550, rue Sherbrooke Ouest, bureau 100
Montréal (Québec) H3A 1B9
Tél. : 514 840-1234; Téléc. : 514 840-1244
Place de la Cité – Tour de la Cité
2600, boul. Laurier, bureau 625
Québec (Québec) G1V 2W1
Tél. : 418 648-8080; téléc. : 418 648-8141
http://www.crim.ca

CRIM - Documentation/Communications

Rapport technique
Détection d'émotions à partir de signal audio

Première version

CRIM-07/06-15

auteur(s)
Dr. Narjès Boufaden, Chercheure
Najim Dekak, Étudiant au doctorat, ETS
Yazid Attabi, Étudiant à la maîtrise, ETS
Groupe Reconnaissance de la parole

28/06/2007

Collection scientifique et technique

ISBN-13 : 978-2-89522-099-2
ISBN-10 : 2-89522-099-9

(Note facultative sur le verso de la page titre)

Pour tout renseignement, communiquer avec:
CRIM Centre de documentation
CRIM
550, rue Sherbrooke Ouest, bureau 100
Montréal (Québec) H3A 1B9

Téléphone : (514) 840-1234
Télécopieur : (514) 840-1244

ISBN-13 :978-2-89522-099-2
ISBN-10 : 2-89522-099-9
Dépôt légal - Bibliothèque et Archives nationales du Québec, 2007
Dépôt légal - Bibliothèque et Archives Canada, 2007

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

DÉTECTION DES ÉMOTIONS À PARTIR DU SIGNAL AUDIO

1. INTRODUCTION

Nous expérimentons la détection des émotions à partir de deux bases de données audio: des dialogues Personne-Machine (DARPA Communicator 2000) et des phrases neutres répétées par des acteurs simulant différents états émotionnels (LDC Emotional Prosody). Notre étude porte sur 15 états émotionnels allant de la colère jusqu'au bonheur tel que défini par *Banse & Scherer (1996),* et de manière plus générale sur les états émotionnels négatif et non-négatif.

Les approches que nous avons développées se basent sur la combinaison de différents traits prosodiques incluant la fréquence fondamentale, l'énergie, la largeur de la bande des deux premiers formants ainsi que le power spectrum, le jitter et shimer. Les valeurs de chaque trait ont été lissées avec un polynôme de Legendre et les coefficients des polynômes respectifs ont été modélisés avec différents algorithmes d'apprentissage supervisé. Les résultats que nous présentons bien que préliminaires nous permettent de tracer quelques conclusions sur la difficulté de l'annotation des émotions, l'importance de l'accord entre annotateurs et son impact sur la performance des systèmes développés pour la détection des émotions. Aussi, dans notre approche, nous proposons de centraliser nos efforts sur la détection des émotions à partir de tours de parole constitué uniquement de mots de rétraction comme première étape pour l'amélioration des systèmes de dialogue Personne-machine dédiés aux application de centres d'appels.

Dans un premier temps, nous explorons la détection des émotions à partir d'un corpus préfabriqué : le LDC Emotional Prosody, composé de phrases prononcé avec différents états émotionnels simulés par des acteurs professionnels. Le but de cette expérience est de dégager un premier ensemble de traits prosodiques qui servira de base pour les expériences effectués sur les corpus réels. L'intérêt de ce corpus préfabriqué réside dans le fait, d'une part qu'il est annoté avec les émotions. D'autre part ce corpus étant disponible publiquement, quelques travaux ont été effectués dessus et serviront de base de comparaison pour nos expériences.

La suite du rapport sera consacré aux expériences menées sur les mots de rétroactions extraits de dialogues Personne-machine réels.

1.1 État de l'art

La détection d'émotion est un champ multidisciplinaire dans lequel se combinent des problématiques qui relèvent du traitement du signal aussi bien que de la psychologie et du trai-

tement du langage naturel. En psychologie, les aspects définitoires des émotions, la façon avec lesquels elles sont exprimées ainsi que leurs manifestations neurophysiologiques font l'objet de plusieurs travaux. Par exemple, (Eckman 99, Scherer 96) proposent des classes d'émotions primaires. Malgré ces efforts, la réification des classes d'émotions reste un problème complet dans la mesure où il n'y a pas de définition unique des émotions. À titre d'exemple, il y a de multiples états qui pourraient être considérés comme étant de la colère et tous sont exprimés de façons différentes.

La communauté du traitement des langues naturelles, se penchent sur un autre aspect du langage qui pourrait aider à la classification des émotions. Ces travaux bien qu'en encore à leur débuts, montre que la prise en compte du sens des mots et de leur contexte discursif peuvent aider à une meilleure classification des émotions.

Toutefois, la plus grande part des travaux en détection des émotions est faite par la communauté du traitement de signal. La tâche principale dans se contexte est d'identifier des caractéristiques acoustiques corrélées avec les différentes classes d'émotions et en isolé les plus discriminantes.

Selon le but de l'application considéré, différentes classes d'émotions primaires ont été définies. Par exemple, Ang et al. (Ang, 2002) ont proposé la contrariété, la frustration, fatiguée et l'amusement comme les émotions primaires utiles pour des applications dédiées aux centres d'appels.

Toutefois pour des raisons pratiques, la plupart de la recherche se fait sur la classification des émotions en deux classes : les émotions négatives versus les émotions dites non-négatives. Ce type de classification semble suffisant pour des applications réelles, dans la mesure où les fournisseurs de télécommunication sont plus intéressés à identifier les moments où les clients éprouvent une émotion négative sans pour autant distinguer si cela est de la frustration ou de la contrariété. Par ailleurs, la classification binaire (négative versus non-négative) est un moyen simple de diminuer le problème d'accord entre annotateurs observé à travers toutes les expériences d'annotation d'émotions.

Un grand nombre de travaux ont porté la détection d'émotion. Différents types de traits linguistiques et prosodiques ont été expérimentés avec différents algorithmes de classification. Cependant, plusieurs de ces expériences utilisaient des corpus privés rassemblés à partir d'application commerciale telle que le système HMIHY de AT&T le système de SpeechWork. D'autres travaux, moins nombreux, ont été faits sur des bases de données disponibles publiquement. De ce fait, il est difficile de présenter une comparaison directe des résultats obtenus.

Parmi les études faites sur les corpus privés, (Lee, 2005 *et al.*) proposent de modéliser une combinaison de traits prosodiques, de mots et les actes de dialogues avec un classificateur linéaire pour classifier les tours de parole extraits de dialogue Personne-Machine dans une des deux classes : émotion négative versus état neutre. Les auteurs ont utilisé un algorithme

de sélection des attributs les plus prometteurs. Le meilleur résultat obtenu sur un corpus généré avec le système de SpeechWork a était de 84% en utilisant les 10 meilleurs traits parmi ceux présentés dans le tableau 2.

Dans le domaine de données publiquement disponibles, trois expériences intéressantes ont été rapportées. Ang (Ang, 2002) a utilisé une combinaison de caractéristiques prosodiques, linguistiques et de style de discours pour classer les tours de parole en deux classes : une classe négative incluant la contrariété et la frustration et une classe positive incluant toutes autres émotions y compris les émotions fatigue et amusement. Le tableau 2 montre les caractéristiques utilisées dans leurs expériences. Pour classer les tours de parole, ils ont utilisé un modèle de langue et un arbre de décision. Le modèle de langue fournit une probabilité des classes d'émotion étant donné des trigrammes de mots. Ensuite, la différence des probabilités obtenues pour chaque classe, les caractéristiques prosodiques et le style de discours étaient utilisés comme les données d'entrée d'un arbre de décision. Les auteurs ont obtenus un taux de classification correcte de 80,2% sur un sous-ensemble du corpus de Communicateur de DARPA préalablement annoté avec les émotions.

