combinaison qui représente le F-score suit la formule suivante :

$$F = \frac{2 * (\text{Précision} * \text{rappel})}{\text{précision} + \text{rappel}}$$

Où

$$rappel_i = \frac{\text{Nombre d'instances correctement attribuées à la classe } i}{\text{Nombre d'instances appartentant à la classe } i}$$

Et

$$précision_i = \frac{\text{Nombre d'instances correctement attribuées à la classe } i}{\text{Nombre d'instances attribuées à la classe } i}$$

C'est cette mesure que nous avons retenue pour la présentation de nos scores et que nous utiliserons donc principalement dans la suite de ce chapitre pour évaluer nos résultats.

Les résultats peuvent être présentés de deux façons i) pondérés en fonction du nombre d'instances par classe, dans ce cas le score de la classe ayant le grand nombre d'instance est pondérée suivant sa proportion d'apparition au sein du corpus (*weighted F-score*) ii) non pondérés, dans ce cas les scores de toutes les classes sont pris en compte de manière égale pour le calcul du score final (*unweighted F-score*)

5.3.1.3 Les résultats obtenus sont ils statistiquement significatifs?

Lorsque nous obtenons des résultats suite à une procédure de classification automatique, il est préférable de calculer un intervalle de confiance permettant d'estimer la validité statistique du résultat. Cet intervalle a pour but d'indiquer qu'il est possible à x% de chance que le résultat obtenu soit statistiquement valable. Nous calculons pour cela l'intervalle de confiance à 95%. Cet intervalle se base sur le taux de reconnaissance correcte (P), le taux d'erreur (Q) et le nombre d'instances (N) impliquées dans le test. La formule suivante permet de calculer cet intervalle (Montacié and Chollet 1987) :

$$r = \mathrm{Z}\alpha \sqrt{\frac{\mathrm{p} * \mathrm{q}}{\mathrm{n}}}$$

Où $\alpha = 0.95$ et $\mathrm{Z}\alpha = 1.96$

5.4Choix de l'ensemble de descripteurs acoustiques pour la reconnaissance des émotions

Afin de sélectionner l'ensemble de descripteur et la méthode d'extraction la plus performante pour un tâche de classification nous avons étudié un sous ensemble de notre corpus et avons testé les modèles basés sur des jeux de descripteurs différents.

5.4.1 Données utilisées

Afin de tester les descripteurs les plus adaptés à une tâche de détection des émotions nous avons établi un ensemble de test composé de 3684 instances parmi celles ayant obtenues une annotation consensuelle (voir chapitre 4 pour plus de détail sur l'annotation du corpus). Les données sont réparties comme présenté dans le tableau suivant :

Positif	Neutre	Négatif
1235	1182	1267

Tableau 5-7 Répartition des instances utilisées pour la construction des modèles de détection

Afin d'éviter les problèmes de sur-apprentissage (over fitting) notamment sur les descripteurs relatifs à la fréquence fondamentale nous avons souhaité entrainer et tester nos modèle en étant indépendant du locuteur. Dans cette configuration un locuteur présent dans l'ensemble d'entrainement ne le sera pas dans l'ensemble de test. Nous nous sommes pour cela basé sur un modèle de type « Leave One Out Cross Validation » (voir paragraphe 5.3.1.1 sur la méthode de validation croisée). Ce sous ensemble comprend deux types de locuteur agents (1745 instances à et client (1939 instance). La classe Neutre est attribuée à hauteur de 30% au client (355 instances) et 70% à l'agent (827 instances). Pour la classe négative l'agent obtient 29% (371 instances) du total de la classe contre 71% pour l'agent (896 instances). Enfin 56% (692 instances) des instances positive sont attribuée à l'agent contre 44% pour le client (543 instances).

5.4.2 Sélection de descripteurs : Sélection de descripteurs grâce au F-score

Afin de déterminer les descripteurs les plus pertinents pour notre étude nous appliquons la méthode de sélection des descripteurs décrite dans (Chen and Lin 2005) et implémentée en tant que plug-in dans LibSVM (Chih-Chung et al. 2001). La méthode repose sur la sélection d'un descripteur en fonction du F-score obtenu par ce dernier. Le F-score de chaque descripteur est calculé de la manière suivante :

Etant donné un vecteur d'entrainement x_k avec k=1,….m si le nombre d'instances des classes positives et négatives sont représentées respectivement par n_+ et n_- alors le f-score du ième descripteur est défini par :

$$F(i) = \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n+}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n-}\left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2}$$

Ou

$x_i, x_i^{(+)}, x_i^{(-)}$ sont respectivement les moyennes du ième descripteur du corpus entier, de la classe positive et de la classe négative du corpus.

Et $x_{k,i}^{(+)}$ est le i-ème descripteur de la k-ième instance positive et $x_{k,i}^{(-)}$ est le i-ème descripteur de la k-ième instance négative. En pratique plus le f-score est élevé et plus le descripteur considéré est important pour la tâche de classification. Les étapes réalisées par l'algorithme sont les suivantes :

1. Calculer le score de chacun des descripteurs
2. Choisir arbitrairement plusieurs seuils au delà desquels conserver ou éliminer un descripteur
3. Pour chaque seuil
    a) Eliminer les descripteurs en dessous du seuil
    b) Séparer aléatoirement deux ensembles pour l'entrainement et la validation
    c) Entrainer les modèles pour effectuer une prédiction sur l'ensemble de validation
    d) Répéter 5 fois l'étape afin de calculer le taux d'erreur moyen
4. Choisir le seuil avec le taux d'erreur le plus bas
5. Eliminer les descripteurs avec des F-score en dessous du seuil sélectionné

Durant la phase d'évaluation nous avons défini manuellement 5 ensembles de descripteurs contenant respectivement i) la totalité des descripteurs ii) les 191 meilleurs iii) les 95 meilleurs iv) les 47 meilleurs et v) les 23 meilleurs. Pour chacun de ces sous ensembles le F-score était calculé. La décision finale a été prise en fonction des F-score les plus élevés.

A l'issu de la phase de sélection nous obtenons les 23 descripteurs suivants comme étant les plus discriminants et donnant les meilleurs scores sur tout le corpus en ce qui concernant l'ensemble de descripteurs calculés avec Praat et les bibliothèques LIMSI :

1. Max_Barkenergy7	11. Range_Loudness	21. Max_Barkenergy11
2. Max_Barkenergy8	12. Max_Barkenergy4	22. Range_Barkenergy5
3. Max_Barkenergy6	13. Range_Barkenergy9	23. Range_Mfcc2
4. Max_Intensity	14. Range_Barkenergy8	
5. Max_Barkenergy9	15. Max_Barkenergy10	
6. Max_Barkenergy3	16. Range_Barkenergy10	
7. Max_Barkenergy2	17. Range_Barkenergy7	
8. Max_Barkenergy	18. Mean_Barkenergy3	
9. Max_Barkenergy5	19. Mean_Barkenergy4	
10. Max_Loudness	20. Mean_Barkenergy6	

Tableau 5-8 Descripteurs retenus dans l'ensemble Praat + bibliothèques LIMSI

Les résultats obtenus sur cet ensemble de descripteurs indiquent une sur représentation des indices liés à l'échelle perceptive Bark (20 sur 23) (Zwicker and Feldtkeller 1981), ces derniers semblent particulièrement discriminants dans notre cas. Etonnamment nous ne retrouvons dans cette liste aucun indice relatif à la fréquence fondamentale. Ces derniers sont pourtant généralement réputés discriminants dans une tâche de reconnaissance des émotions.

Concernant la sélection effectuée sur les descripteurs issus du challenge 2009 et extraits avec l'outil openEAR nous obtenons les résultats suivants, 47 descripteurs ont été retenus dans ce cas :

1. mfcc_sma[11]_range	25. mfcc_sma_de[11]_min
2. mfcc_sma[10]_range	26. mfcc_sma[2]_range
3. mfcc_sma[8]_range	27. pcm_RMSenergy_sma_stddev
4. mfcc_sma_de[11]_range	28. mfcc_sma[11]_max
5. pcm_RMSenergy_sma_de_max	29. F0_sma_max
6. pcm_RMSenergy_sma_de_range	30. F0_sma_range
7. mfcc_sma[12]_range	31. F0_sma_de_range
8. voiceProb_sma_max	32. mfcc_sma[8]_linregerrQ
9. mfcc_sma_de[8]_range	33. mfcc_sma_de[8]_max
10. voiceProb_sma_de_range	34. mfcc_sma_de[2]_range
11. pcm_RMSenergy_sma_range	35. mfcc_sma_de[10]_max
12. mfcc_sma_de[12]_range	36. mfcc_sma[11]_stddev
13. pcm_RMSenergy_sma_max	37. F0_sma_amean
14. pcm_RMSenergy_sma_de_min	38. voiceProb_sma_de_min
15. mfcc_sma_de[11]_max	39. F0_sma_de_stddev
16. mfcc_sma_de[10]_range	40. F0_sma_de_min
17. voiceProb_sma_range	41. mfcc_sma_de[10]_min
18. mfcc_sma[10]_max	42. mfcc_sma_de[8]_min
19. voiceProb_sma_amean	43. mfcc_sma[8]_stddev
20. mfcc_sma_de[12]_max	44. mfcc_sma[4]_range
21. mfcc_sma[12]_max	45. mfcc_sma[6]_range
22. pcm_RMSenergy_sma_de_stddev	46. mfcc_sma_de[12]_min
23. voiceProb_sma_de_max	47. mfcc_sma_de[6]_range
24. mfcc_sma[9]_range	

Tableau 5-9 Descripteurs retenus dans l'ensemble challenge emotion 2009 et openEAR

Dans cette deuxième sous sélection les indices les plus représentés sont ceux relatifs à l'échelle perceptive mel (29 sur 47). Encore une fois les indices relatifs à la fréquence fondamentale ne sont que peu représentés (6 sur 47 descripteurs retenus). Les descripteurs relatifs à l'énergie sont présents à hauteur de 7 occurrences sur 47.

5.4.3 Validation de l'ensemble de descripteur

Nous avons principalement utilisé trois jeux de descripteurs pour construire nos modèles de détection. Les descripteurs obtenus avec la librairie LIMSI ont été calculés en utilisant 2 contextes différents. La première configuration se base sur l'ensemble des points (voisés ou non voisés) d'un segment, la seconde n'effectue le calcul que sur les points voisés.

Les meilleurs résultats ont été obtenus avec un classifieur de type SVM et un noyau de type RBF (Radial Basis Function). Pour chacun des tests les paramètres de C (coût) et G (largeur avec la frontière séparatrice) ont été optimisés à l'aide de l'outil libSVM (Chih-Chung et al. 2001).

		Pos	Neut	Neg	Weighted F-score	Unweigthed F-score
Ensemble 1	Tous les points **(744 descripteurs)**	0.512	**0.661**	**0.664**	**0.624**	**0.612** (±2%)
Ensemble 2	Tous les points voisés **(744 descripteurs)**	0.518	0.598	0.635	0.597	0.583 (±2%)
Ensemble 3	descripteur challenge 2009 **(384 descripteurs)**	**0.539**	0.621	0.597	0.591	0.585 (±2%)

Tableau 5-10 Variation des scores du modèle de détection en fonction de l'ensemble de paramètre utilisé

Comme nous pouvons le voir sur le Tableau 5-10 les scores les plus élevés sont obtenus avec les calculs de descripteurs prenant en compte l'ensemble des points, qu'ils soient voisés ou non voisés. Cet ensemble obtient les meilleures performances pour la détection des états neutres et négatifs. Les états positifs étant quant à eux mieux reconnus si l'on utilise l'ensemble des descripteurs proposés dans le challenge interspeech 2009 et extraits avec l'outil openEAR. Les variations entre le plus haut et le plus bas score pour chacun des états affectifs peux être plus important selon le jeu de descripteur sélectionné. Ainsi elle est de 2.7% pour la classe positive, 6.3% pour la classe neutre et 3.3% pour la classe négative, ce qui met en avant l'importance du choix de l'ensemble de descripteurs à sélectionner. A un niveau global nous observons jusqu'à 2.9% de différence entre le F-score le plus élevé et le F-score le moins élevé.

Nous pouvons voir que si en moyenne les scores prenant tous les points voisés et non voisés d'un segment semblent obtenir de meilleurs résultats, nous remarquons également que le nombre de descripteurs est beaucoup plus élevé (744 contre 384) pour une différence de score de 2.9%. Afin de vérifier si le résultat venait d'un choix plus adapté des descripteurs nous avons ramené le nombre d'indices de l'ensemble 1 à 372 (31 * 12) en supprimant les dérivés de chacun des descripteurs de cet ensemble. Nous conservons ainsi les 31 descripteurs bas niveaux présentés précédemment et les 12 fonctionnelles associées. Nous présentons également les résultats obtenus sur le jeu de descripteur réduit issu de la sélection de descripteur décrite dans le paragraphe 5.4.2. Nous avons, comme pour les autres expériences de ce chapitre, testé cet ensemble de paramètre sur des données entièrement indépendantes du locuteur afin d'observer les performances qu'aurait cette sélection dans une situation plus réelle. Nous présentons les scores obtenus par cet ensemble et le comparons à ceux de l'ensemble 3 dans le Tableau 5-11 ci-dessous :

		Pos	Neut	Neg	Weighted F-score	Unweigthed F-score (intervalle de confiance à 0.95)
Ensemble 3	descripteur challenge 2009 **(384 descripteurs)**	**0.539**	0.621	0.597	0.591	0.585 (±2%)
Ensemble 3'	descripteur challenge 2009 **(47 descripteurs)**	0.458	0.604	0.629	0.572	0.563 (±2%)
Ensemble 4	Tous les points **(372 descripteurs)**	0.513	**0.654**	0.634	**0.61**	**0.60** (±2%)
Ensemble 4'	Tous les points **(23 descripteurs)**	0.416	0.574	**0.656**	0.562	0.548 (±2%)

Tableau 5-11 Variation des scores du modèle de détection en fonction de l'ensemble de paramètre utilisé

Nous constatons une baisse des performances de l'ordre de 1% lorsque nous réduisons le jeu de descripteurs à 372. La baisse la plus significative se situe au niveau des états négatifs où l'on observe une dégradation des performances de 3%. Au niveau global la différence avec l'ensemble 3 apparait désormais comme très réduite (1.5% de variation au total). Bien que globalement les scores obtenus soient proche les différences de score entre les classes sont quant à elles plus notables puisque nous pouvons observer jusqu'à 3.7% d'écart entre les deux ensembles si nous considérons la classe négative. Concernant les tests sur un sous ensemble issu de la sélection de descripteurs nous pouvons constater que l'ensemble 3' obtient des performances plus proches de celles obtenues par l'ensemble 3 d'origine, nous observons un écart de 2.2% contre 5.2% si l'on observe les ensemble 4 et 4'. Deux raisons peuvent expliquer les scores moins élevés concernant les expériences 3' et 4'. La première est que contrairement aux expériences 3' et 4' l'étape de sélection de descripteurs a été effectuées à partir d'un ensemble dépendant du locuteur, ce qui modifie sensiblement la difficulté du problème de classification. De plus Il est possible que la chute constatée des F-score vienne du manque de diversité de l'ensemble 4' puisque les indices retenus appartenaient pour leur très grande majorité aux descripteurs relatifs à l'échelle Mel (indices MFCC). D'une manière générale les ensembles complets semblent obtenir des scores plus stables entre les différents types d'états affectifs. Nous ne considèrerons donc au moment de notre choix de jeu de descripteurs que les ensembles complets (ensemble 3 et 4).

Choix de l'ensemble de descripteurs pour la suite de l'étude :

Nous effectuerons notre choix sur l'ensemble de descripteurs à retenir entre les ensembles 3 et 4 considérés comme les plus performants d'après nos tests. Nous avons vu que les scores, bien que pouvant variér d'une classe à une autre, étaient globalement très proches et ne pouvaient pas constituer à eux seul, un critère de choix pour la sélection d'un jeu de descripteurs en particulier. L'ensemble 3 issu d'une sélection effectuée pour le challenge interspeech 2009 présente l'avantage d'avoir été utilisé par un grand nombre de participants et d'avoir ainsi prouvé sa robustesse sur une grande variété de méthodes et de classifieurs. Nous avons calculé l'écart type en prenant en compte

les résultats obtenus sur chacune des classes pour les ensembles 3 et 4. Nous obtenons respectivement 0.04 pour l'ensemble 3 et 0.07 pour l'ensemble 4, le premier étant ainsi potentiellement plus robuste aux variations. Nous retiendrons donc l'ensemble 4 comme jeu de descripteurs par défaut pour la suite de nos tests.

5.4.3.1 Reconnaissance des émotions en fonction du locuteur (agent/client)

Les données que nous utilisons mettent en jeu au minimum deux locuteurs : un client et un agent. Si le client est libre d'exprimer ses émotions sans contraintes particulières autres que celles des règles sociales, l'agent est pour sa part tenu, de part le contexte commercial de l'échange à une expression beaucoup plus modérée. Afin de déterminer si la différence d'expression d'une émotion entre le client et l'agent pouvait conduire à des diminutions de performance au niveau de l'apprentissage nous avons construit des modèles dépendants du type de locuteur (i.e. un ensemble d'entrainement et de test ne contenant que des clients et un second ne contenant que des agents). Nous utiliserons les scores obtenus avec le jeu de descripteurs du challenge « Emotion 2009 » comme score de référence. Le Tableau 5-12 ci-dessous présente les scores obtenus :

		Pos	Neut	Neg	Weighted F-score	Unweigthed F-score (intervalle de confiance à 0.95)
Expérience 1	**Modèle mixte agent/client**	**0.539**	0.621	0.597	0.591	0.585 (±2%)
Expérience 2	**Modèle Client**	0.483	0.580	**0.685**	0.607	0.582(±2%)
Expérience 3	**Modèle Agent**	0.504	**0.669**	0.452	0.588	0.541(±2%)

Tableau 5-12 Résultats d'une détection des émotions avec des modèles dépendants du type de locuteur (client/agent)

Les résultats que nous obtenons en utilisant des modèles indépendants du type de locuteurs ne montrent pas d'amélioration par rapport au modèle de base. D'une manière globale les f-score obtenus sont soit inférieurs (modèle agent) soit équivalents (modèle client). Si nous regardons les scores de chacun des états affectifs nous remarquons que localement les résultats peuvent être plus élevés que le modèle de base (exemple : négatif pour le client ou neutre pour l'agent). Ces résultats sont cependant au détriment des deux classes restantes qui ont des scores systématiquement plus bas. La conception de modèles dépendants du type de locuteur ne semble donc pas être une source d'amélioration dans notre cas.

5.5 Choix du vocabulaire pour la reconnaissance des émotions

Comme pour les indices acoustiques, il est nécessaire pour la partie textuelle de sélectionner les indices pertinents pour la construction de nos modèles. Nous avons testé différentes configurations afin de choisir la plus pertinente. La première utilise l'ensemble du vocabulaire (1408 mots) sans lemmatisation et en utilisant la totalité des classes grammaticales. La seconde se base sur du vocabulaire lemmatisé afin de réduire le nombre de dimensions lors de la classification (966 mots). La troisième expérience est menée sur un vocabulaire que nous avons filtré selon la catégorie

grammaticale pour ne conserver que les mots étiquetés par le tagger. Nous retenons ainsi seulement les mots de type verbes (sauf auxiliaires être et avoir), adverbes, noms, adjectifs et interjections (1253 mots) mais nous n'appliquons pas de processus de lemmatisation. La quatrième et dernière configuration porte sur un vocabulaire lemmatisé et un filtrage des mots selon leur catégorie grammaticale, nous ne conservons dans ce cas que les mots ayant été étiquetés comme des adverbes, verbes, interjection et noms (619 mots au total). Les balises de type [nom_client_agent], [rire] sont également conservées. Il faut noter que l'étiquetage morphosyntaxique fourni par l'outil « brill tagger » peut être perfectible du fait de la nature de nos données (conversations orales), certains mots pouvant recevoir une annotation erronée. Cette méthode offre cependant l'avantage de permettre des expériences sur les catégories de mots du vocabulaire sans devoir annoter manuellement notre corpus. Les expériences menées par (Garnier-Rizet et al. 2008) montrent des scores d'étiquetage de l'ordre de 91.65% en précision. L'ensemble des expériences ont été menées sur une représentation des tours de parole sous forme de sac de mots. Pour toutes les expériences, nous avons appliqué une normalisation TF-IDF et avons limité le vocabulaire sélectionné aux mots apparaissant au minimum 2 fois dans l'ensemble du corpus. La classification a été faite à l'aide d'un classifieur de type SVM et un noyau RBF avec les paramètres C et G optimisés. Le Tableau 5-13 ci-dessous présente les résultats obtenus :

		Pos	Neut	Neg	Weighted F-score	Unweigthed F-score (intervalle de confiance à 0.95)
Expérience 1	**Vocabulaire complet**	0.577	0.642	0.7	0.640	0.639 (±2%)
Expérience 2	**Vocabulaire lemmatisé**	0.653	0.619	**0.727**	**0.666**	0.665 (±2%)
Expérience 3	**Vocabulaire filtré sur la catégorie grammaticale et non lemmatisé**	**0.656**	**0.626**	0.718	**0.666**	**0.666** (±2%)
Expérience 4	**Vocabulaire lemmatisé et filtre sur les catégories grammaticales**	0.53	0.622	0.674	0.61	0.608 (±2%)

Tableau 5-13 Résultats des différentes configurations pour la classification lexicale

Dans toutes les configurations testées, c'est au niveau de l'expérience 2 que les résultats sont les plus intéressants. Malgré un nombre de mots inférieur de 32% à l'expérience 1 la classification sur le vocabulaire lemmatisé se montre globalement plus performante de 2.5%. L'expérience 3 utilise un vocabulaire filtré sur les catégories grammaticales (noms, adverbe, adjectifs, interjection) mais sans lemmatisation. Avec une taille de vocabulaire inférieure d'environ 11% (1253 mots vs. 1408 mots) à celui de l'expérience 1 nous obtenons un gain de performance de l'ordre de 2.5% également. En appliquant une lemmatisation ainsi qu'un filtrage sur les catégories grammaticales mentionnées ci-dessus, l'expérience 4 obtient une baisse de score de 3% par rapport à l'expérience 1 avec 619 mots de vocabulaire au total.

Nous avons cherché à vérifier la ou les catégories grammaticales porteuses de sens pour une détection des émotions. L'objectif était ici de trouver la combinaison permettant de se rapprocher le plus possible des scores obtenus au cours de l'expérience 2 en conservant un nombre de dimensions inférieur. En partant de l'expérience 4 nous avons successivement ajouté, de manière indépendante pour chaque expérience, des catégories grammaticales. Dans une première expérience nous ajoutons les pronoms et déterminants, dans un second temps les auxiliaires « être » et les prépositions et dans un dernier temps les substantifs. Ces derniers ne constituent pas à proprement parler une classe grammaticale puisque nous pouvons retrouver des natures de mots très variées en leur sein (verbes, déterminants, adjectifs) mais sont porteurs de sens dans la mesure où ceux-ci permettent de désigner ou de faire référence à une cible en particulier. Le Tableau 5-14 ci-dessous présente les résultats obtenus :

		Pos	Neut	Neg	Weighted F-score	Unweigthed F-score (intervalle de confiance à 0.95)
Expérience 4	**Vocabulaire lemmatisé et filtre sur les catégories grammaticales (619 mots)**	0.53	0.622	0.674	0.61	0.608 (±2%)
Expérience 5	**Expérience 4 + pronoms et auxiliaires (639 mots)**	0.537	**0.627**	0.686	0.618	0.616 (±2%)
Expérience 6	**Expérience 4 + déterminants et prépositions (649 mots)**	0.578	0.592	0.667	0.613	0.612 (±2%)
Expérience 7	**Expérience 4 + substantifs (1048 mots)**	**0.653**	0.626	**0.718**	**0.666**	**0.665** (±2%)

Tableau 5-14 Résultats en fonction de l'ajout des classes grammaticales

Les scores obtenus en faisant varier les différentes catégories au cours des expériences 5 et 6 ne permettent pas d'augmenter de manière significative les scores globaux. Nous n'observons que des variations inférieures à 1%. Le nombre de mots utilisés est de +3% pour l'expérience 5 et de +5% pour l'expérience 6. Les résultats obtenus dans l'expérience 5 mettant en jeu les pronoms et les auxiliaires sont décevants sachant que la catégorie grammaticale « pronoms » était très largement représentée comme étant une catégorie discriminante des émotions négatives au cours de l'analyse lexicométrique entreprise dans le paragraphe 5.1.6.1. L'expérience 7 offre quant à elle des scores plus élevés puisque nous arrivons à un gain global de +6%. Nous observons cependant que l'ajout de la classe « substantif » rajoute un nombre de mots importants (+41%). Il est donc difficile de déterminer si le gain obtenu vient de l'augmentation importante de la taille du vocabulaire ou de la sélection d'une classe de mots en particulier ; Dans les faits il est probable que les deux raisons soient à l'origine de cette hausse.

Au moment de faire notre choix concernant l'ensemble de vocabulaire à prendre en compte deux possibilités sont envisageables. La première est celle concernant le vocabulaire mis en œuvre dans l'expérience 4 (vocabulaire lemmatisé + filtre sur les catégories grammaticales) et la seconde sur l'expérience 7 (vocabulaire de l'expérience 4 + substantif). Si les indices lexicaux utilisés dans l'expérience 7 offrent des performances plus élevées il nous est cependant difficile de cerner la cause exacte de l'augmentation des scores (classe grammaticale ou augmentation du lexique). Il est fort probable que les deux facteurs en soient responsables. Le vocabulaire de l'expérience 4 quant à lui, malgré des performances un peu en retrait, offre une base théorique plus solide. Il a en effet été démontré, notamment dans (Turney 2002) que les classes de mots de type verbes, noms et adverbes étaient discriminantes pour la détection d'états affectifs dans le texte.

Partant de ce constat et sachant les très grandes variations de scores pouvant être obtenues en fonction des données utilisées il nous semble préférable de baser la suite de notre étude sur les données issues de l'expérience 4, plus solide méthodologiquement parlant et donc potentiellement plus robustes à la variation de données de l'ensemble de test.

5.6 Conclusion

Nous avons vu dans ce chapitre les différents outils retenus pour la classification et l'évaluation des expériences menées dans les chapitres suivants. Un très large panel d'autres méthodes sont bien entendu existantes, celles présentées dans les lignes ci-dessus sont celles nous ayant permis d'obtenir les résultats les plus concluants étant donné les tests que nous avons effectués et la littérature du domaine parmi laquelle nous pouvons notamment citer (Schuller et al. 2009). Dans une seconde partie de ce chapitre nous avons mené une étude afin de déterminer quels types d'indices acoustiques étaient pertinents pour la construction d'un modèle de détection des émotions basé sur des données naturelles. A l'issue d'une phase de sélection de descripteurs les indices relatifs aux échelles perceptives Bark et Mel semblent avoir été particulièrement performants dans les tâches de classification que nous avons testées. Nous avons pour finir comparé différents ensembles de descripteurs dans le but de choisir le plus approprié pour la suite de notre étude. A l'issue de ces différents tests nous avons choisi de retenir les descripteurs utilisés dans le challenge émotion 2009 (Schuller et al. 2009) du fait de sa potentielle plus grande robustesse face à des données variées. Ce choix nous permet également de nous positionner plus facilement vis-à-vis des différents travaux de la communauté ce qui peut être intéressant pour l'évaluation de nos modèles. Nous avons pour finir expérimenté des modèles dépendants du type de locuteur (agent/client) qui n'apportent pas de gains significatifs, les scores les plus élevés étant atteints par les modèles mixtes.

Basant notre étude sur des indices de type acoustique mais également linguistiques (issus des transcriptions de conversations) nous avons présenté les premiers résultats obtenus à partir de ces derniers. Nous avons vu que la sélection de vocabulaire par catégorie grammaticale pouvait donner des résultats intéressants en terme de rapport « performance/ nombre de dimensions ». Les baisses de scores constatées lorsque nous n'utilisons que certaines classes de mots (verbes, adverbes, noms, interjections) ne sont que de 3% avec un nombre de dimensions inférieur de plus de 50% par rapport à la solution mettant en œuvre l'ensemble du vocabulaire (619 mots vs 1408 mots) . Cet ensemble a été retenu comme jeu de descripteurs lexicaux par défaut pour la suite de l'étude.

Chapitre 6 Détection des émotions real-life en centre d'appels : Fusion d'indices linguistiques et paralinguistiques

Résumé

Ce chapitre s'attache à mettre en avant les spécificités de la détection des émotions réalisée sur des données entièrement automatiques. Nous montrerons ainsi les dégradations de performances que nous obtenons lorsque nous nous plaçons dans un contexte de plus en plus réaliste (utilisation d'une segmentation automatique, d'un système de reconnaissance de la parole pour obtenir une transcription automatique à la place d'une transcription manuelle, conservation des segments mettant en jeu de la parole superposée ou un environnement bruité).

Observant une diminution importante des résultats obtenus nous proposons une méthode de fusion d'indices basée sur les sorties de nos modèles acoustiques et lexicaux de détection des émotions ainsi que sur celles de modèles de détection linguistique des sentiments et des opinions. Nous montrons que ces approches sont complémentaires et nous permettent dans tous les cas d'obtenir des scores de détection supérieurs à ceux obtenus en utilisant une approche acoustique seule. Nous finissons ce chapitre en testant cette méthode sur une sélection de 264 dialogues en éliminant toute intervention non automatisée sur l'ensemble de test.

6.1 Vers plus de réalisme pour les systèmes de détection des émotions

Nous avons pu voir dans les chapitres précédents que les corpus issus de données real-life sont beaucoup plus riches en termes de variété d'expressions émotionnelles. Cette richesse rend cependant leur détection automatique beaucoup plus difficile. Afin d'illustrer ce propos, nous commencerons ce chapitre avec la présentation d'une étude utilisant 3 corpus de centres d'appels différents. Nous verrons que même avec des données mettant en œuvre des types de conversations identiques (conversations téléphoniques en centre d'appels) mais avec des tâches différentes, la généralisation des modèles est une tâche qui reste problématique du fait de la très grande variabilité expressive que peut avoir une émotion. Par la suite afin de pouvoir appréhender les phénomènes affectifs de manière plus globale (et donc potentiellement généraliser plus facilement nos modèles), nous proposons de vérifier l'impact, en terme de taux de reconnaissance, de l'introduction de bruits dans les données afin de se rapprocher le plus possible d'une condition réelle d'utilisation. Ce bruit passe par l'utilisation de tours de parole contenant de la parole superposée mais également de bruits environnementaux (animaux, musique, etc). Nous démarrerons ainsi l'étude en utilisant des données « propres » seulement et des conditions d'entrainement idéales (utilisation de tours de parole issus d'une segmentation manuelle, exclusion des tours de parole bruitée, jeux d'entrainement des données speaker-dépendant) et introduirons successivement du bruit en test dans le but de nous rapprocher le plus possible d'une situation réelle (tours de parole issus de la segmentation automatique, segments non consensuels lors de la phase d'annotation dus à une expression affective complexe, jeux de données speaker-indépendant à 100%). Ces deux premières études nous permettront ainsi de dresser un bilan de la complexité de la tâche ; puis dans un second temps, nous verrons comment, grâce à l'introduction de nouveaux indices représentant des niveaux d'expression variés (lexicaux, syntaxiques, sémantiques, acoustiques) nous avons pu améliorer la qualité de la détection en nous plaçant dans une situation nécessitant le minimum de prétraitement manuel possible. Dans ce cadre, les données utilisées seront composées de tours de parole définis de façon automatique et de transcriptions fournies par un système de reconnaissance de la parole. Les différents indices acoustiques et linguistiques (que nous détaillerons dans les paragraphes suivants) seront issus intégralement de traitements automatiques.

