ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

COMME EXIGENCE PARTIELLE
À L'OBTENTION DE LA
MAÎTRISE EN GÉNIE
M.Ing.

PAR
ATTABI, Yazid

# RECONNAISSANCE AUTOMATIQUE DES ÉMOTIONS À PARTIR DU SIGNAL ACOUSTIQUE

MONTRÉAL, LE 17 FÉVRIER 2008

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

M. Pierre Dumouchel, directeur de mémoire
Département de génie logiciel et des technologies de l'information à l'École de technologie supérieure

M. Éric Granger, président du jury
Département de génie de la production automatisée à l'École de technologie supérieure

Mme Sylvie Ratté, membre du jury
Département de génie logiciel et des technologies de l'information à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 22 JANVIER 2008

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

## REMERCIEMENTS

Au terme de ce travail, je tiens à remercier mon directeur de recherche Monsieur Pierre Dumouchel, Professeur à l'École de technologie supérieure et vice-président scientifique du Centre de recherche informatique de Montréal (CRIM) pour m'avoir donné l'opportunité de faire cette recherche, pour le grand intérêt qu'il a montré ainsi que pour son évaluation de ce manuscrit.

J'aimerai remercier les professeurs Éric Granger et Sylvie Ratté qui ont accepté de participer au jury d'évaluation de ce mémoire. Je remercie également les membres de l'équipe de recherche sur l'émotion au CRIM où le travail de ma maîtrise a été effectué, notamment le doctorant Dehak Najim.

J'adresse mes remerciements à tous ceux qui m'ont enseigné et en particulier à mon Professeur à Alger Monsieur Hammoudi Nadir.

Enfin, je remercie ma famille qui m'a toujours encouragé à poursuivre mes études.

# RECONNAISSANCE AUTOMATIQUE DES ÉMOTIONS À PARTIR DU SIGNAL ACOUSTIQUE

ATTABI, Yazid

## RÉSUMÉ

Nous nous intéressons à la détection automatique des appels problématiques dans un contexte réel de centres d'appels téléphoniques. Nous utilisons l'information sur l'état émotionnel du locuteur, véhiculée par le signal acoustique, pour détecter les problèmes de compréhension entre un locuteur et un système de dialogue humain-machine. Notre contribution se situe à deux niveaux. Au premier niveau, nous avons développé un système de reconnaissance automatique des émotions (RAE) basé sur les traits de type MFCC, avec la célérité et l'accélération, extraits au niveau d'une trame, analysés à l'échelle d'un *énoncé*, et modélisés par un mélange de gaussiennes. Nous avons optimisé les performances de ce système en ajustant trois types de paramètres : le nombre de mélanges de gaussiennes, l'utilisation de coefficients MFCC d'ordre supérieur (20 versus 13 coefficients) et l'utilisation d'un modèle du monde (UBM) pour l'entraînement des modèles GMM. Le système a été entraîné et testé pour reconnaître les classes des émotions du corpus de données LDC Emotional Prosody (LDC). D'après les résultats obtenus, nous avons apporté une amélioration de l'ordre de 11% par rapport aux meilleurs résultats de l'état de l'art utilisant le même corpus de données pour l'expérience *neutre* vs *tristesse* alors que nous avons reproduit les meilleures performances pour l'expérience *neutre* vs *colère* et pour l'expérience avec 15 classes d'émotions.

Notre seconde contribution est l'expérimentation d'un nouveau modèle de système de RAE basé sur l'information prosodique à long terme obtenue par une approximation des courbes de l'énergie et de la fréquence fondamentale par des coefficients de *polynômes de Legendre* sur une échelle d'analyse appelée *pseudosyllabe*. Afin de mesurer l'efficacité de ce type de trait à long terme et de l'unité d'analyse, nous avons réalisé une comparaison de performance entre ce système et un système exploitant l'information prosodique à court terme (niveau de trame) sur l'échelle d'un *énoncé*. Les taux de reconnaissance obtenus avec un système basé sur la *pseudosyllabe* et les coefficients de *polynômes de Legendre et* expérimenté avec le corpus LDC, sont nettement supérieurs à ceux d'un système basé sur l'*énoncé* et l'information à court terme. Le gain relatif réalisé est de l'ordre de 6% pour la reconnaissance des émotions *neutre* vs *colère*, tandis que ce gain est de l'ordre 91% pour *neutre* vs *tristesse*. Enfin, nous avons obtenu une amélioration de l'ordre de 41% pour la détection de 15 classes d'émotions.

**Mots-clés** : reconnaissance automatique des émotions, signal de la parole, détection des dialogues problématiques, prosodie, MFCC, GMM, MAP, énoncé, pseudosyllabe, polynôme de Legendre, LDC Emotional Prosody.

RECONNAISSANCE AUTOMATIQUE DES ÉMOTIONS À PARTIR DU SIGNAL ACOUSTIQUE

ATTABI, Yazid

## ABSTRACT

This research concerns the automatic identification of problematic dialogs in a context of real telephone calls centers. Information on the speaker emotional state conveyed by the acoustic signal is used to detect human-machine communication problems.

Our contribution lies at two levels. At the first level, an automatic emotion recognition system based on MFCC with speed and acceleration features extracted for each frame, analyzed on *utterance* level and modeled by a Gaussian mixture is developed. In order to maximize the performance of our system, we have tuned three types of parameters: the number of Gaussian mixtures, the use of the UBM model to train emotion classes models and the extension of the feature vector dimension to the higher order MFCC coefficients (from 13 to 20 coefficients). The system has been trained and tested using the LDC Emotional Prosody (LDC) corpus. The results show that we obtain a relative improvement of about 11% compared to the best results of the state of the art using the same corpus to recognize *neutral* vs. *sadness* emotions. We have also achieved the best performance for the recognition of *neutral* vs. *anger* emotions and for the experience with fifteen emotion classes.

Our second contribution is the experimentation of a new model of Automatic Emotional Recognizer system based on long-term prosodic information based on approximation of the energy and the fundamental frequency curves by *Legendre polynomials* coefficients computed at *pseudo-syllable* level. To evaluate the efficiency of this type of long-term feature with this analysis unit, we compare the performance of this system to that of a system operating on short term information (at frame level), analyzed on the *utterance* level. The accuracy achieved with a system based on the *pseudo-syllable* and *Legendre polynomials* coefficients and trained and tested using LDC corpus is much higher than the one obtained with a system based on the *utterance* and the short term information. The relative gain obtained is about 6% for the recognition of *neutral* vs. *anger* emotions. We also improved the accuracy by 91% for *neutral* vs. *sadness* experience and, finally, the relative gain is 41 % for the experience with 15 emotions.

## TABLE DES MATIÈRES

Page

## LISTE DES TABLEAUX

Page

## LISTE DES FIGURES

## LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

| | | |
|---|---|---|
| AD | Approche dynamique | |
| AS | Approche statique | |
| CPL | Coefficients de polynôme de Legendre | |
| EM | Expectation Maximization | |
| FFT | Fast Fourier Transform | Transformée de Fourier rapide |
| GMM | Gaussian Mixture Model | Modèle à mélange de gaussiennes |
| HMM | Hidden Markov Model | |
| ICT | Information à court terme | |
| ILT | Information à long terme | |
| CVP t-Test | K-Fold CrossValidation Paired t Test | |
| K-NN | K-Nearest Neighbor | K-plus proches voisins |
| LD | Linear Discriminant | |
| LDC | Linguistic Data Consortium Emotional Prosody | |
| LOSO | Leave-One-Speaker-Out | |
| LFPC | Log Frequency Power Coefficients | |
| LPC | Linear Predictive Coefficients | Codage par prédiction linéaire |
| LPCC | Linear Prediction Cepstral Coefficients | |
| MAP | Maximum A Posteriori | |
| MFCC | Mel-Frequency Cepstral Coefficients | Coefficients cepstraux sur l'échelle Mel |
| ML | Maximum Likelihood | |
| PL | Polynôme de Legendre | |

| | |
|---|---|
| PS | Pseudosyllabe |
| RAE | Reconnaissance automatique des émotions |
| RAP | Reconnaissance automatique de la parole |
| SMEG | Système-MFCC-Énoncé-GMM |
| SPPLG | Système-Prosodie-Pseudosyllabe-polynôme-Legendre-GMM |
| UBM | Universal Background Model |

# CHAPITRE 1

# INTRODUCTION

## 1.1 Problématique

Savoir conserver la satisfaction de ses clients est un facteur déterminant dans le succès de nombreuses entreprises. La mise en place de centres d'appel est un moyen qui permet aux entreprises de prendre en charge la relation client à distance. Auparavant opérés par des téléphonistes, les nouveaux centres d'appel utilisent de plus en plus des systèmes de compréhension automatique de la parole afin de répondre plus rapidement et plus économiquement aux requêtes des clients. Ces systèmes ne sont pas exempts d'erreur et peuvent même provoquer la colère du client. La qualité du système de reconnaissance de la parole, sur lequel se base le système de compréhension automatique de la parole des centres d'appel, engendre souvent des difficultés de communication aux appelants. Comment traiter ces cas?

Dans le cadre de nos travaux de recherche, nous nous intéressons à la détection automatique des appels problématiques dans un contexte réel de centres d'appels téléphonique. Un appel est considéré problématique si le client se trouve dans l'une des deux situations suivantes :

- il n'a pas réussi à exécuter la tâche désirée;
- il a difficilement réussi la tâche faisant l'objet de son appel.

La détection précoce d'un appel problématique d'un client en colère permettrait à la machine d'entreprendre plusieurs stratégies de gestion de l'échec de l'appel, parmi celles-ci :

1. restreindre et guider le dialogue;
2. s'excuser et apaiser l'usager par une voix douce et des remarques plaisantes;
3. traiter l'appel en priorité en le redirigeant vers un opérateur humain dans le cas où le problème persiste ;

4. détecter les bris de communication afin de rappeler les clients mécontents qui, découragés par la tournure du dialogue, ont mis fin prématurément au dialogue.

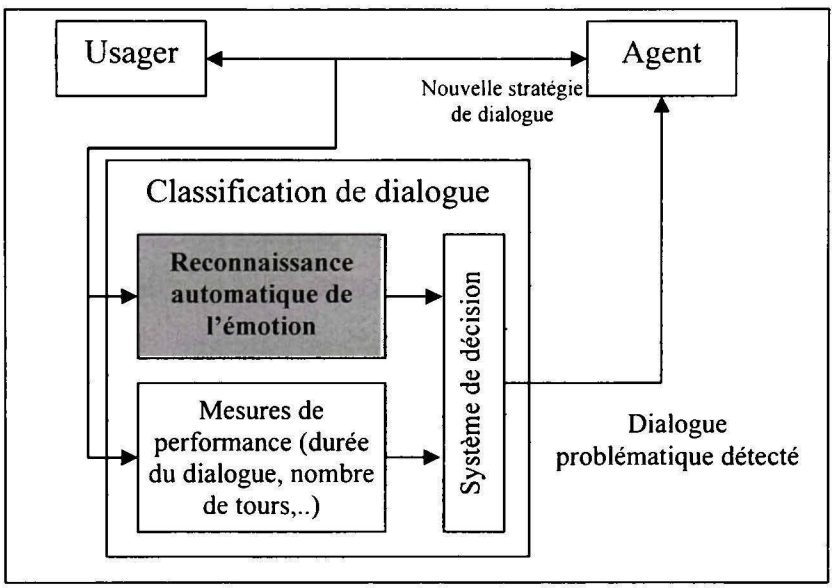


**Figure 1.1 Système de gestion de dialogues humain-machine.**

Une des catégories d'informations que nous pouvons utiliser pour détecter les appels problématiques est l'ensemble des mesures du cadre PARADISE (PARAdigm for DIalog System Evaluation) (Walker et al., 1997), telles que la *mesure du succès de la tâche*, les *mesures d'efficience* comme la *durée du dialogue* et le *nombre de tours de parole* par exemple. Ces mesures ont été étudiées dans la première partie du projet «Gestion des émotions dans les dialogues humain-machine» développé par ÉTS (École de technologie supérieure) et le CRIM (Centre de recherche informatique de Montréal) avec la collaboration de Bell Canada Corp. et de Patrimoine canadien. Dans la deuxième partie de ce projet, objet de ce présent mémoire, nous nous intéressons à l'information sur l'état émotionnel du locuteur véhiculé par le signal acoustique pour détecter les appels problématiques tel qu'illustré par la Figure 1.1. En effet, la frustration ou la colère, par exemple, peut être

considérée comme un avertisseur qui renseigne sur l'état d'un client faisant face à des difficultés induites par l'utilisation du système.

## 1.2 Objectif

Dans ce projet, notre objectif consiste à concevoir un système de reconnaissance automatique d'émotion (RAE) à partir du signal acoustique de parole en utilisant les techniques de traitement du signal et de reconnaissance de formes. Nous nous intéresserons exclusivement à l'information paralinguistique (prosodique et spectrale) pour la reconnaissance de l'état émotionnel du locuteur. Le système sera entraîné et testé pour reconnaître les classes des émotions du corpus de données utilisé (LDC Emotional Prosody). L'état émotionnel du locuteur sera classé dans l'une des deux grandes catégories : négative et non négative. L'émotion détectée du locuteur sera exploitée ultérieurement par le gestionnaire du système de dialogue afin d'appliquer une des stratégies de gestion de l'échec de l'appel.

Notre contribution dans cette recherche se situe à deux niveaux. Au premier niveau, notre objectif est de développer un système de RAE, nommé SMEG, compétitif avec les résultats de l'état de l'art. Ce système est basé sur les traits de type MFCC, avec la célérité et l'accélération, extraits au niveau de la trame, analysés à l'échelle d'un *énoncé* et modélisés par un mélange de gaussiennes. Un tel système n'a pas été encore expérimenté avec le corpus LDC Emotional Prosody. Les performances de ce système sont optimisées en ajustant certains paramètres qui se situent au niveau de l'extraction des caractéristiques acoustiques et celui du paramétrage du classificateur.

Notre seconde contribution consiste à étudier, pour la première fois dans le domaine de la RAE, la pertinence d'un système, nommé SPPLG, basé sur l'information prosodique à long terme obtenue par une approximation des courbes de l'énergie et de la fréquence fondamentale par des coefficients de *polynômes de Legendre* (CPL) sur une échelle d'analyse appelée *pseudosyllabe*. Le choix d'une unité d'analyse autre que *l'énoncé* est motivé, d'une part, par le fait que la charge de l'émotion n'est pas uniformément distribuée sur tout

l'énoncé, c'est-à-dire qu'il existe des segments d'un même énoncé émotionnellement plus saillants que d'autres. D'autre part, nous constatons qu'il existe des énoncés véhiculant plus d'une émotion. Enfin, l'approximation par des coefficients des polynômes de Legendre a pour but de rechercher une information à long terme plus robuste que les valeurs statistiques utilisées par la communauté scientifique.

Après avoir exposé dans le premier chapitre de ce mémoire quelques notions théoriques sur l'émotion dans le domaine de la psychologie, nous passerons en revue les différents travaux effectués sur la reconnaissance automatique des émotions à partir du signal de parole appliquée au domaine de l'ingénierie. Dans le deuxième chapitre, nous présenterons l'architecture typique d'un système de reconnaissance automatique de l'émotion et nous décrirons en détail chacun de ses composants. Les deux chapitres suivants seront dédiés à la présentation de chacune des deux contributions que nous avons tenté d'apporter. Dans le chapitre 3, nous exposerons la méthodologie suivie pour la conception du système de RAE et les résultats des performances obtenus que nous voulions compétitifs avec ceux de l'état de l'art. Dans le chapitre 4, nous proposerons et étudierons un nouveau système de RAE basé sur une nouvelle information à long terme extraite au niveau d'une nouvelle unité d'analyse, avant de conclure et de présenter les travaux futurs dans le dernier chapitre.

classification les plus efficaces qui s'appliquent à ce type de problème pour concevoir un système de RAE le plus robuste.

## 2.2 Définitions et modèles des émotions

Dans cette section, nous aborderons, sous l'angle du domaine de la psychologie, la description de l'émotion et ses modèles théoriques sur lesquels se sont basées les études réalisées dans ce domaine. Nous discuterons notamment des incertitudes qui caractérisent la définition des émotions, ses catégories, ainsi que leurs modèles.

### 2.2.1 Définition et limitations

Un des problèmes majeurs rencontré dans le domaine de la psychologie est l'absence d'un consensus autour de la définition de l'émotion et des différents types d'émotions. L'absence du consensus apparaît déjà dans la distinction entre l'émotion et les autres types d'états affectifs du locuteur que sont l'humeur (mood), attitudes interpersonnelles (interpersonal stances), attitudes et traits de personnalité affectifs. Afin de distinguer entre ces cinq types d'états, Scherer (Scherer 2000) a suggéré une approche caractéristique basée sur un ensemble de critères distinctifs tels que l'intensité, la durée, le degré de synchronisation des sous-systèmes organismiques, le degré de dépendance avec un événement déclencheur et le degré d'impact de l'état affectif sur le comportement. Ainsi à partir du Tableau 2.1, nous constatons que l'émotion se distingue par une plus grande intensité, mais de courte durée, qui possède un grand impact sur le comportement de l'individu et qui est susceptible de changer plus rapidement.

### 2.2.2 Modèles psychologiques des émotions (Scherer, 2000; Scherer, 2003)

Deux théories ont fortement déterminé le passé de la recherche dans le domaine des modèles émotionnels, à savoir la théorie de l'émotion discrète et la théorie dimensionnelle.

**Tableau 2.1**

Caractéristiques de démarcation entre les différents états affectifs
(Tiré de Scherer 2003, traduit de l'anglais)

| Type d'état affectif : brève définition (exemples) | Intensité | Durée | Synchronisation | Dépendance d'un événement | Évaluation de stimulation | Rapidité du changement | Impact sur le comportement |
|---|---|---|---|---|---|---|---|
| **Émotion** : Relativement bref épisode, d'une réponse synchronisée de l'ensemble ou la plupart des sous-systèmes organismiques, en réponse à l'évaluation d'un événement externe ou interne, d'une importance majeure (colère, tristesse, joie, peur, honte, fierté, désespoir) | ++ - +++ | + | +++ | +++ | +++ | +++ | +++ |
| **Humeur** : État affectif diffus, plus prononcé comme changement de sentiment subjectif, mais de faible intensité, d'une durée relativement longue, souvent sans apparente cause (joyeux, triste, irritable, apathique, déprimé, vif) | + - ++ | + + | + | + | + | + + | + |
| **Stances interpersonnelles** : Attitude affective envers une autre personne lors d'une interaction, caractérisant l'échange interpersonnel (distant, froid, chaud, favorable, méprisant) | + - ++ | + - + + | + | + + | + | +++ | ++ |
| **Attitudes** : Relativement durable, conviction teintée d'affection, préférence et prédisposition envers des objets ou des personnes (plaisant, affectueux, détestable, appréciable, désirant) | 0 - ++ | ++-+++ | 0 | 0 | + | 0 - + | + |
| **Traits de personnalité** : Émotionnellement chargé, caractères de personnalité et tendances de comportement stables, typique pour une personne (nerveux, anxieux, imprudent, morose, hostile, envieux, jaloux) | 0 - + | +++ | 0 | 0 | 0 | 0 | + |
| 0 : bas, + : moyen, ++ : élevé, +++ : très élevé, -: indique une étendue | | | | | | | |

**La théorie de l'émotion discrète** se concentre particulièrement sur l'étude de l'expression motrice ou du schème du comportement adaptatif. Les théoriciens de cette tradition

proposent l'existence d'un petit nombre, compris entre 9 et 14, d'émotions de base ou fondamentales caractérisées par des modèles de réponse très spécifiques en physiologie ainsi que dans les expressions faciales et vocales. La plupart des études qui se sont intéressées à l'effet vocal des émotions ont utilisé ce modèle et ont choisi d'examiner les effets de la joie, de la tristesse, de la peur, de la colère et de la surprise.

**La théorie dimensionnelle** s'intéresse principalement à la description verbale des sentiments subjectifs (subjective feeling). Dans cette tradition, les différents états émotionnels sont cartographiés dans un espace de deux ou trois dimensions. Les deux dimensions principales consistent en la dimension de valence (agréable-désagréable) et une dimension de l'activité (actif / passif). La troisième dimension, si elle est utilisée, représente souvent le contrôle ou la puissance intellectuelle.

Dans leurs travaux relatifs à l'effet de l'émotion sur la voix, les partisans de ce modèle se limitent souvent à l'étude des différences qui existent entre l'état émotionnel positif versus négatif et actif versus passif.