D'autres résultats intéressants ont été rapportés dans (Yacoub, 2003) et Huang (Huang, 2006). Les auteurs ont utilisés différents algorithmes de classification sur des données acoustiques extraites d'un corpus publiquement disponible : LDC Emotional Prosody.

Yacoub *et al*. ont utilisé une combinaison de caractéristiques acoustiques montrées (voire Table 2) modélisées avec quatre algorithmes de classification : les réseaux neuronaux, la machine à vecteur de support, les k-proche voisins et un arbre de décision. Le meilleur résultat a été obtenu avec les réseaux neuronaux avec 94% de classification correcte pour une classification en deux classes : colère versus neutre. Avec le même corpus, Huang *et al.* ont aussi utilisé des caractéristiques acoustiques (voire Table 2) mais les ont modélisé avec un modèle de Markov caché utilisant la distribution continue. Avec ce modèle et moins de traits que ceix proposés par Yacoub *et al.*, les auteurs ont obtenus de meilleurs résultats avec un 98% de classification correcte pour le même classes d'émotion (colère versus neutre).

Le tableau 2 récapitule les différents traits proposés dans les travaux présentés ci-dessus.

État de l'art	Traits acoustiques	Traits linguistiques
(Lee, 2005)	Fundamental frequency (F0), energy, duration, formants (first and second) and their bandwidths.	Salient words defined as words that correlate with a particular emotion. Mutual information has been used to detect emotional salience. Dialog acts such as repeating, rejections and ask-start over

État de l'art	Traits acoustiques	Traits linguistiques
(Ang, 2002)	Duration and speaking rate, pauses, F0, Energy, spectral tilt	Trigrams of words Dialog acts including repeats and corrections
(Yacoub, 2003)	The fundamental frequency (contour, first derivative, jitter), Energy (contour, first derivative, shimmer), Duration (min, max, mean and standard deviation of audible segments, min, max, mean and standard deviation of inaudible segments and ratios of audible and inaudible segments).	-
(Huang, 2006)	Fundamental frequency, Energy features, Zero Crossing rate, energy slope	-

Tableau 1: Synthèse des traits utilisés pour la détection des émotions

1.2 Approches développée

Dans les expériences que nous présentons, nous utilisons différents algorithmes de classification sur les mêmes combinaisons de traits auxquels est appliquée la même technique de lissage. Les expériences effectuées sur le corpus LDC Emotional Prosody sont différentes des autres expériences puisqu'elles ont été faites sur un corpus composé de tours de parole qui sont des phrases prononcées avec différents états émotionnels simulés par des acteurs. A cet égard, la technique de segmentation ainsi que les traits utilisés sont différents de ceux utilises pour toutes les autres expériences. Dans ce qui suit, nous décrivons l'ensemble des traits utilisés ainsi que la technique de lissage utilisée dans toutes nos expériences.

1.2.1 Traits utilisés

Le choix des traits se base sur l'étude bibliographique que nous avons effectuée ainsi que sur nos propres recherches.

Plusieurs travaux ont montré la corrélation existant entre la fréquence fondamentale, l'énergie ainsi que la largeur de bande des formants et les différentes classes d'émotion (Razak, 2003), (Vidhyasaharan, 2007). Les figures 5 et 6 illustrent en partie les corrélations potentielles. Les figures sont des captures des résultats de l'analyse de deux phrases identiques dites une fois avec un ton heureux (figure 1) et une seconde fois avec un ton triste (figure 2). Les lignes colorées représentent les formants, la bande en dessous contient

l'évolution de l'énergie tandis que la dernière bande contient l'évolution de la fréquence fondamentale. Nous pouvons remarquer une variation des contours de fréquences fondamentales ainsi que des formants et de l'énergie entre les deux phrases qui sont identiques et prononcées par un même locuteur.

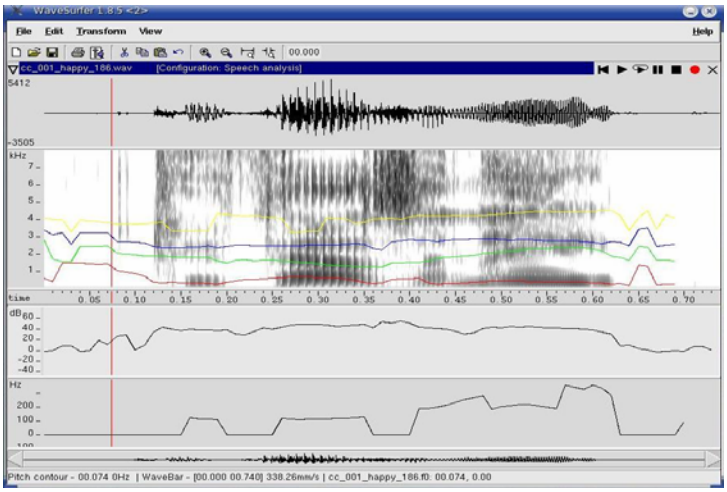

Figure 1: Analyse du signal audio d'une phrase dite avec un ton content

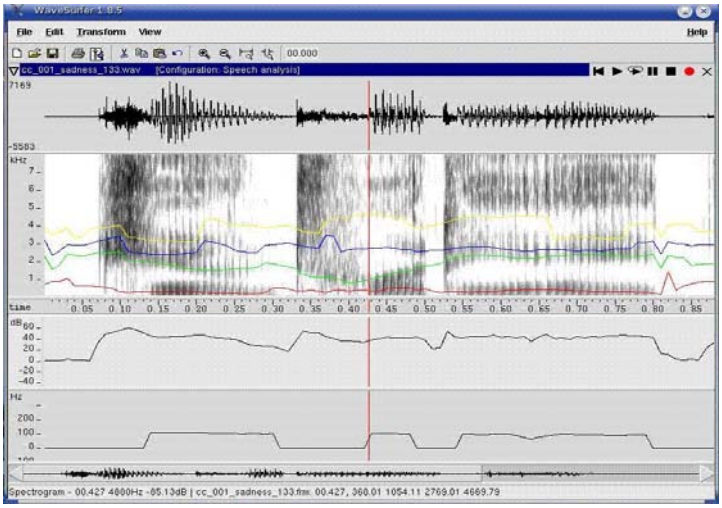

Figure 2 : Analyse du signal audio d'une phrase dite avec un ton triste

En plus de ces traits, nous avons aussi testé des variantes de la fréquence fondamentale et de l'énergie; à savoir le *jitter* et *shimmer* qui modélisent respectivement les variations des

$$= \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - T_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N-1}T_i}$$

contours de la fréquence fondamentale et de l'énergie ainsi que le *power spectrum* qui est la magnitude de la transformée de fourrier du signal original. Ces traits sont définis comme suit :

- Jitter , T_i, représente la fréquence fondamentale

- Shimmer $= \dfrac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - A_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N-1} A_i}$, A_i représente l'énergie

- Power spectrum, $Sxx = \left|X(\omega)\right|^2$, $X(\omega)$ est la transformée de fourrier du signal.

D'autres traits ont été expérimentés toutefois sans montrer d'amélioration de la classification des émotions. En particulier, nous avons comparé les résultats de la combinaison des traits décrits ci-dessus avec les MFCC sans améliorations. En réalité les MFCC combiné à la fréquence fondamentale, l'énergie la largeur de bande du premier et second formant c'est avéré moins performant que la même combinaison dans laquelle nous avions remplacé les MFCC par le power spectrum.