6.2 Détection automatique des émotions naturelles en centre d'appels : Quels sont les challenges à surmonter ?

Nous avons vu dans la section dédiée à la détection des émotions dans les centres d'appels du chapitre 3 qu'un certain nombre d'études existaient. Ces études mettent en jeu un ou plusieurs corpus de taille variable préalablement « nettoyés » à l'aide de prétraitements manuels. Nous retrouvons en règle générale parmi ces tâches de prétraitement :

1) le découpage manuel des conversations dans le but d'obtenir des instances de longueur et de contenu directement exploitable pour une tâche d'entrainement de modèle computationnel ;

2) L'élimination des segments non consensuels lors de la phase d'annotation (n'obtenant pas un nombre de votes pour une classe donnée suffisamment important) ;
3) L'élimination de segments contenant de la parole superposée ou des bruits d'arrière-plan, ces derniers venant fausser le calcul des descripteurs lors de la phase d'extraction des indices et par extension polluer le modèle construit sur cette base.

S'il semble évident que ces prétraitements vont contribuer à améliorer les scores obtenus, ils nous éloignent cependant des conditions réelles qu'aurait à gérer un système de détection des émotions dans le cadre d'une utilisation dans un contexte de production. Partant de ce constat, des auteurs comme (Schuller et al. 2007) ont par exemple introduit artificiellement du bruit dans leurs données afin de tester la robustesse de leurs modèles. Ils constatent des baisses importantes dans les taux de reconnaissance observés au fur et à mesure que le bruit introduit augmente (supérieur à 10% de dégradation dans certains cas). Le problème de la segmentation est également mis en avant ; toutefois la baisse de performance en cas de segmentation non adaptée (coupure de mots par exemple) semble moins pénalisante que dans l'exemple précédent (1,5% en moyenne).

Dans notre cas, nous reprendrons les points 1, 2 et 3 mentionnés ci-dessus pour vérifier leur impact sur les systèmes de détection. Les paragraphes ci-dessous vont donc s'attacher à décrire les variations de performances de ces modèles en fonction des situations de test (des plus artificielles au plus real-life).

6.2.1 Modes de segmentation et données utilisées

Le sous corpus que nous avons sélectionné pour l'entrainement des modèles d'apprentissage comprend 3684 segments issus de la segmentation manuelle ayant obtenu une annotation identique de la part des deux annotateurs (voir chapitre 4 pour le détail concernant les scores d'annotation). Trois classes émotionnelles seront considérées. La classe négative contient des émotions de type colère, déception, surprise négative. La classe positive est composée quant à elle d'émotions de type satisfaction, joie et surprise positive. La classe neutre comprend des segments ou aucun état affectif particulier n'est exprimé.

La répartition des segments peut être observée dans le Tableau 6-1 ci-dessous :

	Positif	Neutre	Négatif
Nombre d'instances retenues	1235	1182	1267

Tableau 6-1 Répartition du nombre d'instances en classes positives, négatives et neutres

Considérant la nature des données utilisées nous retrouvons un nombre important de segments mal formés (segments contenant plusieurs locuteurs mais ne parlant pas en même temps, contenant de la parole superposée, des bruits de fond, de la musique, etc). Ce type de segments représente 13% du corpus segmenté manuellement et passe à 18% pour les données segmentées automatiquement.

6.2.2 Premiers résultats : approches paralinguistique et lexicale

Comme nous le mentionnions précédemment, nous basons notre analyse, d'une part sur des données prétraitées manuellement (segmentation et transcription), et d'autre part sur des données obtenues à l'aide de systèmes automatiques que nous présentons dans le paragraphe ci dessous. Le système de la reconnaissance de la parole nous ayant été fourni par la société Vocapia et l'équipe TLP du LIMSI nous présenterons principalement les performances de ce dernier sur des tâches de reconnaissance de la parole dans des conversations téléphoniques issues de la base de données utilisée dans cette thèse. Les lecteurs désireux de connaître plus de détail sur le système de reconnaissance de la parole utilisés peuvent se référer aux articles suivants : (Gauvain et al. 2003; Garnier et al. 2005; Barras et al. 2006).

6.2.2.1 Le système de reconnaissance de la parole

La reconnaissance de la parole au cours de conversations téléphoniques est une tâche connue pour sa difficulté de part la grande variété de phénomènes liés au support téléphonique pouvant venir perturber la tâche de reconnaissance (bruit d'arrière plan, du combiné téléphonique, etc). Le système utilisé est basé sur celui de (Garnier-Rizet et al. 2008). Le système a été entrainé avec des corpus de conversations téléphoniques et notamment des dialogues provenant du projet CallSurf (Garnier-Rizet et al. 2008) ainsi que de la base de données utilisée dans cette thèse incluant des heures de paroles où de la parole émotionelle est présente (voir chapitre 4 pour plus de détails sur le corpus).

Afin de pouvoir évaluer les performances du système de reconnaissance de la parole sur des données émotionnelles nous présentons les taux d'erreur WER (Word Error Rate) obtenus sur différents type de données. Le Word Error Rate est une mesure couramment utilisée en reconnaissance de la parole. L'idée est de mesurer la distance entre une transcription de référence (usuellement effectuée par un humain) et les hypothèses fournies par le système. Le taux *WER* est défini par l'équation suivante :

$$WER = \frac{S + I + D}{N}$$

où S représente le nombre de substitutions (Un mot est remplacé par un autre), I le nombre d'insertions (Un mot est ajouté dans l'hypothèse alors qu'il n'est pas présent dans la phrase de référence), et D le nombre « *deletion* » (un mot présent dans la phrase de référence est manquant). N représente le nombre de mots dans la phrase de référence.

L'évaluation du système de reconnaissance à été réalisée sur 3 heures de paroles. 2 de ces 3 heures contenant des dialogues émotionnellement marqués « joie » ou « colère » sont issues du corpus présenté en chapitre 4. Une heure de dialogue supplémentaire composée de dialogues neutres est venu compléter cet ensemble de test. Les taux d'erreur ont été calculés en prenant en compte différents critères (toutes les données, données mettant en jeu des émotions positives/négatives/neutre, parole exprimée par le client ou l'agent et par un homme ou une femme). Les résultats obtenus sont présentés dans le Tableau 6-2 ci-dessous (Clavel et al. 2012) :

	Nombres de mots dans les phrases de références	Word Error Rate
Tout type de parole confondue	108k	33.8%
Neutre	36k	29.6%
Colère	38k	30.8%
Joie	34k	41.6%
Agent	58k	32.5%
Client	50k	35.0%
Homme	36k	38.1%
Femme	72k	31.6%

Tableau 6-2 Résultats du système de reconnaissance de la parole en fonction du type de parole considéré

Le système de reconnaissance de la parole obtient un taux d'erreur de reconnaissance de 33.8% tout type de parole confondu. De manière étonnante nous pouvons remarquer que ce taux d'erreur est par ailleurs plus élevé si l'on considère les émotions de type joie que lorsque nous considérons les émotions de type colère (WER de 41.6% pour la joie contre 30.8% pour la colère).

6.2.2.2 Résultats de l'approche paralinguistique

Nous mentionnions dans les paragraphes précédents la grande disparité de résultats que nous pouvions constater selon les données utilisées et le protocole de test mis en œuvre. Afin de rester cohérent, le jeu utilisé dans l'ensemble d'apprentissage ne varie jamais. Les données présentées dans le paragraphe 6.2.1 sont utilisées. Nous présentons dans un premier temps les scores que nous obtenons en fonction de 3 contextes de test différents. Nous partons de la situation la plus favorable (expérience 1) pour arriver à la situation se plaçant dans le contexte le plus réaliste (expérience 3) :

1) La première expérience est considérée comme la plus simple. Elle est réalisée sur 3 classes d'émotions, le mode d'évaluation utilisé est celui de la validation par test croisé. Les données utilisées ont été segmentées manuellement. Dans ce mode, certains locuteurs présents dans l'ensemble de tests peuvent être également présents sur le sous ensemble de données pris pour le test. Ce type de jeu de données est qualifié de *speaker-dependent.*
2) La seconde expérience propose d'utiliser une partie des données d'apprentissage afin de créer un jeu de tests indépendant. Les données utilisées en test sont issues de la segmentation manuelle. Le but ici est de tester le modèle sur un ensemble de données ou les locuteurs présents dans l'ensemble d'apprentissage n'apparaissent jamais dans l'ensemble de tests. Ce type de jeu de données est qualifié de *speaker-independent.*
3) La troisième expérience reprend les conditions utilisées dans l'expérience 2 mais utilise dans l'ensemble de tests des données issues de la segmentation automatique.

Les résultats sont présentés dans le Tableau 6-3 ci-dessous :

Expériences	Résultats sur 3 classes d'émotions positives, négatives et neutres F-score (intervalle de confiance à 0.95)
Expérience 1 (segmentation manuelle / données speaker dépendant)	0.71 (±1%)
Expérience 2 (segmentation manuelle / données speaker indépendant)	0.59 (±5%)
Expérience 3 (segmentation automatique / données speaker indépendant)	0.54 (±6%)

Tableau 6-3 Résultats des expériences 1 à 3 sur le canal Paralinguistique

Les différences de scores observées sont relativement importantes considérant le fait que les mêmes données sont utilisées pour chaque expérience. Nous pouvons voir que la différence de F-score entre l'expérience 1 et l'expérience 3 est de 0.17. La différence de score entre l'expérience 1 et l'expérience 2 peut être expliquée par le fait que certains paramètres comme la fréquence fondamentale par exemple, bien qu'ayant prouvé son efficacité pour la détection d'émotions, est extrêmement dépendante du locuteur. Ainsi, dans un jeu de données speaker- dépendant comme c'est le cas dans l'expérience 1, les paramètres relatifs à des descripteurs comme la F0 par exemple peuvent être utilisés et augmentent de ce fait les scores de détection correcte de manière artificielle. Cet impact est particulièrement visible dans des tests comme celui que nous présentons du fait de la grande variabilité de locuteurs présents dans le corpus.

La différence entre l'expérience 2 et l'expérience 3 vise à mettre en avant l'impact du type de segmentation sur les scores de détection. Nous voyons ici une dégradation de l'ordre de 5% qui peut être expliquée par la longueur moyenne plus importante des segments automatiques par rapport aux segments manuels (voir chapitre 4). Etant donnée la durée de certains segments automatiques (supérieurs à 1 minute dans certains cas), l'émotion exprimée ne serait pas nécessairement homogène sur toute la longueur du segment. Une autre explication possible est celle que (Schuller et al. 2007) mettaient en avant, à savoir la prise en compte de segments ayant été coupés par le système de reconnaissance de la parole avant une partie de segment qui aurait été utile à la reconnaissance d'un état émotionnel (une voyelle par exemple).

6.2.2.3 Approche lexicale

Les mêmes expériences (Expériences 1 à 3) ont été réalisées en utilisant cette fois-ci le canal linguistique comme support. Nous avons réalisé cette expérience en utilisant les indices lexicaux (la méthode d'extraction est décrite dans le chapitre 5). Les scores sont présentés dans le Tableau 6-4 ci-dessous :

Expériences	Résultats sur 3 classes d'émotions positives, négatives et neutres F-score (intervalle de confiance à 0.95)
Expérience 1 (segmentation manuelle / données speaker dépendant)	0.61 (±2%)
Expérience 2 (segmentation manuelle / données speaker indépendant)	0.55 (±6%)
Expérience 3 (segmentation automatique / données speaker indépendant)	0.52 (±6%)

Tableau 6-4 Résultats des expériences 1 à 3 sur le canal linguistique

Sur ce canal nous pouvons observer une variation relativement faible des scores obtenus selon les expériences. Les causes de ces variations semblent différentes de celles que nous avons pu constater sur le canal paralinguistique. Nous constatons dans un premier temps que la dégradation entre une approche indépendante du locuteur et une approche dépendante du locuteur est moins marquée (-6%) contrairement à ce que nous avions pu voir pour le canal paralinguistique. Cette constatation semble logique du fait que l'approche que nous avons adoptée est simplement lexicale (présence ou absence de mots dans un segment) et que seulement des parties de la phrase considérées indépendamment les unes des autres sont retenues pour la constitution des sacs de mots. Dans ce cas, des expressions ou structures syntaxiques spécifiques à un locuteur ont peu de chance d'influencer la classification. La dégradation est également relativement faible lorsque nous changeons de mode de segmentation (de manuelle à automatique). La perte observée peut être expliquée par le fait que des phrases peuvent être coupées inopinément lors d'un tour de parole du fait d'une erreur du système de reconnaissance de la parole. Notre approche se basant sur un étiquetage morpho-syntaxique pour l'extraction des descripteurs lexicaux (adverbe, verbe, nom, interjection) ce mode de sélection peut être problématique et fausser l'étiquetage du fait d'une phrase incomplète par exemple. Il faut également noter que la structure syntaxique de l'oral est souvent déformée par rapport à l'expression écrite ce qui peut amener à un étiquetage défaillant. Nous remarquons par ailleurs dans l'ensemble de tests, et ce quelle que soit l'expérience, un nombre important de vecteurs vides (28%) dû au manque de données annotées, notre dictionnaire ne se composant au total que de 619 mots pleins. Dans la situation la moins favorable le score de détection reste au dessus de celui de la chance mais il existe d'importantes confusions entre les émotions de types positif et neutre.

6.2.2.4 Détection paralinguistique des émotions : première utilisation dans un contexte complètement automatisé :

Nous avons pu avoir un premier aperçu, dans les paragraphes 6.2.2.2 et 6.2.2.3, des résultats et de la dégradation de ces derniers selon le contexte de l'expérience qui est menée (du contexte le plus artificiel au plus réaliste). Nous avons vu que le canal paralinguistique est dans notre cas plus performant, ce que constatent également les auteurs des études (Batliner et al. 2010; Polzehl et al. 2011).

D'après ces constatations, nous avons souhaité évaluer les performances que nous pourrions obtenir en utilisant la totalité des segments d'un dialogue dans l'ensemble de tests comme ce serait le cas si nous nous placions dans un contexte d'exploitation dans la vie réelle. Le protocole employé est le suivant pour réaliser cette expérience : la segmentation automatique fournie par le système de

reconnaissance de la parole. Nous utilisons pour cette fois seulement les indices paralinguistiques. Pour la validation, nous adoptons le protocole « leave one dialog out ». Pour ce faire, nous construisons les modèles à partir de N – 1 dialogues. Le dialogue restant est utilisé comme ensemble de tests. La totalité du dialogue est utilisé ; nous conservons ainsi les segments n'ayant pas reçu une annotation consensuelle et ceux contenant de la parole superposée, du bruit ou des silences.

Les segments que nous utilisons pour l'entrainement sont identiques à ceux utilisés pour les expériences paralinguistiques et lexicales décrites précédemment. Les données de l'ensemble de tests varient partiellement du fait de l'ajout des segments bruités et non consensuels. Afin d'observer les variations de résultats obtenus, trois sous-ensembles de tests ont été construits :

- Sous-ensemble 1 : Il comprend les segments dits « propres » ne contenant qu'un seul locuteur et ayant une annotation consensuelle de la part des deux annotateurs.
- Sous-ensemble 2 : Il prend en compte le sous ensemble 1 plus les segments pouvant contenir plusieurs locuteurs ainsi que du bruit d'arrière- plan mais ayant une annotation consensuelle.
- Sous-ensemble 3 : Celui-ci comprend les sous- ensembles 1 et 2 plus les segments contenant du bruit ou une annotation non consensuelle. La totalité des segments des dialogues est présente dans ce sous- ensemble.

Dans le cas du sous-ensemble 3, la présence de segments non consensuels dans l'ensemble de tests rend l'évaluation par segments impossible. Nous avons donc procédé à une évaluation globale dialogue par dialogue. Pour rappel, lors de la constitution du corpus nous avions sélectionné 41 dialogues : 26 d'entre eux ont été retenus pour leur contenu majoritairement négatif et 15 autres pour leur contenu positif.

Nous avons défini une fonction de décision sous forme de règles. Pour chacun des 41 dialogues, deux règles de décision simple sont définies :

- IF #negative seg > #positive seg THEN dialog = negative
- IF # positive seg > #negative seg THEN dialog = positive

Dans cette configuration, les segments étiquetés « neutres » par le système ne sont pas pris en compte, aucun dialogue n'ayant été sélectionné dans ce but. Les règles présentées ci-dessus sont celles qui ont obtenu les meilleurs résultats. Différentes règles ont été essayées : pondération du poids de chaque segment par le F-score de la classe qui lui est attribué, prise en compte d'un nombre réduit de segments tirés aléatoirement dans le dialogue en fonction de leur durée, regroupement de segments par fenêtre temporelle. Ces dernières se sont révélées plus compliquées à mettre en œuvre avec des résultats équivalents ou inférieurs. Les résultats sont présentés dans le Tableau 6-5 ci-dessous :

	Dialogues positifs	Dialogues négatifs	F-Score (intervalle de confiance à 0.95)
Sous-ensemble 1 : segments avec un seul locuteur et ayant une annotation consensuelle	7/15	22/26	0.66 (±7%)
Sous-ensemble 2 : Sous ensemble 1 + segments comportant de la parole des 2 locuteurs	7/15	23/26	0.68 (±7%)
Sous-ensemble 3 : Sous ensemble 1 +sous ensemble 2 + bruits et segments non consensuels	4/15	21/26	0.51 (±8%)

Tableau 6-5 F-score et intervalle de confiance en fonction des sous-ensembles 1 à 3 (du plus artificiel au plus réaliste)

Nous nous servirons du sous-ensemble 1 comme référence puisqu'il représente la situation expérimentale la plus contrôlée. Aucune dégradation de performance n'est constatée entre les sous-ensemble 1 et 2 ; un léger gain de performance concernant les dialogues négatifs est même à noter. L'introduction de segments erronés comprenant de la parole des deux locuteurs ne semblerait donc pas poser de problème significatif lors de l'évaluation de dialogues à un niveau global.

L'analyse est clairement différente lorsque nous comparons les sous- ensembles 1 et 3. Dans ce cas, l'introduction de segments non consensuels vient considérablement bruiter la classification. Ainsi, si les dialogues négatifs conservent un taux de rappel correct (> 0.8), la classification des dialogues positifs est quant à elle nettement plus problématique avec un score de rappel de 0.26. L'introduction d'une classe destinée à identifier ces segments problématiques (bruités ou ambigus) n'a pas donné de résultats satisfaisants, les scores présentés dans le tableau Tableau 6-5 restant identiques.

Cette série d'expériences montre clairement toute la difficulté d'évaluer des données émotionnelles de manière totalement aveugle (aucun prétraitement manuel de l'ensemble de tests). Si dans la globalité les scores restent au -dessus du niveau du hasard, la variabilité entre les classes positives et négatives est importante et la perspective d'améliorer de manière significative ces scores, en ne se basant que sur un canal de communication (ici paralinguistique), semble compromise. Pour essayer de remédier à ce problème, nous proposons donc dans les paragraphes suivants une détection à plusieurs niveaux (paralinguistique, lexicale, syntaxique, sémantique).

6.3 Fusion d'indices pour la détection d'émotions

Comme nous avons pu le voir dans le paragraphe 6.2.2.4, la détection des émotions de manière complètement automatisée semble difficile en l'état actuel de l'état de l'art. La fusion d'informations, dont un état de l'art est donné dans le chapitre 3, permet de coupler des indices collectés à différents niveaux afin d'avoir une classification plus fine. Dans (Vidrascu and Devillers 2005a) les auteurs mettaient en avant la complémentarité des canaux linguistiques et paralinguistiques pour une tâche de détection des émotions en utilisant des modèles mixtes lexicaux/acoustiques pour améliorer leurs résultats. Nous reprendrons dans cette étude une approche mettant en jeu des indices intervenant à 4 niveaux différents. Deux de ces niveaux

(acoustique et lexical) ont déjà été décrits dans les paraphages 5.1.1 et 5.1.6 . Les deux types de descripteurs supplémentaires qui interviennent au niveau syntaxique et sémantique sont décrits dans les paragraphes 6.3.1.1 et 6.3.1.2 ci-dessous et sont issus des travaux de deux partenaires industriels impliqués dans le projet (Temis et Sinequa). Les paragraphes suivants présentent les méthodes employées pour la combinaison de l'ensemble de ces indices ainsi que les gains obtenus suite à cette fusion.

6.3.1 Descripteurs supplémentaires

6.3.1.1 Descripteurs issus de patrons textuels (Cailliau and Cavet 2010)

Les descripteurs issus de la méthode décrite dans (Cailliau and Cavet 2010) consistent en deux scores associés à une polarité négative ou positive. Plus le score est élevé, plus la polarité est marquée. Un segment de parole se voit ainsi attribuer un score représentant sa polarité positive et un second représentant sa polarité négative.

Afin de calculer la polarité d'une portion de dialogue, cinq axes sont définis. Le cadre théorique adopté est celui défini dans (Charaudeau 1992). Chacun de ces axes peut être exprimé au cours de la conversation de manière émotionnelle ou non. Les cinq axes retenus sont :

- **Appréciation :** Deux polarités favorables ou défavorables sont possibles pour cette modalité. L'appréciation est mise en relation avec une expression de satisfaction ou d'insatisfaction de la part du locuteur envers un fait ou un objet. Exemple d'appréciation favorable : *« c'est **génial** »*. Exemple d'appréciation défavorable : *« je trouve ça vraiment **inadmissible** »*
- **Acceptation/refus :** Cet axe est illustré par l'acceptation ou le refus d'une proposition émise par un locuteur envers un autre. Un exemple d'acceptation : *« Oui, **d'accord**, ça me va comme ça ».* Exemple de refus : *« C'est **hors de question** »*
- **Accord/Désaccord :** Cette modalité se présente comme une gradation comportant quatre niveaux (accord total, accord approximatif, rectification, désaccord). Une phrase comme : *« Oui bien sûr »* correspond par exemple à un accord total. A l'autre bout du continuum, une déclaration telle que : *« Ca n'a **pas de sens** ! »* représente un désaccord total .
- **Opinion :** Cette modalité traduit un degré de certitude de la part du locuteur. Ce degré de certitude est représenté par une échelle allant de la conviction, en passant par la supposition (faible, forte) jusqu'au doute pouvant remettre en cause l'assertion exprimée par l'interlocuteur. Exemple de phrase mettant en œuvre une conviction : *« J'en suis **absolument certain** »*. Exemple de phrase mettant en œuvre du doute : *« J'**espère que** tout ira bien »*
- **Surprise :** Trois types de surprise sont différenciés : positive (*« Quelle **belle surprise** ! »*), négative (*« **Oula !** ça **ne va pas** là ! »*) et neutre (++++).

Les entités mettant en œuvre ces cinq axes sont extraites grâce à des patrons textuels selon la technologie *Text Mining Agent* (TMA) mise au point par la société Sinequa[2]. Cette technologie propose de construire des patrons textuels à l'aide de critères pouvant intervenir à plusieurs niveaux (expressions régulières, lemmes, catégories morpho-syntaxiques). Ces patrons sont représentés par des transducteurs à l'image d'outils comme Unitex (Paumier 2002).

Les scores de polarité (positifs/négatifs) sont calculés à partir des cinq axes que nous avons définis ci-dessus et de leur densité d'apparition pour une fenêtre temporelle donnée. Dans cette méthode

l'une des hypothèses formulées est qu'une présence importante d'une ou plusieurs modalités sur un passage de conversation donné indique des passages de dialogues intéressants (émotionnels). Afin d'éviter des pics isolés pour une modalité ou une autre, une fenêtre temporelle correspondant à 5 segments (2 segments précédents le segment courant, le segment courant et 2 segments suivants le segment courant) est prise en compte dans le calcul. Le Tableau 6-6 ci-dessous présente les poids associés à chacune des modalités des cinq axes :

		Polarité	Poids non émotif	Poids émotif
Appréciation	Favorable	positive	2	4
	Défavorable	négative	2	4
Acceptation/refus	Acceptation	positive	2	4
	Refus	négative	2	4
Accord/désaccord	Accord total	positive	2	4
	Accord approx.	positive	1	2
	Rectificatif	négative	1	2
	Désaccord	négative	2	4
Opinion	(tous)	neutre	1	2
Surprise	Positif	positive	2	4
	Négatif	négative	2	4

Tableau 6-6 Poids attribués aux différentes modalités retenues pour le calcul de la polarité des segments (Cailliau and Cavet 2010)

Chaque segment se voit attribuer ainsi un score de polarité (positif et négatif) qui est obtenu d'après la somme des poids des entités apparaissant dans la fenêtre temporelle examinée. Dans cette méthode le poids cumulé des entités neutres est attribué au score positif ou négatif le plus élevé pour l'unité considérée.

6.3.1.2 Descripteurs sémantiques issus d'analyses à base de cartouches de connaissances (skill cartridge [9])

Les cartouches de connaissances utilisées ici se basent sur les cadres définis dans la théorie de l'évaluation (appraisal theory) tels que décrits dans (Martin and White 2005; Bloom et al. 2007).

Nous pouvons schématiquement représenter l'adaptation de la théorie de l'évaluation telle que définie dans le chapitre 3 par la Figure 6-1 ci-dessous :

[9] « Skill Cartridge » est une marque déposée par la société TEMIS (http://www.temis.com)

[2] Plus d'informations sur la société Sinequa : http://www.sinequa.com/

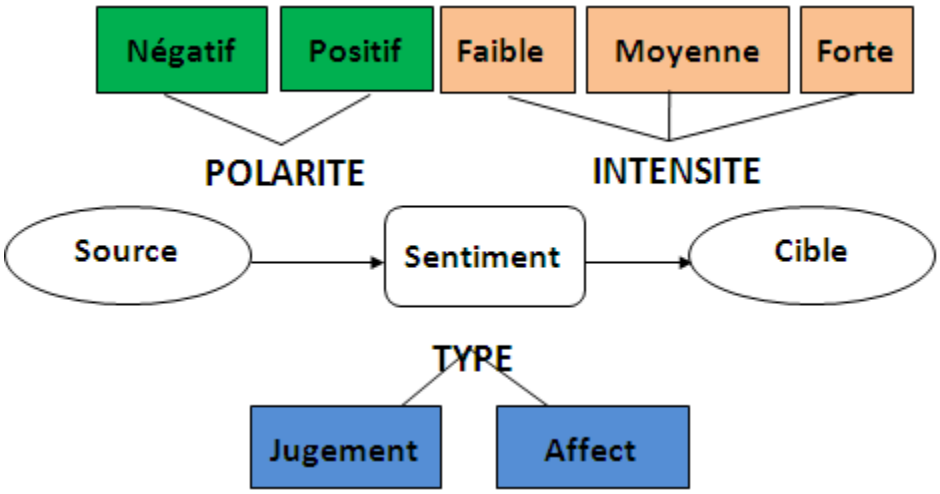


Figure 6-1 Représentation de la théorie de l'évaluation selon (Martin and White 2005)

Les auteurs définissent ainsi une expression évaluative suivant trois points distincts : une source (l'objet de l'évaluation), une évaluation (l'état affectif), une cible. Chaque expression peut être affectée par une polarité de type négatif ou positif, une intensité (faible, moyenne, forte) et un type (pouvant être un jugement, un affect, etc).

Concrètement, la cartouche de connaissances s'appuie sur les éléments suivants pour pouvoir établir une relation entre deux concepts et la polarité associée :

1) Un **lexique de termes et d'expressions évaluatives**, contenant des informations sur la polarité (*negative* ou *positive*), l'intensité (*low* ou *strong*) et le(s) type(s) d'opinion de chacune des entrées (*jugement de personne, jugement de produit, jugement de situation, affect*). Dans ce lexique sont retenus les mots appartenant aux catégories grammaticales recommandées dans (Turney 2002) à savoir les verbes, adverbes, adjectifs, noms auxquels viennent s'ajouter les pronoms ainsi que les interjections. Des expressions évaluatives sont également stockées dans le lexique. Ces expressions peuvent être représentées sous forme d'expressions figées ou de patrons syntaxiques tels que définis dans le Tableau 6-7 ci-dessous :

Expression figée	Patron	Polarité	Enoncés dérivés auxquels le patron s'applique (exemple)
Ne pas valoir la peine	valoir / **(pas|jamais)** / #DET_SG / (peine|détour|coup)	Négative	Ça ne vaut jamais le coup Il ne vaut pas la peine
Valoir la peine	(valoir / #DET_SG|valoir / **!(pas|jamais)*** / #DET_SG) / (peine|détour|coup)	Positive	Ça vaut le coup il vaut vraiment le détour

Tableau 6-7 Exemple de patrons dérivés d'expressions figées et leurs énoncés dérivés associés

2) Un **lexique métier**, contenant les cibles potentielles d'opinion, catégorisées par type *(personne, produit, situation) et* identifiées comme pertinentes dans le cadre de l'analyse de conversations naturelles en centre d'appels ;

3) Des règles d'extraction, décrivant la syntaxe des expressions évaluatives pour former des relations d'opinion. Les règles s'appuient sur des patrons morpho-syntaxiques constitués à partir de l'étude du corpus et des transcriptions orales et qui sont regroupés par syntagmes selon leur comportement syntaxique. Le principe appliqué est que tout élément modifié par un terme ou une expression issu du lexique évaluatif fait partie d'un syntagme à extraire.
Les patrons dits **« certains/safe** » sont des syntagmes comportant au moins un terme issu du lexique évaluatif. Ils constituent le sentiment de l'expression évaluative. Les patrons dits **« incertains/unsafe** » sont des syntagmes ne comportant pas de terme ayant une polarité positive ou négative, c'est-à-dire comportant les termes neutres du lexique évaluatif, ceux issus du lexique du domaine et/ou ceux issus de lexiques annexes. Lorsque ces patrons représentent des syntagmes adverbiaux (i.e. « *très* ») ou adjectivaux (i.e. « *dernier* »), ils sont utilisés pour constituer des syntagmes adjectivaux, nominaux et verbaux plus grands (i.e. « *le dernier numéro de téléphone* »), les syntagmes nominaux et verbaux constituant généralement le contexte ou la cible de l'expression évaluative. Ces patrons peuvent également être associés à des patrons « **certains/safe** » pour former des patrons « **certains/safe** » plus grands.
Afin de constituer des segments de phrases, par l'ajout de syntagmes de plus en plus grands, les règles sont organisées de façon hiérarchique :

- Au premier niveau sont formés les patrons lexicaux
- À un second niveau sont formés les patrons syntagmatiques
- Enfin, à un troisième niveau sont définies les règles d'extraction d'opinion

4) Un composant sémantique calcule pour chaque relation sa polarité, son intensité et son type d'opinion à partir des informations des lexiques et de la fonction des éléments dans les relations d'opinion. Pour cela, le composant sémantique manipule et interprète les rôles de la relation d'opinion (obtenus d'après les règles exprimées dans l'étape précédente) et les attributs issus du lexique dans l'étape précédente. Durant cette étape deux tâches sont principalement effectuées:

- Création de nouvelles relations typées [Jugement] ou [Affect]
    - o Le calcul du type d'opinion (affect ou jugement)
    - o La détermination de la cible (personne, produit ou situation).
- Puis, pour chaque relation typée,
    - o Calcul de sa polarité (négative, positive ou neutre) obtenue d'après les patrons dérivés d'expressions figées et des entités constituées dans l'étape 3
    - o Calcul de son intensité (forte, normale ou faible). La présence « modifieurs » de type « trop », « très » déterminent l'intensité de la relation.