**Le modèle d'émotion à composant (componential model of emotion)** est un nouveau modèle d'émotion qui ne cesse de gagner d'influence. Le champ d'intérêt de ce modèle ne se limite pas à l'étude des sentiments subjectifs (telle que la théorie dimensionnelle) ni au nombre supposé d'émotions de base (comme c'est le cas avec la théorie discrète). Ce modèle met l'accent sur la variabilité des différents états émotionnels tels occasionnés par les différents types de patrons d'évaluation (appraisal patterns en anglais). Ils offrent également la possibilité de modéliser les différences qui existent entre les membres de la même famille d'émotion, telle que la colère forte, la colère froide et le mépris. Aussi d'après (Scherer, 2003), ces approches fournissent une base solide pour une élaboration théorique des mécanismes qui sont censés sous-tendre la relation émotion-voix et permettent de générer des hypothèses très concrètes qui peuvent être testées empiriquement.

### 2.2.3 Type de corpus des émotions

Nous distinguons essentiellement deux catégories de corpus de données d'émotions utilisées dans le domaine de la détection automatique des émotions : les émotions naturelles et les émotions simulées.

**Les émotions naturelles** sont des enregistrements d'états émotionnels vécus naturellement et spontanément. Ce corpus de données est caractérisé par une très haute validité écologique. L'inconvénient est que ces données sont très limitées en nombre de locuteurs, de courtes durées, souvent de piètre qualité, en plus d'être très difficile à étiqueter en classes d'émotions.

**Les émotions simulées** des émotions produites par des acteurs professionnels ou semi-professionnels en se basant sur le nom de la classe d'émotion et/ou des scénarios typiques. Cette méthode représente le moyen préféré pour constituer les données dans ce domaine. Cependant, certains griefs sont adressés à cette méthode. Il est soupçonné, par exemple, que l'émotion simulée est stéréotypée et qu'elle soit caractérisée par une plus grande intensité que l'émotion naturelle.

Notons qu'une troisième catégorie d'émotions, appelée émotions induites, est utilisée dans le domaine de la psychologie afin de déterminer si la stimulation des états émotionnels du locuteur produit les changements acoustiques correspondants. Les émotions de cette catégorie sont induites expérimentalement, par exemple en provoquant le stress par l'exposition du sujet à des tâches difficiles à accomplir sous la contrainte de délai ou par la présentation d'images ou de films émouvants.

## 2.3 État de l'art sur la RAE

La reconnaissance automatique de l'émotion à partir de la parole fait objet d'un intérêt croissant durant ces dernières années en raison de l'étendue du domaine d'application pouvant bénéficier de cette technologie. À titre d'exemple, un système de détection des émotions peut servir au développement de systèmes à interaction humaine-machine efficace,

naturelle et sensible au comportement de l'utilisateur. Utilisé dans un contexte d'enseignement à distance, un tel système tutoriel serait capable de savoir si l'utilisateur est ennuyé, découragé ou irrité par la matière enseignée et pourra par conséquent changer le style et le niveau de la matière dispensée, fournir une compensation et un encouragement émotionnel ou accorder une pause à l'utilisateur (Li, Zhang et Fu, 2007) (Zhu et Luo, 2007).

La RAE peut être utilisée dans beaucoup d'autres contextes applicatifs. Il peut :

- servir à détecter la fatigue et l'influence de l'alcool chez un conducteur automobile afin d'activer des routines de sécurité (Schuller, 2002);
- permettre au système équipant la voiture de fournir au conducteur un support et un encouragement au cours d'une pénible expérience de conduite (Jones et Jonsson, 2007);
- détecter la présence d'émotions extrêmes, principalement la peur, dans le cadre de la surveillance dans les lieux publics (Clavel et al., 2006);
- prioriser automatiquement les messages cumulés dans la boîte vocale selon différents axes affectifs tels que l'urgence, la formalité, la valence (heureux vs triste) et l'excitation (calme vs excité) pour alerter le propriétaire du compte et lui permettre d'écouter les messages importants en premier (Inanoglu et Caneel, 2005);
- utiliser les traits spéciaux véhiculés par les émotions pour le développement de systèmes de vérification automatique du locuteur (VAL) plus robustes et précis (Panat et Ingole, 2008);
- évaluer l'urgence d'un appel et par conséquent prendre une décision, dans le cadre d'un centre d'appel médical offrant un service de conseils médicaux aux patients (Devillers et Vidrascu, 2007);
- améliorer le service à la clientèle lorsque, le système de RAE est intégré aux systèmes de réponse interactive à la voix (Interactive Voice Response) dans les centres d'appels commerciaux (Lee et Narayanan, 2003; Maganti, Scherer et Palm, 2007; Petrushin, 2000; Yacoub et al., 2003).

Dans les différents travaux réalisés, une multitude de méthodes et de techniques ont été expérimentés pour aborder cette problématique. Afin de mieux présenter et analyser les motivations ayant conduit à chacun des choix ou approches, nous proposons de procéder à une classification de cet éventail d'études entreprises, en fonction de quatre critères :

1. le type d'unité d'analyse utilisé pour la reconnaissance des émotions;
2. la nature de l'information paralinguistique employée comme traits caractéristiques qui peut être soit de type prosodique ou spectrale;
3. la portée temporelle des traits utilisée, qui peut être soit de type à court terme ou à long terme;
4. le type d'approche choisi pour concevoir le classificateur (dynamique, statique, floue).

Nous présenterons dans ce qui suit les travaux effectués selon les quatre critères de classification présentés.

### 2.3.1 Travaux selon le type d'unité d'analyse

Un critère que nous pouvons examiner pour la classification des systèmes de RAE est le type d'unité d'analyse utilisée dans la reconnaissance des émotions. L'unité d'analyse représente le segment de données de base extrait de l'énoncé que nous soumettons au classificateur pour déterminer sa classe d'émotion. Dans les travaux réalisés à ce jour, cinq types d'unités ont été expérimentés (Voir Figure 2.1) :

1. **L'énoncé** : unité d'analyse sur laquelle est basée la quasi-totalité des travaux effectués. Les vecteurs de traits de la totalité de l'énoncé sont extraits et soumis en une seule entrée au classificateur pour déterminer la catégorie de l'émotion. Parmi les travaux qui se sont basés sur cette unité, citons (Beritelli et al., 2007; El Ayadi, Kamel et Karray, 2007; Grimm et Kroschel, 2005; Inanoglu et Caneel, 2005; Li et al., 2007; Lin et Wei, 2005; Pao et al., 2005; Petrushin, 2000; Seppänen, Väyrynen et Toivanen, 2003; Sethu, Ambikairajah et Epps, 2007; Vlasenko et al., 2007).

2. **Le phonème** : le *phonème* représente la plus petite unité de son d'une langue. Le choix de cette unité est motivé par l'hypothèse que l'état émotionnel d'un locuteur affecte les *phonèmes* d'un énoncé avec des intensités très variables. L'énoncé est alors segmenté en phonèmes et chaque classe de *phonèmes* est modélisée séparément. Afin de vérifier cette hypothèse, Lee et ses collègues (Lee et al., 2004) ont réalisé deux expériences; dans la première, un classificateur HMM (Hidden Markov Model ou modèles de Markov cachés) émotionnel générique est utilisé. Ce classificateur est entraîné en utilisant les données d'apprentissage de toutes les classes de *phonèmes*. Dans la deuxième expérience, des HMM par classe de *phonèmes* sont utilisés. Le taux de reconnaissance obtenu avec le modèle HMM générique est de 64.77% alors que les résultats pour les modèles HMM par classe de *phonèmes* sont respectivement de 72.16%, 54.86%, 47.43%, 44.89% et 55.11% pour les classes voyelles, semi-voyelles, nasales, consonnes occlusives et fricatives. Les résultats d'une classification basée sur la combinaison des modèles des cinq classes de *phonèmes* atteignent 75.57%. Ces résultats montrent d'une part que les classes des *phonèmes* ne véhiculent pas, dans les mêmes proportions, la même charge émotionnelle. Les voyelles sont émotionnellement plus saillantes que les autres classes. D'autre part, une classification basée sur une modélisation par classe de *phonèmes* offre de meilleures performances que la classification à partir d'un modèle générique.

3. **La syllabe** : pour la même raison ayant conduit à l'expérimentation de l'unité *phonème*, Schuller et ses collègues ont procédé à une segmentation de l'*énoncé* basée sur l'unité *syllabe* (Schuller et al., 2007b). Les performances du système basé sur l'unité syllabique sont inférieures à ceux du système basé sur l'*énoncé*.

4. **Le mot** : l'unité d'analyse *mot* a été testée et comparée avec les performances d'un système basé sur l'unité *énoncé* (Schuller et al., 2007a). D'après les résultats obtenus, l'unité *mot* est préférable à l'*énoncé* à condition qu'un système de segmentation par *mot* efficace soit disponible.

5. **Le fragment** : l'autre unité d'analyse expérimentée est le *fragment*. L'énoncé est segmenté automatiquement en *fragments*, en fonction des propriétés acoustiques de l'énoncé (Schuller et al., 2007b). Les résultats obtenus sur le corpus de données utilisé montrent que les performances du système basé sur l'unité *fragment* sont meilleures que celles obtenues avec l'unité *syllabe*, mais qu'elles restent en deçà de celle de l'*énoncé*.

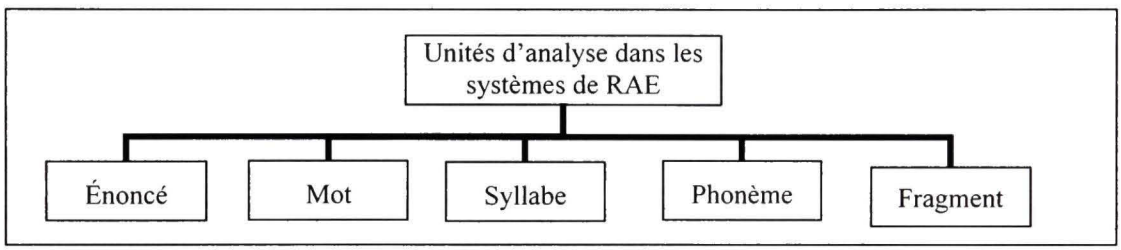


**Figure 2.1 Types d'unités d'analyse étudiés dans le domaine de la RAE**

### 2.3.2 Travaux selon la portée des traits caractéristiques

Par ailleurs, nous pouvons considérer une autre dichotomie dans les travaux réalisés dans le domaine de la RAE, en se basant sur la portée de l'information utilisée comme traits caractéristiques. Cette information est classée dans l'une des deux classes : l'information à court terme ou l'information à long terme.

**Information à court terme (ICT)** : information locale qui s'étale généralement sur un intervalle de temps, appelé trame, allant de 10 ms à 30 ms, cadencé à chaque 10 ms. Chaque trame constitue un vecteur de traits caractéristiques. La séquence des vecteurs de l'énoncé véhicule sa structure temporelle.

Nous distinguons deux types d'information à court terme utilisés dans les systèmes de RAE : l'information spectrale, véhiculée en général par les coefficients spectraux, et l'information

prosodique (temporelle) représentée par l'énergie, la fréquence fondamentale (F0) et le rythme.

Le vecteur de traits à court terme le plus utilisé est composé des coefficients Mel-Frequency Cepstral Coefficients (MFCC), leurs dérivées premières et leurs dérivées secondes. Parmi les travaux qui se sont basés sur ce type de vecteur, nous retrouvons ceux de (Gao, Chen et Su, 2007; Hung, Quénot et Castelli, 2004; Kim et al., 2007b; Neiberg, Elenieus et Laskowski, 2006; Nwe, Foo et De Silva, 2003; Pirker, 2007; Vogt et Andre, 2005; Wahab, Chai et De, 2007).

D'autres paramètres spectraux ont été également expérimentés, en combinaison avec les coefficients MFCC (Nwe, Foo et De Silva, 2003; Pao et al., 2005). Ces paramètres sont issus des techniques de compression de la parole telles que le codage par prédiction linéaire (Linear Predictive Coefficients, LPC), Linear Prediction Cepstral Coefficients (LPCC), Log Frequency Power Coefficients (LFPC) et Perceptual Linear Prediction (PLP).

L'information prosodique a été également utilisée en tant que traits à court terme, parfois seule (Huang et Ma, 2006; Sethu, Ambikairajah et Epps, 2007) ou combiné avec les coefficients cepstraux tels les MFCC et les LPCC (Kwon et al., 2003; Li, Zhang et Fu, 2007; Lin et Wei, 2005).

**Information à long terme (ILT)** : information à long terme caractérise l'énoncé dans sa globalité. Cette information est représentée sous forme de valeurs statistiques pour une séquence temporelle de valeurs. Ce sont les variables de type prosodique qui sont les plus couramment utilisées, telles que la fréquence fondamentale, l'énergie, le débit de la parole, le nombre et la durée des silences, le rapport de la durée de la région voisée à la région non voisée et celles de type phonétique telles que les formants et leurs bandes passantes, ainsi que des variations de paramètres, dont le tremblotement (Shimmer) et le vacillement (Jitter).

La moyenne, l'écart type, le maximum, le minimum, la déviation, la pente, le quartile, le taux de passage par zéro (Zero-Crossing-Rate) sont des exemples de fonctions statistiques généralement utilisées.

Parmi les travaux qui se sont basés sur l'utilisation des statistiques de la prosodie, nous citons (Alvarez et al., 2007; Giripunje et Bawane, 2007; Grimm et Kroschel, 2005; Inanoglu et Caneel, 2005; Lee, Narayanan et Pieraccini, 2002; Petrushin, 2000; Pittermann et Pittermann, 2007a; Schuller et al., 2007b; Seppänen, Väyrynen et Toivanen, 2003; Yacoub et al., 2003; Yu et al., 2001).

Certains auteurs ont combiné, dans leurs vecteurs de traits de type à long terme, entre les statistiques de la prosodie et les statistiques des coefficients spectraux, que nous retrouvons par exemple dans les travaux de (Kwon et al., 2003; Schuller et al., 2007a; Vlasenko et al., 2007). Rares sont les travaux où seules les statistiques des coefficients cepstraux sont utilisées comme traits caractéristiques à l'instar des travaux de (Beritelli et al., 2007).

Les types de traits caractéristiques que nous avons cités jusqu'ici appartiennent tous à la catégorie d'information paralinguistique. Notons que certains auteurs ont introduit, dans leurs modélisations, l'information linguistique, notamment l'information au niveau lexical et au niveau du discours (les actes de dialogue) en combinaison avec l'information paralinguistique (Chen et al., 2006; Chuang et Wu, 2004; Devillers et Vidrascu, 2007; Forbes-Riley et Litman, 2004; Schuller et al., 2005). L'introduction de cette information linguistique a permis de réaliser un certain gain en performance, mais au coût de perdre la propriété d'indépendance du système de RAE de tout langage, qui est une caractéristique des systèmes basés uniquement sur l'information paralinguistique.

### 2.3.3 Travaux selon l'approche de classification

Un autre critère qui distingue les différents travaux effectués est le type d'approche utilisé pour la classification des émotions. La plupart des méthodes utilisées s'inscrivent dans l'une des trois approches suivantes :

**L'approche dynamique** : les classificateurs HMM sont utilisés dans ce type d'approche. L'avantage des modèles HMM est qu'ils permettent de modéliser la structure temporelle des énoncés. Parmi les travaux basés sur cette approche, nous pouvons citer les travaux de (El Ayadi, Kamel et Karray, 2007; Huang et Ma, 2006; Inanoglu et Caneel, 2005; Kwon et al., 2003; Lee et al., 2004; Li et al., 2007; Lin et Wei, 2005; Nwe, Foo et De Silva, 2003; Pao et al., 2005; Pirker, 2007; Pittermann et Pittermann, 2007a; Sethu, Ambikairajah et Epps, 2007; Vlasenko et al., 2007; Wagner, Vogt et Andre, 2007).

**L'approche statique :** dans cette approche des classificateurs de type statique, tels que les GMM, SVM, LDC et les réseaux de neurones sont utilisés pour la modélisation. Avec ce type d'approche, il est possible également de récupérer l'information sur la structure temporelle de l'énoncé en optant pour l'information de portée à long terme comme traits caractéristiques. Quand les performances de plusieurs modèles de classificateurs sont testées par un auteur, c'est en général le classificateur GMM qui offre de meilleures performances. Parmi les travaux où ce type de classificateur a été expérimenté, nous retrouvons (Alvarez et al., 2007; Beritelli et al., 2007; Kwon et al., 2003; Lee, Narayanan et Pieraccini, 2002; Li, Zhang et Fu, 2007; Lin et Wei, 2005; Neiberg, Elenieus et Laskowski, 2006; Pao et al., 2005; Petrushin, 2000; Schuller et al., 2007b; Seppänen, Väyrynen et Toivanen, 2003; Yacoub et al., 2003).

**L'approche logique floue** : la troisième approche utilisée est basée sur un système d'inférence flou. Ce choix est motivé par les incertitudes qui caractérisent les émotions, particulièrement afin de résoudre le problème de l'absence de frontières claires entre les définitions des différentes catégories d'émotions ainsi qu'au problème de chevauchement des classes dans la réalisation acoustique des émotions. Plusieurs travaux ont examiné cette approche (Giripunje et Bawane, 2007; Grimm et Kroschel, 2005; Lee et Narayanan, 2003).

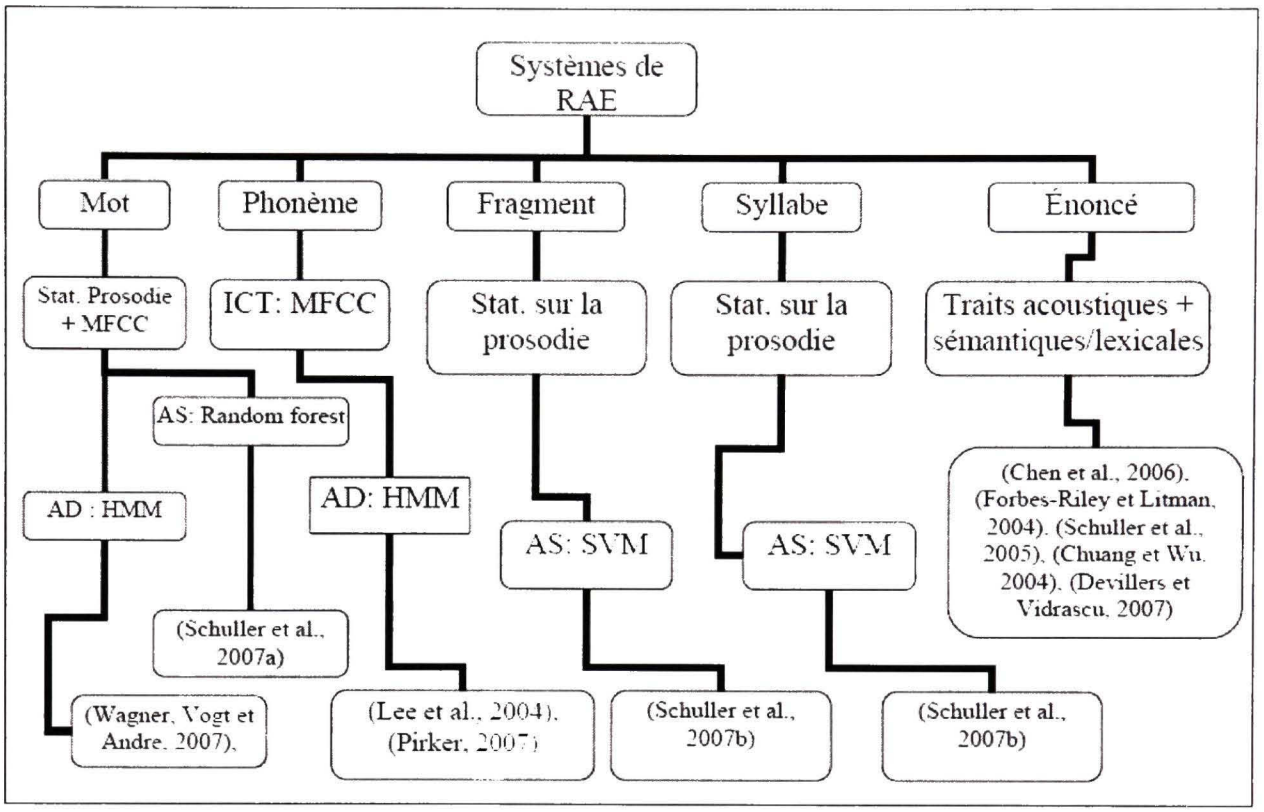


**Figure 2.2 Classification des travaux sur la RAE selon l'unité d'analyse, le type de traits caractéristiques et l'approche utilisée pour la classification des émotions.**

En résumé, nous pouvons caractériser un système de reconnaissance automatique des émotions par le type d'unité d'analyse, la portée et le type de traits caractéristiques utilisées ainsi que le type d'approche utilisé pour la modélisation des classes d'émotions. Nous proposons, à travers le diagramme de la Figure 2.2, une synthèse des différents types de systèmes de RAE rencontrés dans la revue de littérature.