1.2.2 Lissage avec le polynôme de Legendre

Dans chaque segment obtenu, nous réalisons une approximation du contour de la fréquence fondamentale et celui de l'énergie. Cette approximation est réalisée en utilisant le polynôme de Legendre et en se basant sur le critère des moindres carrées. L'approximation d'une fonction f_t par un polynôme de Legendre est définie comme suit :

$$f_t = \sum_{i=1}^{M} a_i . P_i(t)$$

Avec t est un index qui correspond au temps. M est le plus grand ordre du polynôme. a_i et p_i représentent respectivement le coefficient et le polynôme de Legendre d'ordre i. Dans le domaine du traitement de la parole, le polynôme de Legendre est souvent utilisé cependant il est important d'effectuer une normalisation du temps. Tous les segments doivent être interpolés dans un intervalle -1,1]. Dans le cas ou nous possédons un segment de N points, la normalisation du temps consiste à diviser l'intervalle -1,+1] en N-1 parties. Le point -1 correspondra au point de début du segment et le point +1 correspondra à la fin du segment. Cette technique de segmentation et d'approximation des contours de la fréquence fondamentale et celui de l'énergie a déjà été utilisée dans le domaine de la reconnaissance de la

langue et de la reconnaissance du locuteur. La figure 3 montre un exemple d'approximation du contour log pitch en utilisant des polynômes de Legendre d'ordre différents. L'axe des X correspond au temps normalisé et l'axe des Y au log de la fréquence fondamentale.

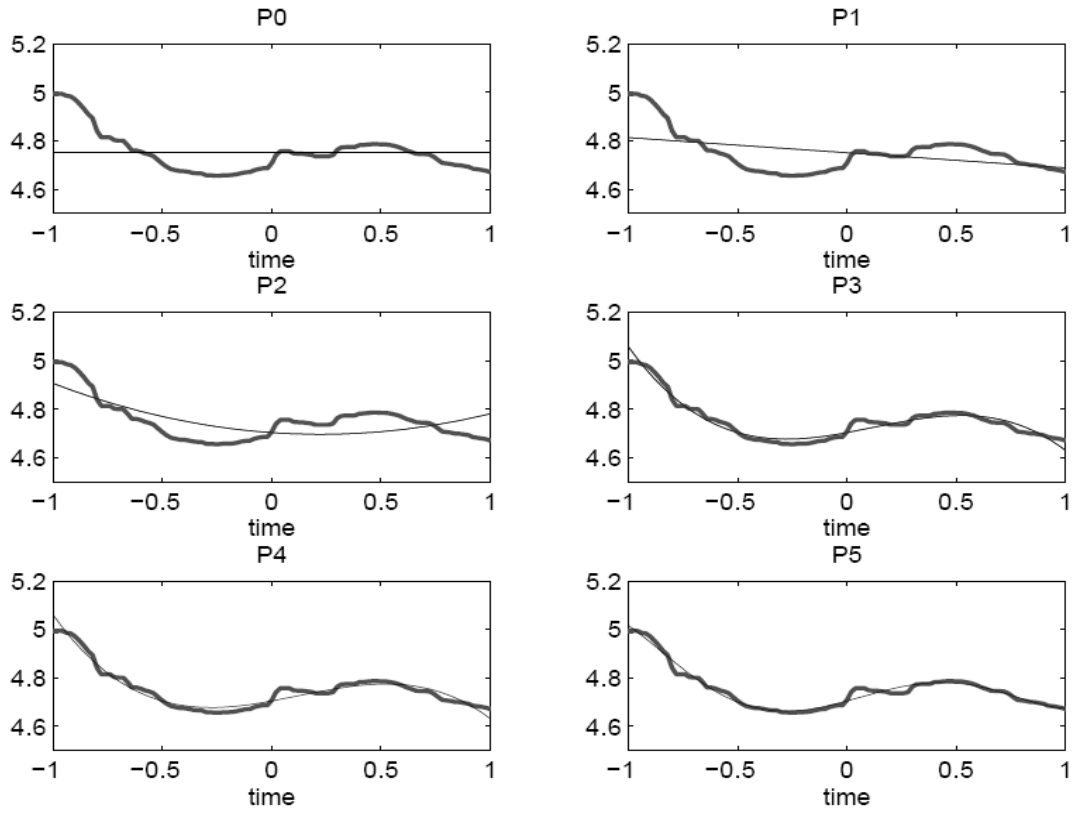

Figure 3: Approximation du contour du log de la fréquence fondamentale en utilisant des polynômes Pn avec un ordre n variant de 0 à 5.

1.3 Préparation des données

Les expériences en détection des émotions ont été faites sur trois corpus :

1. Le corpus Emotional Prosody est un corpus distribué par le Lingustic Data Consortium (LDC). Ce corpus a été développé dans le but d'étudier les manifestations des émotions sur le plan médical et neurophysiologique. De plus amples informations sont fournies

sur le site web suivant http://projects.ldc.upenn.edu/Prosody/[2]. Ce corpus est le seul corpus annoté avec les émotions et disponible publiquement.

2. Le corpus DARPA Communicator est aussi distribué par LDC. La description du corpus est donnée dans la section 2.2.1
3. Le corpus de Bell qui est un ensemble de 1000 tours de parole contenant uniquement des mots de rétroactions.

1.3.1 Conversion des fichiers audio

Tous les fichiers audio provenant de LDC sont en format sphere de NIST. Afin d'extraire les paramètres acoustiques, nous les avons convertit au format wave en utilisant le programme sox[3] et en gardant les paramètres (i.e. la fréquence d'échantillonnage) du fichier d'origine.

L'extraction des traits prosodiques s'est faite en utilisant les outils Snack[4] et Praat[5]. Nous avons développé plusieurs scripts en Perl à cet effet afin de générer les traits présentés dans la section 1.2.1.

1.3.2 Procédure d'annotation

La procédure d'annotation visait à attribué une classe d'émotion à chaque tour de parole considéré. Nous nous sommes limités à deux classes d'émotions :

4. La classe négative qui englobe les émotions : *cold anger*, *hot anger*, *frustration*, *disgust*, *anxiety*, *panic*, *sadness* et *despair*. Par soucis de cohérence avec l'annotation du corpus LDC Emotional Prosody, nous avons gardé certaines classes d'émotion, telles que *disgust* ou *sadness* inappropriée dans le contexte de notre application.
5. La classe non négative contenait les émotions *boredom*, *contempt*, *happiness*, *interest*, *neutral*, *pride*. Ici aussi, pour les mêmes raisons de cohérence, nous avons gardé certaines émotions qui ne sont pas adéquates telles que *happiness* et *pride*.

Afin d'annoter les tours de parole avec les émotions, nous avons constitué un groupe de trois annotateurs. Nous avons fixé le nombre d'annotateurs à trois pour deux raisons :

---

[1] Ce site a été vérifié le 27 juin 2007.
[2] Ce site a été vérifié le 27 juin 2007.
[3] Sox est un logiciel de traitement de signal audio publique qu'il est possible de télécharger à l'adresse suivante: http://sox.sourceforge.net/ (site vérifié le 06/06/2007).
[4] Snack est un logiciel de traitement de signal audio publique qu'il est possible de télécharger à l'adresse suivante: http://www.speech.kth.se/snack/ (site vérifié le 06/06/2007).
[5] Praat est un logiciel de traitement de signal audio publique qu'il est possible de télécharger à l'adresse suivante http://www.fon.hum.uva.nl/praat/ (site vérifié le 06/06/2007).