En sortie l'outil fournit des portions de dialogues de mots à mots ou des types d'entité sont caractérisés à l'aide des différentes étapes décrites ci-dessus. L'expression ci-dessous illustre comment est représenté une sortie de l'outil :

```
<Event name="/Relationship/Sentiment/Analysis/Jugement/Negative/Person/Interlocuteur"
    sid="w411" eid="w414" text="vous criez comme ça" source="auto"/>
```

6.3.2 Fusion des indices paralinguistiques et lexicaux : stratégies et résultats

Nous prendrons comme score de référence le F-score obtenu d'après les indices paralinguistiques et qui a été présenté dans le paragraphe 6.2.2.2.

Comme nous avons pu le constater, un modèle se basant sur un canal unique, qu'il soit acoustique ou lexical, offre des performances modérées vue la complexité des données utilisées. Nous proposons pour cela une fusion des différents descripteurs. Comme il est décrit dans le chapitre 3, deux types de fusion sont principalement utilisés : fusion au niveau des descripteurs et fusion au niveau de la décision.

Protocole de test : Les tests ont été réalisés en nous basant sur une situation correspondant à l'expérience 2 du paragraphe 6.2.2.2 à savoir un jeu de données indépendant du locuteur mais basé sur une segmentation manuelle utilisant le protocole leave one dialog out. Dans un premier temps, nous avons effectué les tests en fusionnant les indices acoustiques et lexicaux d'après les paradigmes de fusion au niveau du descripteur et dans un second temps, au niveau de la décision. Nous avons utilisé pour la fusion au niveau des descripteurs une classification à base de SVM, jugée plus robuste pour la résolution de problèmes impliquant un grand nombre de dimensions. Concernant la fusion au niveau de la décision, nous avons utilisé des classifieurs SVM pour la construction des modèles acoustiques et lexicaux et un classifieur « naïve bayes » pour la prise de décision finale.

Les résultats sont présentés dans le Tableau 6-8 ci-dessous :

Type de fusion	F-score (intervalle de confiance a 0.95)	Gain par rapport au score de référence
Score de référence (paralinguistique seulement)	0.59 (±5%)	-
Au niveau des descripteurs	0.65 (±6%)	+7%
Au niveau de la décision	0.68 (±2%)	+9%

Tableau 6-8 Fusion des indices lexicaux et paralinguistiques

Nous pouvons voir dans chacun des deux cas une augmentation sensible des F-scores avec une amélioration de +7% dans la situation la moins favorable et de +9% dans le meilleur des cas. Outre le fait de présenter des scores plus élevés, la méthode de fusion au niveau de la décision (fusion tardive) présente également l'avantage de ne pas avoir besoin de travailler sur un nombre de dimensions trop élevé. Au moment de la décision, nous avons en effet un problème de dimension n=2 à résoudre contre un problème de dimension n=834 pour la fusion au niveau des descripteurs (fusion précoce) ce qui augmente considérablement le besoin en temps de calcul et en mémoire nécessaire. Si nous observons dans le détail les scores pour chaque émotion nous voyons également que par rapport à la situation de base, nous obtenons des gains pour chacune des 3 classes d'émotions considérées comme le montre le tableau ci-dessous :

	Positif F-score (gain)	Neutre F-score (gain)	Négatif F-score (gain)
Score de référence : Acoustique seulement	0.54 (-)	0.59 (-)	0.63 (-)
Fusion au niveau des descripteurs	0.60 (+0.06)	0.66 (+0.07)	**0.68 (+0.05)**
Fusion au niveau de la décision	**0.64 (+0.10)**	**0.74 (+0.15)**	0.66 (0.03)

Tableau 6-9 Gains obtenus pour chaque classe d'émotion en combinant les indices acoustiques et lexicaux

Pour chacune des classes, un gain est constaté lorsque l'une ou l'autre des méthodes de fusion est adoptée. Le gain le plus important concerne l'état « neutre » qui bénéficie d'une augmentation de 0.15 dans le cas d'une fusion au niveau de la décision. Les états affectifs négatifs et positifs bénéficient respectivement de gains de +0.10 et +0.15 lorsque l'on considère les situations les plus favorables pour chacun de ces deux états. Etant donnés les gains en termes de détection correcte et de complexité en temps et en espace de recherche, nous conserverons uniquement la fusion tardive pour la suite de l'étude.

6.3.3 Fusion des indices paralinguistiques, lexicaux, syntaxiques et sémantiques

Considérant les gains obtenus lors des premiers tests de fusion présentés dans le paragraphe ci-dessous, nous avons souhaité poursuivre l'expérience en enrichissant nos modèles à l'aide de descripteurs sémantiques de l'étude (Cailliau and Cavet 2010) et sémantiques (approche TEMIS). Nous essayons par le biais de cette méthode de nous rapprocher des indices utilisés par un humain lors d'une conversation. Pour cette expérience nous utilisons le paradigme de fusion dite « tardive » (au niveau de la décision) dans laquelle nous combinons les sorties obtenus auprès de chaque classifieur.

Nous avons vu qu'une combinaison simple paralinguistique-lexicale permettait d'obtenir une amélioration intéressante des F-scores. Nous avons également voulu effectuer l'expérience en combinant l'ensemble des indices linguistiques à notre disposition (lexicaux, syntaxique, sémantique). Le score de référence retenu ici est celui que nous avons obtenu avec l'utilisation d'indices lexicaux. Les résultats sont fournis dans le Tableau 6-10 ci-dessous :

	Positif	Neutre	Négatif	Global
Score de référence Indices lexicaux	0.50	0.58	0.61	0.55
Fusion des indices linguistiques	0.51	0.58	0.737	**0.61**

Tableau 6-10 Résultats Gains obtenus après fusion des indices linguistiques (lexicaux, syntaxiques, sémantiques)

La principale amélioration obtenue par rapport aux scores utilisant une seule source d'informations concerne les émotions négatives (+0.12), les émotions neutres et positives restant stables.

Etant donné, la faible amélioration obtenue pour 2 classes d'émotions sur 3, nous avons souhaité tester la complémentarité des indices linguistiques avec les indices paralinguistiques. Nous avons ainsi successivement combiné les 384 indices acoustiques avec : i) les indices syntaxico-sémantique SINEQUA et ii) les indices sémantiques TEMIS. Le score de référence est ici celui obtenu grâce à la fusion paralinguistique/lexicale que nous avons présentée précédemment :

	Positif	Négatif	Neutre	Global
Score de référence Acoustique/Lexical	**0.64**	**0.74**	0.66	**0.68**
Acoustique/syntaxico-sémantique	0.57	0.67	0.66	0.64
Acoustiques/sémantique	0.54	0.66	**0.69**	0.64

Tableau 6-11 Scores obtenus en fonction des différentes combinaisons acoustiques/linguistiques

De manière globale, comme nous le voyons dans le Tableau 6-11 ci-dessus, la combinaison acoustique/lexicale semble être la plus performante. Localement, l'analyse sémantique donne de bons résultats sur le neutre mais présente en contrepartie des résultats en retrait sur les émotions positives. Même si les performances des indices syntaxiques et sémantiques sont en retrait par rapport à la combinaison acoustique/lexicale, nous pouvons cependant noter qu'il arrive qu'une émotion soit localement mieux reconnue si l'on considère un type d'indices en particulier, c'est le cas du neutre et de la combinaison acoustique/sémantique par exemple. Cette constatation laisse ainsi à penser qu'une complémentarité entre les indices est possible. Nous avons donc, pour l'expérience suivante, combiné l'ensemble des indices à notre disposition d'après un paradigme de fusion tardive. Nous retiendrons comme scores de référence respectivement les résultats obtenus grâce aux indices paralinguistiques seuls ainsi que ceux obtenus grâce à la combinaison paralinguistique/lexicale. Les nouveaux scores seront de ce fait comparés aux résultats les plus prometteurs que nous ayons pour le moment présentés.

Les résultats sont présentés dans le tableau ci-dessous :

	Positif	Négatif	Neutre	Global
Score de référence 1 : Acoustique seulement	0.54	0.63	0.59	0.59
Score de référence 2 : Acoustique/Lexical	0.64	**0.74**	0.66	0.68
Fusion : Acoustique/lexicale/syntaxique/sémantique	**0.67**	0.72	**0.75**	**0.71**

Tableau 6-12 Scores obtenus suite à une fusion des descripteurs acoustiques et de l'ensemble des descripteurs linguistiques

Dans chacun des cas, le score de référence 1 est dépassé avec un gain global de +13% en faveur de la fusion. La différence obtenue entre le score de référence 2 et celui de la fusion est quant à lui moins net. Si le gain global est non négligeable, nous observons cependant qu'il n'est pas systématique pour chacune des classes émotionnelles. Le gain le plus important est reçu par la classe neutre (+9%). La classe positive est également améliorée mais dans une moindre mesure (+3%). Au final nous constatons une amélioration globale de +3% par rapport au score de référence 2. L'augmentation modérée que nous obtenons lors la fusion de l'ensemble des indices peut être expliquée par le fait que les descripteurs sémantiques et lexicaux n'ont pas été annotés suivant le même protocole que le nôtre. Ainsi par exemple, les prédictions fournies par le modèle syntaxique sont issues d'une analyse portant sur une fenêtre temporelle de 5 segments dans l'optique de fournir une orientation positive, négative ou neutre à l'échelle d'un dialogue complet, ce qui n'est pas notre cas (nous n'effectuons pas de pondération ou de lissage entre des segments consécutifs présents dans notre corpus).

Même si des améliorations restent à apporter afin d'homogénéiser les performances à travers les différentes classes, nous pouvons cependant considérer cette approche à quatre niveaux comme prometteuse. Les gains par rapport au score de base, obtenus à partir des modèles paralinguistiques et lexicaux seuls, sont notables, passant d'un f-score de 0.59 pour les indices paralinguistiques et de 0.55 pour les indices lexicaux, à un score global final de 0.71. Dans cette dernière situation, le taux d'accord mesuré par le Kappa est de $k = 0.56$, ce qui représente un taux satisfaisant compte tenu de la nature des données.

6.3.4 Test de la combinaison sur un corpus étendu

A la vue de l'amélioration globale que nous avons pu constater lors de la combinaison des indices et dans le but de se placer dans une situation plus réaliste, nous avons étendu le test de notre modèle à une sélection de 264 dialogues supplémentaires que nous avons utilisés comme ensemble de test indépendant. Cette sélection représente plus de 49 heures de parole découpées en 29683 segments contenant au total plus de 270 locuteurs différents et indépendants de ceux de l'ensemble d'entrainement. Au regard de l'important volume de données analysé dans cette partie de l'étude, l'ensemble de test n'a pas pu être annoté de manière fine. Les résultats obtenus ne pourront donc être évalués que d'après des métriques définies par un sous-ensemble qui a été analysé (annoté) de manière précise.

Nous avons utilisé le modèle composé de l'ensemble des indices acoustiques et linguistiques (lexicaux, syntaxiques et sémantiques) et l'avons lancé sur l'ensemble des 29683 segments. A titre de comparaison, nous avons également effectué les tests sur les modèles paralinguistiques et paralinguistiques/lexicaux. La répartition des prédictions fournies par le modèle est présentée dans le Tableau 6-13 ci-dessous :

Répartition des segments Pos, Neg et Neut sur **264 dialogues complets**	Modèle Paralinguistique	Modèle Paralinguistique/Lexical	Modèle Paralinguistique/Lexical /Syntaxique/Sémantique
Pos/Neut/Neg	29%/31%/40%	25%/39%/36%	24%/41%/35%

Tableau 6-13 Répartition des prédictions sur une sélection de 264 dialogues complets et non annotés issus du corpus

Pour comparaison, nous fournissons dans le tableau ci-dessous la répartition des émotions d'après les annotations effectuées sur les corpus « colère, joie » segmentés à l'aide des time-code fournis par le système de reconnaissance de la parole.

Répartition après annotation	Positif	Neutre	Négatif
Corpus Colère	1%	52%	46%
Corpus Joie	18%	76%	6%
Moyenne	9.5%	64%	26%

Tableau 6-14 Répartition des classes émotionnelles au sein des corpus « colère » et « joie »

La première constatation que nous pouvons porte sur la représentation des classes émotionnelles par rapport à la classe « Neutre » dans les prédictions fournies par les trois modèles par rapport à la répartition des annotations sur le corpus d'apprentissage qui est certes pas le même corpus. Si nous considérons les résultats fournis par les modèles, nous obtenons en moyenne une classe neutre représentée à 37%, ce qui est largement en dessous des tendances que nous avons lorsque nous observons les résultats fournis d'après les annotations. Il nous faut de plus faire remarquer que les corpus sélectionnés pour l'annotation l'ont été pour la présence d'expressions affectives dans un volume plus important que la normale, au cours de la conversation qui s'y déroule. Même si nous n'avons pas d'évaluation précise du fait que nous n'avons pas annoté plus de dialogues, nous avons estimé la présence de contenu émotionnel à hauteur de 15% à un niveau global au cours de dialogues sélectionnés au hasard, ce qui est très largement inférieur aux prédictions de nos modèles.

Il faut cependant pondérer cette considération en mettant en avant l'ambiguïté perceptive qu'il existe et qui est souvent un problème lors de l'annotation de contenus émotionnels. Nous avons pour illustrer cela sélectionné au hasard un dialogue parmi les 264 analysés. Ce dialogue a, part la suite, été annoté selon la procédure décrite dans le chapitre 4. Les résultats de cette annotation sont présentés dans le schéma ci-dessous :

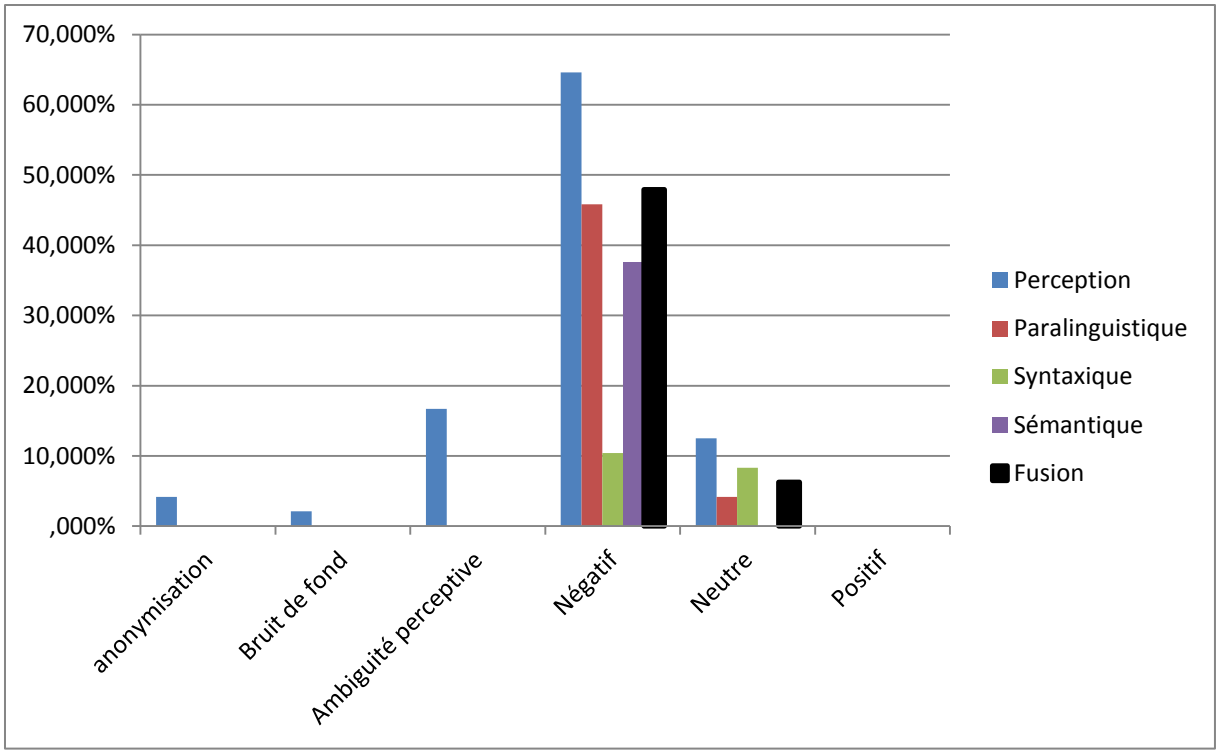


Figure 6-2 Proportion de segments émotionnels, bruités et non consensuels d'un dialogue complet sélectionné au hasard

L'axe des ordonnées représente la proportion en pourcentage d'un type de segments. Nous voyons par exemple pour ce dialogue que 23% des segments ne sont pas utilisables ou le sont difficilement pour diverses raisons (procédure d'anonymisation afin de masquer les noms et numéros de cartes de crédit par exemple, bruits de fond gênant perceptivement l'identification de l'état affectif et le désaccord des deux juges sur la valence d'un segment donné). Nous pouvons noter que le modèle fusion est dans chaque situation plus proche de l'évaluation perceptive que les modèles paralinguistiques et ce, malgré les difficultés mentionnées ci-dessus, ce qui est encourageant. Nous pensons cependant que le faible volume de données annotées (environ 1200 instances pour chaque classe émotionnelle) nuit à la qualité des prédictions fournies du fait d'un pouvoir de généralisation limité. Etant donné le grand volume de temps de parole mis à notre disposition, il nous semble intéressant de nous pencher sur un modèle de sélection automatique des données au sein de notre corpus pour enrichir nos modèles. Une méthode sera ainsi proposée dans le chapitre 7.

6.4 Conclusion

Ce chapitre a été l'occasion de présenter plusieurs difficultés que nous pouvons rencontrer lorsque nous proposons de traiter des données naturelles de façon complètement automatique. Nous avons présenté les résultats obtenus sur différents canaux de communication (linguistiques et paralinguistiques) en faisant varier les conditions de test (partant de conditions artificielles avec des prétraitements manuels pour finir avec des tests ne prenant en compte que des données obtenues de manière automatique). Nous avons dans ce dernier cas constaté des baisses de résultats importantes. Pour essayer d'apporter une réponse à ce problème, nous avons combiné des indices à plusieurs niveaux linguistiques (syntaxiques, lexicaux, sémantiques) et paralinguistiques, dans le but d'aborder le problème de détection d'une façon plus globale. Les résultats obtenus montrent une

amélioration des résultats de 13% ou plus pouvant être obtenus si nous utilisons un modèle fusionné au lieu de modèles linguistiques et paralinguistiques évalués indépendamment. Lorsque nous avons étendu le test des modèles à une sélection plus large de 55 heures de paroles (264 dialogues), nous avons constaté une surreprésentation des phénomènes émotionnels dans les prédictions montrant que, malgré des améliorations notables, l'utilisation de nos modèles dans des contextes d'exploitation totalement réels peut sembler difficile. Par ailleurs nous pointions également le problème des volumes de données disponibles pour la phase d'entrainement. Nous disposons de 13 heures annotées sur 1620, ce qui nous semble faible, surtout lorsque nous commençons à essayer des volumes de données de plusieurs centaines de locuteurs. Considérant que nous disposons d'une base de données d'environ 10.000 appels et 1620 h, nous proposerons donc, dans le chapitre 7, une méthode permettant d'enrichir les modèles existants dans le but d'obtenir un pouvoir de généralisation des modèles plus important que celui dont nous disposons actuellement.

Chapitre 7 Enrichissement automatique des modèles paralinguistiques

Résumé

Nous abordons dans ce chapitre le thème de l'enrichissement automatique de modèles paralinguistiques.

Nous avons vu dans le chapitre précédent que des gains pouvaient être obtenus lorsque nous fusionnions des indices acoustiques et linguistiques. Cependant, nous n'avions que peu d'instances dans l'ensemble d'apprentissage, ce qui limitait le pouvoir de généralisation de nos modèles lorsque nous appliquions la solution proposée à des données obtenues de manière complètement automatisée. Dans ce chapitre, nous conservons l'idée d'une approche mixte acoustique/linguistique, mais appliquée cette fois-ci à une sélection d'instances qui auront pour but de rejoindre l'ensemble d'entrainement afin d'augmenter la taille de nos modèles. Le processus mis en œuvre propose dans un premier temps une sélection à l'aide du modèle linguistique décrit dans (Cailliau and Cavet 2010). Dans un second temps, il est effectué une nouvelle sélection dans ce sous ensemble grâce à un modèle détectant le degré d'activation.

Nous avons effectué les tests en utilisant jusqu'à 100h de conversation supplémentaires. Les résultats montrent des gains sur chacune des 3 classes considérées (positive, neutre, négative). Nous obtenons globalement jusqu'à 6% de gain au niveau global ; 2% sur la classe positive ; 7% sur la classe négative et 11% sur la classe neutre. Nous avons décliné les modèles obtenus suivant différentes modalités afin de vérifier la robustesse des modèles. Nous montrons que ces gains restent stables quelle que soit la segmentation considérée (manuelle ou automatique).

7.1 Introduction

Comme nous l'avons vu dans le chapitre précédent des améliorations notables de performances de détection sont possibles lorsque nous combinons des indices intervenant à différents niveaux entre eux. Cependant nous faisions également remarquer que lorsque nous utilisons les modèles d'une manière réaliste, c'est-à-dire à l'aveugle sur des dialogues complets (pouvant contenir bruits, paroles superposées, processus d'anonymisation) et sur un volume de données important, le corpus que nous avons annoté s'avérait mettre en œuvre un pouvoir de généralisation insuffisant compte tenu de la très grande variabilité du nombre de locuteurs au sein du corpus (1620 heures de parole et 10 000 locuteurs différents). Pour palier ce problème, l'une des solutions est d'augmenter la taille du corpus d'entrainement. L'augmentation du volume d'un corpus d'entrainement passe traditionnellement par l'annotation de données supplémentaires ce qui représente un coût en temps et en argent non négligeable. Considérant ce problème, nous avons décidé de nous tourner vers des méthodes semi-supervisées afin de pouvoir augmenter à moindre coût la taille de notre corpus.

7.2 Apprentissage semi-supervisé et sélection active de classes (active class selection) pour l'entrainement de données sur de grands corpus

Les méthodes d'apprentissage semi supervisées (*semi supervised learning*) ou de sélection de classes actives (*active class selection*) sont pour le moment très peu utilisées dans le domaine de la reconnaissance des émotions dans la voix. Ce constat est principalement dû à la taille réduite des corpus, comme nous le mentionnions dans le chapitre 2. Quelques études comme (Mahdhaoui 2010) ou (Schuller et al. 2011) proposent de telles méthodes dans le but d'enrichir automatiquement des modèles préexistants. Certaines études comme celle proposée par Schuller (Schuller et al. 2006) introduisent la génération automatique de descripteurs à partir de donnée existantes dans le but d'enrichir les modèles avec des informations élargissant l'espace de recherche. Nous allons présenter dans les paragraphes qui suivent quelques méthodes d'apprentissage semi supervisées permettant de sélectionner et d'incorporer de manière automatique des instances nouvelles dans des modèles déjà existants.

7.2.1 Le self training

L'algorithme *«self training »* a été exposé notamment dans (Zhu 2006). Sa mise en œuvre très simple a rendu cette méthode très populaire notamment dans le domaine du traitement automatique des langues. Celle-ci se base sur un ensemble de données étiquetées et sur un ensemble de données non étiquetées. Cette approche propose d'entrainer un algorithme de manière classique, à l'aide des données étiquetées, puis, une fois le modèle construit, ce dernier sera utilisé pour fournir une prédiction sur les instances non étiquetées. Les N % d'instances qui auront été étiquetées avec le plus grand degré de confiance seront sélectionnées et insérées dans l'ensemble d'entrainement aux cotés des données étiquetées. Par la suite, ce processus est répété jusqu'à ce qu'il n'y ait plus de données non étiquetées ou que le degré de confiance soit inférieur au seuil fixé.

L'algorithme de *self learning* est présenté ci-dessous :

Début :
Ensemble L des données étiquetées
Ensemble U des données non étiquetées
Un nombre n des données à ajouter à l'ensemble L à chaque itération
Tant que U !=
Apprendre un classifieur h sur l'ensemble L
Etiqueter l'ensemble U en utilisant le classifieur h
T est l'ensemble des n exemples de U prédits avec la meilleure confiance par h
Ajouter T à L
Supprimer T de U
Fin

Tableau 7-1 L'algorithme self training

Cette méthode a été appliquée dans (Lomsadze et al. 2005) pour l'identification de gènes ou en traitement automatique des langues pour des tâches de désambigüisation comme dans (Yarowsky 1995; Mihalcea 2004).

7.2.2 Le co-apprentissage (co-training)

La méthode dite de co-apprentissage ou *co-training* à été initialement proposée par (Blum and Mitchell 1998). Cette méthode suppose que le problème d'apprentissage considéré peut être décomposé en au moins deux ensembles indépendants permettant de décrire de manière complémentaire le problème étudié. Le but est alors d'entrainer chacun de ces sous-ensembles indépendamment sur des représentations du modèle différentes. A terme l'idée est d'amener chacun des classifieurs à s'améliorer à l'aide d'un important volume de données non étiquetées. Chacun des classifieurs classifie les données non étiquetées et sélectionne un petit nombre d'instances avec le score de confiance le plus élevé dans le but d'enrichir les modèles de l'autre classifieur. Cette méthode propose d'amener chacun des modèles à des prédictions homogènes aussi bien sur les données non étiquetées que sur les données étiquetées. L'algorithme tiré de (Blum and Mitchell 1998) est présenté ci-dessous :

Début

Ensemble L des données étiquetées

Ensemble U des données non étiquetées

Tant que U != ;

Construction d'un sous ensemble U' en choisissant *u* exemples aléatoirement à partir de U

Apprendre un classifieur c_1 sur l'ensemble L

Apprendre un classifieur c_2 sur l'ensemble L

Etiqueter un nombre p des exemples de l'ensemble U' en utilisant le classifieur c_1

Etiqueter un nombre p des exemples de l'ensemble U' en utilisant le classifieur c_2

Ajouter l'ensemble T des exemples étiquetés par h_1 et h_2 à l'ensemble L

Supprimer T de U'

Fin

Tableau 7-2 Algorithme de co-apprentissage

L'algorithme de co-training à été utilisé avec succès dans (Mahdhaoui 2010) où l'auteur entraine séparément différents types de descripteurs acoustiques (cepstraux, prosodiques et perceptifs) et obtient des gains pouvant aller jusqu'à +10% de reconnaissance dans la situation la plus favorable.

7.3 Enrichissement de corpus : la solution proposée

7.3.1 Nombre d'instances dans l'ensemble d'entrainement : variation des résultats

Afin d'introduire le problème que nous souhaitons analyser, nous présentons dans cette première expérience les variations de résultats que nous obtenons en fonction du nombre d'instances sélectionnées par classe émotionnelle. Nous prendrons comme score de référence les résultats obtenus dans le chapitre 5 à partir des indices paralinguistiques. Pour rappel, le corpus était composé de 3684 instances réparties sur 3 classes d'émotions (1235 positif, 1182 neutre, 1267 négatif). Un classifieur SVM avec un noyau RBF a été utilisé après une étape d'optimisation des paramètres réalisée d'après l'algorithme « gridsearch » implémenté dans libSVM (Chih-Chung et al. 2001). Le test s'est déroulé en 3 phases. La première contenant l'ensemble d'entrainement complet (score de référence), le second contenant 50% des instances dans chaque classe et le troisième contenant 25% des instances dans chaque classe. Pour chacune de ces trois étapes, le jeu d'entrainement n'a pas varié. La méthode de test utilisée est le Leave One Dialog Out permettant l'utilisation d'ensembles d'entrainement et de test de manière complètement indépendante du locuteur. Les résultats sont donnés dans le Tableau 7-3 ci-dessous :

	Positif	Négatif	Neutre	Global (F-score)
Score de référence : 3684 instances	0.54	0.59	0.63	0.59
1842 instances	0.52 (-2%)	0.53 (-4%)	0.59 (-4%)	0.55 (-4%)
921 instances	0.50 (-4%)	0.51 (-8%)	0.59 (-4%)	0.53 (-6%)

Tableau 7-3 Evolution des F-score en fonction du nombre d'instances présentes dans l'ensemble d'entrainement

Comme prévu, le score de référence contenant 3684 instances obtient le résultat le plus élevé. Les dégradations sont légères mais progressives au fur et à mesure que le nombre d'instances prises en compte pour l'entrainement diminue. Nous observons ainsi une baisse globale des résultats de -4% lorsque nous retirons 50% du nombre d'instances de l'ensemble d'entrainement et -6% lorsque 75% des instances sont retirées. La classe la plus impactée par ce retrait est la classe négative qui obtient -8% lorsque nous comparons la situation prenant en compte les 3684 instances et celle n'en prenant en compte que 921. Ce recul peut être expliqué par le fait que la classe négative est dans notre étude la classe la plus large. Pour rappel, nous incluons sous l'étiquette « négatif » des états émotionnels aussi variés que la colère (froide et chaude), l'agacement, la surprise négative, l'impatience. Partant de ce constat, les instances sélectionnées se doivent d'être représentatives de l'ensemble de la classe afin de fournir au modèle un pouvoir de généralisation suffisamment fort. A l'inverse, la classe neutre par exemple est moins impactée par ce changement du fait d'une expression en moyenne plus homogène.

Considérant qu'un nombre plus important d'instances en apprentissage aurait pour effet d'augmenter mécaniquement les performances des modèles en fournissant un pouvoir de généralisation plus élevé, notamment lorsque nous utilisons les modèles sur des corpus de grande taille, notre but sera dans ce chapitre de sélectionner de nouvelles instances de manière automatique et de les incorporer au modèle existant.

7.4 La sélection de données pour l'enrichissement de modèle

Afin d'enrichir notre ensemble d'entrainement composé de 3684 instances nous utilisons une sélection de 264 dialogues supplémentaires. Cette sélection de 50 heures de conversations supplémentaires contient plus de 270 locuteurs différents et est répartie sur 29683 segments dont les frontières ont été définies de manière automatique. La durée moyenne d'un segment est d'environ 6 secondes, mais il existe une très grande variabilité à ce niveau : la durée minimum d'un segment est d'environ une seconde, à l'opposé, les segments les plus longs peuvent durer plus de 2 minutes. Les dialogues retenus sont ceux ayant obtenu le score émotionnel positif ou négatif le plus élevé d'après la méthode syntaxico-sémantique que nous avons présentée dans les chapitres précédents.

7.4.1 Protocole appliqué pour la sélection des instances supplémentaires

A la vue des résultats obtenus dans le chapitre précédent, une approche permettant de sélectionner les données suivant des critères acoustiques et linguistiques nous a semblé l'approche la plus prometteuse. Notre sélection s'est donc déroulée en quatre temps :

- Dans un premier temps nous avons utilisé la méthode syntaxico-sémantique présentée dans le chapitre 6. Nous avons sélectionné uniquement les segments ayant un score positif ou négatif supérieur à ou égal à 2.2 comme décrit dans (Cailliau and Cavet 2010).
- Dans un second temps nous avons construit et utilisé un modèle basé sur la dimension activation (arousal) telle que définie dans (Mehrabian 1998) dans le but de ne retenir que les segments considérés comme actifs. Ces derniers ayant plus de chances d'être porteurs d'une émotion (positive ou négative).
- Considérant la longueur extrêmement variable des segments sélectionnés, des redécoupages sont effectués. Les détails seront présentés dans les paragraphes ci-dessous.
- Les segments sélectionnés sont par la suite intégrés à l'ensemble d'apprentissage en veillant toutefois à conserver un équilibre entre les différentes classes émotionnelles considérées.