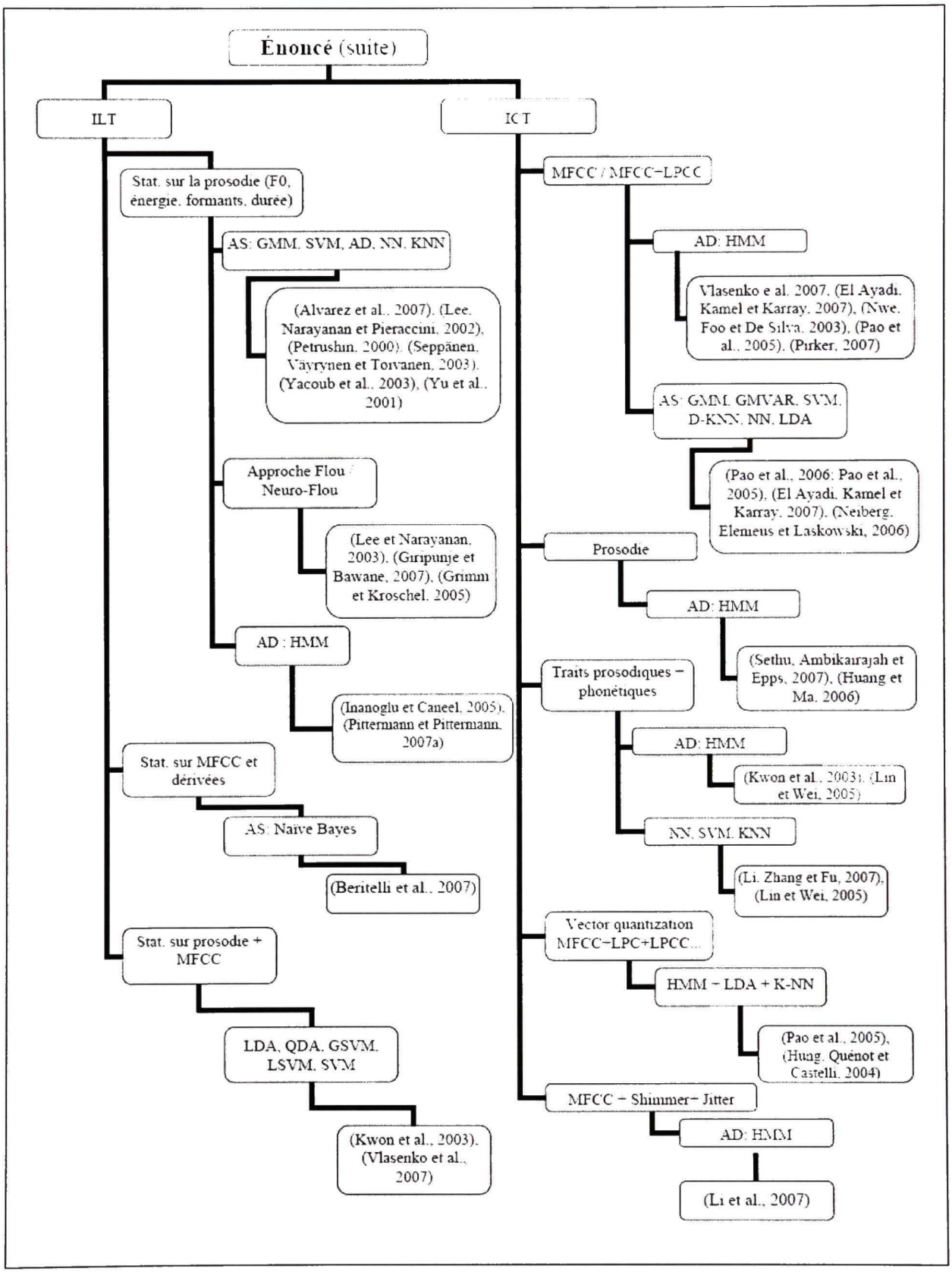


**Figure 2.2 Classification des travaux sur la RAE (suite).**

Une question reste en suspens. Elle concerne le choix du meilleur système. En fait, il est difficile d'établir une comparaison entre les différents systèmes et d'en choisir le meilleur, et ce pour les raisons suivantes. D'abord, notons l'absence d'un corpus de données commun des émotions qui servirait de test de performance (benchmark). Aussi, les différents systèmes dressés dans la Figure 1.1 diffèrent dans la quantité et la qualité (le nom) de classes d'émotions utilisées dans les expérimentations. Enfin, l'utilisation de protocoles d'expérimentation dissemblables (test dépendant du locuteur versus test indépendant du locuteur) rend la comparaison de performance des différents systèmes une opération hasardeuse, voire impossible.

## 2.4 Le corpus LDC Emotional Prosody

Le corpus Emotional Prosody du Linguistic Data Consortium, que nous désignerons sous l'acronyme LDC dans le reste de ce mémoire, est un corpus d'émotions de langue anglaise utilisé par plusieurs chercheurs et dont les énoncés et l'annotation sont plus accessibles en comparaison à d'autres corpus utilisés par la communauté scientifique. L'utilisation de ce corpus facilite donc la comparaison de performances entre systèmes. Nous nous intéresserons dans ce qui suit à la description du corpus LDC ainsi qu'aux performances des systèmes de RAE entraînés à partir de ce corpus. Ces résultats nous serviront de référence pour mesurer les performances de notre système qui sera également entraîné et testé à partir des données LDC. Une description des autres corpus de parole émotionnelle utilisés est donnée au Tableau 2.2

### 2.4.1 Description du corpus LDC

Les données utilisées dans les expériences sont tirées du corpus de données LDC. Ce sont des fichiers WAVE, un par acteur, enregistrés sur les deux canaux et échantillonnés à 22.05 KHz. Les deux micros utilisés sont de type Shure SN94 monté sur un pied de micros et un casque Seinnheiser HMD 410.

**Tableau 2.2**

**Corpus de parole émotionnelle existants**

(Tiré de Humaine 2009)

| Identifier | Emotional content | Emotion elicitation methods | Language |
|---|---|---|---|
| TALKAPILLAR (Beller, 2005) | Neutral, happiness, question, positive and negative surprised, angry, fear, disgust, indignation, sad, bore | Contextualised acting | French |
| Reading-Leeds database (Greasley et al., 1995; Roach et al., 1998, Stibbard 2001) | Range of full blown emotions | Natural | English |
| France et al. (France et al., 2000) | Depression, suicidal state, neutrality | Natural | English |
| Campbell CREST database, ongoing (Campbell 2002; see also Douglas-Cowie et al. 2003) | Wide range of emotional states and emotion-related attitudes | Natural | English Japanese Chinese |
| Capital Bank Service and Stock Exchange Customer Service (as used by Devillers & Vasilescu 2004) | Mainly negative - fear, anger, stress | Natural: call center human-human interactions | English |
| SYMPAFLY (as used by Batliner et al. 2004b) | Joyful, neutral, emphatic, surprised, ironic, helpless, touchy, angry, panic | Human machine dialogue system | German |
| DARPA Communicator corpus (as used by Ang et al. 2002) See Walker et al. 2001 | Frustration, annoyance | Human machine dialogue system | English |
| AIBO (Erlangen database) (Batliner et al. 2004a) | Joyful, surprised, emphatic, helpless, touchy (irritated), angry, motherese, bored, reprimanding, neutral | Human machine: interaction | German |
| Fernandez et al. (Fernandez et al. 2000, 2003) | Stress | Induced | English |
| Tolkmitt and Scherer (Tolkmitt and Scherer, 1986) | Stress (both cognitive & emotional) | Induced | German |

**Tableau 2.2**

**Corpus de parole émotionnelle existants (suite)**

(Tiré de Humaine 2009)

| **Identifier** | **Emotional content** | **Emotion** elicitation methods | **Language** |
|---|---|---|---|
| Iriondo et al. (Iriondo et al., 2000) | Desire, disgust, fury, fear, joy, surprise, sadness | Contextualised acting | Spanish |
| Mozziconacci (Mozziconacci, 1998) Note: database recorded at IPO for SOBUproject 92EA. | Anger, boredom, fear, disgust, guilt, happiness, haughtiness, indignation, joy, rage, sadness, worry, neutrality | Contextualised acting | Dutch |
| McGilloway (McGilloway, 1997; Cowie and Douglas-Cowie, 1996) | Anger, fear, happiness, sadness, neutrality | Contextualised acting | English |
| Belfast structured Database (Douglas- Cowie et al. 2000) | Anger, fear, happiness, sadness, neutrality | Contextualised acting | English |
| Danish Emotional Speech Database (Engberg et al., 1997) | Anger, happiness sadness, surprise neutrality | Acted | Danish |
| Groningen ELRA corpus number S0020 (www.elda.org/catalogue/en/speech/S0020.html) | Database only partially oriented to emotion | Acted | Dutch |
| Berlin database (Kienast & Sendlmeier 2000; Paeschke & Sendlmeier 2000) http://www.expressive-speech.net/ | Anger- hot, boredom, disgust, fear-panic, happiness, sadness-sorrow, neutrality | Acted | German |
| Pereira (Pereira, 2000) | Anger (hot), anger (cold), happiness, sadness, neutrality | Acted | English |
| van Bezooijen (van Bezooijen, 1984) | Anger, contempt disgust, fear, interest joy, sadness shame, surprise, neutrality | Acted | Dutch |
| Abelin (Abelin 2000) | Anger, disgust, dominance, fear, joy, sadness, shyness, surprise | Acted | Swedish |
| Yacoub et al (2003) (data from LDC, www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28 | Neutral, hot anger, cold anger, happy, sadness, disgust, panic, anxiety, despair, elation, interest, shame, boredom, pride, contempt | Acted | English |

À chaque fichier audio est associé un fichier de transcription incluant un alignement temporel des énoncés. Les émotions sont simulées par sept acteurs professionnels (4 femmes et 3 hommes) dans 15 catégories d'émotion[1] : *anxiety, boredom, contempt, cold anger, despair, disgust, elation, happiness, hot anger, interest, neutral, panic, pride, sadness, et shame,* sélectionnées d'après l'étude de *Banse & Scherer* selon l'expression émotionnelle vocale en allemand (Banse et Scherer, 1996). Le texte utilisé pour simuler les classes d'émotion est sémantiquement neutre composé de dates et de nombres.

L'enregistrement des énoncés du corpus LDC est réalisé, en demandant à chacun des sept acteurs de répéter une phrase, en simulant un des états émotionnels, jusqu'à ce qu'il juge que la qualité de la simulation est satisfaisante. C'est pourquoi seule la dernière répétition est considérée comme information utile pour l'apprentissage des modèles. Toutes les répétitions des énoncés, utiles et non utiles, ainsi que d'autres sons et paroles, regroupés dans la catégorie *MISC* dans le fichier de transcription, sont sauvegardés dans le fichier audio original. Nous avons donc, procédé au filtrage des données, pour ne garder que les énoncés correctement simulés.

Les lignes suivantes représentent un extrait du fichier de transcription du locuteur *CC*, prononçant « *Two thousand one* », en simulant l'état émotionnel neutre. Seule la dernière, des cinq répétitions, qui s'étend du temps 60.79 s. jusqu'à 61.97 s. est valide.

---

[1] Nous traduisons ces quinze classes d'émotions par : anxiété, ennui, mépris, colère froide, désespoir, dégoût, exaltation, joie, colère forte, intérêt, neutre, panique, fierté, tristesse et honte.

```
0.00 38.74 A: [MISC]
38.88 40.85 A: (Emotional continuum) Neutral no emotion
42.26 43.28 A: neutral,Two thousand one
49.25 50.23 A: neutral,Two thousand one
52.63 53.53 A: neutral,Two thousand one
56.20 57.16 A: neutral,Two thousand one
60.79 61.97 A: neutral,Two thousand one
62.66 70.91 A: [MISC]
73.16 74.25 A: neutral,no emotion
75.04 75.97 A: neutral,Two thousand two
...
```

L'extrait suivant représente le même fichier de transcription après l'opération de filtrage :

```
60.79 61.97 A: neutral,Two thousand one
80.84 81.95 A: neutral,Two thousand two
...
```

Après l'opération d'épuration, le corpus totalise 1161 énoncés répartis, par acteur et par classe d'émotion, selon les valeurs du Tableau 2.3

Pour nos expériences, nous avons également procédé à l'extraction de l'un des deux canaux des enregistrements audio, pour nous en servir lors de l'extraction des traits.

### 2.4.2 Travaux sur le corpus LDC

Les principaux travaux basés sur les données LDC ont été réalisés par Yacoub (Yacoub et al., 2003), Huang (Huang et Ma, 2006) et Sethi (Sethu, Ambikairajah et Epps, 2007). Tous ces chercheurs ont utilisé le protocole de la validation croisée par locuteur pour tester les performances de leurs systèmes, ce qui nous permet d'effectuer une comparaison de performances entres ces systèmes. Leurs systèmes ont été évalués principalement dans trois expériences. Ces expériences diffèrent par les classes d'émotions participantes. Dans la première expérience, que nous appellerons *Exp I*, il s'agit de reconnaître les énoncées appartenant à deux classes d'émotions : *neutre* et *colère forte*. Dans la deuxième expérience

(*Exp II*), nous retrouvons la classe d'émotion *neutre* et la classe *tristesse*. Dans la troisième expérience (*Exp III*), les classificateurs sont entraînés pour reconnaître quinze (15) classes d'émotion. Les résultats des différentes expériences rapportées par les trois auteurs sont présentés dans le Tableau 2.4, ainsi que des informations sur le type de traits et de classificateurs utilisés.

**Tableau 2.3**

**Répartition des données LDC en fonction du locuteur et de la classe d'émotion**

| | CC | CL | GG | JG | MF | MK | MM | Total |
|---|---|---|---|---|---|---|---|---|
| **Anxiété** | 10 | 12 | 10 | 10 | 11 | 11 | 21 | **85** |
| **Ennui** | 11 | 10 | 11 | 11 | 19 | 10 | 12 | **84** |
| **Mépris** | 11 | 11 | 10 | 11 | 11 | 11 | 10 | **75** |
| **Col. froide** | 11 | 10 | 11 | 11 | 21 | 10 | 11 | **85** |
| **Désespoir** | 11 | 11 | 10 | 11 | 13 | 11 | 11 | **78** |
| **Dégoût** | 11 | 10 | 22 | 13 | 6 | 10 | 11 | **83** |
| **Exaltation** | 10 | 11 | 10 | 10 | 11 | 11 | 10 | **73** |
| **Joie** | 11 | 10 | 11 | 11 | 10 | 20 | 11 | **84** |
| **Colère** | 10 | 11 | 10 | 10 | 11 | 11 | 10 | **73** |
| **Intérêt** | 11 | 10 | 11 | 11 | 10 | 20 | 11 | **84** |
| **Neutre** | 8 | 7 | 7 | 7 | 9 | 7 | 7 | **52** |
| **Panique** | 11 | 10 | 10 | 11 | 10 | 10 | 20 | **82** |
| **Fierté** | 13 | 10 | 11 | 11 | 10 | 10 | 11 | **76** |
| **Tristesse** | 11 | 10 | 11 | 11 | 12 | 10 | 10 | **75** |
| **Honte** | 9 | 11 | 10 | 9 | 11 | 11 | 11 | **72** |
| **Total** | **159** | **154** | **165** | **158** | **175** | **173** | **177** | **1161** |

Note : Les titres de colonnes representent les initiales des locuteurs.

**Tableau 2.4**

Caractéristiques et performances de systèmes de RAE de l'état de l'art, entraînés à partir des données LDC

| Auteurs | Traits | Classificateur | Taux de classification | | |
|---|---|---|---|---|---|
| | | | Exp I | Exp II | Exp III |
| **Yacoub et al.** | Statistiques sur F0, énergie, durée, tremblotement et vacillement | NN | 94 % | 50 % | 9 % |
| **Huang et al.** | F0, énergie, ZC, pente de l'énergie | HMM | 98 % | 69 % | 18 % |
| **Sethu et al.** | F0, énergie, ZC, pente de l'énergie, avec gaussianisation des traits | HMM | 95 % | - | - |

NN : réseau de neurones. ZR (Zero Crossing) : nombre de passage par zéro du signal de parole dans une fenêtre de traitement donnée.

Nous citons également les travaux de Lee (Lee, 2004) et ceux de Yu et ses collègues (Yu et al., 2004). Lee s'est intéressé à la reconnaissance des sept classes d'émotions (ennui, dégoût, joie, colère, neutre, panique, tristesse) en utilisant comme traits les statistiques de la prosodie (F0, énergie, durée), le premier et second formant et leurs bandes passantes. Les expériences ont été réalisées avec deux types de classificateurs : le discriminant linéaire (LD) et les K-plus proches voisins (K-NN). Chacun des classificateurs est entraîné séparément selon le sexe du locuteur (homme ou femme). Les meilleurs résultats ont été obtenus avec le classificateur LD pour les données des hommes (47.04%) et avec le classificateur K-NN pour les données des femmes (47.14%). Contrairement aux travaux des trois auteurs précédents, un protocole de validation croisée à 10 groupes, sélectionnés aléatoirement, a été utilisé dans

ces travaux, un protocole qui n'est pas strictement indépendant du locuteur. Dans les travaux de Yu, un classificateur de type SVM a été utilisé avec les statistiques de la prosodie comme traits pour reconnaître sept classes d'émotions (ennui, intérêt, joie, colère, neutre, panique, tristesse). Le taux de reconnaissance du système, qui fonctionne en mode dépendant du locuteur, est égal à 69%.

## 2.5 Conclusion

Dans ce chapitre, nous avons présenté l'émotion du point de vue psychologique, qui est caractérisée par un manque de consensus autour de sa définition et de son modèle théorique. Nous avons examiné également les différentes méthodes utilisées pour constituer un corpus des émotions. Nous avons par la suite exposé et classifié les différents travaux sur la RAE selon quatre critères. Enfin, nous avons décrit et présenté les résultats de systèmes de RAE basés sur le corpus LDC.

Dans le chapitre suivant, nous décrirons l'architecture d'un système de RAE en général et examinerons en détail le fonctionnement et les méthodes utilisées dans chacun de ses modules.

# CHAPITRE 3

# SYSTÈME DE RECONNAISSANCE AUTOMATIQUE DES ÉMOTIONS

La reconnaissance de l'état émotionnel du locuteur peut être vue comme un problème de reconnaissance de forme. Le système admet en entrée un signal de la parole et produit en sortie la catégorie d'émotion véhiculée par la voix du locuteur.

Nous distinguons deux phases dans le processus de la reconnaissance de forme : l'*apprentissage* et le *test* tel qu'illustré par la Figure 3.1 Lors de la phase d'apprentissage, le modèle associé à chacune des classes d'émotions est créé puis des frontières de décision sont établies pour délimiter ces classes. La phase d'apprentissage se déroule en quatre étapes :

1) le prétraitement, qui consiste à préparer le signal d'entrée avant l'extraction des traits. Comme exemple, nous pouvons citer les opérations d'extraction d'un canal d'un fichier audio stéréo, la segmentation du signal en énoncés selon l'alignement indiqué dans le fichier de transcription ou encore l'élimination du bruit et du silence;
2) l'extraction des caractéristiques qui consiste à mesurer des propriétés du signal qui permettent de distinguer le type du signal;
3) la sélection des caractéristiques les plus pertinentes pour la classification parmi les traits extraits;
4) la modélisation qui permet de fixer les différents paramètres ajustables (les frontières de décision par exemple) du classificateur choisi pour bien représenter le concept selon les données d'apprentissage.

Lors de la phase de test, la position d'une nouvelle donnée par rapport aux frontières de décision établit la classe d'appartenance de cette donnée. Notons que seules les deux premières des quatre étapes précédentes, c'est-à-dire le prétraitement et l'extraction des caractéristiques, sont nécessaires pour la phase de test.