- Avoir au moins deux annotateurs pour être en mesure d'évaluer l'accord entre annotateur. Etant donné que le choix dune émotion est une tâche très subjective, nous devons tenir compte de cette subjectivité pour fournir au système des données d'entrainement le moins subjective possible.
- Le choix d'un troisième annotateur permet de départager les votes différents. C'est aussi un moyen de produire un corpus avec un minimum de consensus en choisissant de manière systématique l'émotion ayant le plus de vote.

La mesure dévaluation de l'accord entre annotateur que nous utilisons est le *pairwise kappa* (Cohen, 1960) qui calcule l'accord entre toutes les paires d'annotateurs et prend la moyenne comme mesure d'accord globale entre les différents annotateurs. La formule du kappa est donnée ci-dessous (Cohen, 1960) :

$$K = \frac{P(A) - P(E)}{1 - P(E)},$$

$$P(A) = \textit{accord observé}$$

$$P(E) = \textit{accord par chance}$$

Cette formule indique qu'une valeur de kappa=0 indique que l'accord obtenu n'est pas différent de celui obtenu par chance. Une valeur de kappa=1 implique un parfait accord entre les annotateurs.

Afin d'interpréter valeurs intermédiaires du kappa, différentes échelles d'interprétation peuvent être utilisées. Nous avons utilisé l'échelle proposée dans (Rietveld, 1993) pour interpréter les mesures d'accord entre annotateurs que nous avons obtenus. Selon les auteurs, un accord est acceptable lorsque 0,41< kappa<0,60. Une valeur supérieure à 0,60 indiquerait un accord substantiel entre les annotateurs. Enfin, un accord en dessous de 0,41 signifierait que les annotations ne sont pas très fiables. Sur un plan pratique cela impliquerait que les données d'entraînement et de test utilisées pour les expériences seront très bruitées.

D'autres échelles haussent le seuil d'acceptabilité à 0.60 (Grove, 1981) et même à 0,80 notamment par (Krippendorff 1960) selon le type de tâche d'annotation considéré. Toutefois, dans notre cas et étant donné la difficulté de la tâche, l'échelle proposée dans (Rietveld, 1993) semblait adéquate. Plus de détails sur les différences entre ses échelles d'interprétation sont donnés dans (Di Eugenio, 2002).

1.4 Détection des émotions à partir du corpus Emotional LDC

Nous présentons les résultats des expériences de détection des émotions effectuées sur le corpus LDC Emotional Prosody. Ce sont les seules expériences qui ont porté sur des tours de parole comportant plus d'un mot.

1.4.1 Description du corpus LDC Emotional Prosody

Les données utilisées dans le cadre des expériences de ce projet proviennent de la base de données LDC Emotional Prosody. C'est un corpus qui a été produit par 7 acteurs professionnels (4 femmes et 3 hommes).

Les phrases prononcées sont exprimes dans 15 états émotionnels différents qui sont : *anxiety*, *boredom*, *contempt*, *cold anger*, *despair*, *disgust*, *elation*, *happiness*, *hot anger*, *interest*, *neutral*, *panic*, *pride*, *sadness*, etc.

Dans ce cas particulier, aucune annotation n'était requise puisque l'émotion était connue.

La détection des émotions est réalisée en utilisant les contours prosodiques (qui sont le pitch et l'énergie) dans des unités appelées pseudo-syllabes. Ces contours prosodiques sont modélisés en utilisant des polynômes de Legendre. Les coefficients des polynômes de Legendre obtenus pour chaque pseudo-syllabe plus le *jitter*, *shimmer* et la durée des pseudo-syllabes, constituent les vecteurs caractéristiques qui modélisent les contours prosodiques. Ces vecteurs sont ensuite modélisés en utilisant le modèle de machine à vecteurs de support avec une sortie probabiliste.

1.4.2 Approche

Dans le système que nous proposons, L'unité de base utilisée représente une pseudo-syllabe. Un contour prosodique peut contenir plusieurs syllabes en même temps. Le but de cette étape est de segmenter les segments longs en des segments plus courts qui peuvent représenter des pseudo-syllabes. Pour effectuer cette segmentation, nous divisons tout d'abord les valeurs de la fréquence fondamentale par leur moyenne dans le fichier audio afin de produire des valeurs indépendante du locuteur par la suite nous réalisons un alignement du contour de pitch avec celui de l'énergie. La seconde étape consiste à utiliser le contour de l'énergie pour segmenter les contours prosodiques. À ce niveau, nous détectons les points minima locaux du contour de l'énergie. Le choix des points minima qui constitueront les frontières des segments courts dépend de la durée des pseudo-syllabes trouvées. Un minimum de 60 ms a été choisi pour la durée des pseudo-syllabes. Nous avons utilisé cette durée minimale parce que chaque contour de segment trouvé va être interpolé par un polynôme de Legendre à six coefficients qui nécessite un contour de 6 points au minimum. Puisque l'extraction de la fréquence fondamentale et l'énergie est effectuée tout les 10ms cela justifie le choix du minimum de la durée de 60 ms. Un exemple de segmentation est représenté dans la figure 4.

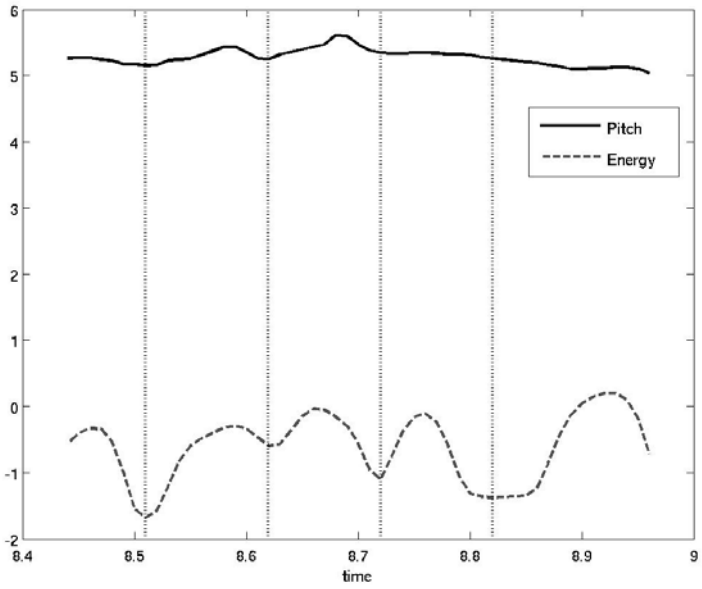

Figure 4: Segmentation des contours de pitch et de l'énergie.

Nous avons utilisé les six coefficients du polynôme de Legendre (PL) utilisé pour le lissage de la fréquence fondamentale, les six coefficients du PL pour le lissage de l'énergie, la durée, le jitter et le shimmer. L'ensemble de ces traits forment un vecteur de caractéristiques de dimension 15.

Nous avons utilisé le modèle de machine à vecteurs de support avec sortie probabiliste pour modéliser les coefficients des polynômes de Legendre qui caractérisent les pseudo-syllabes. La probabilité sur le tour de parole est ensuite obtenue selon une des trois méthodes suivantes :

1. Calculer la somme des logs probabilités de toutes les pseudo-syllabes par classe d'émotion. Ensuite, choisir la classe qui obtient la somme maximale.
2. Calculer le maximum de toutes les pseudo-syllabes par classe d'émotion sur un tour de parole.
3. Calculer la médiane de toutes les pseudo-syllabes par classe d'émotion sur un tour de parole. Ensuite, choisir la classe qui obtient la médiane maximale.