La figure ci-dessous récapitule les différentes étapes de sélection et d'intégration :

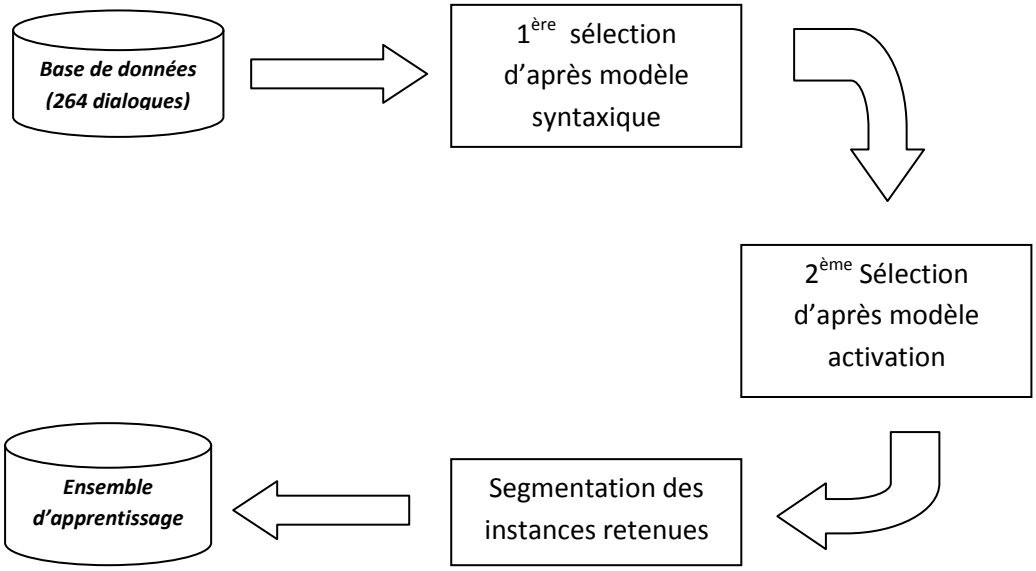


Figure 7-1 Processus de sélection de nouvelles instances pour l'enrichissement de l'ensemble d'apprentissage

La partie « segmentation des instances retenues » représentée sur la Figure 7-1 part du constat que la longueur des segments issus d'un découpage automatique par le système de reconnaissance de la parole est extrêmement variable et peut, dans certains cas, être supérieur à 2 minutes. Dans le cas de segments trop longs, les calculs des valeurs des différents descripteurs, et notamment des fonctionnels, se retrouvent biaisés du fait d'une variation émotionnelle pouvant être importante au cours du segment. Par définition, une émotion étant un évènement bref et ponctuel (Scherer and Peper 2001), son expression acoustique peut par conséquent être complètement noyée au milieu d'autres informations si l'unité observée est de longueur trop importante. Nous avons, pour tenter de palier ce problème, essayé de segmenter chaque instance avec des granularités temporelles différentes (3 secondes, 5 secondes, segments complets). Dans le cas de la segmentation de 3 et 5 secondes, nous avons essayé de ne conserver que les parties émotionnelles. La Figure 7-2 ci-dessous détaille le procédé :

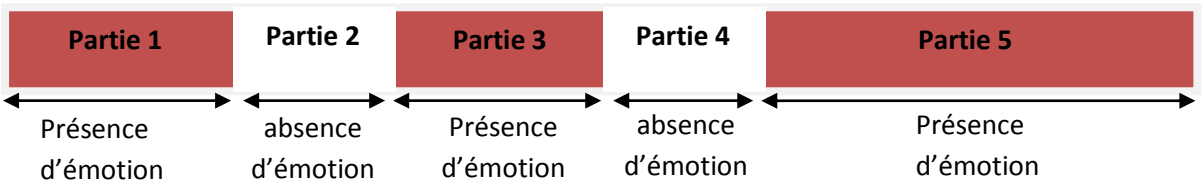


Figure 7-2 Exemple de découpage d'un segment long (d'une durée supérieure à 5 secondes)

Afin d'essayer de palier les problèmes mentionnés ci-dessus concernant la variabilité émotionnelle pouvant survenir au cours d'un segment long, nous avons, comme expliqué précédemment, découpé les segments en plusieurs sous parties de longueurs équivalentes. Nous avons par la suite éliminé les parties ne contenant potentiellement pas d'expression émotionnelle. Cela est représenté sur la Figure 7-2 par les parties blanches.

Comme pour le chapitre précédent, nous avons souhaité nous rapprocher le plus possible de conditions réelles d'utilisation du système de détection. Par conséquent, seule la phase d'annotation de l'ensemble d'entrainement et de test a nécessité une intervention manuelle. Le reste du processus ne met en jeu que des procédés automatiques.

7.4.2 Evaluation du modèle de sélection acoustique (activation)

Nous avons sélectionné, parmi les fichiers annotés, deux types d'instances : celles ayant un score d'activation égal à 2 (sur une échelle allant de -2 à 2) et celles ayant une activation égale à 0. Le but de cette sélection est d'appliquer un filtre sur les instances candidates proposées par le modèle syntaxique. Deux classes d'instances seront identifiées par le modèle. L'une représentant une activation qualifiée de « forte » et l'autre de « neutre ». Partant du constat que, dans des conditions réelles, l'état neutre représente plus de 80% des états affectifs exprimés, le but de ce modèle est de filtrer de manière plus fine les instances à intégrer dans notre ensemble d'entrainement et ainsi éviter de bruiter le modèle de détection.

Les instances des classes fortes et neutres sont réparties de la manière suivante :

Degré d'activation	Nombre d'occurrences	Emotions présentes
Forte	1107	Colère (74%), Négatif (20%), Positif (6%)
Basse	1038	Neutre (90%), Négatif (10%)

Tableau 7-4 Répartition des instances représentant des degrés d'activation faibles et forts ainsi que les émotions associées

Nous pouvons voir que les classes d'émotions relatives à des degrés forts d'activation contiennent principalement des émotions de type négatif (Colère, Irritation). Seulement 6% d'émotions relatives à la classe positive sont présentes dans la classe d'activation « forte ». Concernant la classe d'activation « basse » la classe émotionnelle majoritairement représentée le neutre (90% des instances de la classe sont neutres). Nous avons pris le parti, en rajoutant ce second filtre, de rechercher en priorité les émotions actives. Ce choix s'explique par le fait que traditionnellement les émotions les plus actives sont celles qui sont reconnues avec le taux d'accord le plus élevé par les annotateurs ; elles semblent donc être plus discriminantes. Aussi, étant donné la sensibilité des

modèles au bruit, nous prenons le parti de rajouter des émotions potentiellement plus prototypiques afin d'essayer de limiter le risque d'erreur.

Les performances de notre modèle d'activation ont été testées suivant le protocole « leave-one speaker-out » décrit dans le chapitre 5. Les scores obtenus en utilisant un classifieur SVM et noyau RBF (Radial Basis Function) sont présentés dans le Tableau 7-5 ci-dessous :

Classe	F-score (intervalle de confiance à 0.95)
Forte	0.84 (±2%)
Neutre	0.86 (±1%)

Tableau 7-5 F-score obtenu lors du test des modèles de détection de l'activation

Les résultats obtenus sont, pour chacune des deux classes, très nettement au-dessus du niveau du hasard et laissent penser que le critère actif/non actif peut être déterminé de manière fiable d'après une procédure automatique.

7.4.3 Sélection des instances destinées à enrichir les modèles

Nous avons appliqué la procédure de sélection de nouvelles instances telle que décrite dans le paragraphe 7.4.1. Nous y avons présenté l'ensemble d'instances retenues au fur et à mesure du processus de sélection. Pour rappel, l'ensemble des 264 dialogues dans lesquels nous avons sélectionné les nouvelles instances sont découpés de manière automatique en 29683 segments. Ces 264 dialogues ont été sélectionnés dans une base de données composée des 10 000 appels les plus marqués émotionnellement parlant. Le choix s'est effectué en sélectionnant les dialogues obtenant les scores de polarité négative ou positive les plus élevés parmi les dialogues disponibles. Ces scores ont été calculés en utilisant la méthode décrite dans (Cailliau and Cavet 2010)

A l'issue de cette étape, ***3593 segments*** sur 29683 sont retenus. Cette sélection implique que les 3593 segments ont un score supérieur ou égal à 2,2 en positif ou en négatif selon la méthode de calcul détaillée dans (Cailliau and Cavet 2010). Nous pouvons noter que sur les 3593 segments retenus, 10% (335 segments) ont un score supérieur ou égal au seuil de sélection en positif et négatif. Le ratio segments émotionnels vs non émotionnels est de 12% ce qui place le nombre de segments sélectionnés dans la moyenne généralement constatée lors de la recherche d'émotion dans des corpus naturels.

Les segments retenus sont par la suite filtrés une seconde fois à l'aide du modèle d'activation. Seuls les segments considérés comme actifs, plus à même de contenir des expressions émotionnelles, sont retenus. Ce sont au total **1914 segments** considérés comme actifs qui sont retenus, soit 53% de la sélection initiale. Cette sélection finale représente 417 minutes de temps de parole émotionnelle supplémentaire soit 276 minutes émotionnellement négatives et 141 minutes émotionnellement positives. La longueur moyenne des segments retenus est de 13 secondes (soit environ une durée cinq fois supérieure à des segments découpés manuellement).Ce sont au total 8 heures de paroles supplémentaires qui ont été sélectionnées pour enrichir nos modèles

Comme mentionné dans (Batliner et al. 2010) la segmentation des données est une étape essentielle pour la construction d'un système de détection des émotions. La longueur moyenne d'un segment découpé manuellement étant de 2.7 secondes, nous avons souhaité pouvoir observer les éventuelles

variations de résultats suivant différentes fenêtres temporelles. Trois niveaux de granularité temporelle ont donc été utilisés. Le premier utilise les segments dans leur ensemble ; le second se base sur des segments découpés par tranche de 5 secondes ; et le troisième est quant à lui segmenté en utilisant la longueur moyenne d'un segment découpé manuellement, soit 3 secondes. Nous récapitulons dans le Tableau 7-6 ci-dessous le nombre de segments contenus dans chaque classe suite au découpage et au filtrage à l'aide du modèle d'activation de ces derniers.

Segmentation	Nombre d'instance avant le passage du filtre d'activation	Nombre d'instance après le passage du filtre d'activation	Nombre d'instances positives supplémentaires	Nombre d'instances négatives supplémentaires
3 secondes	8580	3671		
5 secondes	8093	2246	1085	1161
Segments complets	3593	1914	648	1266

Tableau 7-6 Nombre d'instances selon le mode de segmentation et selon la classe émotionnelle

A noter que pour éviter tout biais des résultats avec un nombre de segments par classe très différent, nous avons, dans l'ensemble d'entrainement, équilibré les trois classes utilisées (positive, négative et neutre) en alignant le nombre d'instances retenues sur la classe la plus réduite. Les segments neutres ont été retenus parmi les segments ayant obtenu un score positif et négatif de 0. Leur nombre est également déterminé en fonction de la classe la plus réduite (la classe positive dans notre cas). Cette stratégie a également été utilisée dans (Batliner et al. 2006).

7.4.4 Premiers résultats

Nous avons calculé les premiers résultats en nous basant sur la segmentation de 5 secondes, 3 secondes ainsi que la prise en compte des segments complets. Notre ensemble de tests qui est commun à l'ensemble des configurations présentées suit le protocole « leave one dialog out » et contient des segments obtenus automatiquement. Nous prenons comme score de référence le F-score obtenu avec un jeu de tests indépendant du locuteur sur les classes de données positives, négatives et neutres tel que présenté dans le chapitre 6. Considérant la très grande variation de scores entre les différentes classes, nous avons choisi d'exprimer le score global en nous basant sur un F-score **non pondéré**. Les résultats sont présentés dans le Tableau 7-7 ci-dessous :

	Positif	Négatif	Neutre	Global *unweighted* F-score
Classification aléatoire[10]	0.08	0.32	0.60	0.33
Score de référence : Corpus non enrichi	**0.18**	0.53	0.60	0.44 (±2%)
Corpus Enrichi (segmentation de 3 secondes)	0.11	0.54	0.64	0.43 (±2%)
Corpus Enrichi (segmentation de 5 secondes)	0.16	**0.59**	**0.71**	**0.49 (±2%)**
Corpus Enrichi (segments complets)	0.15	0.51	0.65	0.43 (±2%)

Tableau 7-7 Scores obtenus après enrichissement de l'ensemble d'entrainement avec des instances annotées automatiquement

Les premiers résultats mettent en avant des résultats très variables selon les classes. La classe positive déjà très basse lorsque nous nous plaçons dans cette configuration, enregistre une baisse supplémentaire de 2%. Les classes négative et neutre sont, de leur côté, en hausse avec des gains de respectivement +6% et +11%. Pour essayer de comprendre cette variation entre les différentes classes, nous avons analysé de manière détaillée les instances qui ont été sélectionnées automatiquement.

7.4.4.1 Détail des instances sélectionnées

Contrairement aux résultats intéressants sur les classes neutre et négative, la classe positive est en retrait et enregistre même une baisse de performance. Pour tenter d'expliquer cela, nous avons commencé par rechercher si des phénomènes acoustiques traditionnellement perturbateurs (bruits, paroles superposées) étaient présents dans le jeu d'instances sélectionné de manière automatique.

Nous avons, pour cela, annoté les 2346 segments supplémentaires sélectionnés. D'après cette analyse, nous voyons que **7%** des segments sélectionnés contiennent des perturbations (musique, bruits divers d'arrière-plan) venant bruiter le signal. Par ailleurs, un taux de 52% de segments contenant de la parole superposée est également présent dans cette sélection. Ce taux très élevé de paroles superposées est causé par le mode de segmentation automatique qui a été utilisé. Comme constaté dans le chapitre 4, ce taux est presque doublé par rapport aux données segmentées manuellement. Cette superposition de locuteurs sur une même portion de signal peut venir fausser le calcul de paramètres très dépendants du locuteur, comme c'est le cas pour la fréquence fondamentale par exemple. Lors de cette phase d'annotation, nous avons également vérifié les classes émotionnelles que nous avons ajoutées à l'ensemble d'entrainement. La matrice de confusion ci-dessous présente les résultats de cette annotation :

[10] Le score par classification aléatoire est calculé selon la répartition de chaque classe émotionnelle sur l'ensemble du corpus. Il représente le score pouvant être atteint si les étiquettes étaient attribuées de manière totalement aléatoire à une instance donnée.

	Positif	Neutre	Négatif
Positif	125	431	529
Neutre	15	1080	10
Négatif	30	162	969

Tableau 7-8 Matrice de confusion représentant l'accord entre une sélection automatique et un annotateur humain

Les chiffres présentés semblent cohérents avec les scores obtenus. Si les classes neutres et négatives obtiennent des scores honorables avec respectivement 97% et 84% d'accord, la classe positive est quant à elle très largement confondue avec l'ensemble des autres classes (11% d'accord). La méthode de sélection adoptée semble donc inopérante de ce point de vue. L'une des raisons peut être le filtre activation. Etant donné la nature des données (centre d'appels), la manifestation d'états affectifs positifs est produite de manière beaucoup plus lexicale que dans d'autres contextes (ex : je vous remercie beaucoup de votre aide). Partant de ce constat, nous avons essayé d'améliorer les scores obtenus sur la classe positive en utilisant les indices syntaxico-sémantiques et sémantiques présentés dans le chapitre 6. La segmentation utilisée restera la même que pour les tests précédents, à savoir des segments de longueur équivalente à 5 secondes maximum.

Par ailleurs, les données sélectionnées pour la classe positive de l'étape précédente seront remplacées par la nouvelle sélection effectuée à l'aide des indices sémantiques et syntaxiques. Pour cette nouvelle sélection, les instances positives devront répondre à deux critères :

i) obtenir un score positif supérieur à 2.2 tel que défini dans (Cailliau and Cavet 2010),

ii) avoir été étiqueté « Positive » ou « StrongPositive » par le modèle sémantique.

Nous obtenons avec cette approche un nombre de 840 segments positifs sélectionnés. Les modalités de test du modèle sont inchangées, à savoir un ensemble de test et d'entrainement indépendants du locuteur. Cependant, les instances composant l'ensemble de test sont issues d'une segmentation automatique alors que celles composant l'ensemble d'entrainement sont mixtes, en partie automatique et en partie manuelle.

Les résultats obtenus sont présentés dans le Tableau 7-9 ci-dessous :

	Positif	Négatif	Neutre	Global unweighted F-score
Score de référence : Corpus non enrichi	0.18	0.53	0.60	0.44 (±2%)
Corpus Enrichi (segmentation de 5 secondes)	**0.19**	**0.59**	**0.71**	**0.50 (±2%)**

Tableau 7-9 Scores obtenus après une sélection de la classe positive avec une approche syntaxique et sémantique

Les classes positive et négative n'enregistrent pas de changement par rapport à la précédente sélection. Le but était ici d'essayer d'améliorer les résultats obtenus sur la classe positive. Nous voyons ainsi, avec une variation cette fois-ci positive du score de cette classe par rapport au score de référence (+1%), une amélioration de +3% par rapport aux scores obtenus avec la sélection précédente. Cependant, même si cette tendance est positive, les scores pour la classe positive restent très bas. Nous conserverons pour la suite de cette étude ce dernier ensemble d'entrainement comme score de référence.

7.4.5 Impact de l'emplacement de la fenêtre temporelle sur les résultats

Nous avons vu dans le paragraphe précédent que des gains étaient obtenus lorsque nous sélectionnions de manière automatique de nouvelles données dans une base de données d'appels téléphoniques. Nous avons également vu que ces gains peuvent être variables suivant la manière dont sont segmentées les données (3 secondes, 5 secondes et pas de segmentation). Considérant que la méthode découpant les segments par tranche de 5 secondes donne les meilleurs résultats, nous nous sommes servis de cette base pour la suite des tests. Nous avons ainsi conservé la fenêtre temporelle de 5 secondes mais avons décalé de 2,5 secondes (1/2 segment) le début de la segmentation afin d'observer si les scores restaient stables. Le schéma ci-dessous illustre le procédé :

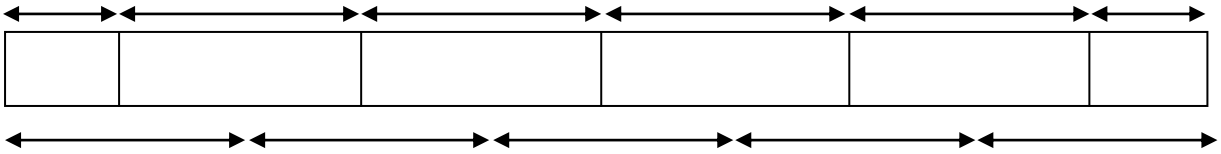


Figure 7-3 Les différentes segmentations des instances supplémentaires utilisées lors de l'enrichissement des modèles

Les scores ont été calculés en utilisant l'un, puis l'autre mode de segmentation. Nous avons ensuite, essayé de combiner les deux modes de segmentations afin d'observer si l'utilisation de segments décalés au cours du temps pouvait permettre un complément d'information lors de la phase de détection. Le score de référence est celui obtenu grâce à la segmentation par intervalle de 5 secondes, à des instances filtrées grâce au modèle d'activation (négatif, neutre) et à la sélection par modèle sémantique/syntaxique (classe positive) présenté dans le Tableau 7-9.
Nous pouvons en voir les résultats dans le Tableau 7-10 ci-dessous :

	Positif	Négatif	Neutre	Global
Score de référence : Segmentation 1	0.19	0.59	0.71	0.50 (±2%)
Segmentation 2	0.19	0.58	0.70	0.49 (±2%)
Combinaison segmentation 1 + 2	0.20	0.58	0.70	0.49 (±2%)

Tableau 7-10 Variation des scores obtenus en fonction des modes de segmentation

Nous voyons d'après les scores obtenus dans le tableau ci-dessus que nous n'obtenons pas de variations significatives des résultats (1% tout au plus) aussi bien en déplaçant la fenêtre temporelle (segmentation 1 vs segmentation 2) qu'en combinant les deux modes de segmentation. Si nous ramenons l'expérience de détection à deux classes, neutre et négative, les constations sont les mêmes avec 0.25% de variation au niveau global.

Pour le moment nous avons vu que les scores variaient suivant la durée de la fenêtre temporelle utilisée, mais que ces derniers restaient stables si nous décalions cette fenêtre afin de découper un segment de différentes façons (voir Figure 7-3 et Tableau 7-10). Afin de valider définitivement cette fenêtre temporelle et ce mode de sélection des instances, nous avons voulu effectuer le test sur des données découpées manuellement. Les dialogues utilisés pour les tests précédents ont été conservés, mais nous avons utilisé, dans l'ensemble de test, des segments découpés manuellement. Nous présentons ainsi dans le Tableau 7-11 ci-dessous les gains obtenus lorsque nous utilisons les

modèles enrichis et les modèles de base sur des segments manuels. Les segments venant enrichir les modèles sont quant à eux les mêmes que ceux utilisés précédemment, c'est-à-dire toujours issus de la segmentation automatique.

	Positif	Négatif	Neutre	Global
Modèle non enrichi	0.51	0.66	0.67	0.61 (±2%)
Modèle enrichi	0.52	0.71	0.71	0.65 (±2%)

Tableau 7-11 Gains obtenus après enrichissement des modèles sur des données segmentées manuellement

Nous obtenons globalement un gain de +4%, soit un gain de 2% inférieur à celui obtenu dans le meilleur des cas lorsque nous avons effectué le test sur des segments obtenus manuellement (voir Tableau 7-9). Le gain est dans ce cas inférieur pour la classe neutre par rapport aux expériences précédentes. Les classes positive et négative obtiennent respectivement des gains de +1% et +5%.

Cette série de tests nous permet de constater des gains dans les F-scores pour chacune des situations dans lesquelles nous avons testé notre modèle. Nous conserverons donc, pour la suite de ce chapitre, le sous ensemble enrichi.

7.5 Sélection fine des instances destinées à enrichir le modèle

Nous avons montré que les segments que nous avons sélectionnés étaient par la suite redécoupés dans le but d'homogénéiser la longueur de chaque instance. Pour rappel, le système de reconnaissance de la parole que nous avons utilisé se base sur (Gauvain et al. 2003) et segmente le flux de paroles selon les silences ou le fait qu'un changement de locuteur soit détecté.
Nous avons, dans ce paragraphe, conservé notre segmentation par tranches de 5 secondes et avons voulu vérifier si une partie du signal d'un segment pouvait être porteur de plus d'informations que d'autres. Pour ce faire, nous n'allons conserver qu'une partie du corpus enrichi présenté dans les paragraphes précédents.

Ainsi, nous avons successivement conservé, d'abord le sous-segment représentant les 5 premières secondes d'un segment, ensuite le sous-segment représentant les 5 dernières secondes d'un segment, puis enfin chacune de ces deux extrémités temporelles. Nous avons, avec cette expérience, cherché à savoir si nous pouvions diminuer le volume de données une fois ces dernières re-segmentées dans le but de réduire la complexité d'apprentissage en temps et en mémoire sans perte de performance. Idéalement, l'idée était de limiter chaque instance ayant dû être redécoupée en un ou deux sous-segments seulement. Le tableau ci-dessous présente les scores obtenus pour la construction du modèle, en utilisant seulement les premiers, puis seulement les derniers, ainsi qu'enfin, ces deux types de segments.

	Positif	Négatif	Neutre	Global	Nombre de segments dans l'ensemble d'entrainement
Score de référence 1 : corpus non enrichi	0.18	0.53	0.60	0.44 (±2%)	3703
Score de référence 2: Corpus enrichi	0.19	0.59	0.71	0.50 (±2%)	6434
Expérience 1 : Premiers sous segments de chaque instance conservés	0.18	0.57	0.70	0.48	4626
Expérience 2 : Derniers sous segments de chaque instance conservés	0.17	0.55	0.70	0.47	4626
Expérience 3 : Premiers + derniers sous segments conservés	0.18	0.56	0.69	0.48	5563

Tableau 7-12 Scores en fonction de la portion de segments conservés pour l'entrainement

Nous observons avec ces résultats, un gain quelle que soit la situation par rapport au score de référence 1 obtenu. A contrario, les scores sont globalement plus bas si nous observons les résultats obtenus par rapport au score de référence 2. En outre, les scores à un niveau global varient pour chacune des 3 expériences. Si nous considérons les expériences 1 et 2, nous pouvons noter que l'expérience 1 où nous avons conservé seulement les sous-segments représentant les 5 premières secondes d'une instance, a tendance à avoir des scores plus élevés. La différence est la plus marquée pour la classe négative (+2% lorsque nous conservons uniquement les sous segments de début). Ce gain peut être expliqué par le fait que le segmenteur utilisé dans le système de reconnaissance de la parole ne crée un nouveau segment que lorsqu'une rupture (silence, portion de signal changeant de manière abrupte) est détectée. De ce point de vue, la classe négative contenant en majorité de la colère, souvent caractérisée par une intensité forte sur une courte période de temps, peut profiter de ce découpage. Ces tendances seraient toutefois à confirmer sur un autre jeu de données.

7.6 Quel jeu de descripteurs ?

Etant donné que l'enrichissement a été fait avec des instances contenant un fort taux de paroles superposées, nous nous sommes posé la question de savoir si certains descripteurs comme la fréquence fondamentale de la voix, fortement liée à un locuteur en particulier, restaient pertinents pour des données comme celles que nous avons ajoutées. Nous avons, pour cela, éliminé les descripteurs relatifs à la fréquence fondamentale, ainsi qu'à la probabilité d'une portion de signal

d'appartenir à une partie voisée et les paramètres de type Zero Crossing Rate (Zcr). Nous conservons ainsi les paramètres relatifs à l'énergie avec leurs dérivées et fonctionnelles associées, ainsi que les 12 premiers MFCC. Le nombre de descripteurs passe ainsi de 384 à 308. Sur cette base de 308 descripteurs nous retirons dans un troisième temps les descripteurs dérivés soit (154 descripteurs en moins). Les résultats sont donnés dans le tableau 1-13 ci-dessous :

	Positif	Négatif	Neutre	Global (F-score)	Nombre de descripteurs
Score de référence : Corpus enrichi	0.19	0.59	0.71	0.50 (±2%)	384
Modèle enrichi	0.21	0.60	0.71	0.51 (±2%)	308
Modèle enrichi sans les deltas	0.21	0.59	0.71	0.50 (±2%)	154

Tableau 7-13 Scores obtenus après une sélection manuelle de descripteurs

L'hypothèse que nous avons émise concernant l'ensemble de descripteurs semble être confirmée. Nous arrivons en effet à obtenir des gains pour deux classes sur trois (positive et négative) avec 80 descripteurs en moins. Au delà du gain de performance qui est à relativiser étant donné la faible évolution, c'est l'économie en termes de temps de calcul et d'extraction des descripteurs qu'il faut considérer ; surtout au regard de l'important volume de données que nous avons à disposition. En ce qui concerne la stagnation de la classe neutre, nous pouvons l'expliquer par le fait que de manière générale, la reconnaissance de la parole ainsi que la segmentation sont en moyenne beaucoup plus « propres » (peu de paroles superposées et longueurs de segments plus homogènes en moyenne) lorsqu'il n'y a pas d'émotion exprimée. De ce fait, la suppression de descripteurs relatifs à un locuteur n'est pas bénéfique en termes de score de détection mais reste par contre intéressants en termes de dimensionnalité du problème de classification.

7.7 Enrichissement d'un modèle lexical

Les résultats obtenus dans les paragraphes précédents ont montré que la méthode d'enrichissement permettait de gagner jusqu'à 6% à un niveau global en enrichissant nos modèles acoustiques avec des données sélectionnées automatiquement grâce à une méthode de sélection hybride acoustique/linguistique. Afin de pousser plus loin nos investigations en matière d'enrichissement automatique de modèles, nous avons essayé d'utiliser la même méthode pour enrichir nos modèles lexicaux. Pour rappel, la méthode retenue à ce niveau se base sur une représentation de type « sac de mots » (*bag of words*). Le lexique utilisé avec le modèle de base contient 493 mots sélectionnés après une étape de lemmatisation et d'annotation morpho syntaxique réalisée à l'aide de l'étiqueteur de Brill (Brill 1995). Les détails sur la conception de modèles lexicaux sont présentés dans le chapitre 5. Les mêmes instances de test et d'entrainement ont été utilisées.

Les résultats obtenus avec un corpus non enrichi serviront de score de référence. Ils sont présentés dans le Tableau 7-14 ci-dessous :

	Positif	Négatif	Neutre	Global
Modèle non enrichi	0.17	0.55	0.65	0.45 (±2%)

Tableau 7-14 Scores obtenus sur un modèle lexical non enrichi

Les résultats sans enrichissement suivent la même tendance en acoustique qu'en lexical. La classe positive obtient un score bas du fait de sa faible représentation (8% dans le corpus de tests) mais reste au-dessus du seuil du hasard.

Nous avons enrichi ces modèles à l'aide des instances utilisées pour construire les modèles dont les scores sont présentés dans le Tableau 7-9. Le nombre de mots de vocabulaire sélectionnés passent ainsi de 493 à 580 grâce à l'ajout de segments des 264 dialogues supplémentaires. Afin de conserver les mêmes instances que les modèles acoustiques, nous avons sous-segmenté les instances lexicales par tranches de 5 secondes grâce au système de reconnaissance de la parole. Ce dernier nous permet d'obtenir les repères temporels (*timecode*) au niveau du mot.

Nous avons donc enrichi le modèle lexical et présentons les résultats obtenus avec ce nouveau modèle dans le Tableau 7-15 ci-dessous :

	Positif	Négatif	Neutre	Global (F-score)
Score de référence : Modèle non enrichi	0.16	0.55	0.65	0.45 (±2%)
Modèle enrichi	0.16	0.57	0.61	0.44 (±2%)

Tableau 7-15 Scores obtenus après enrichissement des modèles lexicaux par des instances sélectionnées automatiquement

Comme nous pouvons le voir sur le Tableau 7-15 ci-dessus, la méthode d'enrichissement appliquée pour sélectionner des instances au niveau acoustique ne semble pas être fonctionnelle pour la sélection de nouvelles instances destinées à enrichir le modèle lexical. Le seul gain obtenu se situe au niveau de la classe négative et est de +2% seulement par rapport à la situation de départ. Les autres classes voient quant à elles leurs scores diminuer (neutre) ou stagner (pos). Cette absence de gain peut être expliquée par notre mode de segmentation des instances supplémentaires. Le fait de ne sélectionner que des portions de phrases est nettement plus problématique pour l'aspect linguistique du fait de la coexistence de mots différents complémentaires entre eux au sein d'un même segment.

7.8 Enrichissement avec 100 heures de dialogues choisis aléatoirement

Notre première sélection de dialogue portait sur les 264 dialogues les plus émotionnels répartis en 29682 instances et à partir desquels nous avons pu enrichir nos modèles de 1914 segments émotionnels. Pour cette dernière étape, nous avons porté le volume de parole à 100 heures. Les 50 heures issues de la première sélection ont été conservées et 50 heures réparties sur 401 dialogues supplémentaires ont été sélectionnées aléatoirement. Ces 401 dialogues sont découpés en 45596 segments. Sur cette sélection de 45596 segments, 13179 ont été retenus comme étant potentiellement porteurs d'une émotion (score du modèle linguistique supérieur au seuil de 2.2). La procédure de sélection reste identique à celle présentée dans le paragraphe 7.4.1.