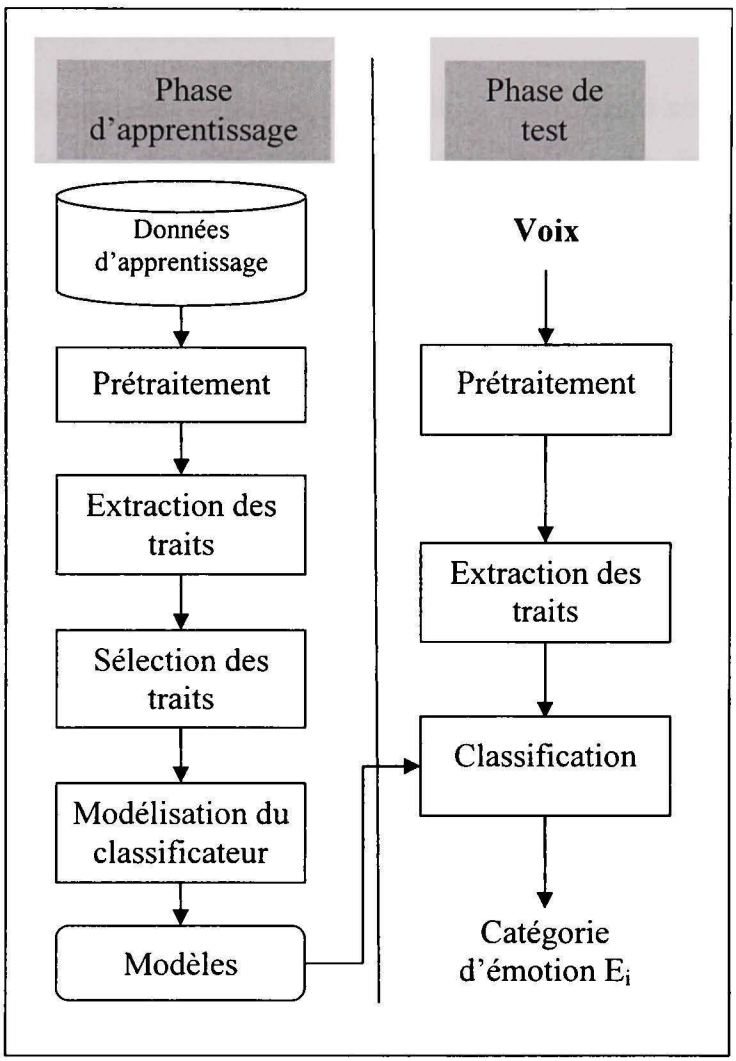


**Figure 3.1 Architecture d'un système de RAE à partir de la parole pour les phases d'apprentissage et de test.**

Dans ce chapitre, nous aborderons avec plus de détails quelques aspects théoriques des méthodes utilisées au cours de la phase d'extraction des caractéristiques, de la sélection des traits et de la conception du classificateur.

## 3.1 Extraction des caractéristiques

Les caractéristiques permettent de réduire la quantité d'information contenue dans un signal de parole échantillonné. Ces caractéristiques sont sélectionnées afin de distinguer une forme appartenant à une classe par rapport aux formes des autres classes. Nous avons vu qu'il existait, dans le domaine de la RAE, deux types d'information pour distinguer les classes d'émotions : l'information linguistique et l'information paralinguistique. Dans ce mémoire nous allons considérer uniquement l'information paralinguistique, étant donné que nous voulons concevoir un système de RAE indépendant de la langue, appliqué pour un centre d'appel où deux langues sont d'usage : le français et l'anglais. Ce choix est motivé également par le résultat des études de Tickle (Tickle, 2000) qui suggère qu'il est possible de réaliser un système pour la détection de l'émotion, indépendant du locuteur, basé uniquement sur l'information acoustique (Huang et Ma, 2006). Ces études établissent qu'un humain arrive presque aussi bien à reconnaître les émotions véhiculées par des énoncés dans une langue étrangère que celles dans sa langue natale. L'expérience est réalisée en utilisant des énoncés dépourvus de sens. C'est pourquoi nous intéresserons dans ce qui suit, à la description de l'information paralinguistique et particulièrement à l'information prosodique et à l'information spectrale, les MFCC en occurrence.

### 3.1.1 La prosodie

« *It isn't **what** you said; it's **how** you said it!* » (Huang, Acero et Hon, 2001)

La prosodie est un canal parallèle au contenu sémantique du message parlé dans les conversations quotidiennes à travers lequel l'auditeur peut percevoir les intentions et l'état

émotionnel de l'orateur[2]. C'est à travers la prosodie également que le locuteur peut donner à l'énoncé un ton de déclaration, d'une question, ou d'une commande (Huang, Acero et Hon, 2001).

Les prononciations d'un même mot peuvent avoir des prosodies substantiellement différentes sans affecter l'identité du mot : les phones peuvent être longs ou courts, forts (loud) ou doux (soft) et avoir des fréquences fondamentales variées. La prosodie s'intéresse à la relation qui lie la durée, l'amplitude et le pitch au son. Les traits prosodiques sont dits suprasegmentaux, dans le sens où leurs domaines d'interprétation sont au-delà de la limite de l'unité du phone (O'Shaughnessy, 2000).

#### 3.1.1.1 Le pitch et la fréquence fondamentale

Le pitch est le phénomène prosodique le plus expressif. Il exprime la hauteur perçue par un humain. Les systèmes de traitement de la parole utilisent la fréquence fondamentale, appelée encore F0, pour estimer le pitch. La fréquence fondamentale représente la cadence du cycle d'ouverture et de fermeture des cordes vocales de larynx durant la phonation des sons voisés. Les cordes vocales peuvent vibrer de 60 cycles par seconde (Hz), pour un grand homme, jusqu'à 300 Hz ou plus pour une jeune femme ou un enfant. En parlant, nous varions systématiquement notre fréquence fondamentale pour exprimer nos sentiments ou pour diriger l'attention de l'auditeur vers un aspect important de notre message parlé. Un paragraphe prononcé avec un pitch constant et uniforme paraîtra peu naturel (Huang, Acero et Hon, 2001).

---

[2] Nous excluons l'accent tonique qui est de niveau prosodique et qui sert à distinguer le type de mots au niveau syntaxique. Par exemple, en anglais, l'accent primaire sur la première syllabe de *export* signale le nom *exportation* en français tandis que l'accent primaire sur la seconde syllabe signale le verbe *exporter* en français.

### 3.1.1.2 L'intensité et l'amplitude

L'intensité est une sensation auditive basée sur la perception de la force du signal acoustique. L'amplitude du mouvement vibratoire est la contrepartie acoustique de l'intensité. L'amplitude est fonction de la pression sonore; plus celle-ci est grande, plus l'amplitude est grande. La pression sonore représente les variations de pression de part et d'autre d'une pression atmosphérique moyenne. L'amplitude est mesurable en watt/$cm^2$ et est proportionnelle au carré de la pression sonore. L'aire d'audition humaine se mesure sur une échelle logarithmique relative de l'intensité dont l'unité est le décibel (dB). Le décibel est égal à 10 log (niveau d'intensité en watts/$cm^2$ / $10^{-16}$ watts/cm2). Le seuil de l'audition varie de 0 à 40 dB selon la fréquence alors que le seuil de la douleur se situe environ à 130 dB (Galarneau, Tremblay et Martin).

### 3.1.1.3 Le rythme et le débit

Le rythme de l'énoncé est déterminé par la durée des silences et la durée des phones (Boite et al., 2000). Nous distinguons, en matière de débit, entre la vitesse d'articulation d'unité comme la syllabe, appelée débit articulatoire, et le débit de la parole qui comprend les hésitations, les interruptions et les pauses. Le débit de la parole se calcule en syllabes, en segments ou en mots. Un débit régulier peut être soit de type lent, moyen ou rapide alors qu'un changement de débit est soit une accélération ou un ralentissement (Galarneau, Tremblay et Martin).

### 3.1.1.4 Perturbation de F0 (vacillement) et perturbation d'intensité (tremblotement)

Les impulsions naturelles glottales ne sont pas réellement périodiques, mais présentent des perturbations appelées tremblotement (Shimmer) et vacillement (Jitter). Le vacillement représente les variations trame par trame dans les périodes de F0. Le tremblotement représente les variations cycle par cycle dans les périodes de l'énergie. Les voix normales ont un vacillement de 0.5 à 1.0% (p. ex. 1 Hz) et un tremblotement de l'ordre de 0.04 à 0.21%, ce qui représente un niveau assez bas pour qu'il soit directement perceptible. Bien que le

tremblotement et le vacillement soient deux concepts différents, ils sont légèrement corrélés (Huang, Acero et Hon, 2001).

Dans ce projet, le vacillement est mesuré par l'estimation du pourcentage de la différence des valeurs de F0 entre les cycles adjacents (voir formule (3.1)) et le tremblotement par l'estimation du pourcentage de la différence des valeurs d'amplitude entre les cycles adjacents (voir formule (3.2)).

$$\text{Vacillement (\%)} = \frac{100\ N \times \sum_{1}^{N-1} |T_i - T_{i+1}|}{(N-1)\sum_{i=1}^{N} T_i} \tag{3.1}$$

où $T_i$ représente l'inverse de la fréquence fondamentale de la i$^{ième}$ trame et $N$ le nombre total de trames voisées dans l'énoncé.

$$\text{Tremblotement (\%)} = \frac{100\ N \times \sum_{1}^{N-1} |A_i - A_{i+1}|}{(N-1)\sum_{i=1}^{N} A_i} \tag{3.2}$$

où $A_i$ représente la valeur du sommet de l'amplitude de la i$^{ième}$ trame et $N$ le nombre total de trames voisées.

### 3.1.2 Coefficients cepstraux sur l'échelle Mel (MFCC)

Les coefficients cepstraux sur l'échelle Mel (MFCC, Mel-Frequency Cepstral coefficients) ont été intensivement utilisés comme vecteur de traits caractéristiques dans les systèmes de reconnaissance de la parole et du locuteur.

L'utilisation des MFCC est motivée par les deux propriétés suivantes (Rose, 2006) :

- **Déconvolution** : les MFCC découplent les caractéristiques du conduit vocal (qui véhicule la plus grande partie de l'information disponible sur les traits distinctifs de la parole) des caractéristiques générées par l'excitation (information prosodique et l'information dépendante du locuteur).
- **Décorrélation** : La transformée en cosinus discrète possède un effet de décorrélation entre les éléments du vecteur de traits.

Les MFCC sont une représentation définie comme étant la transformée cosinus inverse du logarithme du spectre de l'énergie du segment de la parole. L'énergie spectrale est calculée en appliquant un banc de filtres uniformément espacés sur une échelle fréquentielle modifiée, appelée échelle *Mel*. L'échelle *Mel* redistribue les fréquences selon une échelle non linéaire qui simule la perception humaine des sons.

### 3.1.2.1 Étapes de calcul du vecteur caractéristique de types MFCC

Dans ce qui suit, nous décrivons chacune des étapes nécessaires pour l'obtention d'un vecteur caractéristique tiré des coefficients MFCC, tel qu'illustré par la Figure 3.1

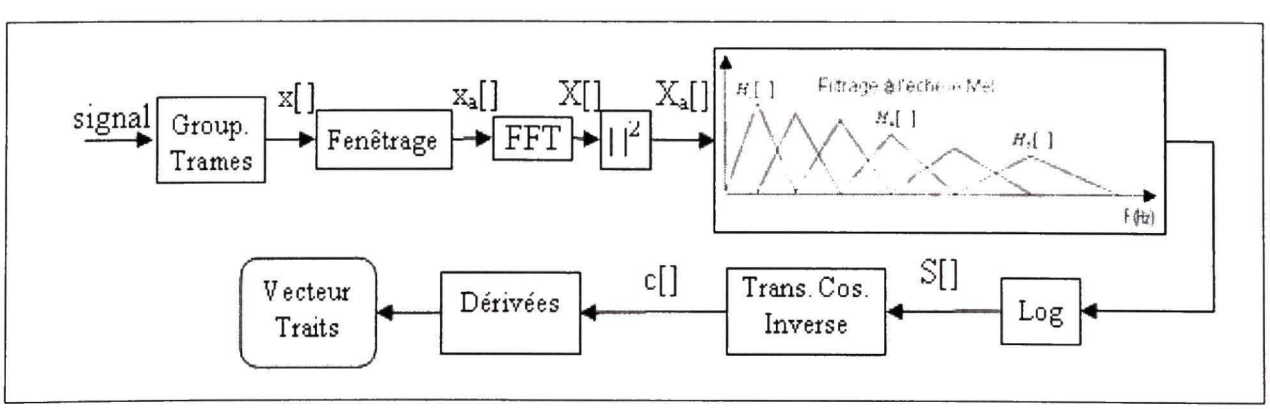


**Figure 3.2 Étapes de calcul d'un vecteur caractéristique de type MFCC.**

1. **Groupement en trame (Frame blocking)** : Le signal acoustique continu est segmenté en trames de $N$ échantillons, avec un pas d'avancement de $M$ trames $(M<N)$, c'est-à-dire que deux trames consécutives se chevauchent sur $N$-$M$ échantillons. Les valeurs couramment utilisées pour *M et N sont respectivement* 10 et 20. Comme prétraitement, il est d'usage de procéder à la préaccentuation du signal en appliquant l'équation de différence du premier ordre (3.3) aux échantillons $x(n)$, $0 \le n \le N-1$.

$$\boxed{x'(n) = x(n) - k\,x(n-1),\ 0 \le n \le N-1} \qquad (3.3)$$

k représente un coefficient de préaccentuation qui peut prendre une valeur dans l'étendue $0 \le k < 1$.

2. **Fenêtrage** : Si nous définissons $w(n)$ comme fenêtre où $0 \le n \le N-1$ et $N$ représente le nombre d'échantillons dans chacune des trames, alors le résultat du fenêtrage est le signal $x_a$, donné par la formule (3.4).

$$\boxed{x_a(n) = x(n)w(n),\ 0 \le n \le N-1} \qquad (3.4)$$

.

C'est la fenêtre de *Hamming* qui est généralement utilisée et dont la forme est donnée par la formule suivante (3.5) :

$$\boxed{w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right),\ 0 \le n \le N-1} \qquad (3.5)$$

3. **Calcul de la transformée de Fourier rapide (Fast Fourier Transform, FFT)** : Au cours de cette étape chacune des trames, de $N$ valeurs, est convertie du domaine temporel au domaine fréquentiel. La FFT est un algorithme rapide pour le calcul de la transformée

de Fourier discret (DFT) et est définie par la formule (3.6). Les valeurs obtenues sont appelées le spectre.

$$X[k]=\sum_{n=0}^{N-1} x_a[n]e^{-j2\pi nk/N}, \quad 0 \le k \le N \tag{3.6}$$

En général, les valeurs X[k] sont des nombres complexes et nous nous considérons que leurs valeurs absolues (énergie de la fréquence).

4. **Filtrage sur l'échelle Mel** : Le spectre d'amplitude est pondéré par un banc de *M* filtres triangulaires espacés selon l'échelle *Mel*. Dans l'échelle de mesure *Mel*, la correspondance est approximativement linéaire sur les fréquences au dessous de 1 kHz et logarithmique sur les fréquences supérieures à 1 kHz. Cette relation est donnée par la formule (3.7) (O'Shaughnessy, 2000) :

$$m = 2595 \log_{10}\left(1+{}^{f}\!/_{700}\right) \tag{3.7}$$

Le logarithme de l'énergie de chaque filtre est calculé selon l'équation (3.8) :

$$S[m]=\ln\left[\sum_{k=0}^{N-1} X_a[k]H_m[k]\right], \quad 0 < \mathrm{m} \le \mathrm{M} \tag{3.8}$$

5. **Calcul du cepstre sur l'échelle Mel :** Le cepstre sur l'échelle de fréquence *Mel* est obtenu par le calcul de la transformée en cosinus discrète du logarithme de la sortie des *M* filtres (reconversion du log-Mel-spectre vers le domaine temporel).

$$c[n]=\sum S[m]\cos(\pi n(m-1/2)/M), \quad 0 \le \mathrm{n} < \mathrm{M} \tag{3.9}$$

Le premier coefficient, $c_0$, représente l'énergie moyenne dans la trame de la parole; $c_1$ reflète la balance d'énergie entre les basses et hautes fréquences; pour i >1, $c_i$ représente des détails spectraux de plus en plus fins (O'Shaughnessy, 2000). La valeur de M utilisée est égale à 13.

6. **Calcul des caractéristiques dynamiques des MFCC :** Les changements temporels dans le cepstre (c) jouent un rôle important dans la perception humaine et c'est à travers les dérivées des coefficients (Δc, *coefficients delta* ou vélocité) et les dérivées secondes (ΔΔc, *coefficients delta du second ordre* ou accélération) des MFFC statiques que nous pouvons mesurer ces changements.

En résumé, un système de parole typique de l'état de l'art effectue premièrement un échantillonnage à une fréquence de 16 kHz et extrait les traits suivants (Huang, Acero et Hon, 2001) :

$$x_k = \begin{pmatrix} c_k \\ \Delta c_k \\ \Delta\Delta c_k \end{pmatrix} \tag{3.10}$$

où

- $\mathbf{c}_k$, est le vecteur MFCC de la k$^{ième}$ trame;
- $\Delta\mathbf{c}_k = \mathbf{c}_{k+2} - \mathbf{c}_{k-2}$, dérivée première des MFCC calculée à partir des vecteurs de MFCC de la k$^{ième}$ +2 trame et k$^{ième}$ – 2 trame;
- $\Delta\Delta\mathbf{c}_k = \Delta\mathbf{c}_{k+1} - \Delta\mathbf{c}_{k-1}$, seconde dérivée des MFCC.

## 3.2 Sélection des caractéristiques

Les caractéristiques permettent de distinguer une forme appartenant à une classe par rapport aux formes des autres classes. Dans certains cas, il existe des caractéristiques redondantes ou non pertinentes donc nuisibles en termes de calculs ou carrément inutiles dans la classification de l'objet d'où l'importance de la phase de sélection des caractéristiques.

Dans les sections suivantes, nous décrirons les méthodes de sélection de traits en général et la méthode *RELIEF-F* et *SEQUENTIAL FORWARD SELECTION* en particulier. Nous donnerons par la suite un aperçu de l'importance de la sélection des caractéristiques ainsi que son utilisation dans le domaine de la RAE.

### 3.2.1 Méthodes de sélection des caractéristiques pour la classification (Dash et Liu, 1997; Langley, 1994)

L'objectif principal de la sélection des traits est de réduire au plus petit nombre possible le nombre de caractéristiques utilisées tout en essayant d'améliorer ou du moins sans détériorer significativement la performance de la classification du système de reconnaissance.

Cet objectif est défini par une fonction objective à optimiser. Nous relevons deux critères essentiels dans la sélection des caractéristiques : les *procédures de recherche* et les *fonctions d'évaluation.*

#### 3.2.1.1 Les procédures de recherche

La procédure de recherche permet de générer un sous-ensemble de traits caractéristiques pour son évaluation à partir de l'ensemble des candidats possibles. Le nombre total de sous-ensembles possibles est $2^L$ où $L$ représente le cardinal de l'ensemble des caractéristiques.

Afin de trouver une solution, parmi ce nombre très élevé de candidats, trois approches de recherche ont été développées : complète, heuristique et aléatoire.

1. **Complète**

Pour rechercher le sous-ensemble optimal, selon la fonction d'évaluation, la procédure de génération effectue une recherche complète. Cependant, différentes fonctions sont utilisées afin de réduire l'espace de recherche sans compromettre les chances de trouver ce sous-ensemble optimal. Branch and Bound (B&B), Best First Search (BFF), Beam Search (BS) et Minimum Description Length Method (MDLM) sont des exemples de procédures de recherche complètes.

**2. Heuristique**

Avec cette approche de recherche, toutes les caractéristiques non sélectionnées sont examinées à chaque itération pour une sélection. Ces procédures sont très simples à implémenter et très rapides en temps d'exécution, car l'espace de recherche est seulement quadratique en termes de nombre de traits. Parmi les exemples de procédures heuristiques : Sequential Forward Selection (SFS), Sequential Backward Selection (SBS) et RELIEF.

**3. Aléatoire**

Dans la catégorie de procédures de recherche aléatoire, la recherche des sous-ensembles se fait en effectuant un maximum d'itérations. Chaque procédure de recherche de type aléatoire exige la détermination d'un certain nombre de paramètres, une étape cruciale qui détermine la qualité des résultats obtenus. Les algorithmes génétiques (AG) et le recuit simulé (Simulated Annealing (SA)) sont deux exemples de procédures de recherche aléatoire.