Les expériences ont été réalisées en utilisant la librairie LIBSVM[6]. LIBSVM est une librairie qui supporte la classification multi-classes.

1.4.3 Expériences et résultats

L'évaluation a été faite par validation croisée sur les locuteurs; c'est à dire les données de 6 locuteurs sont utilisées pour l'apprentissage et celles du 7ième pour le test. Les expériences ont porté sur la classification avec deux classes d'émotions et sur les 15 classes d'émotions telles que définies dans le corpus LDC Emotional Prosody.

Le tableau 2 montre les moyennes des résultats sur toutes les classes d'émotions confondues et cela par locuteur. Ces résultats ont été obtenus en prenant le max des probabilités (méthode 1). La colonne de gauche montre le taux de classification correcte par locuteur.

Acteur	Taux
CC	10.17%
CL	8.39%
GG	12.5%
MM	7.69%
MK	7.57%
JG	11.25%
MF	6.12%
Moyenne	9.09%

Tableau 2: Taux de reconnaissance des émotions par locuteur

Les résultats montrent une variation importante dans le taux de classification entre les différents locuteurs, et qui atteint le double entre les locuteurs MF et GG. Ceci peut être expliqué par le fait que l'expression d'un même état émotionnel peut différer d'une personne à une autre, et qu'il n'existe pas une manière universelle pour l'expression de certains types d'émotion.

Le tableau 3 montre les résultats de la classification des émotions en deux classes : *hot anger* et *neutral*.

Locuteur	Méthode 1 (%)	Méthode 2 (%)	Méthode 3 (%)
CC	47,36	52,63	63,15

[6] Cette librairie est publiquement disponible à l'adresse suivante http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (vérifiée le 27/06/07)

CL	88,89	83,33	88,89
GG	76,47	64,71	70,59
MM	58,82	64,71	58,82
MK	94,44	88,88	88,89
JG	76,47	70,59	76,47
MF	95,83	91,67	95,83
MOY	76,89	73,78	77,52

Tableau 3 : Classification de deux classes d'émotion *Hot anger* versus *neutral*, selon la somme, le maximum ou la médiane des logs probabilités des pseudo-syllabes constituant le tour de parole.

Tel qu'illustré par le tableau 3, une classification basée sur le choix de l'état émotionnel de la pseudo-syllabe dont la probabilité est en position médiane, offre les meilleures performances suivie par la somme des probabilités.

Enfin, le tableau 4 montre le résultat d'une expérience similaire à la précédente, toutefois la composition de la classe négative, dans ce cas-ci, inclus *hot anger* et *cold anger* versus *neutral*.

Locuteur	Méthode 1(%)	Méthode 2(%)	Méthode 3(%)
CC	70	70	70
CL	75,86	79,31	79,31
GG	74,07	74,07	66,67
MM	74,07	74,07	74,07
MK	93,1	89,66	89,66
JG	72,41	72,41	65,52
MF	81,58	81,58	81,58
MOY	77,29	77,3	75,26

Tableau 4: Classification de deux classes d'émotion *anger* versus *neutral*, selon la somme, le maximum ou la médiane des probabilités des pseudo-syllabes constituant le tour de parole.

Contrairement au cas de classification entre deux classes (*hot anger* vs *neutral*), Les résultats des expériences entre les deux classes *hot anger* + *cold anger* vs. *neutral*, illustrés dans le tableau 4, c'est le maximum des probabilités des pseudo-syllabes qui donnent de meilleurs résultats. Cependant, globalement, c'est la somme des probabilités qui présentent des performances élevées et non fluctuantes.

1.5 Détection des émotions à partir du corpus DARPA Communicator

Pour les besoins de détection des émotions à partir du corpus DARPA Communicator, nous avons annotés une partie du corpus qui se limite aux mots de rétroactions. Ce choix n'est pas fortuit, il tient compte des points suivants :

- Nous avons montré dans la partie détection des dialogues problématiques que les marques de rétroactions permettaient d'améliorer les résultats.
- Annoter un tour de parole comprenant plusieurs mots prend plus de temps qu'annoter un seul mot.
- De manière générale, la tâche d'annotation des émotions étant une tâche difficile, nous avons observé un meilleur accord entre annotateur sur les tours de parole court (un mot) que sur les tours de parole long contenant plusieurs mots. Nous pensons que le contenu d'un tour de parole influence le choix de l'émotion. Dans le cas des mots de rétroaction, cette influence est réduite.

Dans ce qui nous décrivons les expériences de détection des émotions sur les mots de rétroactions extraits de DARPA Communicator.

1.5.1 Description du corpus DARPA Communicator : *mots de rétroactions*

Nous avons collecté 1554 mots de rétroaction extraits des 1242 dialogues du DARPA Communicator 2001 : 870 "*yes*" et ses variantes et 684 "*no*" et ses variantes. En plus des mots de rétroaction, nous avions aussi l'information sur le sexe du dans un fichier à part.

Les mots de rétroactions sont composés de mots "*ok*", "*okay*", "*yes*", "*yep*", "*yeah*", "*alright*", "*good*", "*no*", "*nop*", "*wrong*". Ces mots sont les mots de rétroaction les plus fréquemment observés dans le corpus.

1.5.1.1 Annotation des mots de rétroactions

Les mots de rétroactions ont été annotés selon le procédé décrit dans la section 3.4.1 par trois annotateurs : deux hommes et une femme. Afin de garantir une annotation le moins subjective possible, nous avons constitué un corpus par consensus, dans lequel l'émotion choisie est celle qui jouit d'un consensus d'au moins deux votes sur trois.

Chaque annotateur avait reçu un fichier Excel contenant cinq colonnes :

1. La première colonne contenait le lien hypertexte de chaque tour de parole contenant un mot de rétroaction.
2. La deuxième colonne contenait l'emplacement pour mettre un score de 0 ou 1 selon que l'émotion attribuée était une émotion non-négative.
3. La troisième colonne contenait l'emplacement pour mettre un score de 0 ou 1 selon que l'émotion attribuée était une émotion négative.
4. Une quatrième colonne pour indiquer un choix indécis.

Nous avons compté à peu près 20 secondes par mot de rétroaction pour un total approximatif de huit heures.

1.5.1.2 Accord entre annotateurs

Le calcul de l'accord entre annotateurs a été fait en utilisant le pairwise kappa (voir section 3.4.1). Le tableau 5 montre l'accord entre toutes les paires d'annotateurs. Le kappa global obtenu pour ce corpus est de 0,34. Selon l'échelle proposée par (Rietveld, 1993), il est clair que l'accord entre annotateurs est faible. Cela confirme aussi que la tâche d'annotation est une tâche ardue.

Annotateurs	A1	A2	A3
A1	1	0.28	0.27
A2	0.28	1	0.45
A3	0.27	0.45	1

Tableau 5: Accord entre paires d'annotateurs selon la formule de kappa définie dans (Cohen, 1960)

Néanmoins, plusieurs auteurs ont soulignés la sensibilité de la mesure kappa quand à la distribution des classes puisque P(E) dépend de la distribution des classes. En effet (Grove, 1981) entre autres explique qu'une distribution biaisée peut baisser la valeur maximal du kappa vers une valeur considérablement inférieur à 1 (accord parfais). En partant de cette constatation et sachant que la proportion de la classe négative versus non-négative est très biaisée dans notre application, nous avons décidé que l'accord était suffisent pour expérimenter cette tâche.