La sélection des instances d'après les scores obtenus grâce au modèle sémantique nous permet d'observer les résultats suivants :

Segmentation	Nombre d'instances avant le passage du filtre d'activation	Nombre d'instances après le passage du filtre d'activation	Nombre d'instances positives supplémentaires	Nombre d'instances négatives supplémentaires
Instances supplémentaires	13179	983	408	575

Tableau 7-16 Nombre d'instances selon le mode de la segmentation et la classe émotionnelle de la seconde sélection

Le taux de sélection est ici très bas (moins de 10% de segments considérés comme émotionnels) au regard du nombre de segments de départ. Ceci peut être expliqué par le fait que, contrairement aux premières sélections, celle-ci a été faite de façon entièrement aléatoire et est donc plus représentative du taux de passages émotionnels réellement présents dans des dialogues récoltés en centre d'appels. Nous avons incorporé les instances nouvellement sélectionnés dans le modèle précédent.

Etant donné le très faible nombre d'instances ajoutées lors de cette deuxième sélection, nous n'observons pas de gain significatif (en moyenne +0.5%) sur les trois classes. Une sélection beaucoup plus large de 500 heures supplémentaires aurait probablement été nécessaire pour faire augmenter les résultats de manière intéressante. Nous pouvons cependant remarquer que la solution proposée semble être robuste ; nous n'observons à aucun moment une baisse de score lorsque nous introduisons des instances supplémentaires sélectionnées de façon automatique à l'aide de notre approche mixte linguistique/acoustique.

7.9 Conclusion

Partant du constat qu'un plus grand volume de données nous permettrait d'obtenir un pouvoir de généralisation plus important pour nos modèles, nous avons proposé dans ce chapitre une méthode d'enrichissement basée sur une approche mixte acoustique/linguistique. Nous avons pour cela procédé en plusieurs étapes. La première consiste en une sélection à l'aide du modèle linguistique utilisé dans le chapitre 6. Ce modèle attribue un score émotionnel positif ou négatif à un segment donné. Nous avons, dans notre méthode, sélectionné uniquement les segments dont le score dépasse un certain seuil, comme préconisé dans (Cailliau and Cavet 2010). La deuxième étape consiste, du point de vue des émotions négatives, en l'utilisation d'un filtre d'activation ; seuls les segments ayant été considérés comme ayant une activation « forte » sont retenus. En revanche, concernant les segments positifs, après avoir observé des faiblesses lorsque nous appliquions la procédure décrite ci-dessus, nous avons décidé de ne retenir que ceux qui obtenaient une annotation « positive » par les deux modèles linguistiques considérés. Ainsi, les segments retenus sont ceux obtenant un score émotionnel égal à 0 et dont l'activation est considérée comme « faible ».

A l'issue de ce processus, nous avons réussi à doubler la taille de notre corpus de base et à obtenir des gains globaux allant jusqu'à +6% dans la situation la plus favorable. Nous avons obtenu des gains

pour chacune des 3 classes émotionnelles, de +2% pour la classe positive à +11% pour la classe neutre, ce qui semble encourageant. Le fait de prendre en compte les modalités, à la fois linguistiques et acoustiques, est une des originalités de notre approche. Cela nous permet d'appréhender le problème de la détection d'états affectifs sous un angle plus global, et ce, avec une approche, concernant l'enrichissement des modèles, entièrement automatique.

Chapitre 8 Evaluation de la qualité d'une conversation : vers une signature émotionnelle d'un dialogue

Résumé

Dans ce dernier chapitre nous proposons une évaluation de la qualité d'une conversation. Notre but est la détermination de la satisfaction client à travers l'expression émotionnelle en cherchant à différencier les dialogues ayant une valence globalement positive ou neutre de ceux étant négatifs émotionnellement parlant.

Lors d'une conversation téléphonique, l'interaction entre chacun des locuteurs va conditionner l'expression émotionnelle émise. Nous avons par exemple vu, dans le chapitre 4, comment un locuteur, dans un état affectif « neutre », pouvait soudainement évoluer vers un état de type « colère », en fonction d'un élément déclencheur (refus d'étalage du paiement, délai d'attente trop long). Cette particularité, liée aux données conversationnelles, nous a donc incité à observer une conversation comme un tout et non comme une suite d'éléments isolés.

Aussi, étant donné que la durée d'une conversation dans notre corpus peut être importante (jusqu'à 30 minutes), nous choisissons d'analyser deux passages clés d'une discussion, à savoir le début et la conclusion de celui-ci. Nous montrons ainsi qu'il est possible avec des données issues d'une annotation manuelle, d'établir la valence positive ou négative d'un dialogue, avec précision, simplement d'après ces deux extraits.

Lorsque nous nous basons sur des données issues de classification automatique, le problème est plus complexe et des erreurs peuvent apparaitre du fait des confusions dans les prédictions des modèles. Aussi, afin d'obtenir des prédictions plus fiables nous testons successivement deux nouveaux types d'indices dialogiques et structurels dans le but de fournir aux modèles construits un ensemble d'informations plus complet sur la conversation étudiée. Avec l'ajout de ces indices, nous arrivons à des scores proches de ceux obtenus lorsque nous nous basions sur des sorties issues d'une annotation manuelle (83% contre 89% pour l'annotation manuelle).

8.1 Introduction

Le concept de signature émotionnelle propose de qualifier une conversation comme étant négative ou positive à l'aide d'indices multi-niveaux (i.e. obtenus à partir de tours de paroles mais aussi de la conversation dans sa globalité). Il convient avant toute chose de mentionner qu'une telle qualification est par nature très subjective et dépend des besoins et attente de l'utilisateur final. Si nous nous plaçons du point de vue d'une société fournissant un service client dans un centre d'appel, une conversation se finissant de manière « positive » (client satisfait) ne sera pas vu comme un dialogue à problèmes. Si nous nous plaçons à présent du point de vue de l'annotateur cherchant des dialogues « émotionnels » une conversation se terminant de manière positive mais contenant une majorité de tours de parole d'un client exprimant une insatisfaction sera considérée comme une discussion négative. C'est dans cette seconde vision des choses que nous plaçons notre approche. D'une manière générale l'une ou l'autre des deux approches reste possible avec les propositions que nous formulons dans ce chapitre, une simple ré-annotation des dialogues d'entraînement en « positifs » ou « négatifs » est nécessaire pour passer d'un point de vue à l'autre. L'idéal étant d'avoir une annotation donnée par les end-users.

Nous proposons également une première approche reprenant les méthodes décrites dans les chapitres 5 (modèles purement acoustiques), 6 (modèles linguistiques-acoustique) et 7 (modèles enrichis), en les appliquant à l'évaluation de la modalité positive ou négative d'un dialogue complet. Au-delà de l'approche globale , comme nous l'avons déjà mentionné dans le chapitre 4, une discussion est un ensemble de tours de parole organisés séquentiellement de manière à rendre un discours intelligible pour un ou plusieurs interlocuteurs (Sacks 1992). Nous avons essayé ici de souligner l'importance de ce principe. En réutilisant les dimensions d'implication affective présentées dans le chapitre 4 et en observant leur évolution au cours de dialogues complets.

Nous prenons également en compte dans ce chapitre différents types d'indices, fondés sur la structure du dialogue ainsi que sur les thématiques abordées qui nous permettrons de compléter les modèles acoustiques pour la qualification globale d'un appel. Ces indices peuvent être extraits à l'aide d'un traitement automatique ou manuel, le but étant essentiellement dans ce dernier chapitre de proposer une ouverture vers un nouveau mode de qualification des conversations.

8.2 Données utilisées

Nous avons utilisé, pour tester nos modèles, 44 dialogues issus du corpus EDF que nous avons décrits dans le chapitre 4. Ces 44 dialogues sont ceux utilisés pour l'entrainement des modèles automatiques présentés dans les chapitres précédents. Ils sont répartis en 17 dialogues positifs sélectionnés pour leur contenu globalement positif et 27 dialogues sélectionnés pour leur contenu globalement négatif. Chaque dialogue a été segmenté et annoté par deux linguistes de formation expérimentées dans le domaine de l'annotation émotionnelle (voir chapitre 4 pour plus de détail). Le score de référence a été établi à partir des annotations manuelles.

8.2.1 Les variables d'un dialogue émotionnel :

Nous avons souhaité insister sur les différents types de dialogues émotionnels que nous avons eu l'occasion d'examiner dans le but de dégager des variables spécifiques à une forme d'expression émotionnelle positive ou négative. Des études portant sur l'incorporation d'indices dialogiques (emphase, coupure, reprise) dans des modèles automatiques ont montré qu'il était possible d'améliorer les scores de détection (Lee and Narayanan 2005). Les résultats obtenus par les auteurs sur des données issues de centre d'appels restaient cependant meilleurs lorsque la combinaison se limitait aux indices acoustiques et linguistiques. Dans (Devillers et al. 2003) les auteurs mettent en avant la corrélation entre des émotions de type colère et des événements dialogiques (nommés *accident*) bloquant la progression de la conversation. Au cours d'une étude ayant pour but de repérer les dialogues conflictuels durant des débats politiques (Kim et al. 2012) proposent la prise en compte d'indices à un niveau dialogique, tels que le nombre de tours contenant de la parole superposée (overlap), la durée d'un tour de parole, ou la modélisation de l'enchainement des tours de parole sous forme de bi-gramme, dans le but de dégager des pattern spécifiques au conflit, à un instant précis de la conversation. Les indices conversationnels sont couplés à des indices acoustiques et permettent d'obtenir des gains de l'ordre de 8% par rapport aux modèles utilisant des indices acoustiques seuls. Les résultats montrent que les indices conversationnels relatifs à la parole superposée et aux interruptions volontaires d'un premier locuteur par un second permetraient de repérer le conflit au cours d'une discussion. Dans (Galley et al. 2011) les auteurs utilisent 3 types d'indices dialogiques différents afin de détecter les séquences d'accord/désaccord dans un corpus composé de conversations enregistrées lors de réunions de travail. Les indices sont les suivants :

i) structurels (nombre de prises de parole ou d'essais de prise de parole durant le discours d'un autre locuteur, d'overlap, nombre de tours de parole de chaque locuteur).
ii) Indices de durées (temps de parole des locuteurs, durée des overlap, taux de parole).
iii) les indices lexicaux (nombre de mots utilisés, nombre de mots communs entre deux locuteurs, nombre de n-gram communs pour les deux locuteurs).

Ces 3 types d'indices supplémentaires permettent une amélioration des résultats de l'ordre de 4%.

8.2.2 Les variables de durées et les variables structurelles des conversations téléphoniques

Une série de mesures relatives à la longueur des dialogues, des tours de parole et des phases du discours sera présentée dans les paragraphes suivants. Nous cherchons par ce biais à dégager à l'aide d'indices structurels des éléments pouvant être discriminants quant à la valence globale d'un dialogue. Les indices retenus seront par la suite combinés aux modèles automatiques élaborés dans le paragraphe 8.4 , afin de construire un modèle automatique enrichi avec les indices structurels que nous aurons calculés.

La durée du dialogue semble être une variable discriminante des appels à problème. Nous avons ainsi mesuré dans notre corpus que les dialogues à valence globalement négative étaient, en moyenne,

plus longs que les dialogues neutres ou positifs. Nous présentons dans le Tableau 8-1 ci-dessous la durée moyenne pour les dialogues négatifs et positifs :

Dialogues non-négatifs	Dialogues négatifs
10 minutes	13 minutes

Tableau 8-1 Durée moyenne des dialogues non-négatifs et négatifs

Nous voyons qu'un dialogue considéré comme contenant du discours négatif est statistiquement plus long qu'un discours positif (d'environ 30%). Un dialogue très nettement plus long que la moyenne peut ainsi fournir une piste quant à la nature des émotions exprimées. Nous poursuivons nos investigations dans cette voie en mesurant la longueur moyenne d'un tour de parole neutre et celle d'un tour de parole perçu comme étant positif. Ces résultats sont présentés dans le Tableau 8-2 ci-dessous :

Longueur moyenne d'un tour de parole neutre	Longueur moyenne d'un tour de parole négatif
2.21 secondes	4.25 secondes

Tableau 8-2 Durée moyenne des tours de paroles négatifs et neutres

Comme pour la longueur globale du dialogue, les tours de parole négatifs sont en moyenne deux fois plus longs que les tours de parole négatifs. Ainsi, la présence dans une proportion donnée de ce type de tours de parole pourrait indiquer une conversation potentiellement négative.

	Longueur moyenne des tours de paroles d'un agent	Longueur moyenne des tours de paroles d'un Client	Différence moyenne
Dialogues Négatifs	2.6 secondes	2.8 secondes	0.2 seconde
Dialogues Positifs	1.8 secondes	2.47 secondes	0.67 secondes

Tableau 8-3 Longueur moyenne d'un tour de parole agent / client sur un corpus de dialogues choisis pour leurs contenu émotionnellement marqués

La différence de longueur entre les tours de parole de clients et d'agents issus de dialogues positifs ou négatifs ne semble quant à elle pas fournir d'indication discriminante. Cet indice ne sera donc pas conservé lors de la phase d'élaboration du modèle automatique.

Afin d'analyser plus en détails les dialogues que nous avons utilisés comme base pour nos modèles d'apprentissage dans les dialogues précédents, nous définirons 4 phases principales traditionnellement mises en œuvre lors d'un dialogue agent/client dans un centre d'appels EDF :

- Ouverture : salutations, formules de politesse
- Identification : identité du client, formulation du motif de l'appel
- Traitement : Mise en œuvre de procédures pour le traitement de la demande ou la résolution du problème mentionné par le client
- Clôture : Formules de politesse, fin de la discussion

D'après une étude réalisée pour le projet voxFactory (Balague et al. 2011) sur les centre d'appels EDF que nous étudions, il apparait que ces différentes phases se répartissent le temps de conversation dans les proportions suivantes si nous observons le corpus avec une sélection aléatoire de 200 conversations :

	Moyenne	Ecart-type	Minimum	Maximum
% Phase d'ouverture	1.1	0.9	0.1	9.4

% Phase d'identification	25.4	18.9	2.2	89.7
% Phase de traitement	71.2	19.7	6.1	96.6
% Phase de clôture	2.1	1.7	0.2	8.7

Tableau 8-4 Proportion moyenne des différentes phases de conversation (Balague et al. 2011)

Etant donnée la très grande variabilité existant entre le minimum et le maximum de chacune des phases du dialogue (notamment la phase d'identification), nous proposons de concentrer notre attention sur les valeurs moyennes de chaque phase. Nous présentons dans le tableau ci-dessous les valeurs moyennes de chacune des phases des dialogues de notre corpus d'apprentissage, composé de 44 conversations (voir 8.2 pour le détail). Les valeurs présentées dans le Tableau 8-4 serviront de valeurs de référence.

	Dialogues Négatifs	Dialogues non négatifs
% Phase d'ouverture	*Impossible à calculer du fait de l'anonymisation des 20 premières secondes de parole*	
% Phase d'identification	14%	14%
% Phase de traitement	84%	81%
% Phase de clôture	1%	4%

Tableau 8-5 Proportion moyenne des différentes phases de conversations dans les 44 dialogues de notre corpus

La différence entre les deux types de dialogue se situe au niveau de la phase de clôture, nettement plus longue dans le cas des dialogues positifs. Cette différence se traduit en règle générale par des remerciements et des formules de politesse pouvant durer nettement plus longtemps dans le cas des dialogues positifs (23 secondes en moyenne contre 6 pour les dialogues négatifs). L'identification automatique d'une phase de clôture anormalement courte, inférieure ou égale à 6 secondes par exemple, pourrait être le marqueur d'un dialogue négatif et d'une rupture nette dans la conversation. La durée plus courte que nous pouvons constater pour la phase de traitement va également dans le sens de l'étude décrite dans (Balague et al. 2011) au cours de laquelle les auteurs corrèlent la longueur de la phase de traitement avec une valence négative. Ainsi, plus ladite phase de traitement est longue, plus le dialogue se passe mal.

Nous avons par la suite voulu savoir dans quelle phase d'un dialogue les émotions négatives étaient le plus exprimées et par quel locuteur. Nous poursuivons donc l'étude en corrélant les états affectifs par locuteur, obtenus d'après les annotations manuelles et les différentes phases des dialogues de notre corpus.

	Proportion d'émotions négatives exprimées par l'agent	Proportion d'émotions négatives exprimées par le client
Phase d'identification	6%	21%
Phase de traitement	14%	32%
Phase de clôture	9%	13%

Tableau 8-6 Proportion de tours négatifs exprimés par les locuteurs durant chacune des phases des dialogues

Sans surprise, les émotions négatives sont en moyenne plus exprimées par le client ; il peut par contre être intéressant de voir que dans le cadre d'un dialogue à problème, les agents expriment un ressenti négatif dès la phase d'identification. Cette expression a la particularité d'être presque exclusivement véhiculée par canal acoustique et avec une intensité très faible. La situation la plus prototypique est le cas d'un client démarrant la conversation sur un ton particulièrement offensif ex :

« non mais vous me prenez pour un con ! ». Bien que peu courante, une expression négative produite par l'agent en début de dialogue pourrait être un marqueur intéressant à prendre en compte.

8.3Test sur les débuts et fins de dialogues : résultats d'après une annotation manuelle des données

Pour tester l'hypothèse d'importance des débuts et fins de dialogue, nous avons choisi de ne conserver que les N premiers et N derniers segments d'un dialogue pour la classification. Outre l'économie en temps de calcul, le choix de segments à ces deux moments particuliers de la conversation a été fait en partant du principe qu'ils étaient représentatifs du déroulement d'une conversation et permettait potentiellement d'en inférer son orientation. Nous avons ainsi représenté un dialogue sous la forme d'un vecteur. Chaque segment sélectionné représente un élément de ce vecteur. Une étiquette émotionnelle « positif », « négatif » ou « neutre » est attribuée à un segment donné, d'après l'annotation qui a été fournie par un humain ou par un des modèles de détection. Si on considère N=15, un segment ayant une longueur moyenne de 5 secondes c'est environ 1 minute 30 de dialogue de début et de fin d'un appel qui sera prise en compte pour chaque évaluation. Nous avons fait varier N à 10, 15, 20 segments en début et fin des dialogues.

Dans un premier temps, nous établissons le score de référence d'après les annotations manuelles données par les annotateurs du corpus sur chaque segment automatique. Cette segmentation étant souvent difficile (20% segments comportent des données des deux locuteurs), les deux annotateurs n'étaient pas toujours d'accord. Comme nous voulions une annotation sur le dialogue entier, nous présentons ci-dessous les scores obtenus d'après les annotations d'un seul des deux experts.

Nous utilisons un algorithme de type bayesien naif (*naive bayes*) pour la classification. Les données d'entrée sont les annotations manuelles pour les N premiers segments et N derniers segments. Les résultats sont présentés dans le Tableau 8-7 ci-dessous :

Annotation manuelle	Début de dialogue	Fin de dialogue	Début et fin de dialogue
10 segments utilisés	0.72	0.75	0.79 (±9%)
15 segments utilisés	0.83	0.79	0.89 (±9%)
20 segments utilisés	0.86	0.79	0.91 (±9%)

Tableau 8-7 Résultats d'après une annotation manuelle.

Logiquement, la situation dans laquelle nous utilisons un nombre de segment le plus important obtient les meilleurs résultats. Cependant, étant donné la durée de certains dialogues (inférieurs à 5 minutes) l'utilisation de 20 segments en début et fin de dialogue reviendraient à prendre en compte la quasi-totalité du dialogue, ce qui n'est pas notre but ici. Nous retiendrons donc pour la suite de l'étude le cas d'utilisation mettant en œuvre une sélection de 15 dialogues.

Les scores obtenus sont conformes aux attentes avec des scores élevés de classification correcte. Si nous prenons une fenêtre temporelle de 15 segments ce sont au total 39 dialogues sur 44 qui ont été classés correctement. Cela nous permet de valider notre hypothèse selon laquelle la détermination de l'orientation d'un dialogue est possible seulement d'après une petite portion jugée représentative de ce dernier. Il faut cependant noter que cinq dialogues n'ont pas reçu une prédiction correcte, du

fait d'une très forte présence de l'état « neutre » ou d'une mixité importante des états positif et négatif au sein d'une même portion de dialogue. Ce problème est essentiellement imputable au choix des portions de dialogue, celui-ci n'étant pas adapté à ces 5 dialogues en particulier. Deux cas de figure sont principalement à l' origine de ces problèmes. Dans le premier cas de figure, l'essentiel de l'expression affective est réalisée dans le corps du dialogue comme c'est le cas dans l'exemple présenté sur les Figure 8-1, Figure 8-2, Figure 8-3 ci-dessous. Nous présentons les frises obtenues pour chacun des annotateurs, montrant ainsi des divergences entre les 2 annotations :

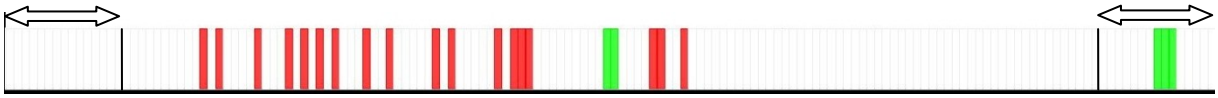


Figure 8-1 Exemple de dialogue problématique 1 (annotateur 1) (chaque segment a la même taille)

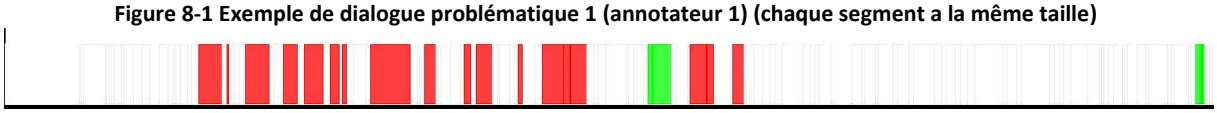


Figure 8-2 Exemple de dialogue problématique 1 (annotateur 1) (la taille des segments est fonction de leur durée)

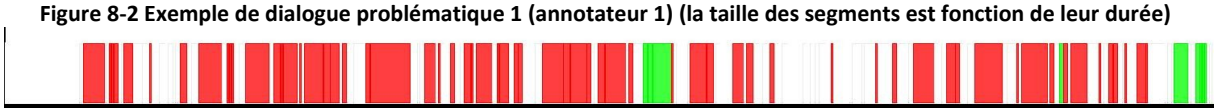


Figure 8-3 Exemple de dialogue problématique 1 (annotateur 2) (la taille des segments est fonction de leur durée)

La Figure 8-1 ci-dessus représente le dialogue et l'annotation reçue pour chaque segment de ce dernier. Un segment rouge représente une annotation relative à une émotion négative (colère, exaspération, etc) ; un segment vert représente une annotation relative à une émotion positive (joie, satisfaction, etc). Les segments blancs sont quant à eux utilisés pour représenter un état neutre. Sur les Figure 8-2 et 8-3 la taille des barres est liée à la durée du segment.

L'une des pistes pour palier ce problème serait, dans ce cas, de ne sélectionner que des portions de dialogues contenant un taux minimum de segments émotionnels retrouvés dans d'autres parties du dialogue.

Le deuxième cas de figure est celui où un dialogue a été sélectionné en raison de son contenu émotionnel mais n'a pas d'orientation particulière (autant de contenu négatif que positif) comme c'est le cas sur l'exemple ci dessous (Figure 8-4). Ce dialogue est également issu des 5 conversations qui n'ont pas reçu de prédiction correcte.

Figure 8-4 Exemple de dialogue problématique 2 (annotateur 1) (chaque segment a la même taille)

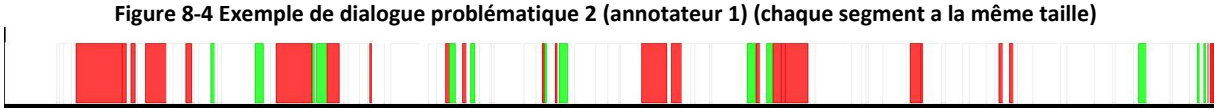


Figure 8-5 Exemple de dialogue problématique 2 (annotateur 1) (la taille des segments est fonction de leur durée)

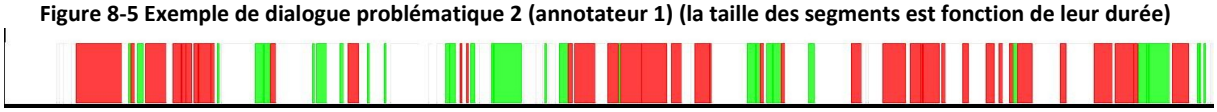


Figure 8-6 Exemple de dialogue problématique 2 (annotateur 2) (la taille des segments est fonction de leur durée)

Ce cas de figure est plus problématique car même si une autre portion de dialogue est sélectionnée, il semble difficile de définir l'orientation globale de ce dialogue. L'une des pistes pourrait être alors

d'associer le locuteur à un segment donné et de ne baser l'analyse de la conversation que sur le client ou l'agent afin d'obtenir un contenu émotionnel potentiellement plus homogène. Ce cas de figure serait valable dans les situations ou l'agent positive de manière exagérée une situation considérée comme très négative par le client.

8.4 Test sur les débuts et fins de dialogues : résultats d'après une annotation automatique

Nous avons réalisé l'expérience dans les mêmes conditions que celles citées précédemment pour N=15 mais en remplaçant les données annotées manuellement par des informations obtenues de manière automatique à l'aide des différentes méthodes que nous avons présentées dans les chapitres précédents (modèle acoustique seul (expérience 2), fusion des informations au niveau acoustique et linguistique (expérience 3), modèle enrichi de manière automatique (expérience 4) et modèle fusionné (expérience 5). Nous avons utilisé le même algorithme bayésien naïf que dans l'expérience 1 citée précédemment. Les résultats sont présentés dans le Tableau 8-8 ci-dessous :

		Début de dialogue (15 premiers segments)	Fin de dialogue (15 derniers segments)	Début et fin de dialogue
Expérience 1	Score de référence : Annotation manuelle	0.83	0.82	0.89 (±9%)
Expérience 2	Score de référence 2 : Modèle acoustique (segmentation manuelle)	0.67	0.75	0.82 (±9%)
Expérience 3	Modèle acoustique (segmentation automatique)	0.62	0.57	0.63 (±14%)
Expérience 4	Modèle enrichi (segmentation automatique)	0.65	0.63	0.69 (±14%)
Expérience 5	Modèle fusionné (segmentation automatique)	0.68	0.59	0.61 (±14%)

Tableau 8-8 Score de détection au niveau dialogique à partir d'une annotation automatique

Nous pouvons observer que de manière générale, la segmentation manuelle est plus performante que la segmentation automatique. Nous ne dépassons que dans de très rares cas, les scores obtenus d'après un découpage manuel des tours de parole. La seule hypothèse dans laquelle un modèle utilisant une segmentation automatique dépasse le manuel est celle dans laquelle est utilisé un modèle de type « fusion » pour l'analyse des débuts de dialogue. Nous pouvons constater que les analyses portant sur les débuts de dialogue semblent être plus discriminantes, au niveau global, que celles portant sur la fin d'une conversation. Cela peut sembler logique, puisqu'un dialogue globalement négatif peut voir sa tendance s'inverser en fin de conversation, si l'agent a répondu de manière satisfaisante à la demande du client. C'est le cas, par exemple, de la conversation illustrée sur la Figure 8-1 du paragraphe précédent. D'une manière générale, nous constatons que la prise en

compte à la fois les portions comprises au début et à la fin du dialogue donne des scores plus élevés que lorsque nous prenons chacun des passages de dialogue de manière individuelle. Cette constatation n'est cependant pas valable dans le cas du modèle par fusion ce qui peut paraître étonnant. Nous voyons également que les résultats des modèles enrichis et des modèles de fusion linguistique/acoustique offrent dans la plupart des cas des scores plus élevés que ceux du modèle acoustique simple. Ces résultats sont cohérents dans le sens ou les résultats obtenus sur ces modèles obtiennent eux même des scores de détections plus élevés que ceux du modèle acoustique simple.

8.5 Ajouts d'informations concernant le rôle du locuteur pour la classification au niveau global

Constatant que la méthode proposée dans le paragraphe précédent s'avère viable, nous décidons d'enrichir les modèles présentés ci-dessus. Nous ajoutons 3 descripteurs supplémentaires au vecteur relatif au locuteur. Ces nouveaux descripteurs représentent le pourcentage de segments faisant intervenir le client, l'agent ou les deux (parole superposée). Nous nous sommes en effet aperçu au cours de l'étude du corpus que l'essentiel de l'expression affective négative était produite par le client. Nous avons émis l'hypothèse que la prise en compte de cette information pourrait ainsi potentiellement améliorer les scores de classification. Nous présentons les scores obtenus dans le ci-dessous :

	Début de dialogue (15 premiers segments)	Fin de dialogue (15 derniers segments)	Début et fin de dialogue (intervalle de confiance)
Score de référence : Annotation manuelle	0.83	0.82	0.89 (±9%)
Annotation manuelle + information relative au locuteur	0.86 (+3%)	0.84 (+2%)	0.89 (±9%)

Tableau 8-9 Score de détection au niveau dialogique à partir d'une annotation manuelle et d'informations relatives au locuteur

A un niveau individuel, nous observons des gains lorsque nous prenons en compte le temps de parole de chaque locuteur. D'une manière générale, c'est essentiellement le descripteur relatif au client qui apporte le gain observé. La prise en compte des portions de dialogue de début et de fin simultanément n'apporte pas de gain par rapport au score précédemment observé. Voyant que des gains étaient possibles nous avons effectué la même opération sur les modèles acoustiques avec, en ensemble de test, des données issues de la segmentation automatique. Les informations dialogiques concernant le locuteur sont cette fois-ci obtenues de manière automatique et non plus annotées manuellement comme cela était le cas dans la configuration précédente.

	Début de dialogue (15 premiers segments)	Fin de dialogue (15 derniers segments)	Début et fin de dialogue
Modèle acoustique (segmentation automatique)	0.60 (-2%)	0.55 (-2%)	0.60 (-3%)

Tableau 8-10 Extrait des scores de détection au niveau dialogique à partir d'une annotation automatique et d'informations relatives au locuteur

Les résultats obtenus avec des informations relatives aux locuteurs, extraits de façon automatique, ne sont pas probants dans ce cas. Dans cette situation, les descripteurs ajoutés viennent bruiter le modèle, ce qui entraine une chute des performances entre -2 et -3%. Cette baisse vient du fait que les segments découpés de façon automatique sont très bruités et que de la parole superposée est fréquemment présente (voir chapitre 4). Les autres tests, dont nous n'avons pas présenté le détail et que nous avons effectués sur les différentes configurations présentes dans le Tableau 8-8, vont dans le même sens, à savoir des baisses de score comprises entre 2% et 4%.

Le constat sur l'ajout d'informations relatives au locuteur est donc en demi-teinte dans notre cas. Si nous pouvons constater des gains lorsque nous nous basons sur des données fiables, tel n'est plus le cas lorsque nous utilisons des informations obtenues automatiquement ; ceci du fait de la difficulté des données considérées (présence de bruit, segmentation automatique ne différenciant les locuteurs que de façon approximative, etc.).

8.6 Prise en compte d'une séquence de prédictions émotionnelles

Nous avons vu, au cours des expériences précédentes, que lorsque des données sont obtenues de manière entièrement automatique, des dégradations de performances apparaissent, même lorsque nous cherchons à qualifier la valence d'un dialogue d'une façon globale. Au regard de la nature des données que nous utilisons, il semble intéressant de chercher à modéliser l'interaction, chaque tour de parole n'étant pas indépendant des tours précédents.