#### 3.2.1.2 Les fonctions d'évaluation

Les fonctions d'évaluation mesurent la qualité du sous-ensemble de traits candidats généré par les procédures de recherche en vue de trouver l'ensemble de traits optimal. L'ensemble de traits optimal est toujours relatif à la fonction d'évaluation utilisée, car l'utilisation d'une autre fonction d'évaluation peut générer un autre ensemble optimal.

Dash et Liu ont proposé une classification des fonctions d'évaluation en cinq catégories en fonction du type de mesure utilisé :

1. la mesure de distance, telle la distance euclidienne, permet de mesurer la capacité de discrimination entre les classes;
2. la mesure d'information, qui permet d'estimer le gain d'information à partir d'une caractéristique;
3. la mesure de dépendance qui permet de quantifier la corrélation qui existe entre les caractéristiques et leurs classes ou entre les caractéristiques elles-mêmes;
4. la mesure de consistance de l'ensemble des caractéristiques;
5. la mesure du taux d'erreur de la classification.

Le Tableau 3.1 dresse une comparaison entre les différentes catégories de fonctions d'évaluation indépendamment du type de la procédure de génération utilisée. La colonne généralité décrit la fiabilité du sous-ensemble de traits sélectionné dans le cas d'un changement de classificateur.

**Tableau 3.1**

Comparaison entre les fonctions d'évaluation

| Fonction d'évaluation | Généralité | Complexité en temps | Précision |
|---|---|---|---|
| Mesure de distance | Oui | Faible | - |
| Mesure d'information | Oui | Faible | - |
| Mesure de dépendance | Oui | Faible | - |
| Mesure de consistance | Oui | Modérée | - |
| Taux d'erreurs de classification | Non | Élevée | Très élevée |

Les fonctions d'évaluation peuvent aussi être classées en deux catégories : la catégorie filtre (Filter) ou la catégorie enveloppante (Wrapper) selon que les sous-ensembles de traits sont évalués dépendamment ou indépendamment de l'algorithme d'apprentissage.

Dans ce qui suit, nous décrirons deux méthodes de sélection de traits, *RELIEF-F* et *Sequential Forward Selection* (SFS), utilisées en RAE et dont les fonctions d'évaluation appartiennent à deux catégories différentes (filtre et enveloppante). Ces deux méthodes

différent également dans leurs complexités en temps de calcul (faible pour la première et élevée pour la seconde) et dans leurs précisions (précision de SFS élevée par rapport à *RELEIF-F*).

#### 3.2.1.3 La méthode Sequential Forward Selection (SFS)

La méthode de recherche ascendante (SFS) est classée dans la catégorie des procédures de recherche heuristique et utilise une fonction d'évaluation de type enveloppante (Wrapper).

SFS débute avec un ensemble de caractéristiques vide et, à chaque itération, une nouvelle caractéristique est ajoutée à l'ensemble de départ. La caractéristique sélectionnée parmi l'ensemble des traits restants est celle qui obtient le meilleur taux de classification lorsqu'elle est ajoutée à l'ensemble des traits constitué au cours de l'itération précédente.

La méthode de recherche descendante (Sequential Backward Selection, SBS) est une méthode de sélection des caractéristiques semblable à la méthode SFS, excepté que l'ensemble de recherche est initialisé avec la totalité des caractéristiques et qu'à chaque itération, la caractéristique la moins pertinente est éliminée de l'ensemble de départ.

#### 3.2.1.4 La méthode RELIEF et RELIEF-F

La méthode *RELIEF*, développée par Kira (Kira et Rendell, 1992), est une heuristique utilisée pour l'estimation de la qualité des attributs dont la fonction d'évaluation est de type filtre.

*RELIEF* traite les attributs discrets et continus et est limité aux problèmes avec deux classes. Il évalue les attributs sur la base de la qualité discriminatoire des valeurs des instances voisines.

*RELIEF-F* est une variante de *RELIEF* développée par Kononenko (Kononenko, 1994) pour les données multi-classes. Cet algorithme est aussi une amélioration par rapport à *RELIEF* en ce sens qu'il traite les cas des données bruitées ou incomplètes. Pour une instance donnée, *RELIEF-F* recherche :

- Les K-plus proches voisins qui appartiennent à la même classe (appelés *nearest hits*).
- Les K-plus proches voisins M(C) pour chacune des différentes classes (appelés *nearest misses*).

**Algorithme RELIEF-F**

L'algorithme admet en entrée les attributs de chaque instance des données d'apprentissage ainsi que leurs classes d'appartenance et retourne en sortie le vecteur W qui représente une estimation de la qualité des attributs.

**Début**
Mette tous les poids *W[A]* : = 0.0;
**Pour** i : = 1 **à** *m* **Faire**
Sélectionner aléatoirement une instance $R_i$;
Rechercher les *k* plus proches hits $H_j$;
**Pour** chaque classe $C \neq \text{class}(R_i)$ **Faire**
À partir de la classe *C* trouver les *k* plus proches misses $M_j(C)$;
**Pour** *A* : = 1 **à** *a* **Faire**

$$W[A] := W[A] - \sum diff(A, R_i, H_j)/(m.k) + \sum_{C \neq class(R_i)} \left[ \frac{P(C)}{1 - P(class(R_i))} \sum_{j=1}^{k} diff(A, R_i, M_j(C)) \right] /(m.k);$$

**Fin**;

où *W[A]* est une estimation de la qualité de l'attribut *A* et représente une approximation de différence de probabilité; *diff(Attribute, Instance1, Instance2)* représente la fonction qui calcule la différence entre les valeurs des attributs de deux instances. Pour les attributs

discrets, cette différence est égale à zéro si les deux valeurs sont égales et vaut un dans le cas contraire. Pour les attributs continus, la différence est normalisée à l'intérieur de l'intervalle [0,1].

La normalisation par *m* garantit que tous les poids sont à l'intérieur de l'intervalle [-1,1]. La fonction *diff* sert aussi bien au calcul de la distance qui sépare les instances que pour la recherche des *k* plus proches voisins.

### 3.2.2 Sélection des caractéristiques dans le domaine de la RAE

La recherche de l'ensemble de traits acoustiques les plus pertinents en RAE est un domaine qui reste toujours ouvert et nécessite encore des recherches en vue de trouver l'ensemble des caractéristiques optimal (golden set) pour la reconnaissance des émotions (Neiberg, Elenieus et Laskowski, 2006) (Lee, 2004).

Des études où un large nombre de paramètres ont été utilisés ont montré que les paramètres spectraux jouent un rôle majeur dans la distinction entre les classes d'émotions (Banse et Scherer, 1996; Klasmeyer, 2000). Par conséquent, il est absolument essentiel qu'une large variété de paramètres soit mesurée dans les futures recherches (Scherer, 2003).

Une approche qui a été adoptée par plusieurs auteurs consiste à expérimenter le maximum de traits caractéristiques possible et procéder par la suite à l'estimation et au classement de ces traits afin de déterminer le sous-ensemble optimal. Oudeyer (Oudeyer, 2002) est parti de 200 caractéristiques pour en sélectionner les six meilleures. Dans les travaux de Batliner (Batliner et al., 1999), 276 caractéristiques sont extraites au départ pour réduire cet ensemble à 11 ou six selon le type de traits utilisés. Dans les travaux de Vogt (Vogt et Andre, 2005), 1280 caractéristiques sont extraites puis réduites à 90 traits par l'utilisation des méthodes de sélection.

Afin de rechercher le plus petit ensemble efficient de traits, différentes méthodes de sélection sont proposées dans la littérature de la reconnaissance des émotions. Nous dressons ci-après, la liste des méthodes utilisées dans certains travaux :

- **Algorithme Relief-F** (Petrushin, 2000; Yu, Aoki et Woodruff, 2004);
- **Forward selection** (Bhatti, Yongjin et Ling, 2004; Fujie et al., 2004a; Huang, Luo et Zhu, 2008; Kwon et al., 2003; Lee, 2004; Lin et Wei, 2005; Liu et al., 2007a; Pao et al., 2005; Sim, Jang et Park, 2007; Xie et al., 2007) (Xiao et al., 2007; You et al., 2006; Zhu et Luo, 2007);
- **Backword selection** (Fujie et al., 2004a; Kwon et al., 2003; Pao et al., 2005; Xie et al., 2007);
- **Algorithme génétique** (Beritelli et al., 2007; Casale, Russo et Serrano, 2007; Noda et al., 2007; Oudeyer, 2002; Scherer, 1996); (Sim, Jang et Park, 2007);
- **Facteur de corrélation** (Vlasenko et al., 2007);
- **Standardized Canonical Discriminant Function Coefficients** et **Structure Coefficients** (Batliner et al., 1999) : cette méthode consiste en l'évaluation des attributs à travers deux coefficients, en utilisant l'analyse discriminante linéaire (LDA);.
- **Estimation of Distribution Algorithm** (EDA) (Alvarez et al., 2007);
- **Méthode de la variance inexpliquée (unexplained variance** en anglais) (Seppänen, Väyrynen et Toivanen, 2003);
- **Best-First Search** et **Correlation-based Feature Selection** (CFS) (Vogt et Andre, 2005)

## 3.3 Modélisation

Différentes stratégies sont utilisées pour concevoir un classificateur de type statistique en reconnaissance de formes. Selon la disponibilité d'information sur les densités conditionnelles des classes, les classificateurs s'inscrivent soit dans une approche paramétrique ou non paramétrique. Dans l'approche paramétrique, la forme des densités

conditionnelles des classes est supposée connue alors que certains paramètres de ces densités sont inconnus. Une estimation des paramètres inconnus des fonctions de densité permettent de résoudre ce type de problème. L'estimation de ces paramètres se fait à partir d'échantillons appelés *données d'apprentissage* ou *données d'entraînement*.

Dans l'approche non paramétrique, la forme des densités n'est pas connue et dans ce cas nous procédons, soit à une estimation de la fonction de densité, soit à l'utilisation de certaines règles non paramétriques telles que la méthode des *K-plus proches voisins*.

Le mode d'apprentissage constitue un autre critère de distinction entre les modèles de classificateurs statistiques. Dans le mode d'apprentissage supervisé, l'entraînement se base sur l'utilisation des données étiquetées avec la catégorie à laquelle la forme appartient. Dans le cas de l'apprentissage non supervisé, l'information *a priori* sur l'appartenance des données d'entraînement est absente.

La Figure 3.3 est une illustration d'un modèle de classification pour la reconnaissance des émotions *neutre* versus *colère forte* à partir des données d'apprentissage appartenant à un seul locuteur. Les données sont caractérisées par deux traits : la fréquence fondamentale (F0) et l'énergie. La frontière de décision, pour cet exemple, possède une forme linéaire. Nous constatons également que pour cet exemple simplifié, la fréquence fondamentale, à elle seule, permet de distinguer entre l'émotion neutre et la colère forte pour ce locuteur. Dans une application réelle, avec l'augmentation du nombre de locuteurs et du nombre de classes d'émotions, les frontières de décision sont beaucoup plus complexes que celle présentée dans la Figure 3.3 et l'utilisation d'un modèle de classification plus robuste est indispensable.

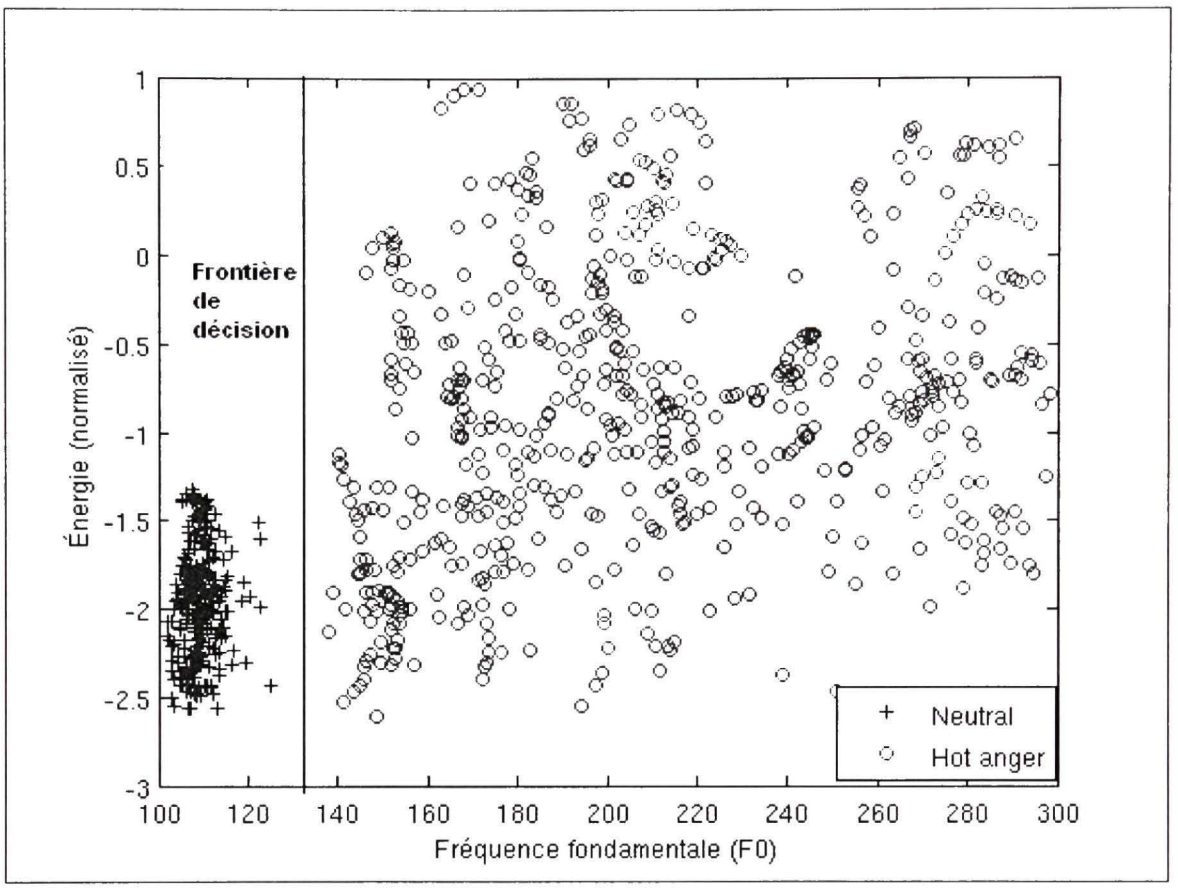


**Figure 3.3 Exemple de classification des données de la classe neutre vs colère appartenant à un seul locuteur et caractérisées par les deux traits F0 et l'énergie.**

Dans ce qui suit, nous étudierons en détail un modèle de classification très employé, les GMM en occurrence, que nous utiliserons pour la modélisation de notre système de RAE.

## 3.4 Le modèle GMM

La modélisation par mélange de gaussiennes (GMM – Gaussian Mixture Model) est une méthode statistique qui a été utilisée dans des domaines aussi variés que celui de l'identification du locuteur (Gish et Schmidt, 1994; Reynolds et Rose, 1995), la compression d'images (Aiyer, 2001), la classification des sons respiratoires en vue de détecter automatiquement des sibilants ou les crises d'asthme (Pelletier, 2006) ou celui des finances et

de l'économie pour la prévision de la bourse et du taux de change (Lindemann, Dunis et Lisboa, 2004). Récemment les GMM ont été utilisés dans la reconnaissance automatique des émotions à partir de la parole dans (Fujie et al., 2004b; Hung, Quénot et Castelli, 2004; Neiberg, Elenieus et Laskowski, 2006).

L'utilisation des GMM dans le domaine du traitement du signal de la parole en général et celui de la RAE en particulier est motivée par la notion intuitive que chaque densité de composante d'un mélange de gaussiennes permet de modéliser une ou un certain nombre de classes acoustiques telles les voyelles ou les fricatives par exemple. Ces classes acoustiques reflètent un aspect général de la configuration du système de la production de la parole (poumons, conduit vocal et cordes vocales) sous l'effet de l'état émotionnel éprouvé. Étant donné que les données d'apprentissage et de test ne sont pas « phonétiquement » annotées, les classes acoustiques sont considérées comme cachées dans le sens où la classe des données observées est inconnue. Par conséquent, la densité des vecteurs de traits générés de ces classes acoustiques cachées prête bien à un mélange de gaussiennes.

### 3.4.1 Propriétés et définition

Les données dont la fonction de densité de probabilité est unimodale et symétrique peuvent être convenablement modélisées par une seule courbe gaussienne. Cependant, dans plusieurs cas de problèmes réels, les données ne peuvent être modélisées adéquatement par un seul paramètre de variance et de moyenne d'où l'intérêt de l'utilisation d'un modèle à mélange de gaussiennes. Les GMM permettent de réaliser une approximation d'une fonction de densité de probabilité, de complexité quelconque, en choisissant un nombre suffisant de composantes gaussiennes avec un choix éclairé des valeurs de ses paramètres.

La **Figure 3.4**-(b) montre un exemple de fonction de densité de probabilité de mélange de gaussiennes obtenue par la combinaison de trois gaussiennes pondérées par $w_1$, $w_2$ et $w_3$ de la **Figure 3.4**-(a).

Les GMM peuvent être considérés comme une approche hybride entre les modèles de densité paramétriques et non paramétriques. À l'instar des modèles paramétriques, les GMM possèdent une structure et des paramètres qui contrôlent le comportement de la densité d'une manière connue sauf qu'ils sont libres de toute contrainte relative à une distribution spécifique des données. De la même manière que pour les modèles non paramétriques, les GMM possèdent plusieurs degrés de liberté, ce qui offre la possibilité de modéliser des données ayant une densité quelconque sans demander une capacité exorbitante de calcul ou de stockage (Reynolds, Quatieri et Dunn, 2000).

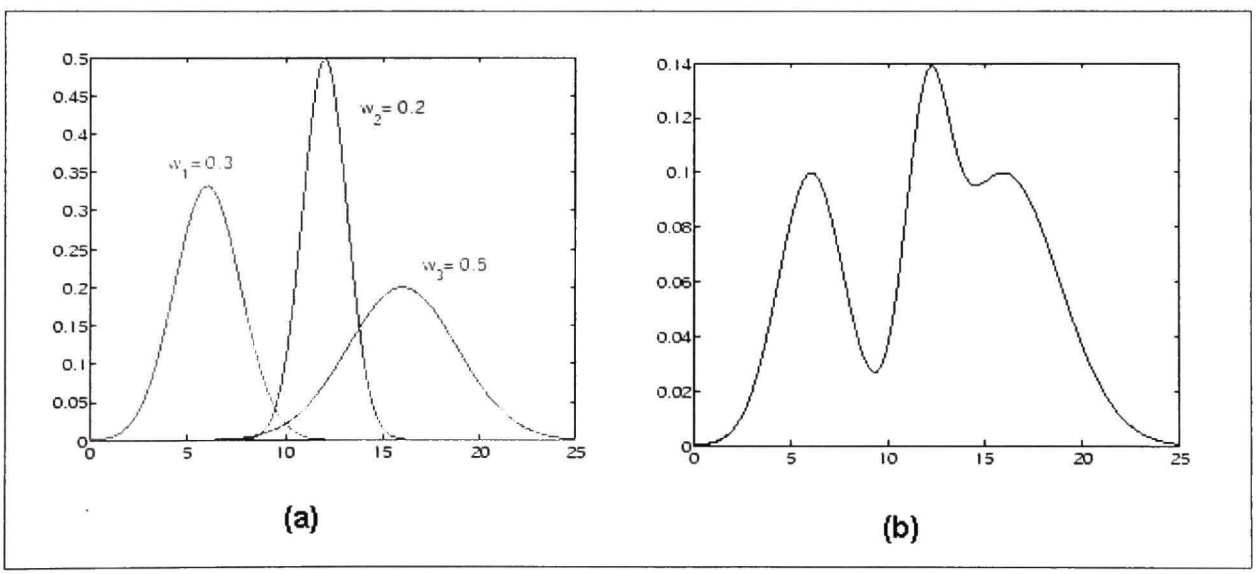


**Figure 3.4 Exemple de mélange de trois gaussiennes (b), obtenue par la combinaison de trois gaussiennes pondérées par $w_1$, $w_2$ et $w_3$ (a)**
Tiré de Resch (2008).