1.5.2 Approche

Dans cette expérience, nous avons utilisé une combinaison de quatre traits que nous avons décrits dans la section 3.3.1.2 ; à savoir, la fréquence fondamentale, le log de l'énergie, la largeur de bande du premier et second formant ainsi que le *power spectrum*. Comme les

tours de parole sont de courtes durées puisque ils sont constitués uniquement d'un mot, nous avons calculé les paramètres au niveau des trames et non au niveau des pseudo-syllabes comme c'était le cas pour le corpus LDC Emotional Prosody. Aussi ici, nous avons lissé ces paramètres avec des polynômes de Legendre d'ordre 7 et nous avons pris les trois derniers coefficients, c'est-à dire c_0, c_1, c_2.

Dans cette expérience, nous avons entraîné différents modèles implémentés dans la librairie weka[7].

1.5.3 Expériences et résultats

Nous avons comparé les résultats de quatre algorithmes en effectuant 10 validations croisées. L'utilisation de la validation croisée avec 10 partitions est la méthode la plus commune qui donne une estimation proche de la performance réelle d'un système. Le choix de modifier la distribution réelle des données dans le corpus d'entraînement donne la chance au système de mieux d'apprendre les traits caractéristiques des différentes classes lorsque celle-ci ont une même distribution. Ce choix est particulièrement important dans cette application, puisque les émotions négatives sont observées moins fréquemment que les émotions non-négatives. En effet, sur le corpus avec annotation obtenue par consensus la distribution réelle sur les 1447 mots de rétroactions était de 6 instances de la classe non-négative pour une seule instance de la classe négative, soit un baseline de 85,71%.

Classifieur	kappa	Correct	Précision		Rappel	F-score
Réseau de neurones	0,386	86,04	Neg	58,8	38,2	46,3
			N-Neg	89,1	95,0	92,0
Naïve Bayes continu	0,247	84,17	Neg	49,5	24,1	32,4
			N-Neg	87,1	95,4	91,0
Arbre de décision	0,230	84,53	Neg	52,2	21,1	30,0
			N-Neg	86,7	96,4	91,3
k-proche voisins	0,353	85,55	Neg	56,8	34,6	43,1
			N-Neg	88,6	95,1	91,7
Machine à support de vecteur	0	84,2	Neg	0	0	0
			N-Neg	0,84	1	91,4

Tableau 6 : Kappa, taux de classification correcte, précision, rappel et F-score obtenus pour le corpus DARPA Communicator

[7] Cette librairie est publiquement disponible à l'adresse suivante : http://www.cs.waikato.ac.nz/ml/weka/ (site vérifié le 26/07/2007)

Nous remarquons que les résultats tournent autour du baseline avec une petite amélioration avec les réseaux de neurones. Ce résultat était prévisible dans la mesure où le kappa sur ce corpus était assez faible (0,34) indiquant que les données étaient considérablement bruitées.

Toutefois, plusieurs améliorations sont à apporter, en particulier pour le rappel au niveau de la détection des émotions négatives. Plusieurs pistes sont envisageables, en particulier nous voudrions expérimenter des classifieurs qui permettent de pondérer les erreurs sur les classes de manière à attribuer un plus grand coût aux erreurs effectuées sur la classe négative. Enfin, d'autres traits prosodiques sont encore à expérimenter afin d'améliorer la détection des émotions négatives.

1.6 Détection des émotions à partir du corpus BELL

Les expériences que nous présentons dans cette section sont sensiblement les mêmes que celles décrites dans la section précédente. Ces expériences sont conduites sur le corpus de Bell constitué aussi uniquement de tours de parole composé de mots de rétroactions. Cependant contrairement au corpus précédent, nous avons séparés le *yes* et ses variantes ainsi que le *no* et ses variantes dans deux corpus séparés.

Le but de cette expérience est de comparer le tâche de détection des émotions négatives et non-négatives sur chacun de classes de mots *yes* et *no*. Cette distinction permettra de se concentrer sur la classe de mots dans laquelle la détection des émotions donne les meilleurs résultats. Cela permettra aussi de vérifier la corrélation des émotions négatives

1.6.1 Description du corpus de Bell : *mots de rétroactions*

Le corpus de Bell est composé de 1000 mots de rétroactions dont 500 pour le mot *yes* et ses variantes et 500 autres pour le mot *no* et ses variantes. Dans ce cas si, l'information sur le sexe des locuteurs n'était pas fournie. Afin de récupérer cette information, nous avons procédé à l'annotation du sexe en plus de l'annotation de l'émotion.

1.6.1.1 Annotation des mots de rétroactions

Nous avons répétée la même procédure d'annotation sur chacun des corpus de Bell. Cependant afin d'attester la qualité de nos annotations nous avons demandé à trois annotateurs du personnel de Bell à effectuer leur propre expérience d'annotation avec les mêmes conditions que nous avons définis. Le but de cet exercice était de :

- S'assurer que notre perception de ce que constitue une émotion négative était la même pour les deux groupes d'annotateurs.

- La comparaison des annotations des deux groupes nous permettrait de décider si nous pourrions nous-mêmes procéder à l'annotation d'autres corpus de manière cohérente avec la vision des annotateurs de Bell.

Cette évaluation des résultats des annotations est importante dans la mesure où les performances du système de détection d'émotions en dépendent directement.

1.6.1.2 Accord entre annotateurs

Nous avons analysé l'accord entre annotateurs au sein de l'équipe de Bell ainsi qu'au sein de l'équipe de l'ETS sur chacun des corpus de *yes* et *no*. Les figures 5 et 6 montrent les pairwise kappa entre les annotateurs de chaque équipe et ce pour chaque corpus.

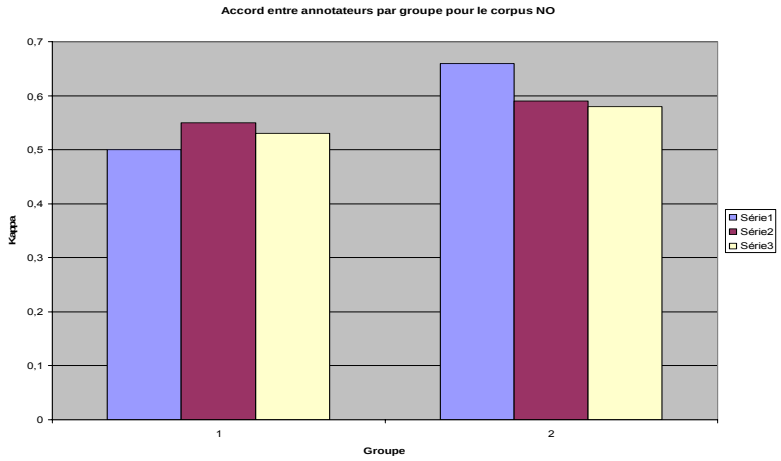

Figure 5 : Comparaison des pairwise kappa entre les paires d'annotateurs pour chaque groupe pour le corpus de no.

Le graphique de droite représente les pairwise kappa pour l'équipe de Bell tandis que le graphique de gauche représente ceux obtenus dans l'équipe de l'ETS. Nous remarquons que globalement l'accord entre les annotateurs du groupe de Bell est supérieur à celui de l'équipe de l'ETS. Par ailleurs, dans les deux cas, nous observons une bonne cohésion des pairwise au sein de chaque équipe.