En effet, dans les expériences précédentes, chaque dialogue est représenté sous la forme d'une suite d'étiquettes émotionnelles, celles-ci correspondant aux prédictions fournies par les modèles de détection présentés dans les chapitres précédents.

Aussi, nous faisons évoluer ces représentations en adoptant, comme cela se fait classiquement en traitement automatique du langage, une représentation sous forme de sacs de N-GRAM (ou n = 1 à 3). Nous cherchons ainsi à retrouver des enchainements émotionnels pouvant concorder avec des paires adjacentes utilisées pour modéliser l'enchainement conversationnel (Clark and Schaefer 1989). Pour cette expérience nous utilisons les informations issues des modèles enrichis, ces dernières présentant le score de détection le plus élevé qui nous servira de score de référence (expérience 4 présentée sur le Tableau 8-8)

	Début de dialogue (15 premiers segments)	Fin de dialogue (15 derniers segments)	Début et fin de dialogue
Expérience 4	0.65	0.63	0.69 (±14%)
Expérience 4' (N-GRAM)	**0.76**	**0.63**	0.59 (±14%)
Expérience 5	0.68	0.59	0.61 (±14%)
Expérience 5' (N-GRAM)	0.71	0.59	**0.72**(±14%)

Tableau 8-11 Résultats d'après une modélisation du dialogue sous forme de « sac de NGRAM »

Les résultats obtenus avec cette forme de modélisation laissent apparaitre des gains sur certaines expériences (notamment celles relatives au début de dialogue). La prise en compte des informations de début et de fin de dialogue simultanément n'est quant à elle guère concluante. Ceci peut s'expliquer par le fait qu'une partie de dialogue (entre les 15 premiers et 15 derniers segments) n'a pas été prise en compte. Ainsi la mise en relation des segments 15 et 16 (respectivement le dernier et le premier de chaque partie) est artificielle et peut venir fausser le processus de classification.

Si nous nous concentrons sur l'expérience donnant les meilleurs résultats (début du dialogue de l'expérience 4 avec NGRAM) et cherchons les séquences les plus représentatives pour les dialogues positifs et négatifs, nous retrouvons les enchainements suivants :

Dialogues Positifs	**Dialogues Négatifs**
Pos – Pos - Neut	Neg – Neg - Neg
/	Neg – Neut - Neg
/	Neut – Neg - Neg
/	**Pos – Pos - Neut**

Tableau 8-12 Séquences ayant une spécificité positive pour la détermination de la valence globale d'un dialogue

Concernant les dialogues négatifs, c'est sans surprise que nous retrouvons des motifs mettant en jeu une forte présence de l'émotion de type « négatif » au sein des schémas. Plus étonnante est la présence du schéma « pos – pos – neut ». Cette occurrence met en relief la difficulté qu'ont nos modèles à différencier les émotions à valence positive et négative. De leur coté, les dialogues positifs n'ont qu'un seul schéma avec une spécificité positive, probablement en raison d'un manque de données, 17 dialogues seulement ayant été sélectionnés comme étant globalement positifs.

8.7 Validation des indices structurels avec un modèle automatique

En ce qui concerne le paragraphe 8.2.2, nous avons proposé une série de descripteurs basés sur la structure d'un dialogue. Pour rappel, ces nouveaux indices comprennent la longueur d'un dialogue, la longueur moyenne d'un tour de parole de la conversation, la durée moyenne de chacune des phases du dialogue ainsi que le pourcentage de parole pour chacun des acteurs (agent, client) et de parole superposée. Afin d'évaluer l'efficience de ces descripteurs, nous avons réalisé des tests avec des modèles automatiques pour la détection de la valence globale d'un dialogue. Les expériences ont été menées en plusieurs itérations. Dans la première, nous considérons l'ensemble des indices structurels retenus dans le chapitre précédent, soit 9 descripteurs au total, puis nous retirons tour à tour les groupes d'indices décrits dans le Tableau 8-13 ci-dessous :

Indices structurels (9 indices)	Groupe d'indices
Longueur du dialogue Longueur moyenne des segments	Indices de longueur
Pourcentage de temps de parole dans la phase d'indentification, de traitement et de clôture	Indices sur le temps de parole dans les phases
Pourcentage de temps de paroleclient/agent/overlap	Indices sur le temps de parole des locuteurs
Pourcentage d'overlap	Indice sur la proportion d'overlap dans le dialogue

Tableau 8-13 indices et groupements d'indices structurels retenus pour la classification globale de la valence d'un dialogue

Nous choisissons un classifieur bayésien naïf afin de déterminer de manière automatique la valence d'un dialogue à l'aide de nos différents indices structurels. Les résultats sont présentés dans le Tableau 8-14 ci-dessous :

	F-score Dialogues positifs	F-score Dialogue Négatifs	unweighted Global F-score
Classification aléatoire	*0.39*	*0.61*	*0.50*
Tous les indices structurels	0.571	0.8	0.69
Tout sauf les indices de longueur	0.6	0.79	0.70
Tout sauf les indices de temps de paroles sur les phases	0.4	0.69	0.55
Tout sauf le temps de paroles des locuteurs	**0.61**	**0.83**	**0.72**
Tout sauf le nombre de tour par dialogue	0.5	0.767	0.63

Tableau 8-14 Récapitulatif des gains apportés par chacun des types d'indices structurels considérés

D'après les résultats obtenus dans le Tableau 8-14, nous pouvons constater que deux types d'indices sont principalement utiles pour la détection d'une valence à un niveau global. Il s'agit des proportions de temps de parole dans chacune des phases de discours (identification, traitement, clôture) et le nombre de tours de parole d'un dialogue. Nous voyons en effet que lorsque nous retirons ce type d'indices, les scores chutent de manière importante (de -2% à -9%). Les indices prenant en compte les longueurs en secondes du dialogue, ainsi que la longueur moyenne des tours de parole ne semblent pas apporter d'information complémentaire, le score restant en effet inchangé par rapport à l'ensemble du jeu de descripteurs. Ainsi, contrairement à nos attentes, des variables comme le temps moyen de parole des locuteurs (agent/client/overlap) viennent bruiter le modèle, de telle sorte que les scores obtenus sont plus élevés lorsque nous retirons ces informations du jeu de descripteurs. Nous ne retiendrons donc, pour la suite de l'étude, que les groupes de descripteurs relatifs au temps de parole des locuteurs au cours des différentes phases du discours et le nombre de tours de parole d'un dialogue.

8.7.1 Combinaison avec les sorties des modèles acoustiques

Pour achever la validation des indices structurels retenus, nous avons choisi de combiner ces derniers avec les sorties des modèles acoustiques testés dans le paragraphe 8.4. Nous retenons ainsi, pour ces tests, les modèles acoustiques prenant en compte les 15 premiers et derniers segments de chaque dialogue, ceux-ci donnant en moyenne les résultats les plus élevés. Pour la classification, nous utilisons un algorithme bayésien naïf et utilisons un paradigme de fusion précoce (early fusion) afin de combiner les indices acoustiques et structurels. Les données acoustiques utilisées restent les mêmes que celles présentées dans le paragraphe 8.2 :

	F-score Dialogues positifs	F-score Dialogue Négatifs	unweighted Global F-score
Indices structurels seuls	0.61	0.83	0.72
Modèle acoustique + informations structurels	0.64	0.83	0.74
Modèle enrichi + informations structurels	**0.68**	0.83	**0.76**
Modèle fusionné + Informations structurels	0.66	0.82	0.74

Tableau 8-15 Scores obtenus après fusion des indices structurels et acoustiques

Les gains obtenus lorsque nous rajoutons les informations acoustiques aux indices structurels sont modérés (au plus +4%). Comme lors des tests effectués dans le paragraphe 8.4, le modèle enrichi fourni les meilleurs résultats. Si nous comparons ces résultats avec ceux présentés dans le Tableau 8-8 pour le modèle enrichi, nous remarquons une amélioration de +7% des résultats, ce qui représente une augmentation non négligeable. De manière intéressante nous pouvons remarquer que l'essentiel des gains obtenus lors des tests l'a été sur les dialogues positifs.

8.8Indices supplémentaires pour l'évaluation de la modalité négative ou positive d'un dialogue

Comme nous le mentionnions en début de ce chapitre, une discussion n'est pas seulement une série de tours de parole enchainée les uns après les autres de façon indépendante. Il nous semble indispensable, si nous souhaitons pouvoir utiliser des systèmes automatisés de manière naturelle de prendre en compte cette séquentialité. Nous avons pour cela sélectionné 3 dialogues (2 négatifs et 1 positif) que nous avons analysés manuellement à l'aide des indices d'implications affectives décrits dans le chapitre 4. Les informations supplémentaires que nous avons prises en compte sont donc :

- Locuteur : Agent ou Client. Nous avons vu que l'ajout de cette information, lorsqu'elle est fiable, pouvait être bénéfique aux modèles automatiques ;
- Dimensions induction/réaction : ces dimensions traduisent l'impact d'un tour de parole sur un ou plusieurs autres tours du dialogue ;

- Lien thématique entre segments : nous avons indiqué si le tour de parole courant faisait référence à un ou plusieurs tours de parole précédents. Nous souhaitons, avec ce référencement, pouvoir repérer les points bloquants du dialogue.
  Exemple : Si une référence est constamment faite, au cours du dialogue, à un même segment ou à une même zone dialogique, ces liens devraient permettre d'identifier un problème dans la conversation.

Ces trois nouvelles informations couplées à celles déjà obtenues grâce à la reconnaissance des émotions, constitueraient une signature permettant d'identifier les dialogues à problème. Comme pour les dialogues du paragraphe précédent, nous avons considéré deux valences possibles pour les dialogues sélectionnés : positive et négative.

8.8.1Modélisation de l'interaction client / télé-conseiller pour la détermination de la valence globale d'un dialogue

D'après les annotations obtenues sur 4 dialogues de test, nous avons cherché des mesures pouvant être discriminantes concernant la valence globale d'un dialogue. Le premier cas de figure que nous analyserons, concerne le taux de « réaction » du client et le taux d'induction de l'agent. Pour rappel, l'axe « réaction » a pour but la représentation du degré de réponse (de faible à fort) vis-à-vis d'un ou plusieurs tours de paroles précédents (« ex : Ce n'est pas possible ce que vous me dites là !»). L'axe induction, quant à lui, est utilisé pour représenter la manière dont un locuteur espère provoquer une réaction particulière chez son partenaire de conversation (ex : «Répondez à ma question ! »).

Dialogues négatifs

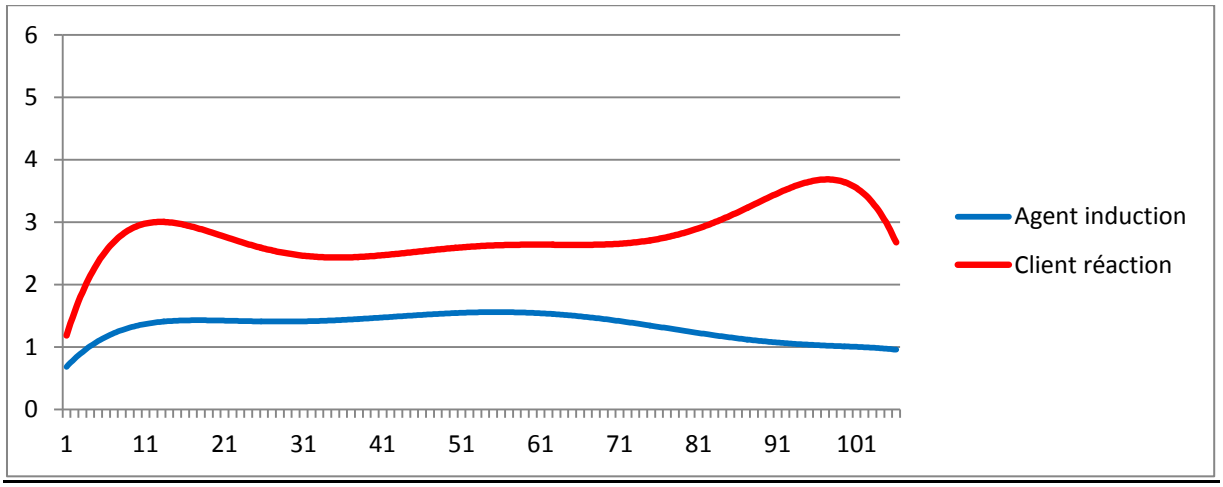


Figure 8-7 Courbes d'induction et de réaction pour un dialogue globalement négatif

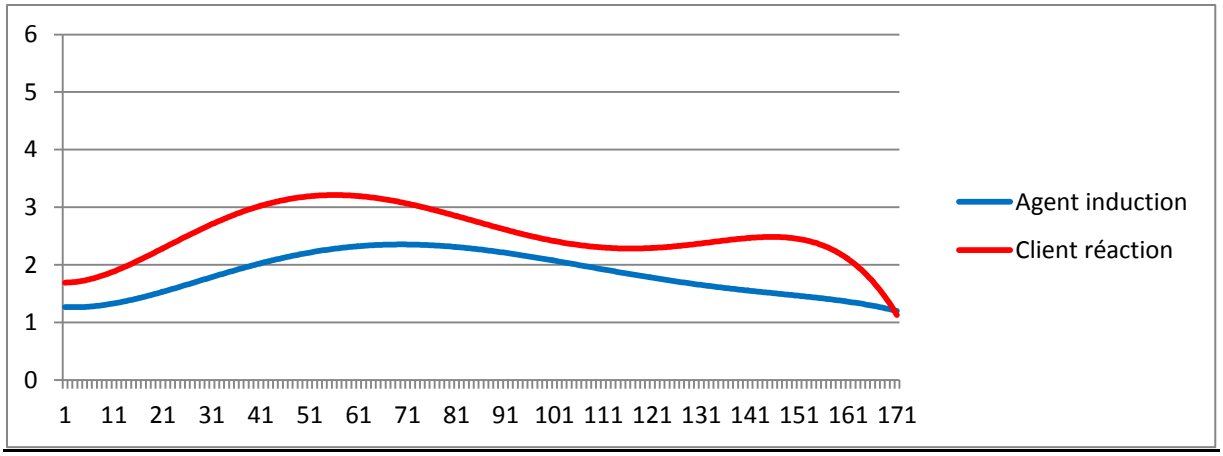


Figure 8-8 Courbes d'induction et de réaction pour un dialogue globalement négatif

Dialogues positifs

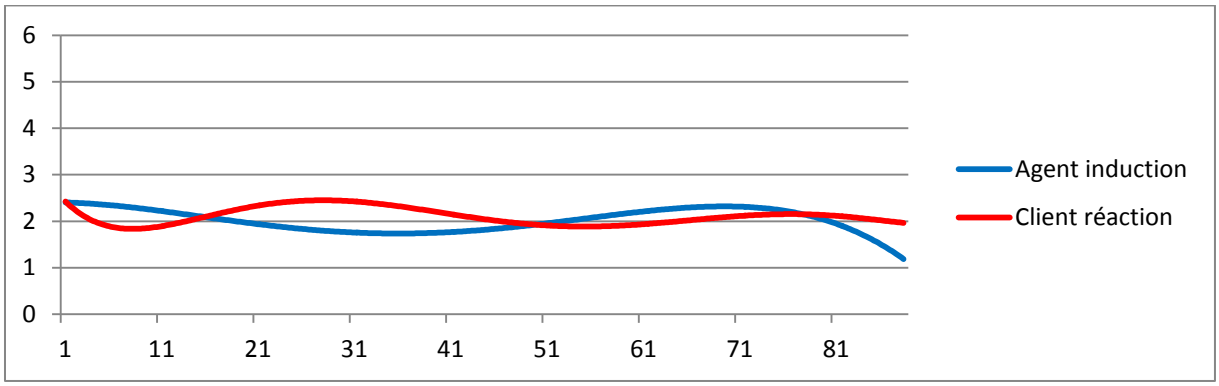


Figure 8-9 Courbes d'induction et de réaction pour un dialogue globalement positif

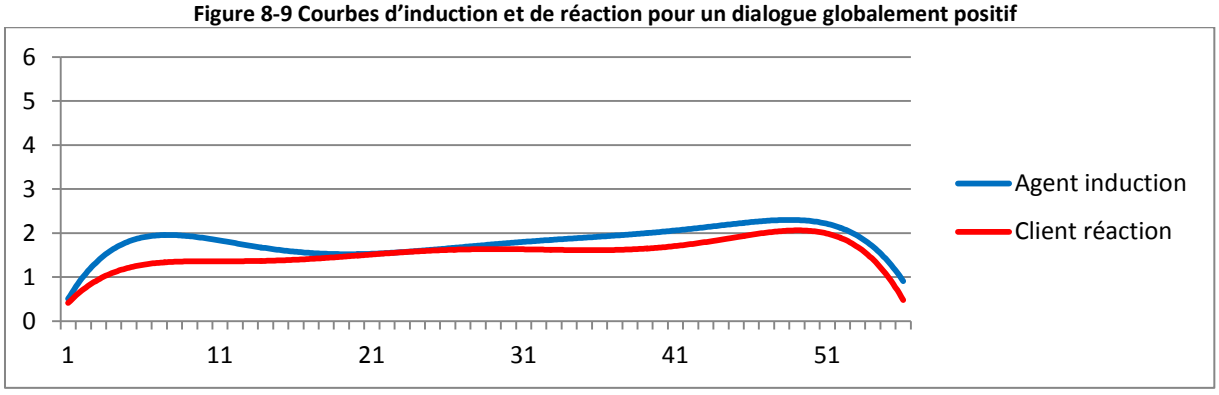


Figure 8-10 Courbes d'induction et de réaction pour un dialogue globalement positif

Les courbes de tendance présentées ci-dessus ont été obtenues à partir des annotations réalisées pour chacun des deux locuteurs. Lorsque nous cherchons à comparer les dialogues ayant une valence globalement négative aux dialogues ayant une valence globalement positive, l'un des premiers éléments que nous pouvons noter est l'amplitude moyenne entre les deux courbes. Cette amplitude se traduit concrètement par une réaction non attendue à une injonction ou une demande particulière de l'un des participants envers l'autre. Nous notons cette amplitude dans le tableau ci-dessous :

	Dialogues positifs	Dialogues négatifs
Amplitude (sur une échelle allant de 1 à 5)	**0.62**	**1.3**

Tableau 8-16 Amplitudes moyennes entre les axes induction et réaction de l'agent et du client pour des dialogues globalement positifs ou négatifs

Afin d'illustrer notre propos, nous analyserons plus en détail les passages des conversations représentées dans les Figure 8-7 et Figure 8-10.

Concernant cette première situation nous prendrons en compte les tours de paroles entre 93 et 103. Nous observons une amplitude moyenne de 2.3 sur cette portion du dialogue. L'amplitude de la courbe est maximum durant ce passage du dialogue. Nous pouvons voir les détails des tours dans le Tableau 8-17 ci-dessous :

Numéro du tour	Locuteur	Emotion	Transcription
93	Mixte	Négatif Négatif	Client : Je n'ai droit à aucune aide en France Agent : Je suis désolée
94	Agent	Négatif	Moi le délai que je peux vous proposer c'est en deux fois
95	Mixte	Négatif	Agent : Sinon je peux pas, je peux pas Client : Faites un petit effort
96	Agent	Colère	Je peux pas, il y a pas d'effort à faire
97	Mixte	Neutre	Client : En 3 mois au moins, en 3 mois. En deux mois c'est impossible Agent : non
98	Agent	Neutre	Non je peux pas monsieur
99	Mixte	Négatif	Client : En 3 mois madame Agent : Non, je peux pas
100	Client	Négatif	Et comment je fais ?
101	Agent	Neutre	Je vous l'ai dis, allez demander au service social
102	Client	Négatif	Moi j'ai pas besoin d'assistante sociale
103	Client	Négatif	Je t'aurais devant moi ma pute, je t'éclaterais la tête

Tableau 8-17 Détail des tours de parole 93 à 103

Si nous considérons le client qui est l'intervenant le plus susceptible d'exprimer ces émotions au cours d'une conversation de ce type, nous remarquons que sur les 7 tours de parole dans lesquels ce dernier intervient, 6 (soit 85% des tours de parole sur la période considérée) ont été annotés avec des émotions négatives. Nous pouvons résumer le Tableau 8-17 à l'aide de la Figure 8-11 ci-dessous :

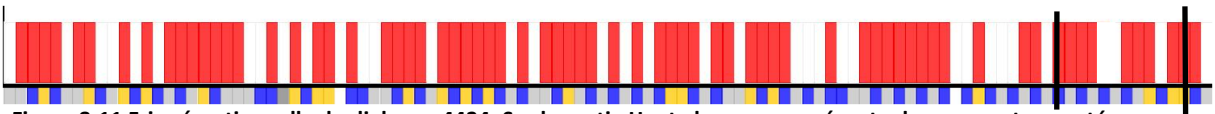


Figure 8-11 Frise émotionnelle du dialogue 4434. Sur la partie Haute le rouge représente des segments annotés comme étant négatifs. Sur la partie basse le gris représente la paroles superposé, le bleu l'agent et le jaune le client

Si nous analysons à présent le dialogue pris en compte sur la Figure 8-10, nous remarquons une amplitude moyenne beaucoup plus faible. Afin d'analyser les spécificités de ce dialogue et tenter d'apporter des éléments discriminants pour la construction de la signature émotionnelle, nous analyserons les segments 24 à 34 du dialogue. En effet, contrairement aux segments sélectionnés précédemment, l'amplitude est minimale sur cette portion de dialogue (0.36 pour cette séquence contre 0.63 en moyenne pour un dialogue globalement positif).

Nous retrouvons dans le Tableau 8-18 ci-dessous le détail des 10 tours de paroles sélectionnés :

Numéro du tour	Locuteur	Emotion	Transcription
24	Agent	Neutre	Pour les tenir tirés [les câbles], pour le vent, pour éviter qu'ils…
25	Client	Neutre	Ha c'est ceux là alors qui étaient cassés
26	Overlap	Neutre	Agent : Ha il était cassé Client : oui, il était cassé
27	Client	Neutre	Il était fendu en deux
28	Overlap	Négatif	Agent : et il vous faisait un bruit Client : oui
29	Client	Neutre	Et alors, donc, il était plus éclairé le soir, bien sûr on le voyait plus
30	Client	Joie	Et là maintenant, ça y est
31	Overlap	Négatif	Agent : Ca doit être aussi s'il est éclairé, pour les avions hélicoptères et tout ça Client : Inaudible
32	Agent	Neutre	Voilà, hélicoptères avions
33	Overlap	Neutre	Client : Comment ? Agent : Pour que ce soit vu par rapport à l'aérien Client : Ha voilà, d'accord, d'accord
34	Client	Joie	Moi j'y connais pas grand-chose, mais en tous cas je sais que c'est réparé et qu'on est bien content

Tableau 8-18 Détail des tours de parole 24 à 34

Ce passage a été retenu pour la particularité qu'il offre lors de la mise en œuvre des tours de paroles pouvant paraître perceptivement négatifs et d'autre perceptivement positifs. Ce phénomène s'explique facilement lorsque nous observons le contenu sémantique du dialogue (les passages négatifs sont des évènements rapportés qui ne sont plus d'actualité). Ainsi a posteriori, même une séquence contenant autant d'éléments positifs que négatifs peut être annotée sans ambigüité (positif dans notre cas). L'une des possibilités que nous proposons, lorsque nous cherchons à dégager cette valence globale, serait l'observation de l'amplitude des courbes actions / rétro actions de chacun des participants. Ainsi la valence globale ne serait plus définie seulement par les émotions exprimées mais également par rapport au décalage observé entre les deux interlocuteurs. Cette observation est d'autant valable dans notre situation que l'agent est tenu, de part la nature commerciale du sujet, à un comportement courtois, mettant en jeu le minimum d'émotions négatives possibles. Nous pouvons voir le résumé du Tableau 8-18 sur la frise émotionnelle/locuteur de la frise ci-dessous :

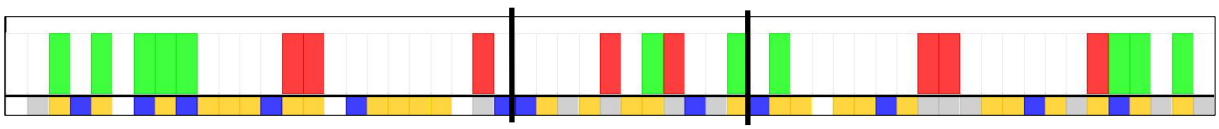
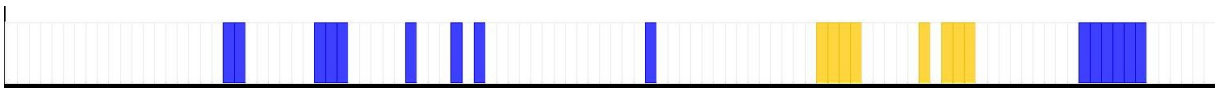


Figure 8-12 Frise émotionnelle du dialogue 9846. Sur la partie haute le rouge représente des segments annotés comme étant négatifs. Sur la partie basse le gris représente de la parole superposée, le bleu l'agent et le jaune le client

8.8.2 Les variables conversationnelles : Le modèle d'interaction dynamique (Luzzati 1995)

Le modèle d'interaction dynamique défini dans (Luzzati 1995) présente deux axes « incident » et « régissant » mis en œuvre dans le cadre d'un dialogue homme-machine. L'axe régissant est défini comme celui permettant au dialogue de progresser pour atteindre un but donné. L'axe incident représente lui les erreurs de compréhension ou d'interprétation qui ne font pas progresser la tâche courante. Cet axe prend en compte des « séquences d'incidents » qui doivent être terminées le plus rapidement possible avec le risque de voir s'instaurer une « incommunication » qui mènerait le dialogue dans une impasse. Afin de rendre l'interaction mesurable, trois variables interactionnelles, incrémentées en fonction des évènements dialogiques, sont introduites (V_r dans le cas d'échanges régissants, V_i dans le cas d'échanges incidents et V_e qui a pour but de mesurer la complexité de l'échange). Ce modèle a donné lieu à deux implémentations dans des systèmes de dialogue DIALORS (Luzzati 1989) et MINIDIAL (Lehuen and Luzzati 1994).

Nous nous sommes inspirés de ce modèle en prenant en compte les relations thématiques des différents tours de parole au cours d'une conversation. Si nous effectuons un parallélisme avec le modèle d'interaction dynamique défini plus haut, nous pouvons définir une séquence incidente lorsqu'une même thématique est abordée à plusieurs reprises au cours du dialogue. Cette séquence incidente serait donc considérée comme porteuse d'une valence particulière au sein du dialogue. Ainsi, dans l'exemple ci-dessous, nous avons annoté chaque segment de manière thématique (représentant en règle générale des sous thèmes de la cause de l'appel). Cette annotation fine, qu'il n'est pas possible de faire à grande échelle étant donné le coût en temps et en argent, a pour simple but d'illustrer notre propos. Il s'agit ici de la représentation thématique d'un dialogue négatif sélectionné dans notre corpus :



Dans cette conversation, deux thèmes principaux (impossibilité de payer la somme due et demande de mise en place d'un délai de paiement), sont mis en œuvre, provoquant chacun au fur et à mesure, la répétition d'évènements incidents. Cette répétition entraine dans ce cas précis la rupture de la conversation et donc un échec de la négociation. Nous pouvons visualiser ces « incidences thématiques » comme des groupements de segments ayant un thème commun. Nous émettons l'hypothèse que si ces ilots se multiplient, la conversation présente une probabilité élevée de se terminer de manière négative. Les thèmes n'appartenant pas aux incidences thématiques sont, quant à eux, attribués à la catégorie « régissant » du modèle d'interaction dynamique de Luzzati.

Pour ce faire, nous utilisons une liste associant des thèmes (« business concept ») à des portions de dialogues. Cette liste obtenue de manière automatique nous a été fournie par la société EDF suivant la méthodologie décrite dans (Danesi and Clavel 2010).

La liste des thèmes considérée sur notre corpus de 44 dialogues est la suivante :

Thèmes	Fréquence d'apparition en fonction de la valence globale du dialogue	
	Positif	Négatif
Courrier	24%	76%
Coordonnées bancaires	40%	60%
Options Internet	100%	0%
Prélèvement	11%	89%
Options Tarifaires Conso	42%	58%
Puissance Phase	40%	60%
Résiliation	29%	71%
Mise en attente	33%	67%
Adresse-code postal	60%	40%
Rendez-vous Intervention	12%	78%
Type Chauffage	67%	33%
Options Tarifaires	66%	34%
Carte bancaire	34%	66%
Aides sociales **(1 seule occurrence)**	100%	0%
Modification Contrat	43%	57%
Moyen de Paiement	25%	75%
Adresse	80%	20%
Numéro Compteur	67%	33%

Tableau 8-19 Répartition des fréquences d'apparition des thèmes selon la valence d'un dialogue

Comme prévu, les thèmes relatifs au paiement (prélèvement, carte bancaire, moyen de paiement) ont tendance à appartenir majoritairement à la classe négative. La mise en attente se retrouve également plus fréquemment au cours des dialogues à problème. En pratique on s'aperçoit que ce thème est très souvent corrélé avec des problèmes techniques difficiles à résoudre et où le télé-conseiller doit demander de l'aide ou un conseil à un autre opérateur. A contrario, d'autres thèmes apparaissent quant à eux plus spontanément dans les dialogues choisis pour leur contenu globalement positif. C'est le cas du thème « option internet » où le télé-conseiller propose au client des outils de suivi en ligne. Les différentes « options tarifaires » sont également présentées dans les dialogues plutôt positifs. Enfin, les demandes d'informations telles que « l'adresse » sont également plus largement présentes dans les dialogues positifs.

Les informations conversationnelles dont nous disposons se présentent sous la forme suivante :

```
<EventList>
    <Event name ="TopicSegment" sid="w9" eid="w551" structure="motif" dur="8.13"
        theme="Resiliation" theme2="Prelevement"/>
</EventList>
```

Sur cet exemple, nous voyons qu'un thème est attribué pour une période donnée (du mot 9 au mot 551) et peut être mentionné à plusieurs reprises au cours d'un même dialogue. La durée en minutes (ici 8 minutes 31) est également mentionnée. Deux thèmes principaux peuvent être attribués à une même portion de dialogue. Nous utiliserons ainsi l'ensemble de ces indices conversationnels (thème1, thème2, durée (mots et minutes)) afin de déterminer si ceux-ci peuvent contribuer à l'amélioration des modèles permettant la classification automatique globale d'un dialogue. Nous gardons bien à l'esprit que le seul thème d'une conversation ne permet pas, en tout état de cause, la

détermination d'une valence, mais nous partons de l'hypothèse que ces derniers fournissent des informations contextuelles pouvant compléter les indices acoustiques utilisés dans les chapitres précédents. C'est dans le but de confirmer ou d'infirmer cette hypothèse que nous commençons par évaluer de manière individuelle les indices conversationnels (thématiques) que nous avons présentés ci-dessus. Nous utiliserons les 44 dialogues qui composent notre corpus pour réaliser les expériences correspondantes.