Un GMM peut être également vu comme étant un HMM, à un seul état, ayant un mélange de gaussiennes comme densité d'observation. Une densité de probabilité d'un modèle de mélange de gaussiennes est une somme pondérée de $M$ composantes de densités et s'écrit sous la forme mathématique suivante :

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i b_i(\mathbf{x}) \tag{3.11}$$

où $\mathbf{x}$ est un vecteur de données de dimension $d$, $\lambda$ est le modèle GMM, les $w_i$ représentent les pondérations des mélanges de gaussiennes avec les contraintes $\sum_{i=1}^{M} w_i = 1$ et $w_i \geq 0$ pour $i=1,\ldots, M$, et $b_i(\mathbf{x})$, $i=1,\ldots, M$, sont les densités normales multidimensionnelles données par :

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right) \tag{3.12}$$

$\boldsymbol{\mu}_i$ et $\boldsymbol{\Sigma}_i$ représentent respectivement le vecteur de la moyenne et la matrice de covariance de la $i^{ième}$ gaussienne, et l'exposant $[.]^T$ désigne la transposée du vecteur ou de la matrice.

Le modèle GMM $\lambda$ est défini par :

$$\lambda = \{w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\} \tag{3.13}$$

où $w_m$, $\boldsymbol{\mu}_m$, $\boldsymbol{\Sigma}_m$ représentent respectivement la pondération, le vecteur de la moyenne et la matrice de covariance de chacune des $M$ composantes gaussiennes constituant le mélange de gaussiennes $\lambda$.

Les matrices pleines et les matrices diagonales sont les deux formes de matrice de covariance les plus largement utilisées dans la modélisation avec les GMM. Le modèle GMM avec une matrice de covariance pleine est le modèle le plus puissant, car il permet de mieux ajuster les données. L'inconvénient de ce type de matrice de covariance est qu'il nécessite un grand volume de données pour l'estimation des paramètres et dépend du schéma de régularisation pour obtenir des estimations précises. Le nombre de paramètres à estimer lors de la phase

d'apprentissage est égal à $M/2*\left(d^2+3d+3\right)$ où $d$ représente la dimension du vecteur de traits caractéristiques. D'autre part, la matrice de covariance diagonale est largement utilisée et permet d'obtenir des performances semblables aux matrices de covariance pleines en utilisant un nombre plus élevé de mélanges de gaussiennes (Reynolds et Rose, 1995). Le nombre de paramètres à estimer pour le cas d'un modèle avec une matrice de covariance diagonale est égal à $M*\left(2d+1\right)$.

### 3.4.2 Estimation des paramètres du GMM

Avec les GMM, l'objectif de la phase d'apprentissage est d'estimer les paramètres λ pour l'ensemble des données d'entraînement, c.-à-d. trouver les valeurs des paramètres qui modélisent le mieux la distribution des données d'apprentissage. Il existe plusieurs techniques pour la l'estimation des paramètres d'un GMM et la méthode la plus populaire et bien établie est la méthode de l'estimation du maximum de vraisemblance (ML, Maximum Likelihood estimation).

Le but de la méthode ML est de trouver les paramètres du modèle qui maximisent la vraisemblance du GMM étant donné les données d'apprentissage (McLachlan, 1988; Reynolds et Rose, 1995). En supposant l'indépendance des vecteurs d'entraînement $\mathbf{X}=\{\mathbf{x}_1,...,\mathbf{x}_n,...,\mathbf{x}_N\}$, la vraisemblance du modèle λ, s'écrit comme :

$$\boxed{p(\mathbf{X}|\lambda)=\prod_{n=1}^{N}p(\mathbf{x}_n|\lambda)} \tag{3.14}$$

Malheureusement, il n'existe pas de méthode analytique connue pour résoudre le problème de maximisation de cette fonction non linéaire du paramètre λ. Cependant, nous pouvons choisir $\lambda=\{w_m,\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m\}$ telle que la vraisemblance $p(\mathbf{X}|\lambda)$ est un maximum local en utilisant une méthode itérative telle que la méthode Estimation-Maximisation connue aussi sous le nom *Baum-Welch* pour les *HMM* ou en utilisant les techniques du gradient.

### 3.4.3 L'algorithme Estimation-Maximisation (EM)

Introduit initialement par Baum (Baum, 1972; Baum et Petrie, 1966; Dempster, Laird et Rubin, 1977), l'algorithme EM est la méthode la plus utilisée pour l'apprentissage statistique faisant intervenir des variables manquantes (missing variables). Il permet de déterminer, suivant un processus itératif, les paramètres du modèle $\lambda$ en maximisant dans l'espace des paramètres $\lambda$ la fonction de la vraisemblance $p(\mathbf{X}|\lambda)$ de l'ensemble des observations $\mathbf{X}$ conditionné sur l'ensemble des paramètres $\lambda$.

Pour des raisons analytiques, il est plus facile de travailler avec le logarithme de la vraisemblance qu'avec la vraisemblance elle-même. Étant donné que le logarithme est croissant et monotone, la valeur de $\lambda$ qui maximise le logarithme de la vraisemblance maximise également la vraisemblance.

$$\log p(\mathbf{X}|\lambda) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\lambda) = \sum_{n=1}^{N} \log \sum_{i=1}^{M} w_i b_i(\mathbf{x}) \tag{3.15}$$

Souvent la valeur moyenne du logarithme de la vraisemblance, obtenu par la division du $\log p(\mathbf{X}|\lambda)$ par $N$ est utilisée. Ceci a pour effet de normaliser le logarithme de la vraisemblance par rapport à la durée.

Soit :

- $\mathbf{X} = \{\mathbf{x}_1,...,\mathbf{x}_n,...,\mathbf{x}_N\}$ l'ensemble des observations;
- $Y = \{y_1,...,y_n,...,y_N\}$ les variables manquantes supposées connues afin de simplifier le problème. Les valeurs $y_n$ peuvent représenter la composante gaussienne qui se réalise pour la donnée observable $\mathbf{x}_n$ dans le cas des GMM, ou la séquence d'états cachés associée aux observations $\mathbf{x}_n$ dans le cas des modèles de Markov cachés;

- $Q(\lambda, \hat{\lambda})$ est une fonction auxiliaire incluant les paramètres $\lambda = \{w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}$ du modèle courant et leurs valeurs estimées $\hat{\lambda} = \{\hat{w}_m, \hat{\boldsymbol{\mu}}_m, \hat{\boldsymbol{\Sigma}}_m\}$ à l'itération $t$. Elle est définie comme étant l'espérance mathématique du logarithme de la vraisemblance jointe des variables observées et des variables cachées :

$$\boxed{Q(\lambda, \hat{\lambda}) = \sum_{Y} P\left(Y \middle| \mathbf{X}, \hat{\lambda}\right) \log p\left(\mathbf{X}, Y \middle| \lambda\right)} \tag{3.16}$$

Maximiser la fonction $Q(\lambda, \hat{\lambda})$ est équivalent à maximiser (le logarithme de) la vraisemblance des données observées, étant donné que :

$$\boxed{Q(\lambda, \hat{\lambda}) \geq Q(\lambda, \lambda) \quad \Rightarrow \quad \log P(\mathbf{X} \middle| \hat{\lambda}) \geq \log P(\mathbf{X} \middle| \lambda)} \tag{3.17}$$

C'est-à-dire que nous avons trouvé un nouveau modèle $\hat{\lambda}$, plus probable que $\lambda$, à partir duquel, il est plus probable que la séquence d'observation soit générée.

En se basant sur cette procédure, si nous procédons au remplacement de $\lambda$ par $\hat{\lambda}$ d'une manière itérative, et que nous répétions le calcul de ré-estimation, nous pouvons alors améliorer la probabilité que **X** soit observée à partir du modèle, et ce, jusqu'à ce qu'un point limite soit atteint.

**Algorithme EM**

1. **Initialisation**: Choisir une estimation initiale λ.
2. **Étape Estimation**: Calculer la fonction auxiliaire $Q(\lambda,\hat{\lambda})$ - qui représente une estimation du $\log p(\mathbf{X}|\lambda)$, en se basant sur les données observables.
3. **Étape Maximisation**: Calculer $\hat{\lambda} = \arg\max_{\lambda} Q(\lambda,\hat{\lambda})$ afin de maximiser la fonction auxiliaire $Q$ sur λ.
4. **Itération**: Mettre $\lambda = \hat{\lambda}$, répéter étape 2 et 3 jusqu'à ce qu'il y ait convergence.

Dans le cas d'un mélange de gaussiennes, l'algorithme EM réalise un apprentissage non supervisé des paramètres de la densité du GMM, c'est–à-dire les moyennes, les matrices de covariances et les coefficients de pondération, à travers les vecteurs de données $\mathbf{X} = \{\mathbf{x}_1,...,\mathbf{x}_n,...,\mathbf{x}_N\}$. Aucune donnée $\mathbf{x}_n$ n'est associée exclusivement à une gaussienne unique, mais plutôt sera considérée comme étant générée par chacune des gaussiennes avec une certaine vraisemblance. Les valeurs des paramètres λ sont données par les formules suivantes :

La probabilité a posteriori :

$$\hat{P}\left(j \middle| \mathbf{x}_n, \hat{\lambda}\right) = \frac{P\left(\mathbf{x}_n \middle| j, \hat{\lambda}\right) P\left(j \middle| \hat{\lambda}\right)}{P\left(\mathbf{x}_n \middle| \hat{\lambda}\right)} \tag{3.18}$$

La pondération d'une gaussienne :

$$\overline{p}_j = \frac{1}{N}\sum_{n=1}^{N} \hat{P}\left(j \middle| \mathbf{x}_n, \hat{\lambda}\right) \tag{3.19}$$

La moyenne :

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{n=1}^{N} \mathbf{x}_n P\left(j \middle| \mathbf{x}_n, \hat{\lambda}\right)}{\sum_{n=1}^{N} P\left(j \middle| \mathbf{x}_n, \hat{\lambda}\right)} \tag{3.20}$$

La covariance :

$$\hat{\boldsymbol{\Sigma}}_j = \frac{\sum_{n=1}^{N} P\left(j \middle| \mathbf{x}_n, \hat{\lambda}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_j\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_j\right)^T}{\sum_{n=1}^{N} P\left(j \middle| \mathbf{x}_n, \hat{\lambda}\right)} \tag{3.21}$$

Avant l'utilisation de l'algorithme EM, il est nécessaire de déterminer deux facteurs importants pour l'apprentissage des modèles GMM : 1) l'ordre $M$ des mélanges de gaussiennes et 2) l'initialisation des paramètres du modèle. Il n'existe pas de bons moyens théoriques qui peuvent guider la sélection de ces deux paramètres. Le meilleur choix demeure donc la solution empirique. Pour l'initialisation des paramètres du modèle de départ, diverses méthodes sont possibles. Indépendamment de la méthode utilisée, EM garantit de trouver le maximum local de la vraisemblance du modèle. Cependant, l'équation de la vraisemblance du GMM possède plusieurs maximums locaux et différents modèles de départ mènent vers différents maximums locaux (McLachlan, 1988). Parmi les méthodes d'initialisation, nous retrouvons la méthode LBG qui réalise des regroupements d'un ordre égal à une puissance de deux, ou encore la méthode *k-moyennes* (k-means) qui permet d'obtenir des regroupements d'un ordre quelconque. Dans les travaux de Reynolds (Reynolds et Rose, 1995) sur l'identification du locuteur en utilisant des modèles à mélange de gaussiennes, trois méthodes différentes pour l'initialisation du modèle ont été expérimentées. La première méthode consiste à sélectionner d'une manière aléatoire un ensemble de vecteurs (50) à partir des données d'apprentissage pour le calcul de la moyenne des modèles et l'utilisation de la matrice d'identité comme matrice de covariance de départ. Pour la deuxième méthode, les données d'apprentissage sont segmentées en 50 classes phonétiques annotées et les moyennes des classes et les variances globales sont utilisées comme modèle initial. La troisième méthode utilisée est la méthode k-moyennes binaire. Contrairement à ce qu'on pouvait s'attendre, aucune différence significative n'est enregistrée dans les performances des classifieurs initialisés avec les trois méthodes précédentes.

Dans la section suivante, nous décrirons l'algorithme LBG que nous avons utilisé pour l'initialisation des modèles GMM de notre système.

## 3.5 L'algorithme LBG

L'algorithme *LBG* est une extension de l'algorithme k-moyennes (k-means) proposé par Linde, Buzo et Gray (Linde, Buzo et Gray, 1980). LBG est un processus itératif qui permet de réaliser des groupements de vecteurs de données en $M$ cohortes (cluster) distinctes où $M$ étant une puissance de 2. Chaque groupe est représenté par son centre, appelé *code de mot* (codeword en anglais), et par la collection de ces codes de mot qui constitue le dictionnaire (codebook). L'algorithme commence par calculer un dictionnaire à un seul vecteur et procède par la suite à la subdivision des codes de mots pour obtenir un dictionnaire à deux vecteurs et continue le processus de division jusqu'à l'obtention du dictionnaire à $M$ vecteurs.

L'algorithme est implémenté comme suit :

**Étape 1 – Initialisation**

Concevoir un dictionnaire à un seul code de mot, représentant le centroïde de toutes les données.

**Étape 2 - Dédoublement de la taille du dictionnaire**

Doubler la taille du dictionnaire en scindant le dictionnaire courant $y_n$ selon la règle suivante :

$$\boxed{\begin{aligned} y_n^+ &= y_n(1+\varepsilon) \\ y_n^- &= y_n(1-\varepsilon) \end{aligned}} \tag{3.22}$$

où $n$ varie de un jusqu'à la taille courante du dictionnaire et ε représente le paramètre de division. Mettre $M = 2*M$.

**Étape 3 - Recherche du plus proche voisin**

Classer chaque vecteur de données dans la cohorte $C_i$ possédant le code de mot le plus proche à ce vecteur.

**Étape 4 - Mise à jour des centroïdes**

Mettre à jour le code de mot de chaque cohorte par le calcul du nouveau centroïde du groupe en utilisant les vecteurs de données associés à cette cohorte.

**Étape 5 – Itération 1**

Répéter les étapes 3 et 4 jusqu'à ce que la distance moyenne soit au dessous d'un certain seuil prédéterminé.

**Étape 6 - Itération 2**

Répéter les étapes 2, 3 et 4 jusqu'à ce que $M$ soit égal à la taille du dictionnaire requis.

## 3.6 L'adaptation MAP

Un des problèmes rencontrés au cours du développement de systèmes de RAE à partir de la parole est l'absence de données d'apprentissage en quantité suffisante pour une modélisation adéquate des caractéristiques de chaque modèle associé à une catégorie d'émotion et plus particulièrement la matrice de covariance $\hat{\Sigma}_j$ de chacune des composantes gaussiennes du modèle GMM. Étant donné que la matrice $\hat{\Sigma}_j$ doit être inversée pour le calcul de la vraisemblance d'un vecteur de données $\boldsymbol{X}$ (voir équation (3.11) et (3.12)), il est important que $\hat{\Sigma}_j$ ne soit pas singulière ou proche de singulière. Ce problème s'accentue davantage lorsque la taille des données d'apprentissage est inférieure au nombre de paramètres à estimer et dans ce cas certains de ces paramètres ne sont pas identifiables à partir de ces données et on parle alors d'un problème « *ill-posed* ». Dans le cas où la taille des données dépasse légèrement le nombre de paramètres à estimer le problème est dit « *poorly posed* ».

Afin de remédier à ce problème, les modèles des émotions sont générés à la fois à partir d'un modèle initial bien entraîné, appelé modèle du monde (UBM, Universal Background Model) et à partir d'une quantité limitée de données d'apprentissage via la méthode du maximum a posteriori ou MAP.

Les paramètres du modèle UBM sont entraînés à partir des données de toutes les classes d'émotions via l'algorithme EM. L'adaptation MAP permet d'ajuster les paramètres du

modèle préentrainé (UBM) de manière à ce que de nouvelles données, en quantité limitée, modifient les paramètres du modèle, guidé par la connaissance *a priori*.

En utilisant les données observées **X**, l'estimation MAP peut être formulée par :

$$\hat{\lambda} = \arg\max_{\lambda}\left[p(\lambda|\mathbf{X})\right] = \arg\max_{\lambda}\left[p(\mathbf{X}|\lambda)p(\lambda)\right] \tag{3.23}$$

En absence de l'information *a priori*, c'est-à-dire aucune connaissance sur λ et si $p(\lambda)$ possède une distribution uniforme, alors l'estimation MAP devient identique à l'estimation ML. Nous pouvons utiliser l'algorithme EM pour estimer les paramètres du GMM de la même façon que nous l'avons fait pour la méthode ML. La *Q*-fonction correspondante est définie par la formule (3.24) (Huang, Acero et Hon, 2001) :

$$\underset{MAP}{Q}(\lambda,\hat{\lambda}) = Q(\lambda,\hat{\lambda}) + \log p(\lambda) \tag{3.24}$$

Étant donné un modèle *UBM* et les vecteurs de données d'apprentissage d'une classe d'émotion $\mathbf{X} = \{\mathbf{x}_1,...,\mathbf{x}_n,...,\mathbf{x}_N\}$, l'adaptation est réalisée selon le processus suivant (Reynolds, Quatieri et Dunn, 2000) :

1. Déterminer l'alignement probabiliste des vecteurs de données d'apprentissage dans les composantes des mélanges de l'UBM (voir Figure 3.5-(a)), en calculant $P(j|\mathbf{x}_n)$ :

$$P(j|\mathbf{x}_n) = \frac{w_j p_j(\mathbf{x}_n)}{\sum_{i=1}^{M} w_i p_i(\mathbf{x}_n)} \tag{3.25}$$

2. Calculer les fonctions exhaustives des observations (*sufficient statistics* en anglais) pour les paramètres de la pondération de la gaussienne, la moyenne et la variance, d'une manière similaire à l'étape Estimation de l'algorithme EM, en utilisant $P(j|\mathbf{x}_n)$ et $\mathbf{x}_n$ :

$$\boxed{s_j = \sum_{n=1}^{N} P(j|\mathbf{x}_n)} \tag{3.26}$$

$$\boxed{E_j(\mathbf{x}) = \frac{1}{s_j}\sum_{n=1}^{N} P(j|\mathbf{x}_n)\mathbf{x}_n} \tag{3.27}$$

$$\boxed{E_j(\mathbf{x}^2) = \frac{1}{s_j}\sum_{n=1}^{N} P(j|\mathbf{x}_n)\, diag(\mathbf{x}_n \mathbf{x}_n{}^T)} \tag{3.28}$$

3. Utiliser ces nouvelles fonctions exhaustives des observations, obtenues à partir des données d'apprentissage, pour mettre à jour les fonctions exhaustives des observations de l'ancien UBM, pour créer les paramètres adaptés de la mixture *j* (voir Figure 3.5-(b)) et ce au moyen des équations suivantes :

$$\boxed{\hat{w}_j = \left[\alpha_j^w s_j / N + (1-\alpha_j^w) w_j\right]\gamma} \tag{3.29}$$

$$\boxed{\hat{\boldsymbol{\mu}}_j = \alpha_j^m E_j(\mathbf{x}) + (1-\alpha_j^m)\boldsymbol{\mu}_j} \tag{3.30}$$

$$\boxed{\hat{\boldsymbol{\sigma}}_j^2 = \alpha_j^v E_j(\mathbf{x}^2) + (1-\alpha_j^v)(\boldsymbol{\sigma}_j^2 + \boldsymbol{\mu}_j^2) - \hat{\boldsymbol{\mu}}_j^2} \tag{3.31}$$

où $\gamma$ représente un facteur de pondération assurant que la somme des pondérations des mélanges de gaussiennes adaptés est égale à l'unité; $\alpha_j^w, \alpha_j^m, \alpha_j^v$ représentent les coefficients d'adaptation qui assurent le contrôle d'équilibre entre les anciennes et les nouvelles estimations pour la pondération, la moyenne et la variance d'un mélange, respectivement. Le paramètre $\alpha_j^\rho, \rho \in \{w, m, v\}$ est défini par la formule suivante :

$$\boxed{\alpha_j^\rho = \frac{s_j}{s_j + r^\rho}} \tag{3.32}$$

où $r^\rho$ est un facteur de pertinence fixe pour le paramètre $\rho$.

La MAP est une adaptation dépendante des données, par conséquent les paramètres du mélange de gaussiennes de l'UBM sont adaptés avec différentes grandeurs. C'est à travers les coefficients $\alpha_j^w, \alpha_j^m, \alpha_j^v$ qu'une adaptation dépendante des données est contrôlée. Ainsi, si une gaussienne est bien représentée par les nouvelles données à utiliser pour l'adaptation, alors ces données auront un poids plus important dans l'estimation des nouveaux paramètres. Dans le cas contraire, c'est-à-dire quand une gaussienne est mal représentée par les nouvelles données, les nouveaux paramètres estimés seront plus influencés par les anciennes valeurs, qui représentent les paramètres du modèle UBM qui eux sont mieux entraînés.