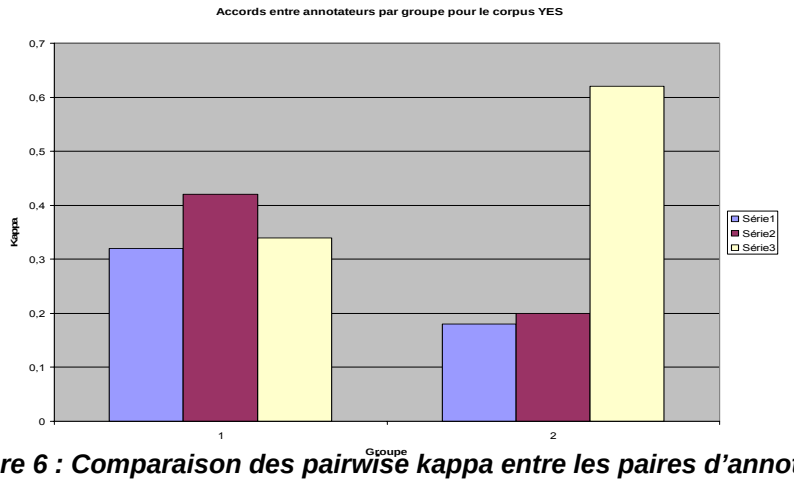

Figure 6 : Comparaison des pairwise kappa entre les paires d'annotateurs pour chaque groupe pour le corpus de yes.

Nous remarquons dans le cas du corpus de *yes* et ses variantes des valeurs de pairwise kappa très variées. En particulier au sein de l'équipe de Bell, nous remarquons que un des annotateurs est montre un très taux d'accord avec les deux autres annotateurs, ce qui explique les deux colonnes de droites nettement plus basses que celle la plus à droite. Dans une proportion moins prononcée, le même phénomène est observé au sein de l'équipe de l'ETS. De prima bord, nous pouvons conclure que pour les deux équipes la tâche de classification des émotions pour le corpus de *yes* est plus difficile que celle pour le corpus de *no*.

En comparant les kappas obtenus sur chaque corpus formé par consensus; c'est-à dire par vote au sein de chaque équipe, nous observons un net meilleur accord entre les deux équipes d'annotateurs pour le corpus de no et nettement plus faible pour le corpus de *yes*. Le tableau 7 montre les accords entre les deux équipes obtenus pour chaque corpus.

Équipe	Kappa
Corpus *yes*	0,24
Corpus *no*	0,49

Tableau 7: Accord entre le groupe de Bell et de lETS pour les corpus de *yes* et *no* obtenus par consensus

Une analyse plus détaillée des annotations par consensus pour chacune des équipes et chacun des corpus montre que le majeur désaccord entre les deux équipes se situe au niveau de l'annotation des émotions négatives. En particulier, l'analyse de la distribution des classes dans chacun des corpus de *yes* et *no* pour les deux équipes montre que l'équipe de l'ETS est plus biaisée sur les émotions négatives que l'équipe de Bell. Le tableau 8 montre la distribution dans chaque corpus pour chaque équipe.

	Négative	Non-négative
Corpus yes	Bell	1 : 69
	ETS	1 : 8
Corpus no	Bell	1 : 20
	ETS	1 : 4

Tableau 8 : Comparaison de la distribution des classes sur les corpus *yes* et *no* annoté par consensus par chaque équipe

Enfin, nous pensons que le meilleur accord obtenu sur le corpus de *no* pourrait être imputable à l'influence du sens du mot *no* qui en soit indique une négation. Il est possible que ce biais influence le choix des annotateurs dans leur choix des émotions négatives. Cette hypothèse est corroborée par la différence moins importante entre les distributions de la classe négative observée dans le corpus de *no* annoté par chaque équipe. Le ratio de la distribution des classes négatives pour le corpus de *yes* entre les deux équipes est d'un peu plus de 8 (69÷8), tandis que pour le corpus de *no*, il est de 5 (20÷4).

1.6.2 Approche

Nous avons effectuées ces expériences avec les mêmes traits et conditions d'évaluation que l'expérience effectuée sur le corpus DARPA Communicator. Les résultats que nous présentons ont été obtenus sur les corpus annotés par consensus par l'équipe de l'ETS.

1.6.3 Expérience et résultats

Les résultats que nous présentons sont ceux obtenus sur le corpus de *yes* et ses variantes et le corpus de *no* et ses variantes. Les expériences ont été effectuées avec la librairie weka. L'évaluation a été faite avec 10 validations croisées et les résultats sont la moyenne sur les 10 tests.

1.6.3.1 Corpus de *no* et ses variantes

Le corpus que nous avons utilisé est composé de 447 instance du mot *no* et ses variantes avec une distribution de 1 instance de classe négative pour 4 instances de la classe non-négative. Ainsi le baseline est de 80%.

Classifieur	kappa	Correct	Précision		Rappel	F-score
Réseau de neurones	0,463	85,07	Neg	63,1	48,8	55,0
			N-Neg	88,8	93,4	91,1

Naïve Bayes continu	0,524	87,30	Neg	73,7	50,0	59,6
			N-Neg	89,3	95,9	92,5
Arbre de décision	0,451	85,96	Neg	71,4	41,7	52,6
			N-Neg	87,8	96,2	91,8
k-proche voisins	0,392	84,63	Neg	66,0	36,9	47,3
			N-Neg	86,8	95,6	91,0
Machine à support de vecteur	0,523	87,30	Neg	73,7	50,0	59,6
			N-Neg	89,39	95,9	92,5

Tableau 9 : Kappa, taux de classification correcte, précision, rappel et F-score obtenus pour le corpus de *no* et ses variantes.

Nous remarquons que les taux de classifications sont sensiblement supérieurs au baseline. Toutefois les meilleurs kappas sont obtenus avec le Naïve Bayes utilisant un kernel gaussien et les machines à support de vecteur. Nous remarquons aussi que la détection de la classe négative est moins bonne que la classe non-négative. Plusieurs éléments peuvent être en causes :

- Le taux d'accord entre annotateurs était acceptable et non pas excellent ce qui laisse entendre que les données étaient un peu bruitées.
- La faible représentation de la classe négative dans les données. Plus une classe est sous représentée dans un corpus moins la généralisation est bonne pour cette classe.
- Les paramètres acoustiques proposés caractérisent mieux les instances de la classe non-négative que les instances de la classe négative.

1.6.3.2 Corpus de *yes* et ses variantes

Le corpus que nous avons utilisé est composé de 475 instance du mot *yes* et ses variantes avec une distribution de 1 instance de classe négative pour 8 instances de la classe non-négative. Nous calculons un baseline de 88,8% sur la classe dominante qui est en occurrence la classe non-négative. Toutefois dans la mesure où c'est plutôt la détection d'émotion négative qui nous intéresse, nous porterons plus d'attention sur la performance des classifieurs pour la détection de cette classe. Pour ce corpus, une performance de base serait de 20% de détection correcte d'émotions négatives.

Classifieur	kappa	Correct	Précision		Rappel	F-score
Réseau de neurones	0,308	88	Neg	43,6	32,7	37,4
			N-Neg	92,0	94,8	93,4
Naïve Bayes continu	0,188	88,84	Neg	47,1	15,4	23,2
			N-Neg	90,4	97,9	94,0
Arbre de décision	0,04	88,42	Neg	28,6	3,8	6,8
			N-Neg	89,3	98,8	93,8
k-proche voisins	0,161	87,78	Neg	36,4	15,4	21,6

			N-Neg	90,3	96,7	93,4
Machine à support de vecteur	0,109	89,07	Neg	0	0	0
			N-Neg	89,1	1	94,2

Tableau 10 : Kappa, taux de classification correcte, précision, rappel et F-score obtenus pour le corpus des *yes* et ses variantes.