Nous débuterons les expériences avec les thèmes. Dans cette phase, chaque dialogue sera représenté comme un vecteur mettant en jeu un nombre d'occurrences pour chacun des thèmes présentés dans le Tableau 8-19. Cette représentation sera effectuée en utilisant successivement seul le thème 1, puis dans un second temps en utilisant le thème 1 + le sous thème. L'expérience a été réalisée en utilisant l'algorithme DTNB (decision table and naive bayes) décrit dans le chapitre 5. Le Tableau 8-20 ci-dessous présente les résultats obtenus :

		F-score Dialogues positifs	F-score Dialogue Négatifs	Unweighted Global F-score
Expérience 1	Thème	0.571	0.847	0.709
Expérience 2	Thème + sous thème	0.571	0.847	0.709

Tableau 8-20 Détermination de la valence d'un dialogue en fonction du thème abordé

Nous voyons qu'il est possible, dans le cadre de notre sous sélection de dialogue émotionnel, d'obtenir des résultats sur la valence globale d'un dialogue en prenant les thèmes abordés au cours de la conversation comme indice. L'ajout d'un deuxième thème pour une même portion de dialogue ne produit aucun effet sur les résultats de la classification. Le nombre d'erreurs ainsi que les instances à problème restent les mêmes. Afin d'essayer d'améliorer les résultats, nous avons incorporé la durée estimée de chacun des thèmes de la conversation.

Pour ce faire, nous multiplions le chiffre représentant le nombre d'occurrences d'un thème donné par sa durée au dialogue en pourcentage.

Exemple : le thème « Prélèvement » apparait 8 fois au cours de la conversation mais ne représente que 20% du temps total de dialogue ; le descripteur relatif au thème « prélèvement » sera représenté par le résultat de l'opération 8 * 0.2 = 1.6.

Cette opération a pour but de normaliser les valeurs entre elles en fonction de leur durée de présence au sein d'un dialogue. Les résultats de cette normalisation sont donnés dans le Tableau 8-21 ci-dessous, où nous reprenons les scores de l'expérience 1 comme score de référence :

		F-score Dialogues positifs	F-score Dialogue Négatifs	Unweighted Global F-score
Expérience 1	Thème	0.571	0.847	0.709
Expérience 3	Thème et normalisation en fonction de la durée d'apparition	0.421	0.82	0.620

Tableau 8-21 Détermination de la valence d'un dialogue en fonction du thème abordé

Les résultats obtenus après normalisation des descripteurs thématiques par leur durée d'apparition dans le dialogue, ne semblent pas bénéfiques pour ce qui est de la classification émotionnelle. Concernant les indices conversationnels, nous retiendrons donc seulement les thèmes principaux abordés dans chaque dialogue, ces derniers semblant être suffisamment discriminants pour apporter des améliorations potentielles aux modèles précédemment conçus.

8.9 Combinaisons des indices, structurels et conversationnels pour la classification globale d'un dialogue

Nous avons, pour compléter cette étude, combiné l'ensemble des indices et modèles décrits dans les paragraphes ci-dessus (sorties des modèles acoustiques, indices structurels et conversationnels validés). Nous avons déjà montré dans les paragraphes précédents que la combinaison entre indices acoustiques et indices structurels (relatifs à la durée de différentes unités du discours) apportaient des informations conduisant à une amélioration de nos modèles. Nous testerons donc ici les possibilités offertes par une combinaison des indices acoustiques + conversationnels et des indices acoustiques + structurels + conversationnels. Nous utilisons pour l'élaboration de ces modèles l'algorithme DTNB (decision table and naive bayes) décrit dans le chapitre 5 qui a donné les meilleurs résultats sur cette série d'expériences.

		F-score Dialogues positifs	F-score Dialogue Négatifs	*Unweighted* Global F-score
Expérience 1	Indices conversationnels	0.571	0.847	0.709
Expérience 2	Indices conversationnels + sorties acoustiques	0.667	0.8	0.734
Expérience 3	Indices conversationnels + indices structurels	0.667	0.828	0.747
Expérience 4	Indices conversationnels + indices structurels + sorties acoustiques	**0.799**	**0.893**	**0.846**

Tableau 8-22 Evolution des scores en fonction des indices inclus

La combinaison de l'ensemble des indices de l'expérience 4 (acoustiques, structurels et conversationnels) apportent un gain important (+13.7%) par rapport à l'expérience 1 ou seuls les indices conversationnels étaient inclus. La combinaison d'indices structurels et conversationnels de l'expérience 3 permet de faire progresser les scores concernant les dialogues positifs mais n'apporte de manière globale qu'une amélioration limitée. Ainsi, la combinaison de l'ensemble des indices retenus semble à la fois permettre leur complémentarité et à la fois apporter des précisions intéressantes lorsque nous cherchons à connaitre de manière globale la valence d'un dialogue. En début de chapitre, nous recherchions des indices permettant de déterminer ce que pourraient être des dialogues à problème. Nous sommes conscients que le nombre de dialogues utilisés (44 au total, 27 positifs et 17 négatifs) limite la portée de nos résultats, mais ceux-ci permettent tout de même de

constater que, malgré une chaine de traitement en très grande partie automatique, il était possible d'augmenter de manière significative les scores en incluant des informations dialogiques.

8.10 Conclusion

Nous avons proposé une approche originale d'identification de dialogues à problème. Dans un premier temps, conscient qu'en situation réelle il ne serait pas possible de traiter l'ensemble des dialogues pour une classification, nous avons sélectionné deux parties clés des conversations téléphoniques à notre disposition :

- Les 15 premiers tours de parole mettant en jeu principalement la phase d'ouverture (environ 1minute30 de conversation). Ces premiers tours de parole permettent d'évaluer la tonalité globale de ce que sera la conversation si le problème mentionné par le client n'est pas résolu. Cela permet à un éventuel superviseur de centre d'appels de garder un œil sur ce qui sera potentiellement une conversation à problème.
- Les 15 derniers tours de parole principalement situés dans la phase de clôture de la conversation. L'utilité de cette seconde partie est de permettre la détermination d'appels à problème pour un examen plus approfondi de ce qui a amené la conversation, soit à ne pas aboutir à la résolution d'un problème existant, soit d'une manière plus générale, au mécontentement du client.

Nous avons ainsi vu dans la première partie de l'étude que lorsque ces analyses étaient basées sur un modèle acoustique simple, issu d'une segmentation automatique, le score obtenu (environ 60% de bonne classification) pourrait rendre cette tâche d'analyse très fastidieuse du fait du grand nombre de faux positifs. Toujours en ne nous basant que sur la sortie de nos modèles acoustiques, nous avons réussi à réduire le taux d'erreurs, grâce aux modèles enrichis présentés dans le chapitre précédent, en ramenant notre taux de classification correct à 69% dans le meilleur des cas. Nous restions cependant encore éloignés des performances obtenues lorsque nous basions notre analyse sur des annotations manuelles, ces dernières obtenant une classification correcte de 89%.

Des études ayant, par le passé, montré qu'il était possible d'améliorer les résultats à l'aide d'indices dialogiques (Lee and Narayanan 2005; Galley et al. 2011), nous avons successivement ajouté des indices structurels basés principalement sur les proportions en temps qu'occupaient différentes unités comme les tours de parole, les phases d'ouverture et de clôture, le nombre de tours de parole en fonction du locuteur, etc. Ces informations structurelles ont permis, dans la situation la plus favorable, d'augmenter nos scores de 69% de classification correcte à 76%. Si dans un dernier temps l'utilisation d'indices conversationnels permettant d'identifier le thème globale de portion de dialogue ne donnait pas de résultat supérieur à ceux déjà obtenus, nous avons vu qu'en couplant ces derniers aux indices acoustiques et structurels nous arrivions à un taux de classification correcte de 85%, donc très proche des classifications basées sur des annotations manuelles.

Conclusions et perspectives

8.11 Conclusions

Le sujet de ma thèse portait sur l'analyse et la reconnaissance des émotions lors de conversations collectées dans des centres d'appels. Les axes de recherche ont été principalement la combinaison d'indices linguistiques et paralinguistiques pour construire des modèles plus performants pour la détection des émotions, l'annotation automatique de données émotionnelles pour enrichir les modèles et enfin la classification des appels complets en fonction d'une mesure de la satisfaction du client montrant la qualité de l'interaction. Dans notre étude, nous sommes intéressés à la détermination de la valence émotionnelle positive et négative des dialogues.

La plupart des travaux menés dans la communauté utilisent des données artificielles, nous avions accès à un grand corpus de données collectées dans des situations écologiques dans deux centres de réclamations EDF. Ce volume de données disponible était à la fois un atout et un challenge. D'un côté le volume de données disponibles (1620 heures) dont nous avons extrait un sous-ensemble (100 heures) nous a permis de construire des modèles à partir de données d'une très grande richesse tant en terme de nombres de locuteurs que d'expression émotionnelle. D'un autre côté le traitement manuel (sélection, segmentation, annotation) de connaissances était rendu difficile de part la quantité de données à analyser. Compte tenu de la nature de nos données et de notre but (qualification de la qualité d'une conversation) deux challenges majeurs étaient posés.

Le premier challenge portait sur la façon de détecter de la manière la plus efficace possible des émotions issues de données naturelles. L'expression émotionnelle collectée dans un contexte interactionnel est en effet soumise à des règles de vie en société telles que la politesse ou des enjeux commerciaux. De ce fait chaque locuteur ne peut exprimer complètement librement son état affectif. Dans ce contexte des stratégies conversationnelles ou des moyens d'expressions plus complexes (recours à une expression émotionnelle de la colère uniquement par la voie sémantique par exemple) sont utilisés. La prise en compte de plusieurs canaux expressifs (acoustique mais aussi lexical et sémantique) s'est alors révélée indispensable. Les résultats obtenus sur les modèles de fusions d'indices acoustiques et linguistiques nous ont permis d'obtenir des résultats sensiblement plus élevés ; Cependant lorsque nous nous sommes placés dans un contexte plus proche du réel, la grande variabilité du nombre de locuteurs nous a obligé à considérer l'importance d'entrainer les modèles sur une quantité plus grande données disponibles. En conservant une approche hybride nous avons donc décidé d'utiliser des informations acoustiques et linguistiques pour enrichir nos modèles à l'aide de données supplémentaires, sélectionnées de manière complètement automatique dans notre corpus. Cette technique nous a ainsi permis de doubler la taille de notre corpus d'apprentissage et d'obtenir des gains intéressants sur l'ensemble des classes émotionnelles considérées. L'utilisation de la totalité des données disponibles dans le corpus, soit 1520 heures supplémentaires permettraient potentiellement un gain en ce qui concerne les scores de détection grâce à la diversification des types de locuteurs pris en compte. Il faut noter que la courbe de gains obtenus suite à des augmentations du corpus d'apprentissage aurait très probablement tendance à suivre une évolution logarithmique, tout en augmentant de manière très significative les besoins en temps de calculs et en espace mémoire. Cette approche serait donc, par exemple, à coupler avec une

méthode efficiente de sélection de descripteurs dans le but de compenser l'augmentation du volume de données à traiter par une réduction de la dimensionnalité du problème de classification.

Le second challenge est la prise en compte de l'interaction entre les participants de la conversation. Les dimensions d'implications affectives que nous avons proposées avaient pour but de modéliser cette interaction en prenant en compte les réactions mises en œuvre par chacun des locuteurs. En accord avec les théories issues de l'analyse conversationnelle cette méthode nous a permis de mettre en lumière des liens forts entre interaction et expression affective. Finalement, nous avons opté pour la qualification globale d'un appel à travers l'analyse de portions de dialogues que nous considérions comme représentatives. Les résultats issus des modèles de détections nous ont permis de représenter un dialogue comme une succession de tours de paroles positifs ou négatifs et de rechercher d'éventuelles combinaisons prototypiques pour un dialogue considéré comme globalement positif ou négatif. L'ajout d'informations issues de la structure de la conversation (la durée, le nombre de tours de parole de chaque locuteur, etc.) a également permis d'ajouter des informations affinant notre approche et nous permettant d'obtenir des gains intéressants en termes de score de reconnaissance. Il faut cependant mentionner la limite de notre approche. Certains indices dialogiques utilisés étaient obtenus manuellement (la délimitation des différentes phases du dialogue ou la mesure des dimensions d'implication affective par exemple). Des recherches supplémentaires seraient nécessaires pour obtenir l'ensemble de ces indices sémantiques et dialogiques de manière totalement automatique.

8.12 Perspectives

Plusieurs prolongements des travaux effectués dans le cadre de cette thèse peuvent être effectués. D'une manière générale la prise en compte de l'interaction est un enjeu majeur pour le traitement du type de données que nous avons utilisé dans cette thèse. Cette prise en compte implique, afin de pouvoir être obtenue, la résolution d'un ensemble de problèmes sous-jacents :

i) La parole superposée

A l'heure actuelle la plupart des échanges téléphoniques sont réalisés à partir d'un canal unique. Ce mode de communication implique des portions de conversations contenant de la parole superposée. Ce type de contenu dialogique est couramment rencontré au cours d'une interaction naturelle. A défaut de pouvoir traiter ce type de contenu (transcription par exemple) il pourrait être intéressant de pouvoir le détecter afin de définir un éventuel point de conflit au cours du dialogue.

Ce problème pourrait être partiellement résolu avec la démocratisation des systèmes de téléphonie sur IP. L'utilisation de deux canaux de communication permettrait non seulement la détection de segments contenant de la parole superposée mais également l'obtention une transcription automatique plus propre et plus simple à exploiter (Barras 2011).

ii) Ajout du contexte à la détection émotionnelle

La prise en compte de la séquentialité des segments à travers i) la détection de patrons émotionnels agent/client spécifiques et ii) la partie du dialogue dans laquelle intervient ce patron pourrait apporter des indices intéressants et permettrait une détermination plus précise des dialogues à problème. Un dialogue commençant de manière négative mais se terminant avec un patron agent/client notoirement positif permettrait une classification plus aisée des types de conversations.

iii) Utilisations de systèmes multi-classes acoustiques-linguistiques.

Les données naturelles mettent en jeu des émotions complexes liées à des stratégies conversationnelles. Du côté du client les buts à atteindre comme la nécessité de l'intervention de la part de l'entreprise mais également les normes sociales comme la politesse viennent modifier l'expression émotionnelle. Des émotions telles que la colère froide (expression négative de faible activation) pourraient être une classe distincte de la classe colère chaude. Dans le même ordre d'idée des prédictions sur des combinaisons opposées (colère/joie) pourraient aider à mettre en lumière des phénomènes émotionnels complexes comme l'ironie par exemple.

Au delà des pistes mentionnées dans les points ci-dessus il semble que des recherches sont encore nécessaires avant de pouvoir incorporer les systèmes de détection des états affectifs dans des dialogues spontanés dans un contexte opérationnel. Si chaque type d'indice utilisé de manière séparée (acoustiques, lexicaux, sémantiques, dialogiques) n'est pas suffisant pour obtenir des résultats fiables sur la qualité d'un dialogue, la combinaison de l'ensemble de ces informations semble être la voie la plus prometteuse à l'heure actuelle. L'utilisation de méthodes mettant en jeux des indices issus d'analyses à un niveau pragmatique seraient également, à notre avis, un complément d'informations à ne pas négliger ; Les données disponibles publiquement (*open data*) constituent d'hors et déjà une source d'information importante pouvant être exploitée par ces méthodes. La caractérisation d'éléments propres à un locuteur donné (traits de personnalité notamment) qui est un thème ayant récemment attiré l'attention d'un nombre croissant de chercheurs pourrait venir compléter de manière intéressante les travaux préalablement menés sur le thème des émotions.

Quoi qu'il en soit, et c'est encore plus vrai lorsque nous essayons de formaliser des phénomènes touchant à l'humain comme les états affectifs, il semble qu'une approche pluridisciplinaire mêlant l'ensemble des sources de données disponibles est indispensable pour l'élaboration des systèmes futurs.

Table des Figures

Figure 1-1 : *Feeltrace (Cowie et al. 2000)* 21
Figure 1-2 La roue des émotions (Scherer 2005) utilisant une approche mixte dimensions/étiquette 22
Figure 1-3 : Types d'attributs de l'Appraisal Theory et options du type appréciation avec exemples d'adjectifs 24
Figure 1-4 : Analyse du groupe d'évaluations *not very happy* 24
Figure 3-1 Le double encodage (Fonagy 1983) ou l'utilisation conjointe des canaux de communication linguistique et acoustique 38
Figure 3-2 Description de la notion de qualité vocale (Garnier et al. 2005) 47
Figure 3-3 Présentation des différents niveaux d'information impliqués dans une communication verbale (partie droite). Le message parlé peut être décomposé par différents supports présentés dans la partie gauche. (Beller 2008) 50
Figure 3-4 Illustration des concepts de fusion tardive (haut) et de fusion précoce (bas) 51
Figure 3-5 Profil de décision établi d'après les valeurs de confiance obtenues en sortie des classifieurs 56
Figure 0-1 Les différents types d'indices utilisés pour la suite de l'étude 63
Figure 4-1 Espace émotionnel construit par un utilisateur se basant sur une série d'étiquette et sur les dimensions activation et valence (emoTool) 69
Figure 4-2 Les deux dimensions d'implication affective utilisées pour la mise en relation des segments d'une conversation 72
Figure 4-3 Interface d'annotation d'emoTool pour l'annotation des conversations du corpus emoVox 74
Figure 4-4 Interface d'annotation continue destinée à des extraits de dialogues 78
Figure 4-5 Interface pour l'annotation discrète des émotions 79
Figure 4-6 Représentation graphique des annotations continues (obtenu sur un extrait complet) 80
Figure 4-7 Discrétisation du flux continu obtenu d'après l'annotation continue en segment (d'après les times code des segments utilisés pour l'annotation discrète). Chaque point représente le début d'un segment 81
Figure 4-8 Répartition des segments après annotation de segments issus d'un découpage automatique des conversations 87
Figure 4-9 Répartition des segments après annotation de segments issus d'un découpage manuel des conversations 87
Figure 4-10 Durée moyenne d'un segment selon la catégorie émotionnelle et le type de segmentation 88
Figure 4-11 Répartition des émotions exprimées en fonction du rôle du locuteur 89
Figure 4-12 Répartition par âge et par sexe pour les locuteurs du corpus 90
Figure 4-13 Répartition des tours de paroles de paroles consensuels par sexe 90
Figure 5-1 Briques procédurales impliquées dans un système de détection des émotions 94
Figure 5-2 Exemple d'un hyperplan optimal(Manning et al. 2008) 103
Figure 5-3 Exemple de matrice de confusion. Les instances correctement classifiées apparaissent sur la diagonale de la matrice 107
Figure 6-1 Représentation de la théorie de l'évaluation selon (Martin and White 2005) 130

Figure 6-2 Proportion de segments émotionnels, bruités et non consensuels d'un dialogue complet sélectionné au hasard 137
Figure 7-1 Processus de sélection de nouvelles instances pour l'enrichissement de l'ensemble d'apprentissage 144
Figure 7-2 Exemple de découpage d'un segment long (d'une durée supérieure à 5 secondes) 145
Figure 7-3 Les différentes segmentations des instances supplémentaires utilisées lors de l'enrichissement des modèles 150
Figure 8-1 Exemple de dialogue problématique 1 (annotateur 1) (chaque segment a la même taille) 163
Figure 8-2 Exemple de dialogue problématique 1 (annotateur 1) (la taille des segments est fonction de leur durée) 163
Figure 8-3 Exemple de dialogue problématique 1 (annotateur 2) (la taille des segments est fonction de leur durée) 163
Figure 8-4 Exemple de dialogue problématique 2 (annotateur 1) (chaque segment a la même taille) 163
Figure 8-5 Exemple de dialogue problématique 2 (annotateur 1) (la taille des segments est fonction de leur durée) 163
Figure 8-6 Exemple de dialogue problématique 2 (annotateur 2) (la taille des segments est fonction de leur durée) 163
Figure 8-7 Courbes d'induction et de réaction pour un dialogue globalement négatif 170
Figure 8-8 Courbes d'induction et de réaction pour un dialogue globalement négatif 171
Figure 8-9 Courbes d'induction et de réaction pour un dialogue globalement positif 171
Figure 8-10 Courbes d'induction et de réaction pour un dialogue globalement positif 171
Figure 8-11 Frise émotionnelle du dialogue 4434. Sur la partie Haute le rouge représente des segments annotés comme étant négatifs. Sur la partie basse le gris représente la paroles superposé, le bleu l'agent et le jaune le client 172
Figure 8-12 Frise émotionnelle du dialogue 9846. Sur la partie haute le rouge représente des segments annotés comme étant négatifs. Sur la partie basse le gris représente de la parole superposée, le bleu l'agent et le jaune le client 173

Listes des tableaux

Tableau 1-1 Classification des différents états affectifs (Scherer and Peper 2001) 19
Tableau 1-2 : Les émotions primaires (élémentaires) 20
Tableau 2-1 Exemple de corpus émotionnels disponibles (Schuller et al. 2009) 32
Tableau 2-2 Les types de données rassemblées et annotées pour la base de données HUMAINE (Douglas-Cowie et al. 2011) 33
Tableau 2-3 Récapitulatif de corpus « call center » utilisés pour la construction de modèles de détection d'émotions 34
Tableau 3-1 Synthèse de la revue de résultats obtenus par (Juslin and Laukka 2003) pour la caractérisation des émotions. Les indices sur fond gris sont les plus couramment utilisés pour la détection d'émotions 40
Tableau 3-2 Récapitulatif des corpus et indices utilisés lors d'études portant sur les centres d'appels 60
Tableau 4-1 Récapitulatif des macro-classes / micro-classes utilisées pour l'annotation de emoVox. 70
Tableau 4-2 Récapitulatif des champs d'applications des dimensions et labels utilisés pour le schéma d'annotation 75
Tableau 4-3 Signification des taux d'accord obtenus après le calcul du Kappa de Cohen 76
Tableau 4-4 Exemples de coefficients linéaires calculés pour un problème à 3 classes 77
Tableau 4-5 Exemple de discrétisation d'une annotation continue 80
Tableau 4-6 Score d'accord intra annotateur pour la dimension activation obtenue sur les 2 phases d'annotations continues 81
Tableau 4-7 Résultats des annotations discrètes/continues obtenus sur un sous ensemble du corpus emoVox 82
Tableau 4-8 Scores inter-annotateurs obtenus d'après une annotation continue et discrète 82
Tableau 4-9 Scores d'accord pour les dimensions d'implications affectives 83
Tableau 4-10 Scores obtenus sur les axes d'induction et de réaction en fonction de l'état émotionnel exprimé 83
Tableau 4-11 Corrélation entre un haut score d'induction et les valeurs des dimensions plaisir, activation, contrôle 84
Tableau 4-12 Corrélation entre un haut score de rétroaction et les valeurs des dimensions plaisir, activation, contrôle 84
Tableau 4-13 Exemple d'alignement émotionnel du client sur l'agent 85
Tableau 4-14 Tentative de maintien de contrôle de la part de l'agent en ne répondant pas à d'éventuelles provocations 85
Tableau 4-15 Tentative de maintien de contrôle de l'agent en positivant la situation 86
Tableau 4-16 Taux d'accord en fonction du type de segmentation 86
Tableau 4-17 Nombres de segments retenus pour la construction des modèles automatiques 89
Tableau 5-1 Récapitulatif des descripteurs bas niveau (LLD), de leur dérivée et des fonctionnelles associées (Schuller et al. 2009) 98
Tableau 5-2 Récapitulatif des descripteurs bas niveau (LLD) et des fonctionnelles associées 98
Tableau 5-3 Termes ayant le plus haut taux de spécificité positive en fonction du locuteur 100
Tableau 5-4 Termes ayant le plus haut taux de spécificité positive en fonction de l'émotion 101

Tableau 5-5 Termes ayant le plus haut taux de spécificité positive en fonction de l'émotion et d'une sélection du vocabulaire en fonction de la catégorie grammaticale 102
Tableau 5-6 Fonctions noyaux utilisées pour des problèmes utilisant des classifieurs de type SVM. 104
Tableau 5-7 Répartition des instances utilisées pour la construction des modèles de détection..... 109
Tableau 5-8 Descripteurs retenus dans l'ensemble Praat + bibliothèques LIMSI..... 110
Tableau 5-9 Descripteurs retenus dans l'ensemble challenge emotion 2009 et openEAR 111
Tableau 5-10 Variation des scores du modèle de détection en fonction de l'ensemble de paramètre utilisé 112
Tableau 5-11 Variation des scores du modèle de détection en fonction de l'ensemble de paramètre utilisé 113
Tableau 5-12 Résultats d'une détection des émotions avec des modèles dépendants du type de locuteur (client/agent) 114
Tableau 5-13 Résultats des différentes configurations pour la classification lexicale..... 115
Tableau 5-14 Résultats en fonction de l'ajout des classes grammaticales 116
Tableau 6-1 Répartition du nombre d'instances en classes positives, négatives et neutres..... 121
Tableau 6-2 Résultats du système de reconnaissance de la parole en fonction du type de parole considéré..... 123
Tableau 6-3 Résultats des expériences 1 à 3 sur le canal Paralinguistique..... 124
Tableau 6-4 Résultats des expériences 1 à 3 sur le canal linguistique..... 125
Tableau 6-5 F-score et intervalle de confiance en fonction des sous-ensembles 1 à 3 (du plus artificiel au plus réaliste) 127
Tableau 6-6 Poids attribués aux différentes modalités retenues pour le calcul de la polarité des segments (Cailliau and Cavet 2010) 129
Tableau 6-7 Exemple de patrons dérivés d'expressions figées et leurs énoncés dérivés associés..... 131
Tableau 6-8 Fusion des indices lexicaux et paralinguistiques 132
Tableau 6-9 Gains obtenus pour chaque classe d'émotion en combinant les indices acoustiques et lexicaux..... 133
Tableau 6-10 Résultats Gains obtenus après fusion des indices linguistiques (lexicaux, syntaxiques, sémantiques)..... 134
Tableau 6-11 Scores obtenus en fonction des différentes combinaisons acoustiques/linguistiques 134
Tableau 6-12 Scores obtenus suite à une fusion des descripteurs acoustiques et de l'ensemble des descripteurs linguistiques..... 134
Tableau 6-13 Répartition des prédictions sur une sélection de 264 dialogues complets et non annotés issus du corpus 136
Tableau 6-14 Répartition des classes émotionnelles au sein des corpus « colère » et « joie »..... 136
Tableau 7-1 L'algorithme self training..... 141
Tableau 7-2 Algorithme de co-apprentissage 142
Tableau 7-3 Evolution des F-score en fonction du nombre d'instances présentes dans l'ensemble d'entrainement 143
Tableau 7-4 Répartition des instances représentant des degrés d'activation faibles et forts ainsi que les émotions associées..... 145
Tableau 7-5 F-score obtenu lors du test des modèles de détection de l'activation 146
Tableau 7-6 Nombre d'instances selon le mode de segmentation et selon la classe émotionnelle . 147
Tableau 7-7 Scores obtenus après enrichissement de l'ensemble d'entrainement avec des instances annotées automatiquement 148

Tableau 7-8 Matrice de confusion représentant l'accord entre une sélection automatique et un annotateur humain 149
Tableau 7-9 Scores obtenus après une sélection de la classe positive avec une approche syntaxique et sémantique 149
Tableau 7-10 Variation des scores obtenus en fonction des modes de segmentation 150
Tableau 7-11 Gains obtenus après enrichissement des modèles sur des données segmentées manuellement 151
Tableau 7-12 Scores en fonction de la portion de segments conservés pour l'entrainement 152
Tableau 7-13 Scores obtenus après une sélection manuelle de descripteurs 153
Tableau 7-14 Scores obtenus sur un modèle lexical non enrichi 153
Tableau 7-15 Scores obtenus après enrichissement des modèles lexicaux par des instances sélectionnées automatiquement 154
Tableau 7-16 Nombre d'instances selon le mode de la segmentation et la classe émotionnelle de la seconde sélection 155
Tableau 8-1 Durée moyenne des dialogues non-négatifs et négatifs 160
Tableau 8-2 Durée moyenne des tours de paroles négatifs et neutres 160
Tableau 8-3 Longueur moyenne d'un tour de parole agent / client sur un corpus de dialogues choisis pour leurs contenu émotionnellement marqués 160
Tableau 8-4 Proportion moyenne des différentes phases de conversation (Balague et al. 2011) 161
Tableau 8-5 Proportion moyenne des différentes phases de conversations dans les 44 dialogues de notre corpus 161
Tableau 8-6 Proportion de tours négatifs exprimés par les locuteurs durant chacune des phases des dialogues 161
Tableau 8-7 Résultats d'après une annotation manuelle. 162
Tableau 8-8 Score de détection au niveau dialogique à partir d'une annotation automatique 164
Tableau 8-9 Score de détection au niveau dialogique à partir d'une annotation manuelle et d'informations relatives au locuteur 165
Tableau 8-10 Extrait des scores de détection au niveau dialogique à partir d'une annotation automatique et d'informations relatives au locuteur 166
Tableau 8-11 Résultats d'après une modélisation du dialogue sous forme de « sac de NGRAM » ... 167
Tableau 8-12 Séquences ayant une spécificité positive pour la détermination de la valence globale d'un dialogue 167
Tableau 8-13 indices et groupements d'indices structurels retenus pour la classification globale de la valence d'un dialogue 168
Tableau 8-14 Récapitulatif des gains apportés par chacun des types d'indices structurels considérés 168
Tableau 8-15 Scores obtenus après fusion des indices structurels et acoustiques 169
Tableau 8-16 Amplitudes moyennes entre les axes induction et réaction de l'agent et du client pour des dialogues globalement positifs ou négatifs 172
Tableau 8-17 Détail des tours de parole 93 à 103 172
Tableau 8-18 Détail des tours de parole 24 à 34 173
Tableau 8-19 Répartition des fréquences d'apparition des thèmes selon la valence d'un dialogue . 175
Tableau 8-20 Détermination de la valence d'un dialogue en fonction du thème abordé 176
Tableau 8-21 Détermination de la valence d'un dialogue en fonction du thème abordé 176
Tableau 8-22 Evolution des scores en fonction des indices inclus 177

Bibliographie

Abrilian S. (2007). Représentation de Comportements Emotionnels Multimodaux Spontanés : Perception, Annotation et Synthèse. Computer Science. Orsay, Paris XI University. **Thèse de doctorat**.

Abrilian S., Devillers L. and Martin J.-C. (2006). Annotation of Emotions in Real-Life Video Interviews: Variability between Coders. 5th Int. Conf. Language Resources and Evaluation (LREC'2006). Genoa, Italy.

Argaman O. (2010). "Linguistic Markers and Emotional Intensity." Journal of Psycholinguistic Research **39**(2): 89-99.

Arunachalam S., Gould D., Anderson E., Byrd D. and Narayanan S. (2001). Politeness and frustration language in child–machine interactions. Eurospeech. Aalborg, Denmark**:** 2675–2678.

Aubergé V. (2002). Prosodie et émotion. Deuxièmes assises nationales du GDR I3

Avanzi M., lacheret-Dujour A., Obin N. and Victorri B. (2011). "Vers une modélisation continue de la structure prosodique: le cas des proéminences syllabiques." French Language Studies **21**: 53-71.

Avanzi M., Obin N., Lacheret-Dujour A. and Victorri4 B. (2011). Toward a Continuous Modeling of French Prosodic Structure: Using Acoustic Features to Predict Prominence Location and Prominence Degree. InterSpeech 2011. Florence**:** 2033-2036.

Balague C., Ducourant H. and Eloire F. (2011). Description et analyse des données empiriques : Qu'est-ce qu'une conversation en centre d'appel ? Lille, lem-cnrs.

Barras C. (2011). Reconnaissance automatique des locuteurs pour la structuration en tours de parole. Informatique. Orsay, Université Paris XI. **Habilitation à Diriger des Recherches:** 80.