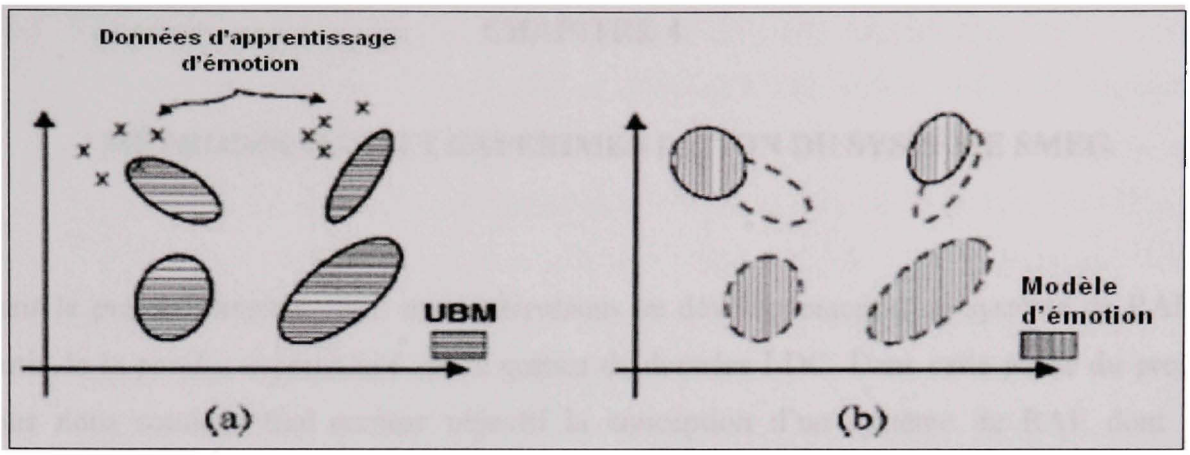


**Figure 3.5 Illustration d'un exemple de l'adaptation d'un modèle d'émotion. Dans (a) Les vecteurs de données d'apprentissage (x) sont mappés dans les mélanges de l'UBM. Dans (b) Les paramètres du mélange de l'UBM sont ajustés en utilisant les nouvelles données.**
Adaptée de Reynolds, Quatieri et Dunn (2000)

## 3.7 Conclusion

Dans ce chapitre, nous avons présenté l'architecture d'un système de reconnaissance de forme en général et décrit par la suite chacun de ses composants. Nous avons examiné en particulier les traits les plus couramment utilisés en RAE à savoir les éléments de la prosodie, le tremblotement, le vacillement et les MFCC calculés au cours de la phase d'extraction des caractéristiques. Nous avons donné une description des méthodes de sélection des caractéristiques employées pour la classification notamment la méthode *RELIEF-F* et *Forward Selection*. Dans le domaine de la modélisation, nous avons décrit le modèle GMM, un modèle de classification puissant, ainsi qu'une autre méthode d'apprentissage, appelée *MAP*, utilisée pour atténuer le problème du manque de données d'entrainement.

Dans le chapitre suivant, nous allons présenter notre premier système de RAE que nous avons développé à partir du corpus LDC ainsi qu'une comparaison de performances entre notre système et ceux de l'état de l'art.

# CHAPITRE 4

# MÉTHODOLOGIE ET EXPÉRIMENTATION DU SYSTÈME SMEG

Dans le présent chapitre, nous nous intéressons au développement d'un système de RAE à partir de la parole, expérimenté sur le corpus de données LDC. Dans cette partie du projet, nous nous sommes fixé comme objectif la conception d'un système de RAE dont les performances seront compétitives avec celles de l'état de l'art affichées au Tableau 2.4. Dans ce qui suit, nous décrirons la méthodologie suivie pour concevoir le système *SMEG* et nous présenterons les résultats des expérimentations.

## 4.1 Méthodologie

L'objectif de cette partie du projet de la reconnaissance des émotions est de concevoir et d'expérimenter un système de RAE utilisant les MFCC comme traits caractéristiques, l'énoncé comme unité d'analyse et les GMM comme classificateur. Nous nous réfèrerons à ce système par *SMEG* (**S**ystème-**M**FCC-**É**noncé-**G**MM). À notre connaissance, un tel système n'a pas été encore expérimenté sur le corpus de données LDC alors qu'il représente l'état de l'art des systèmes de reconnaissance automatique du locuteur.

À travers ce choix, nous visons reproduire les résultats de l'état de l'art obtenus avec les données du corpus LDC et éventuellement améliorer les taux de classification en explorant trois pistes différentes, à savoir le domaine de l'extraction et la sélection des traits caractéristiques, celui de l'optimisation des paramètres du classificateur et, finalement, celui de l'atténuation du problème de la représentativité des échantillons des données d'apprentissage. Dans ce qui suit, nous détaillons chacun de ces trois points :

1. pour le domaine de la sélection et l'extraction des traits, nous envisageons l'utilisation d'une plus large gamme d'information spectrale du signal acoustique,

susceptible de mieux caractériser les classes d'émotions, en augmentant la dimension du vecteur de traits, des 13 premiers coefficients MFCC aux 20 premiers, augmentés par la vélocité et l'accélération. Notons que, dans le domaine de la reconnaissance automatique de la parole, l'information véhiculée par les treize premiers coefficients MFCC, leurs vélocités et leurs accélérations, fournit toute l'information saillante et aucun gain dans les performances n'est enregistré par l'augmentation du nombre des coefficients (Huang, Acero et Hon, 2001);

2. dans le domaine de l'optimisation des paramètres du classificateur, nous allons chercher une meilleure modélisation des catégories des émotions en variant le nombre de mélanges de gaussiennes du modèle afin de trouver le nombre de mélanges qui permettent de mieux modéliser la large gamme des sons phonétiques caractérisant chacune des classes des émotions. Les valeurs possibles du nombre de mélanges de gaussiennes expérimentées appartiennent à l'ensemble {2, 4, 8, 16, 32, 64, 128, 256};
3. dans le domaine de la représentativité des échantillons des données d'apprentissage nécessaires à l'entrainement des modèles, le problème du manque de données peut être atténué par l'utilisation d'un UBM. Nous verrons dans quelle mesure un UBM peut améliorer les performances d'un système de RAE entraîné à partir du corpus *LDC Emotional Prosody*.

Afin de trouver la meilleure combinaison de ces trois critères, nous avons procédé à une exploration complète des combinaisons possibles, en effectuant 32 expériences pour chacun des trois groupes d'expériences *Exp I, II et III* décrits plus loin dans ce chapitre.

Dans nos expériences, nous avons utilisé les GMM comme modèle de classification. Deux systèmes, basés sur des méthodes d'apprentissage différentes, ont été réalisés; un système de type GMM et un autre de type GMM-UBM. Dans le premier type de système, chaque classe d'émotion est modélisée par un GMM, entraîné directement à partir des données de la classe d'émotion correspondante, tirée du corpus de données LDC. Les opérations de la phase d'apprentissage et de test du système GMM sont illustrées à la Figure 4.1.

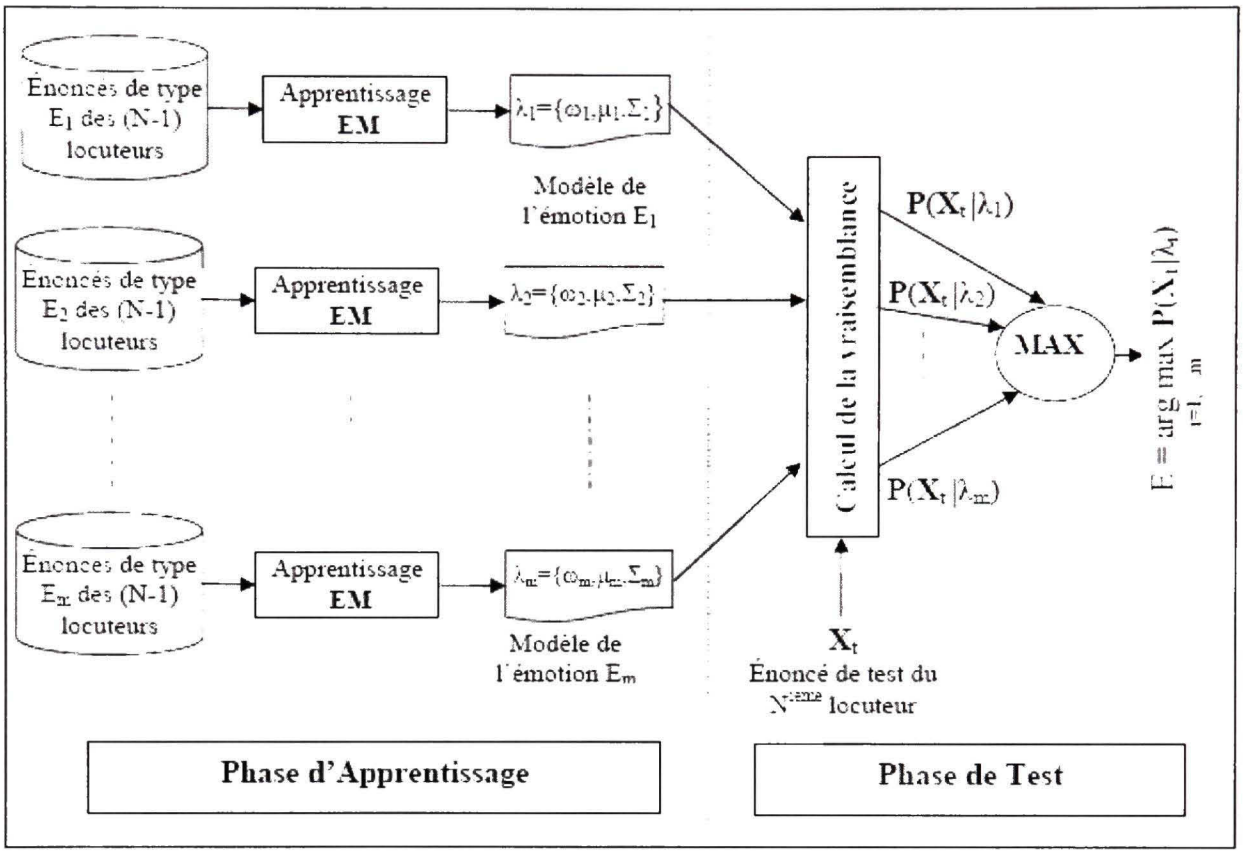


**Figure 4.1 Diagramme bloc de la phase d'apprentissage et de test du système de RAE, à *m* classes d'émotions, basé sur le modèle GMM.**

Dans le second type de système, un modèle du monde UBM est d'abord généré à partir des données de toutes les classes des émotions. Par la suite, chacun des modèles des classes d'émotions est obtenu par l'adaptation du modèle du monde avec la méthode MAP, décrite dans la section 3.6, en utilisant les données associées à la même classe d'émotion. Le processus d'apprentissage du système GMM-UBM est donné à la Figure 4.2, alors que la phase de test est identique au système GMM (voir Figure 4.1).

Tel que nous l'avons mentionné, chacun des deux types de systèmes, GMM et GMM-UBM, a été implémenté avec différents nombres de gaussiennes allants de 2 à 256 et avec des

matrices de covariance diagonales. Les paramètres des modèles GMM sont initialisés avec l'algorithme LBG décrit dans la section 3.5.

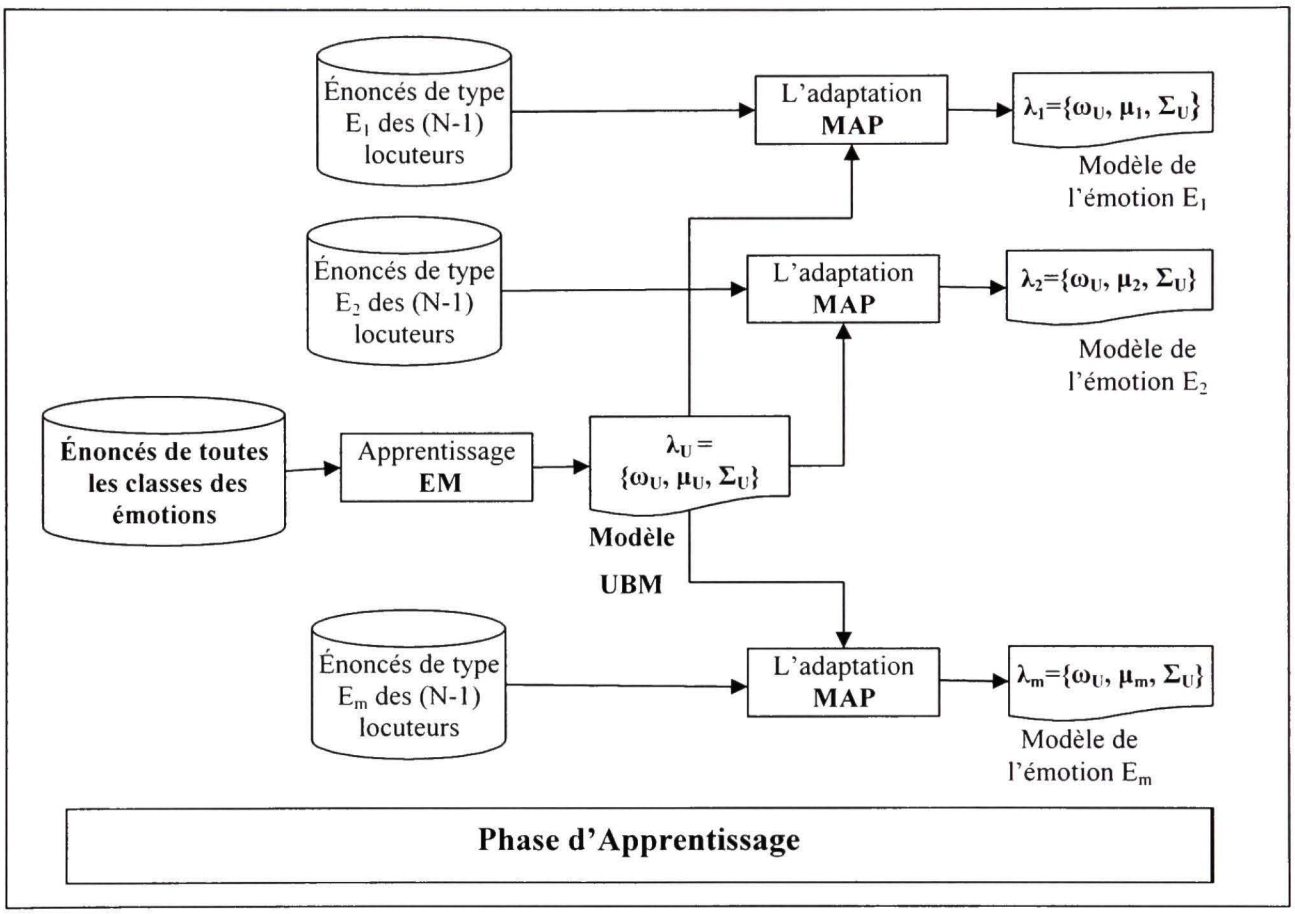


**Figure 4.2 Diagramme bloc de la phase d'apprentissage du système de RAE, à *m* classes d'émotions, basé sur le modèle GMM-UBM.**

Pour la phase de test, la classe d'émotion associée à un énoncé est obtenue en calculant le logarithme de la vraisemblance de l'énoncé pour chacun des *m* modèles des classes d'émotions et en sélectionnant la classe *E* dont le modèle offre la valeur maximale du logarithme de la vraisemblance, selon l'équation (4.1). Le logarithme de la vraisemblance de l'énoncé pour le modèle d'une classe est calculé selon la formule (4.2) :

$$E = \arg\max_{i=1,..,m} \log p(\mathbf{X}|\lambda_i) \tag{4.1}$$

$$\log p(\mathbf{X}|\lambda) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\lambda) \tag{4.2}$$

## 4.2 Le corpus de données

Pour l'apprentissage et le test du système SMEG, nous avons utilisé le corpus d'émotions LDC décrit à la section 2.4.1.

## 4.3 Extraction des traits caractéristiques

Deux vecteurs de types MFCC ont été utilisés comme traits caractéristiques. Le premier, de dimension égale à 39, auquel nous référerons par V(39), est constitué des 13 premiers coefficients MFCC augmentés de leurs premières et secondes dérivées. Le second type de vecteur est composé de 60 traits (V(60)), constitué par les 20 premiers coefficients MFCC leurs premières et secondes dérivées. Ces traits caractéristiques ont été extraits en utilisant l'utilitaire *HTK* (Young, Woodland et Byrne, 1993). Les paramètres que nous avons utilisés pour l'extraction des MFCC sont donnés dans l'ANNEXE I.

## 4.4 Protocole d'expérimentation

Il existe deux méthodes populaires pour l'évaluation d'un système d'apprentissage :

### 4.4.1 Méthode « Holdout »

Avec la méthode « Holdout », les données sont aléatoirement divisées en deux parties; une de taille plus large est utilisée pour l'apprentissage et l'autre partie est réservée pour l'estimation du taux d'erreur. Cependant, cette méthode n'est pas fiable ou pertinente si nous sommes en présence d'un échantillon de taille limitée. Une autre version de cette méthode, appelée «data shuffle», consiste à répéter $L$ fois la division aléatoire des données en deux parties; une pour l'apprentissage et l'autre pour le test et calculer par la suite la moyenne des $L$ estimations des taux d'erreur évalués sur les parties des données de test.

### 4.4.2 Méthode « *K*-fold cross-validation »

Avec la méthode de la validation croisée, les données sont partitionnées en $K$ sous échantillons. Un de ces $K$ sous échantillons est utilisé pour le test du modèle et les $K$-1 restants pour l'apprentissage du modèle. Ce processus de la validation croisée est répété $K$ fois et chacun des $K$ sous-échantillons est utilisé exactement une fois pour le test. L'avantage de cette méthode est que toutes les données sont utilisées à la fois pour l'entraînement et le test. La « 10-fold cross validation » est la méthode la plus couramment utilisée.

Cependant, un inconvénient se présente avec cette méthode, quand elle est appliquée au domaine de la RAE, quand le nombre de locuteurs qui participent à la constitution du corpus d'émotion est limité. La partition aléatoire des données en k sous-échantillons fait, que les $K$-1 sous-échantillons utilisés pour l'apprentissage ainsi que le $K^{ième}$ sous-échantillon utilisé pour le test contiennent tous des données qui proviennent des mêmes locuteurs, ce qui fait perdre au système la propriété d'être strictement indépendant du locuteur. Une variante à « $K$-fold cross-validation », appelée « Leave-One-Speaker-Out » a été utilisée par certains auteurs et que nous avons adopté comme protocole d'expérimentation sur les données LDC, un corpus généré par sept locuteurs.

### 4.4.3 Méthode de la validation croisée par locuteur

La méthode de la validation croisée par locuteur (Leave-One-Speaker-Out, LOSO) permet de garantir une indépendance stricte du système de RAE du locuteur et par conséquent, d'évaluer les performances du système dans le cas le plus défavorable. À chaque itération d'expérience, les données de six locuteurs du corpus LDC sont utilisées pour l'apprentissage des modèles alors que les données du septième sont utilisées pour le test. Ce processus est répété pour chacun des sept locuteurs. Les performances du système sont obtenues par le calcul de la moyenne des performances des sept expériences.

## 4.5 Critères d'évaluation du système de RAE

Différents critères peuvent être utilisés pour évaluer les performances d'un système de reconnaissance. Les critères *exactitude*, *précision*, *rappel* et *F-mesure* sont souvent utilisés par les auteurs.

L'*exactitude* reflète la justesse du classificateur en général. La *précision* représente le nombre de données classées correctement, comme positives par exemple, par rapport au nombre de données totales reconnues comme positives. Le *rappel* représente le nombre de données classées correctement, comme positive par exemple, au regard du nombre de données positives qui existent dans le corpus de données. Les valeurs de la précision et du rappel ne sont pas discutées séparément en général. Plutôt, soit les valeurs de l'une des deux mesures sont comparées à un niveau fixe de l'autre mesure ou soit les deux valeurs sont combinées en une seule mesure. La *F-mesure*, appelée encore *F-score*, est une mesure populaire qui combine la précision et le rappel ainsi que leurs pondérations.

Afin de présenter les formules de calcul des différents critères définis ci-devant, nous introduisons à travers le **Tableau 4.1** la notion de la *matrice de confusion* pour le cas d'un problème de RAE à deux classes d'émotions : positive et négative. La matrice de confusion est aussi utilisée par plusieurs auteurs pour visualiser, pour chaque classe de modèle, les

vraies classifications versus les classifications prédites notamment dans le cas de problèmes à plusieurs classes.