Sans trop de surprise, nous remarquons que les taux de classification sont très proches du baseline. Ce résultat était prévisible dans la mesure où il n y avait à la base pas de bon accord entre annotateurs; c'est à dire que les données étaient très bruitées. Par ailleurs, nous remarquons que le meilleur accord entre la classification générée et les données de tests est obtenu avec les réseaux de neurones qui sont connus pour être robustes aux données bruitées. Enfin, c'est aussi l'algorithme qui détecte le mieux les émotions négatives avec un F-score de 37,4 supérieure minimum de 20% obtenu si le système générait systématiquement la classe non-négative.

1.7 Conclusion et travaux futurs

Nous avons présenté les résultats de plusieurs expériences de détection des émotions. Les premiers résultats obtenus sur le corpus LDC Emotional Prosody montrent que la tâche de classification des tours de parole contenant plusieurs mots est plus difficile que la classification des émotions à partir des tours de parole contenant uniquement un mot de rétroaction. Ce résultat est important dans la mesure où les systèmes de dialogues Personne-Machine sont essentiellement conçus pour minimiser les dialogues libres afin d'assurer un meilleur contrôle du contenu du dialogue. Aussi dans ce contexte les stratégies dialogiques développées pour ces systèmes se basent beaucoup sur les *yes*/*no* questions ainsi que les questions à choix multiples. Dans ces travaux nous avons montré qu'il était possible de détecter les émotions de manière satisfaisante à partir des mots de rétroaction contenant le *no* et ses variantes. Toutefois, les analyses de données ont bien montré que l'amélioration des résultat passe non seulement par la recherche d'autres traits prosodiques et de meilleurs algorithmes de classification, mais surtout par l'amélioration des annotations fournies au système au moment de l'apprentissage.

Dans nos travaux futurs, nous proposons de continuer la recherche de nouveaux traits prosodiques pour améliorer la reconnaissance des émotions négatives en particulier. Nous projetons d'annoter plus de données de *no* et ses variantes afin d'obtenir un meilleur corpus avec un bon accord entre annotateurs. Enfin, nous voudrions élargir ses travaux pour détecter les émotions à partir des mots clés du domaine utilisés pour répondre aux questions à choix multiples.

BIBLIOGRAPHIE

LIVRE

Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.Jerry M. Mendel. Fuzzy Logic Systems for Engineering: A Tutorial. *In Proceedings of the IEEE, 83(3). 345-377. March 1995.*

CHAPITRE DE LIVRE

Eckman. « Basic Emotions ». In T. Dalgleish and T. Power (Eds.) *The Handbook of Cognition and Emotion* Pp. 45-60. Sussex, U.K.: John Wiley & Sons, Ltd., 1999.

ACTES DE CONFÉRENCES : DÉTECTION DES ÉMOTIONS

Cohen J. “A coefficient of agreement for nominal scale”. *Educational and Psychological Measurement*, 20:37-46. 1960.

Di Eugenio B. “On the usage ok Kappa to evaluate agreement on coding tasks”. Proceedings of the conference LREC2000, the Second International Conference on Language Resources and Evaluation, Athens, Greece, 2000.

Lin V and H-C. Wang, “Language identification using pitch contour information” in Proc. ICASSP 2005, Philadelphia, PA, march 2005, pp. 601–604.

Dehak N., P. Dumouchel and P. Kenny. “Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification” to appear in IEEE Transactions on Audio, Speech and Language Processing, 2007.

Razak, A.A., Abidin, M.I.Z., Komiya, R. “Emotion pitch variation analysis in Malay and English voice samples”. Communications APCC 2003. The 9th Asia-Pacific Conference on Volume 1, Issued 21-24 Sept. 2003 Page(s): 108 - 112 Vol.1

Grove W., N.N Andreasan, P. McDonald Scott *et al.* “Reliability Studies of Psychiatric Diagnosis”. Theory and Practice., 38:408-413.

Krippendorff K. “Content analysis: an Introduction to its Methodology”. Beverly Hills: Sage Publications. 1980.

Rietveld T., and R. van Hout. *Statistical Techniques for the study of Language and Language Behaviour*. Mouton de Gruyter. 1993.

Ang 2000] Jeremy Ang, Radjip Dhillon, Ashley Krupsky, Elisabeth Shriberg, and Andreas Stolcke. "Prosody based automatic detection of annoyance and frustration in human computer dialog". In Proceedings of the ICSLP, 2002.

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. N{\"o}th. "Desperately Seeking Emotions: Actors, Wizards, and Human Beings". In Proceedings of ISCA ITRW Speech and Emotion, pages 195-200, 2000.

R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W Fellenz. "What a neural net needs to know about emotion words". In Proceedings of the CSCC'99, pages 5311-5316, 1999.

Eckman, 1999. Basic Emotions. In T. Dalgleish and T. Power (Eds.) The Handbook of Cognition and Emotion Pp. 45-60. Sussex, U.K.: John Wiley & Sons, Ltd., 1999.

N.M. Fraser and G.N Gilbert. ``Simulating Speech Systems``. *Computer Speech and Language*, 5:81-99. 1991.

T. Gordon. "PET: Parent Effectiveness Training". Plume Books, New York, 1970.

Rongqing Huang, Changxue Ma, "Toward a Speaker-Independant Realt-time Affect Detection System ". Proceedings of the 18th International Conference on Pattern Recognition. 2006

H.-B. Kang. " Affective content detection using HMMs". In Proceedings of the eleventh ACM international conference on Multimedia Berkeley. pages 259--262, CA, 2003.

Kallirroi Georgila, K.; Lemon, O. & Henderson, J. (2005): Automatic annotation of COMMUNICATOR dialogue data for learning dialogue strategies and user simulations, Ninth Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL: DIALOR), June 2005, also: DIALOR conference proceedings.

L. Kuncheva and C. Whitaker. "Measure of diversity in classifier ensembles". *Machine Learning*, vol51, pp 181-207, 2003.

C. M. Lee and S. S. Narayanan. "Toward detecting emotions in spoken dialogs". IEEE Transactions on Speech and Audio Processing, 13:293--303, 2005.

Chul Min lee, Shrikanth S. Narayanan, and R. pieraccini. "Combining prosodic and language information for emotion recognition". In Proceedings of the international conference on spoken language processing 2002. 2002.

M.A. Rahurkar, J. H. L. Hansen, J. Meyerhoff, and G. Saviolakis amd M. Koenig. "Frequency B and Analysis for Stress Detection using a Teager Energy Operator based Feature. In Proceedings of the ICSLP 2002. 2002.

Jerry M. Mendel. "Fuzzy Logic Systems for Engineering: A Tutorial. In Proceedings of the IEEE, 83(3). 345-377. March 1995.

K. Forbes-Riley and D. Litman. "Predicting emotion in spoken dialogue from multiple knowledge sources". In Proceedings of the 4th Meeting of HLT/NAACL. Boston 2004.

Mihai Rotaru and Diane J Litman. "Using word-level pitch features to better predict student emotions during spoken tutoring dialogs". In Proceedings of the INTERSPEECH-2005. 2005.

I. Shafran and M. Mohri. "A comparison of classifiers for detecting emotion from speech". In Proceedings of the Prosodics, Speech, and Signal Processing. 2005.

Glossaire

Rappel : nombre de réponses correctes du système/nombre de réponse correcte

Précision : nombre de réponses correctes du système/nombre de réponses du système

F-score : moyenne harmonique pondérée qui combine la précision et le rappel selon la formule suivante $(1+ \beta^2)$ (Précision x Rappel)/(β^2 x Précision + Rappel)