Barras C., Zhu X., Meignier S. and Gauvain J.-L. (2006). "Multi-stage speaker diarization of broadcast news." IEEE Trans. Audio, Speech & Language Processing **14**(5): 1505-15012.

Batliner A., Fisher K., Huber R., Spilker J. and Noth E. (2003). "How to Find Trouble in Communication." Speech Communication **40**: 117-143.

Batliner A., Schuller B., Seppi D., Steidl S., Devillers L., Vidrascu L., Vogt T., Aharonson V. and Amir N. (2011). The Automatic Recognition of Emotions in Speech. Emotion-Oriented Systems Springer**:** 71-99.

Batliner A., Seppi D., Steidl S. and Schuller B. (2010). "Segmentic into adequate units for automatic recognition of emotio-related episodes : A speech-based Approach." Advances in human-computer interaction **10**.

Batliner A., Steidl S., Schuller B., Seppi D., Laskowksi K., Vogt T., Devillers L., Vidrascu L., Amir N., Kessous L. and Aharonson V. (2006). CEICES : Combining Efforts for Improving automatic Classification of Emotional user States: a « forced co-operation » initiative. fifth slovenian and first international language technologies conference. Slovenia.

Beller G. (2008). Transformation of expressivity in speech. The Role of Prosody in the Expression of Emotions in English and in French. P. Lang, Peter Lang.

Beller G. and Rodet X. (2007). Content-based transformation of the expressivity in speech. ICPhS. Germany.

Benveniste E. (1970). "L'appareil formel de l'énonciation." Langages **5**(17): 12-18.

Bloom K., Garg N. and Argamon S. (2007). Extracting Appraisal Expressions. NAACL HLT. Rochester, New York, USA.**:** 308-315.

Bloom K., Stein S. and Argamon S. (2007). Appraisal extraction for news opinion analysis. Sixth NTCIR workshop meeting on evaluation of information access technologies: Information retrieval, question answering, and cross-lingual information access Tokyo, Japan**:** 279-285.

Blum A. and Mitchell T. (1998). Combining labeled and unlabeled data with co-training. Workshop on Computational Learning Theory, Madison, Wisconsin, USA.

Boersma P. and Weenink D. (2009). Praat: doing phonetics by computer.
Borda J. C. (1781). Mémoire sur les élections au scrutin. Histoire de l'académie Royale des Sciences. Paris.
Boulakia G. (1979). Phonosyntaxe: études psychoaccoustiques et linguistiques sur l'intonation et ses rapports avec la syntaxe, notament l'ordre des mots en français. Linguistique. Paris, Université Paris VII.
Breazeal C. (2003). "Emotion and sociable humanoid robots." International Journal of Human-Computer Studies **59**: 119-155.
Breck E., Choi Y. and Cardie C. (2007). Identifying expressions of opinion in context. 20th international Joint Conference on Artifical intelligence Hyderabad, India.
Brendel M., Zacharelli R. and Devillers L. (2010). A Quick Sequential Forward Floating Feature Selection Algorithm for Emotion Detection from Speech. InterSpeech 2010. Makhuari.
Brill E. (1995). "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging." Computational Linguistics.
Bulut M., Shrikanth S., Narayanan S. and Syrdal A. K. (2002). Expressive speech synthesis using a concatenative synthesizer. ICSLP.
Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W. and Weiss B. (2005). A database of german emotional speech. Interspeech 2005. Lissabon, Portugal**:** 1517-1520.
Burkhardt F., Polzehl T., Stegmann J. and Metze F. (2009). Detecting real life anger. International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. Taipei.
Busso C., Deng Z., Yildirim S., Bulut M., Lee C. M., Kazemzadeh A., Lee S., Neumann U. and Narayanan S. (2004). Analysis of Emotion Recognition using Facial Expressions. 6th International Conference on Mutlmodal Interfaces (ICMI 2004). State College, PA, USA.
Cailliau F., Bouffier A., Kucharski M., Guillemin-Lanne S., Devillers L. and Vaudable C. (2010). "Etats de l'art sur la détection des émotions, des opinions et des sentiments dans des interactions clients agents des centres d'appels." Livrable du projet FUI voxFactory.
Cailliau F. and Cavet A. (2010). "Analyse des sentiments et transcription automatique : modélisation du déroulement de conversations téléphoniques." TAL **51**(3): 131-154.
Cailliau F. and Giraudel A. (2008). Enhanced Search and Navigation on Conversational Speech. Searching Spontaneous Conversational Speech Singapour.
Campbell N. and Mokhtari P. (2003). Voice quality : the 4th prosodic dimension. In 15th International Congress of Phonetic Sciences. Barcelone, Espagne**:** 2417-2420.
Castellano G., Kessous L. and Caridakis G. (2008). "Emotion recognition through multiple modalities: face, body gesture, speech." Emotion **4868**: 92-103.
Charaudeau P. (1992). Grammaire du sens et de l'expression. H. Edition**:** 576-629.
Chen Y.-W. and Lin C.-J. (2005). Combining SVMs with Various Feature Selection Strategies. Feature extraction, foundations and applications.
Chih-Chung, Chang and Lin C.-J. (2001). LIBSVM : a library for support vector machines.
Clark H. H. and Schaefer E. F. (1989). " Contributing to Discourse." Journal of Cognitive Science **13**: 259-294.
Clavel C. (2007). Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales. Traitement du Signal et des Images, ENST. **Thèse de doctorat**.
Clavel C. (2007). Analyzes and recognition of the acoustic demonstrations of the emotions of the fear type in abnormal situations. Orsay, Paris XI.
Clavel C., Adda G., Cailliau F., Garnier-Rizet M., Cavetc A., Chapuis G., Courcinous S., Danesi C., Daquo A.-L., Deldossif M., Guillemin-Lanne S. and Suignard P. (2012). "Spontaneous Speech and Opinion Detection: Mining Call-centre Transcripts." Computer Speech and Language.
Clavel C., Vasilescu I., Richard G. and Devillers L. (2007). Voiced and Unvoiced Content of fear-type emotions in the SAFE Corpus. Proceedings of Speech Prosody
Cohen J. (1960). "A coefficient of agreement for nominal scales." Educational and Psychological Measurement **20**: 27-46.
Cornuéjols A., Miclet L. and Kodratoff Y. (2002). Apprentissage artificiel : Concepts et algorithmes.

Cowie R., Douglas-cowie E., Apolloni B., TaylorJ, Romano A. and Fellenz W. (1999). What a neural net needs to know about emotion words, Word Scientific Engineering Society, Society Press.

Cowie R., Douglas-Cowie E., Savvidou S., McMahon E., Sawey M. and Schröder M. (2000). Feeltrace: an instrument of recording perceived emotion in real-time. ISCA ITRW on Speech and Emotion. Newcastle**:** 19-24.

Craggs R. (2004). Annotating emotion in dialogue – Issues and Approaches. 7th Annual CLUK Research Colloquium University of Birmingham.

D'Mello S. and Graesser A. (2010). "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features." User Modeling and User-adapted Interaction **20**: 147-187.

Danesi C. and Clavel C. (2010). Impact of Spontaneous Speech Features on Business Concept Detection: a Study of Call-centre Data. ACM Multimedia Workshop on Searching Spontaneous Conversational Speech. Firenze, Italy.

Darwin C. (1872). The Expression of the Emotions in Man and Animals,. London.

Devillers L., Cowie R., Matin J.-C., Douglas-Cowie E., Abrilian S. and Mcrorie M. (2006). Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches. 5th international conference on Language Resources and Evaluation (LREC 2006). Genoa

Devillers L. and Martin J.-C. (2008). Coding Emotional Events in Audiovisual Corpora. 6th Int. Conf on Language Resources and Evaluation (LREC'2008). Marrakech, Morocco.

Devillers L., Vasilescu I. and Lamel L. (2003). Emotion Detection in a task-oriented Dialog Corpus. IEEE International Confererence on Multimedia, ICME. Batimore.

Devillers L., Vasilescu I. and Lamel L. (2003). Emotion detection in task-oriented dialogs corpus. ICME. IEEE. Baltimore.

Devillers L., Vaudable C. and Chastagnol C. (2010). Real-life emotion-related states detection in call centers. InterSpeech 2010. Makhuari, Japan.

Devillers L., Vidrascu L. and Lamel L. (2005). "Emotion detection in real-life spoken dialogs recorded in call center." Journal of Neural Networks, numéro spécial **volume 18**.

Devillers L., Vidrascu L. and Layachi (2010). Automatic detection of emotion from vocal expression. A blueprint for an affectively competent agent, Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing. K. Scherer, T. Bänziger and E. Roach, Oxford University Press**:** 232-244.

Di Cristo A. (2007) "De la métrique et du rythme de la parole ordinaire : l'exemple du français ".

Dini L. and Mazinni G. (2002). Opinion classification through information extraction. CELI. Turin, Italy.

Douglas-Cowie E., Campbell N., Cowie R. and Roach P. (2003). "Emotional speech: Towards a new generation of databases." Speech Communication **40**(1-2): 33-60.

Douglas-Cowie E., Cowie R., Sneddon I., Cox C., Lowry O., McRorie M., Martin J.-C., Devillers L., Abrilian S., Batliner A., Amir N. and Karpousis K. (2007). The HUMAINE database: Adressing the collection of annotation of naturalistic and induced emotional data. Affective computing and intelligent interaction (ACII 2007).

Douglas-Cowie E., Cox C., Martin J.-C., Laurence Devillers, Cowie R., Sneddon I., McRorie M., Catherine Pelachaud, Peters C., Lowry O., Batliner A. and Hönig F. (2011). The HUMAINE Database. Emotion Oriented Systems. P. Petta, C. Pelachaud and R. Cowie.

Dubreil E., Vernier M., Monceaux L. and Daille B. (2008). Annotating opinion – evaluation of blogs LREC Workshop on Sentiment Analysis: Metaphor, Ontology and Terminology Marrakech.

Ekman P. (1999). Basic Emotion. Handbook of Cognition and Emotion. T. D. a. M. Power. *San Francisco*.

Ekman P. and Friesen W. V. (1975). Unmasking the face - A guide to recognizing emotions from facial clues, Englewood Cliffs.

Engbert I. S. and Hansen A. V. (2007). Documentation of the danish emoemotional speech database DES. Center for PersonKommunikation. Aalborg University, Denmark.

Erden M. and Arslan L. M. (2011). Automatic Detection of Anger in Human-Human Call Center Dialogs. InterSpeech 2011. Florence.

Eyben F., Wollmer M. and Schuller B. (2009). openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009). Amsterdam, IEE. **1:** 576-581.

Eyben M., Reiter F., schuller B., B. Cox, Douglas-Cowie E. and Cowie R. (2008). Abandoning Emotion Classes - Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies. INTERSPEECH 2008 Brisbane, Australia.

Fang R.-Y., Chen B.-W., Wang J.-F. and Wu C.-H. (2011). Emotion Detection Based on Concept Inference and Spoken Sentence Analysis for Customer Service. InterSpeech 2011. Florence.

Fleury S., Lamalle C., Martinez W. and Salem A. (2003). Lexico 3.

Fonagy I. (1983). La vive voix. Paris, Payot.

Fridja N. H. (1987). The Emotions Cambridge , UK, Cambridge University Press.

Galley M., McKeown K., Hirschberg J. and Shriberg E. (2011). Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies. InterSpeech. Florence, Italy.

Garnier-Rizet M., Adda G., Cailliau F., Gauvain J.-L., Guillemin-Lanne S., Lamel L., Vanni S. and Waast-Richard C. (2008). CallSurf - Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. LREC. marrakech.

Garnier M., Heinrich N., Dubois D., Castellengo M., Poitevineau J. and Sotiropoulos D. (2005). "Etude de la qualité vocale dans le chant lyrique." Scolia **20**: 151-169.

Gauvain J.-L., Lamel L., schwenk H., Adda G., chen L. and Lefevre F. (2003). Conversational telephone speech recognition. ICASSP. Hong Kong**:** 212-215.

Georgiou P. G., Black M. P., Lammert A. C., Baucom B. R. and Narayanan S. S. (2011). "That's Aggravating, Very Aggravating": Is It Possible to Classify Behaviors in Couple Interactions Using Automatically Derived Lexical Features? Affective Computing and Intelligent Interaction (ACII). Memphis, USA.

Goffman E. (1981). Forms of Talk. Philadelphia, University of Pennsylvania Press.

Goffman E. (1982). Interaction Ritual - Essays on Face-to-Face Behavior Pantheon; 1st Pantheon Books ed edition.

Goodwin C. and Goodwin M. H. (1992). Assessment and the construction of context. Rethinking Context, language as an interactive phenomenon, Duranti A. and C. Goodwin**:** 147-189.

Goodwin M. H. and Goodwin C. (2000). Emotion within Situated Activity. Linguistic Anthropology: A Reader. A. Duranti**:** 239-257.

Gratch J. and Marsella S. (2004). "A Domain-independent Framework for Modeling Emotion." Journal of Cognitive Systems Research **5**(4): 269-306.

Grossberg J. M. and Wilson H. K. (1968). "Physiological changes accompanying the visualization of fearful and neutral situations." Journal of Personality and Social Psychology **10**: 124-133.

Gupta P. and Rajput N. (2007). Two-Stream Emotion Recognition For Call Center Monitoring. Interspeech 2007. Antwerp, Belgium.

Hall M. and Frank E. (2008). Combining Naive Bayes and Decision Tables. Twenty-First International FLAIRS Conference. Coconut Grove, Florida.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. and Witten I. H. (2009). "The WEKA Data Mining Software: An Update;." SIGKDD Explorations **11**(1).

Halliday M. (1994). Introduction to Functional Grammar.

Hermansky H. and Morgan M. (1994). "RASTA processing of speech." Transactions on Speech and Audio Processing **2**(4): 578-589.

Hirst D. J., Di Cristo A. and Nishinuma Y. (2001). "Prosodic parameters of French : a cross language approach." Contrastive Studies of Japanese and Other Languages Series: 7-20.

Hollard S., Tomokiyo M. and Tuffelli D. (2005). Une Approche de l'expression orale des émotions : étude d'un corpus réel. Quatrièmes Journées de la Linguistique de Corpus. Lorient.

Hsu C.-W., Chang C.-C. and Lin C.-J. (2010). A Practical Guide to Support Vector Classication.

Hussain M. S., Calvo R. A. and Pour P. A. (2011). Hybrid Fusion Approach for Detecting Affects from Multichannel Physiology. Affective Computing and Intelligent Interaction (ACII). Memphis, USA. **Volume Part I**

Ilieva L. (2004). Combining Pattern Classifiers : Methods and Algorithms.

Jackiewicz A., Charnois T. and Ferrari S. (2009). Jugements d'évaluation et constituants périphériques. TALN. Senlis, France.

James W. (1884). "What is an emotion." Mind **9**: 188-205.

Joachims T. (1998). Text categorization with support vector machines: learning with many relevant features. ECML-98, 10th European Conf. on Machine Learning, Chemnitz, Germany.

Jones C. M. and Troen T. (2007). Biometric valence and arousal recognition. 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces ACM Press**:** 191-194.

Juslin P. N. and Laukka P. (2003). "Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?" Psychological Bulletin **129**(5): 770-814.

Kim J. (2007). Bimodal emotion recognition using speech and physiological changes. Robust Speech Recognition and Understanding. M. G. a. K. Kroschel.

Kim S., Valente F. and Vinciarelli A. (2012). Automatic detection of conflicts in spoken conversations: ratings and analysis of broadcast political debates. ICASSP. Kyoto, Japan.

Kittler J. V., Hatef M., Duin R. P. W. and Matas J. (1998). "On combining classifiers." Transaction on Pattern Analysis and Machine Intelligence **20**: 226-239.

Kleinginna P. R. and Kleinginna A. M. (1981). "A Categorized List of Emotion Definitions with Suggestions for a Consensual Definition." Motivation and Emotion **5**: 345-379.

Kuncheva L. I. (2004). Combining Pattern Classifiers: Methods and Algorithms.

Kuncheva L. I., Bezdek J. C. and Duin R. P. W. (2001). "Decision templates for multiple classifier fusion: an experimental comparison." Pattern Recognition **34**: 299-314.

Lacheret-Dujour A. (2011). La prosodie au coeur du verbal. Rééducation orthophonique, 49ème année**:** 87-104.

Lacheret-Dujour A., Kahane S., Pietrandrea P., Avanzi M. and Victorri B. (2011). "Oui mais elle est où la coupure, là ? Quand syntaxe et prosodie s'entraident ou se complètent " Langue Française **170**: 61-80.

Lacheret A. (2011). "Le corps en voix ou l'expression prosodique des émotions " Evolutions psychomotrices **23**(90).

Lam L. and Suen C. Y. (1995). "Optimal combinations of pattern classifiers." Pattern Recognition Letters **16**: 945-954.

Lambrecht K. (2008). Contraintes cognitives sur la syntaxe de la phrase en français parlé. Colloque International "Modèles Syntaxiques", Brussels.

Landis J. R. and Koch G. G. (1977). "The measurement of observer agreement for categorical data." Biometrics **33**(1): 159-174.

Laver J. (1991). The gift of speech. Edinburgh, UK.

Lee C., Narayanan S. and Pieraccini R. (2001). Recognition of Negative Emotions from the Speech Signal. ASRU. Trento, Italy**:** 240-243.

Lee C. M. and Narayanan S. (2005). "Toward detecting emotions in spoken dialogs." IEEE Transactions on Speech and Audio Processing **13**(2): 293-303.

Lee C. M., Narayanan S. S. and Pieraccini R. (2002). Combining acoustic and language information for emotion recognition. InterSpeech. Denver, Colorado, USA.

Lehuen J. and Luzzati D. (1994). Un modèle dynamique de représentation des dialogues finalisés. l'action GENE du PRC-IA : Modélisation d'explications sur un corpus de dialogues, ENST-Paris.

Liscombe J., Riccardi G. and Hakkani-T¨ur D. (2005). Using Context to Improve Emotion Detection in Spoken Dialog Systems. Interspeech. Lisbon, Portugal.

Littlestone N. and Warmth M. (1994). "Weighted majority algorithm." Information and Computation **108**: 212-261.

Lomsadze A., Ter-Hovhannisyan V., Chernoff Y. O. and Borodovsky M. (2005). "Gene identification in novel eukaryotic genomes by self-training algorithm." Nucleic Acids Research **28**(33): 6494-6506.

Losson O. (2000). Modélisation du geste communicatif et réalisation d'un signeur virtuel de phrases en langue des signes française. Lille, Université des sciences et technologies de Lille. **Thèse de doctorat**.

Lu C. Y., Hong J. S. and Cruz-Lara S. (2006). Emotion Detection in Textual Information by Semantic Role Labeling and Web Mining Techniques. Third Taiwanese-French Conference on Information Technology.

Luzzati D. (1989). Recherches sur le dialogue homme-machine : modèles linguistiques et traitements automatiques. Paris, Sorbone. **Thèse de Doctorat**.

Luzzati D. (1995). Le dialogue verbal homme-machine. Paris.

Luzzati D. (2005). Corpus d'hier et d'aujourd'hui : progrès quantitatifs ou progrès qualitatifs Congrès Français fondamental, corpus oraux, contenus d'enseignement, 50 ans de travaux et d'enjeux. ENSLSH, Lyon.

Mahdhaoui A. (2010). Analyse de signaux sociaux pour la modélisation de l'interaction face à face. ISIR. Paris, l'Université Pierre et Marie Curie – Paris 6. **Thèse de doctorat**.

Manning C. D., Raghavan P. and Schütze H. (2008). Support vector machines and machine learning on documents. Introduction to Information Retrieval. Cambridge, Cambridge University Press

Martin J. R. and White P. R. R. (2005). The Language of Evaluation: Appraisal in English. London.

Martin P. (2009). Intonation du français. Paris, Armand Colin.

Maurel S. and Curtoni P. (2007). Classification d'opinions par méthodes symbolique, statistique et hybride. DEFT'07. France.

Mehrabian A. (1998). "Correlations of the PAD emotion scales with self-reported satisfaction in marriage and work." Genetic, Social, and General Psychology Monographs.

Mehrabian A. and Russel J. A. (1974). An Approach to Environmental Psychology, The MIT Press.

Merhabian A. (1996). "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament." Current Psychology: Developmental, Learning, Personality, Social **14**: 261-292.

Metze F., Batliner A., Eyben F., Polzehl T., Schuller B. and Steidl a. S. (2010). Emotion Recognition using Imperfect Speech Recognition. InterpSeech 2010. Makuhari, Japan**:** 478-481.

Mihalcea R. (2004). Co-training and Self-training for Word Sense Disambiguation HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning. Boston, Massachusetts, USA, Association for Computational Linguistics**:** 33-40.

Montacié G. and Chollet G. (1987). Système de références pour l'évaluation d'applications et la caractérisation de base de données en reconnaissance automatique de la parole. 16ème journées d'étude sur la parole (JEP). Hammamet.

Morel M. and Bäzinger T. (2004). "Le role de l'intonation dans la communication vocale des émotions : test par la synthèse." Cahiers de l'Institut de Linguistique de Louvain **30**(13): 207-232.

Mower E. (2010). Affective interactions: analysis and recognition. Los Angeles. **PhD thesis**.

Ortony A., Clore G. L. and Collins A. (1988). The Cognitive Structure of Emotion. Cambridge, UK, Cambridge University Press

Osgood C. E., May W. S. and Miron M. S. (1975). Cross-Cultural Universals of Affective Meaning University of Illinois Press.

Pang B. and Lee L. (2002). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Association for Computational Linguistics (ACL)**:** 271-278.

Pang B. and Lee L. (2008). "Opinion mining and sentiment analysis." Foundations and Trends in Information Retrieval **2**(1-2).

Paumier S. (2002). Manuel d'utilisation d'Unitex. Université de Marne-la-Vallée.

Petrushin V. A. (1999). Emotion in speech : recognition and application to call centers. In Engr**:** 7-10.

Petta P., Pelachaud C. and Cowie R. (2011). Emotion-Oriented Systems: The Humaine Handbook Springer.
Picard R. W. (1997). Affective Computing
Planalp S. (1999). Communicating emotion: Social, moral and cultural processes. Cambridge.
Plutchick R. (1984). "A General Psychoevolutionary Theory." Approaches to Emotion: 293-317.
Polzehl T., Schmitt A., Metze F. and Wagner M. (2011). "Anger recognition in speech using acoustic and linguistic cues." Speech Communication **53**(9-10).
Porter M. (1980). "An algorithm for suffix stripping." Program **14**(3): 130-137.
Quirk R., Greenbaum S., Leech G. and Svartvik J. (1985). A comprehensive grammar of the English language, Longman.
Rabiner L. and Juang B. H. (1993). Fundamentals of Speech Recognition.
Sacks H. (1992). Lectures on conversation Blackwell Publishing.
Sacks H., Schegloff E. and Jefferson G. (1974). "A simplest systematics for the organization of turn-taking for conversation." Language **50**(4): 696-735.
Sander D. and Scherer K. R. (2009). Traité de psychologie des émotions.
Savoy J. (2006). Light stemming approaches for the French, Portuguese, German and Hungarian languages. symposium on Applied computing. Dijon, France.
Scherer K. (1981). Speech and emotionnal states. Speech Evaluation in psychiatry. J. Darby. New York, Grune and Stratton.
Scherer K. (1985). "Vocal affect signalling: A comparative approach." Advances in the study of behavior **15**.
Scherer K., Banse R., Wallbott H. and Goldbeck T. (1991). "Vocal cues in emotion encoding and decoding." Motivation and Emotion **15**: 123-148.
Scherer K., Johnstone T. and Klasmeyer G. (2003). Vocal expression of emotion. Handbook of Affective Sciences. G. editors. Oxford, New York, Oxford University Press.
Scherer K. and Peper M. (2001). Psychological theories of emotion and neuropsychological research. Emotionnal behavior and its disorders. Elsevier. Amsterdam. **5:** 17-48.
Scherer K. R. (1984). "Emotion as multicomponent process : A model of some cross-cultural data." Review of Personnality and social Psychology **5**: 37-63.
Scherer K. R. (1984). On the nature and function of emotion : A component process approach. Approaches to emotion. London, Lawrence Erlbaum Associates**:** 293-317
Scherer K. R. (1999). Appraisal theory. Handbook of congnition and emotion. T. D. a. M. Power**:** 638-663.
Scherer K. R. (2005). "What are emotions? And how can they be measured? ." Social Science Information **44**.
Scherer K. R., Banse R. and Wallbott H. G. (2001). "Emotion inferences from vocal expression correlate across languages and cultures." Journal of Cross-Cultural Psychology **32**: 76-92.
Scherer K. R., Banziger T. and Roesch E. (2010). A Blueprint for Affective Computing. Oxford Oxford University Press.
Scherer K. R., Koivumaki J. and Rosenthal R. (1972). "Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech." Journal of Psycholinguistic Research **1**.
Schmid H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. ACL SIGDAT-Workshop. Dublin, Ireland.
Schröder M. (2001). "Experimental study of affect bursts." Speech communication **40**.
Schröder M., Devillers L., Karpouzis K., Martin J.-C., Pelachaud C., Peter C., Pirker H., Schuller B., Tao J. and Wilson I. (2007). What should a generic emotion markup language be able to represent ? ACII 2007. Lisbon.
Schröder M., Pirker H. and Lamolle M. (2006). First Suggestions for an Emotion Annotation and Representation Language. LREC. Genoa, Italy**:** 88-92.

Schuller B., Batliner A., Steidl S. and Seppi D. (2011). "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge." Speech Communication **53**(February): 1062-1087.

Schuller B., Muller R., Lang M. and Rigoll G. (2005). Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. Interspeech. Lisbon, Portugal, ISCA**:** 805-808.

Schuller B., Reiter S. and Rigoll G. (2006). Evolutionary feature generation in speech emotion recognition. IEEE International Conference on Multimedia and Expo. Toronto, Canada**:** 5-8.

Schuller B., Seppi D., Batliner A., Maier A. and Steidl S. (2007). Toward more reality in the recognition of emotional speech. ICASSP. Honolulu, Hawaii.

Schuller B., Steidl S. and Batliner A. (2009). The INTERSPEECH 2009 emotion challenge. INTERSPEECH**:** 312-315.

Schuller B., Steidl S., Batliner A., Burkhardt F., Devillers L., Mueller C. and Narayanan S. (2010). The INTERSPEECH 2010 paralinguistic challenge. InterSpeech. Makuhari, Japan.

Schuller B., Steidl S., Batliner A., Schiel F. and Krajewski J. (2011). The INTERSPEECH 2011 Speaker State Challenge. InterSpeech. Florence, Italy.

Schuller B., Vlasenko B., Eyben F., Rigoll G. and Wendemuth A. (2009). Acoustic Emotion Recognition: A Benchmark Comparison of Performances. ASRU 2009. IEEE. Merano**:** pp. 552-557

Schuller B., Zhang Z., Weninger F. and Rigoll G. (2011). Selecting Training Data for Cross-Corpus Speech Emotion Recognition: Prototypicality vs. Generalization. Afeka-AVIOS Speech Processing Conference. Tel Aviv.

Schuller B., Zhang Z., Weninger F. and Rigoll G. (2011). Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote? InterSpeech. Florence.

Shafer G. (1976). A Mathematical Theory of Evidence.

Somasundaran S., Wiebe J., Hoffmann P. and Litman D. (2006). Manual annotation of opinion categories in meetings. Workshop on Frontiers in Linguistically Annotated Corpora 2006 Sydney, Australia.

Steidl S. (2009). Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech. **PhD thesis**.

Stern R. (2008). Constitution d'un lexique de sentiment. Paris, Université Paris 7. **Master thesis**.

Stickel C., Ebner M., Steinbach-Nordmann S., Searle G. and Holzinger A. (2009). Emotion detection: Application of the valence arousal space for rapid biological usability testing to enhance universal access. International Conference on Universal Access in Human Computer Interaction. San Diego**:** 615-624.

Tahon M., Delaborde A., Barras C. and Devillers L. (2010). A corpus for identification of speakers and their emotions LREC. Malta.

Tomokiyo M., Chollet G. and Hollard S. (2005). Studies of emotional expressions in oral dialogues: towards an extension of Universal Networking Language. Universal Networking Language: advances in theory and applications. Mexico City.

Turney P., D (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. 40th Annual Meeting of the Association for computationnal Linguistics. Philadelphia.

Utthara M., Suranjana S., Sukhendu D. and Pinaki C. (2010). "A Survey of Decision Fusion and Feature Fusion Strategies for Pattern Classification." IETE Technical Review **27**: 293-307.

Van-Erp M. and Schomaker L. (2000). Variants of the Borda count method for combining ranked classifieur hypotheses. 7th International Workshop on Frontiers in Handwriting Recognition Amsterdam**:** 443-452.

Vapnik V. (1998). The Nature of Statistical Learning Theory, Springer.

Vaudable C., Balagué C. and Devillers L. (2009). Study of consumer's emotion during product interviews. ACII 2009. Amsterdam.

Vaudable C., Rollet N. and Devillers L. (2010). Annotation of Affective Interaction in Real-life Dialogs Collected in a Call-center LREC. Elda. Valetta, Malta.

Vernier M., Ferrari S. and Legallois D. (2007). Suivi d'opinions dans le discours. Journées de la Linguistique de Corpus. France.

Vernier M., Monceaux L., Daille B. and Dubreil E. (2009). "Catégorisation des évaluations dans un corpus de blogs multi-domaine." Revue des nouvelles technologies de l'information (RNTI).

Ververidis D. and Kotropoulos C. (2003). review of emotional speech databases. 9th Panhellenic Conference on Informatics. Thessaloniki, Greece, PCI**:** 560-574.

Vidrascu L. (2007). Analyse et détection des émotions verbales dans les interactions orales. Traitement du Langage Parlé. Orsay, Université Paris XI. **Thèse de doctorat:** 203.

Vidrascu L. and Devillers L. (2005a). Detection of Real-Life Emotions in Call Centers. InterSpeech. Lisbon.

Vidrascu L. and Devillers L. (2005b). Annotation and detection of blended emotions in real human-humain dialogs recorded in call center. ICME. Amsterdam.

Vidrascu L. and Devillers L. (2006). "Représentation et Détection des émotions dans des données issues de dialogues enregistrés dans des centres d'appels : des émotions mixtes dans des données réelles." Revue des Sciences et Technologies de l'Information, serie Revue d'Intelligence Artificielle.

Whitelaw C., Garg N. and Argamon S. (2005). Using appraisal groups for sentiment analysis. 14th ACM international Conference on information and Knowledge Management Bremen, Germany.

Wiebe J. and Riloff E. (2005). Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-05).

Wiebe J., Wilson T. and Cardie C. (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation.

Wundt W. (1913). Grundriss der Psychologie.

Yacoub S., Simske S., Lin X. and Burns J. (2003). Recognition of Emotions in Interactive Voice Response Systems. EUROSPEECH-2003. Geneva, Switzerland**:** 729-732.

Yarowsky D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. 33rd Annual Meeting of the ACL. Cambridge**:** 189-196.

Zara A., Maffiolo V., Martin J.-C. and Devillers L. (2007). Collection and Annotation of a Corpus of Human-Human Multimodal Interactions: Emotion and Others Anthropomorphic Characteristics. Proceedings of the 2nd International Conference on Affective Com-puting and Intelligent Interaction (ACII 2007). Lisbon**:** 464-475.

Zhu X. (2006). Semi-supervised learning literature survey. Madison, Univ. of Wisconsin.

Zwicker E. and Feldtkeller R. (1981). Psychoacoustique.