Tableau 4.1
Matrice de confusion d'un problème à deux classes

| | **Classe prédite positive** | **Classe prédite négative** |
|---|---|---|
| **Vraie positive** | **V**raie **P**ositive (VP) | **F**ausse **N**égative (FN) |
| **Vraie négative** | **F**ausse **P**ositive (FP) | **V**raie **N**égative (VN) |

Les critères *exactitude*, *précision*, *rappel* et *F-mesure* sont calculés comme suit :

$$exactitude = \frac{VP + VN}{VP + FP + FN + VN} \times 100\% \tag{4.3}$$

$$précision = \frac{VP}{VP + FP} \tag{4.4}$$

$$rappel = \frac{VP}{VP + FN} \tag{4.5}$$

$$F\text{-}mesure = \frac{2 \times précision \times rappel}{précision + rappel} \tag{4.6}$$

Les critères *précision*, *rappel* et *F-mesure* varient entre 0 et 1. Plus la valeur de ces critères est élevée plus le critère est meilleur.

Les performances du système de RAE que nous développerons seront mesurées avec les deux critères *exactitude* et la *F-mesure* pour les expériences à deux classes d'émotions, et nous utiliserons *exactitude* et la *matrice de confusion* pour les expériences à 15 classes d'émotions.

Pour des fins de comparaison entre différents systèmes, nous avons utilisé un test d'hypothèse pour affirmer ou infirmer qu'un classificateur est significativement meilleur qu'un autre. Étant donné que nos expériences sont basées sur un protocole LUSO, une variante de la validation croisée, nous avons donc utilisé le *K-Fold CrossValidation Paired t-Test* (Kuncheva, 2004) qui nous parait le plus approprié pour le cas du protocole d'expérimentation que nous avons utilisé.

**K-Fold CrossValidation Paired t-Test *(CVP t-Test)***

Considérons deux modèles de classificateurs $A$ et $B$, et soit $\boldsymbol{X}$ l'ensemble des données divisées en $K$ parties selon les $K$ locuteurs. Chaque partie, c.-à-d. les données associées à un locuteur, est utilisée pour tester les classificateurs $A$ et $B$ entrainés avec les $K$-$1$ parties restantes. Les taux de classification des données de test sont notés par $P_A$ et $P_B$ respectivement. Ce processus est répété $K$ fois (7 fois dans le cas de notre corpus) et les taux de classification sont étiquetés avec l'exposant $(i)$, i = 1,..., $K$. L'ensemble des $K$ différences est calculé par la suite, $P^{(1)} = P_A^{(1)} - P_B^{(1)}$ jusqu'à $P^{(K)} = P_A^{(K)} - P_B^{(K)}$. Comme supposition, nous présumons que cet ensemble de différences représente des valeurs indépendantes d'un échantillon qui suit une distribution normale. Sous l'hypothèse nulle ($H_0$ : taux de reconnaissance identiques), la statistique suivante possède une $t$-distribution avec $K$-1 degrés de liberté :

$$t = \frac{\overline{P}\sqrt{K}}{\sqrt{\sum_{i=1}^{K}(P^{(i)} - \overline{P})^2 / (K-1)}} \tag{4.7}$$

où $\overline{P} = (1/K)\sum_{i=1}^{K} P^{(i)}$ . Si la valeur calculée $t$ est supérieure à la valeur tabulée pour le niveau de signification choisi et pour $K$-1 degrés de liberté, $H_0$ est alors rejetée et nous acceptons qu'il existe une différence significative entre les deux modèles de classification.

Pour un niveau de signification de 0.05 et un degré de liberté égal à 6 (paramètres représentant notre protocole d'expérimentation), la valeur tabulée de la distribution $t$ est égale à 2.447.

## 4.6 Description des groupes d'expériences

Afin de tester les performances du système de RAE que nous avons développé, nous nous sommes intéressés à trois groupes d'expériences qui diffèrent par les classes d'émotions participantes aux différentes expériences. Elles sont définies comme suit :

**Expérience I** (*EXP I*) : Dans ce groupe, nous procédons à la classification des données de la classe émotionnelle *neutre* (*Neutral*) versus la classe *colère forte* (*Hot Anger*), désignée par *colère* tout court dans le reste de ce mémoire, sauf pour lever l'ambiguïté.

**Expérience II** (*EXP II*) : Les classes d'émotions participantes à ce groupe d'expérience sont la classe *neutre* et la classe *tristesse* (*sadness*).

**Expériences III** (*EXP III*) : Dans ce groupe d'expériences, il s'agit de reconnaître les quinze classes d'émotions du corpus *LDC*.

Le choix de ces groupes d'expériences est motivé par les deux considérations suivantes :

1. La première est liée au domaine d'application de notre système. Comme expliqué dans la partie problématique, nous nous intéressons à reconnaître les émotions positives des émotions négatives pour détecter les appels problématiques des clients d'un centre d'appel. Ainsi, la classe *neutre* jouera le rôle de la classe positive et les classes *colère* et *tristesse* constitueront chacune d'elles une classe négative.

2. De plus, nous considérons ces expérimentations comme expériences de référence, qui serviront à des fins de comparaison de performance entre notre système et ceux des autres auteurs qui ont travaillé sur le même corpus de données et les mêmes groupes d'expériences.

## 4.7 Résultats et discussion

Les résultats des trois groupes d'expériences sont présentés à travers les **Figure 4.3** à **Figure 4.5**. Chacune de ces figures illustre les performances de classification pour les quatre types de système de RAE; GMM-V(39), GMM-UBM-V(39), GMM-V(60), GMM-UBM-V(60) en fonction du nombre de mélanges de gaussiennes du modèle. Nous rappelons que V(i) désigne un système de RAE basé sur un vecteur de traits de taille égale à i. Ces graphiques ont été tracés à partir des résultats affichés dans les tableaux II.1, II.2 et II.3 de l'ANNEXE II. Dans chacun de ces trois tableaux, le résultat de chaque expérience est décrit par la valeur du taux de classification correct, la variance et la valeur de F-mesure.

### 4.7.1 Effet de l'utilisation d'un UBM

Les résultats de ces expériences montrent que l'utilisation d'un modèle du monde pour l'entraînement des classes des émotions permet d'améliorer les taux de classification pour l'expérience *Exp II* (**Figure 4.4**) et ce quel que soit le type de vecteurs de traits utilisé (V(39) ou V(60)). Nous avons enregistré un gain relatif moyen de 2% pour le système V(39) et 14% pour V(60). Pour l'expérience *Exp III* (**Figure 4.5**), nous relevons une amélioration de 4% pour le système à V(39) et de 5% pour le système à V(60). Nous constatons également que l'adaptation MAP a plus d'effet sur les systèmes à V(60) que sur les systèmes à V(39), car en absence de données de large volume, le système utilisant des vecteurs de données de grande dimension, V(60) dans notre cas, sont plus susceptibles de subir l'effet de la malédiction de la dimension, et par conséquent, c'est sur ces systèmes que l'utilisation d'un UBM a plus d'impact.

Cependant pour l'expérience *EXP I* (**Figure 4.3**), les meilleurs résultats ont été obtenus sans l'utilisation d'un UBM, et ce, indépendamment également du type de vecteurs de traits utilisé. Les performances des systèmes à base d'un UBM sont plus basses de 6% et de 9% pour les systèmes V(39) et V(60) respectivement. Nous expliquons ceci, d'une part, par le fait que les données des classes *colère* et *neutre* (*EXP I*) sont suffisamment représentatives de leurs classes respectives, dans le sens où les sept acteurs simulaient de la même manière ces catégories d'émotions.

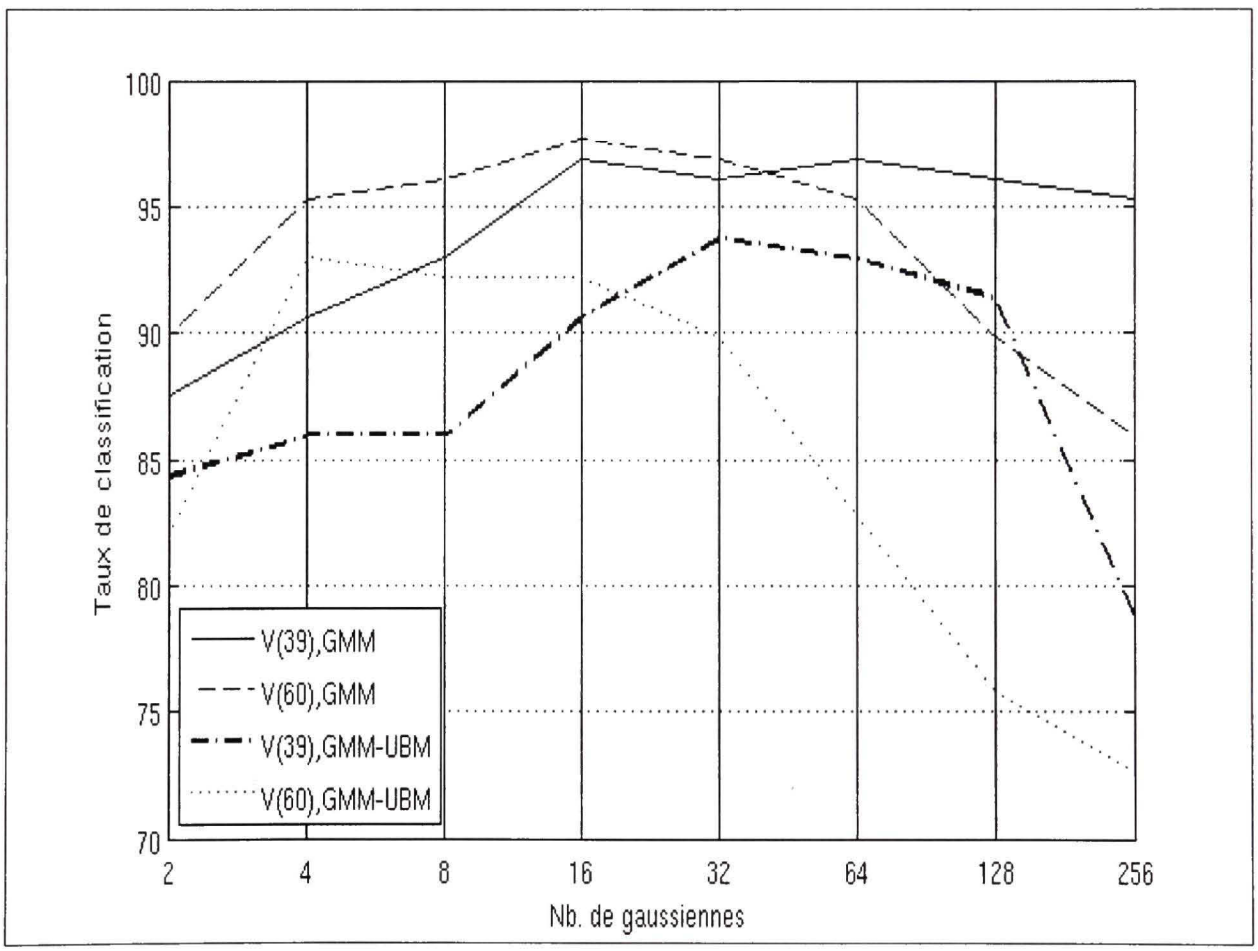


**Figure 4.3 Résultats de la classification de la classe d'émotion *neutre* vs *colère* (EXP I) du système SMEG en fonction du nombre de gaussiennes, du vecteur de traits et du type de classificateur.**

D'autre part, la classe *colère* est caractérisée par une dispersion de données qui est beaucoup plus large que la classe *neutre* (voir **Figure 3.3**); une information dont bénéficie le système GMM (qui utilise une matrice de covariance propre à chaque modèle de classe d'émotion) contrairement au système GMM-UBM où nous avons utilisé une matrice de covariance partagée pour toutes les classes d'émotions.

Tableau 4.2

Résultats du 7-Fold CrossValidation Paired t-Test pour déterminer l'apport d'un UBM sur les performances des systèmes SMEG

| Groupe expérience | Type de système | Système | Taux | t (CVP t-Test) | Décision |
|---|---|---|---|---|---|
| EXP I | V(39) | GMM | 96.88 ± 4.52 | 1.9674 | A |
| | | GMM-UBM | 93.75 ± 10.14 | | |
| | V(60) | GMM | 97.66 ± 2.92 | 2.4808 | R |
| | | GMM-UBM | 92.97 ± 7.48 | | |
| EXP II | V(39) | GMM | 71.65 ± 19.6 | 3.7362 | R |
| | | GMM-UBM | 73.23 ± 17.03 | | |
| | V(60) | GMM | 66.93 ± 14.58 | 9.0645 | R |
| | | GMM-UBM | 76.38 ± 18.02 | | |
| EXP III | V(39) | GMM | 17.18 ± 2.18 | 5.4648 | R |
| | | GMM-UBM | 17.87 ± 2,94 | | |
| | V(60) | GMM | 16.41 ± 4.57 | 4.8697 | R |
| | | GMM-UBM | 17.27 ± 3.82 | | |
| Dans la colonne décision A et R désignent respectivement acceptation et rejet (si $t > 2.447$) de l'hypothèse nulle (différence non significative). | | | | | |

En conclusion, d'après les résultats de tests d'hypothèses du **Tableau 4.2**, l'utilisation d'un UBM permet d'améliorer significativement les performances du système SMEG pour les expériences *EXP II* et *EXP III*. Pour les systèmes V(60) de l'expérience *EXP I* où les résultats du système GMM sont meilleurs que ceux du système GMM-UBM, les deux systèmes sont significativement différents alors que pour le type de système V(39) nous ne pouvons conclure qu'il existe une différence entre les systèmes GMM et GMM-UBM.

### 4.7.2 Effet de la dimension du vecteur de coefficients MFCC

D'après les résultats des expériences *EXP I* (**Figure 4.3**) et *EXP II* (**Figure 4.4**), nous constatons que les meilleurs taux de classification sont atteints avec un vecteur de traits composé des 20 premiers coefficients spectraux (plus les premières et secondes dérivées) versus les 13 premiers coefficients (plus les premières et secondes dérivées). Ces performances sont réalisées par les systèmes V(60)-GMM pour *EXP I* et V(60)-GMM-UBM pour *EXP II*. Notons que la différence de performance entre les systèmes V(39) et V(60) est statistiquement significative pour les deux types de systèmes GMM et GMM-UBM pour le groupe d'expérience *EXPII*, alors qu'elle ne l'est significativement que pour le système GMM dans le cas du groupe *EXP I* (voir résultats du test d'hypothèse du **Tableau 4.3**).

Une comparaison plus approfondie des différents systèmes : V(39)-GMM vs V(60)-GMM et V(39)-GMM-UBM vs V(60)-GMM-UBM sur le plan de la tendance des courbes de performances, permet de relever que la supériorité des vecteurs V(60) sur V(39) se produit quand le nombre de mélanges de gaussiennes n'est pas très élevé ( inférieur ou égal à 32 pour les systèmes GMM et inférieur ou égal à 16 pour les systèmes GMM-UBM) pour les expériences *EXP I*. Pour les expériences *EXP II*, nous constatons que la prévalence des vecteurs V(60) sur V(39) est moins évidente et se produit avec le système GMM-UBM pour un certain nombre de mélanges de gaussiennes. Nous expliquons ceci par le fait que les vecteurs à V(60) nécessitent une plus grande quantité de données d'apprentissage par rapport aux vecteurs V(39) à cause de la dimension plus élevé du vecteur de traits. Par conséquent, les systèmes à V(60) sont plus assujettis au phénomène de la malédiction de la dimensionnalité qui s'accentue encore plus avec l'augmentation du nombre de mélanges de gaussiennes du modèle et dans le cas d'absence d'un modèle UBM.

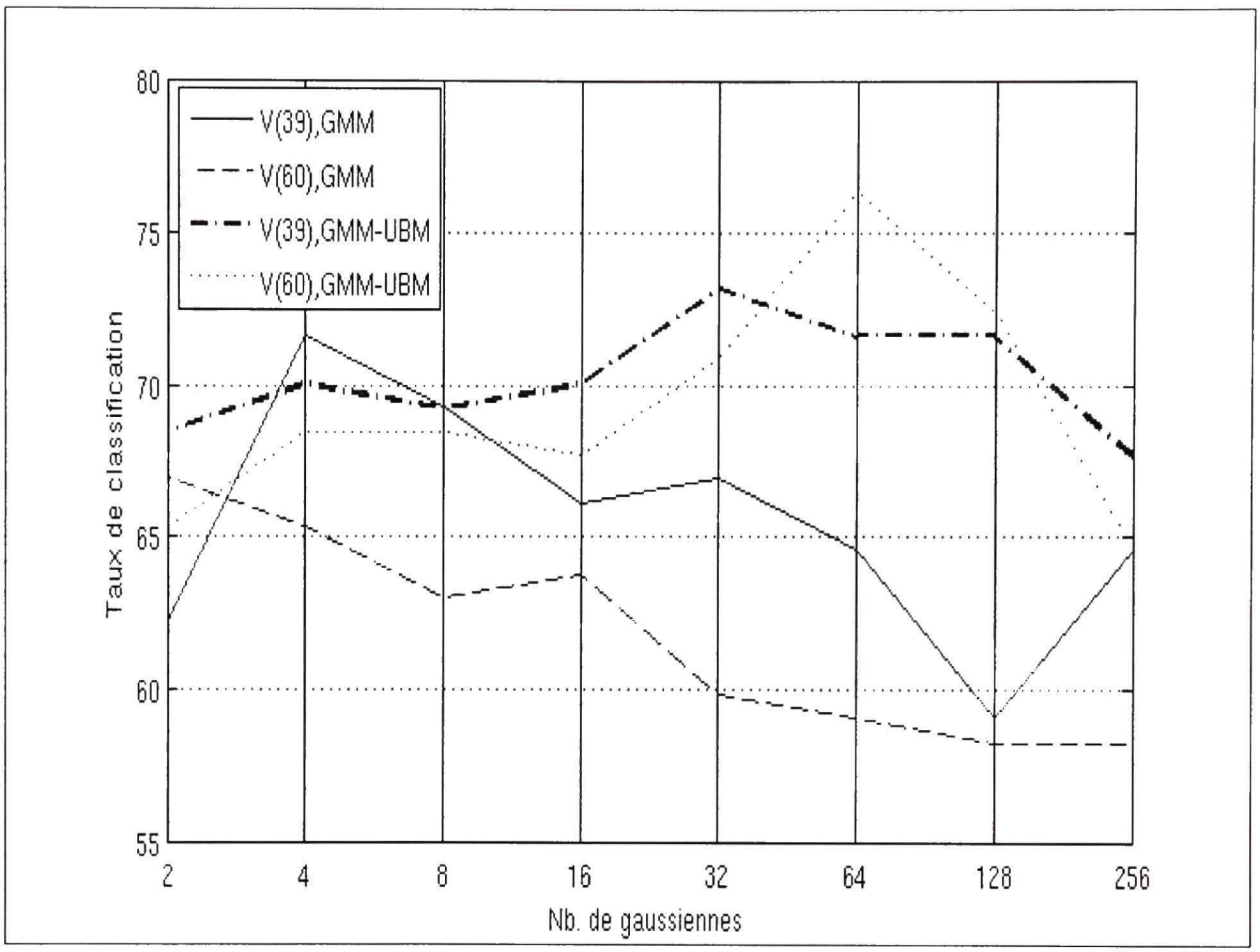


**Figure 4.4 Résultats de la classification de la classe d'émotion *neutre* vs *tristesse* (EXP II) du système SMEG en fonction du nombre de gaussiennes, du vecteur de traits et du type de classificateur.**

Nous pouvons donc penser qu'il existe dans les coefficients spectraux supérieurs une certaine information utile pour la discrimination de certaines classes d'émotions dans le domaine de la RAE contrairement à celui de la RAP, et qu'un gain en performance par l'utilisation de cette information est possible à condition que les données d'apprentissage soient en quantité suffisante.

Tableau 4.3

Résultats du 7-Fold CrossValidation Paired t-Test pour tester l'effet de la dimension du vecteur MFCC sur les performances des systèmes SMEG

| Groupe expérience | Type de système | Système | Taux | t (CVP t-Test) | Décision |
|---|---|---|---|---|---|
| EXP I | GMM | V(39) | 96.88 ± 4.52 | 1 | A |
| | | V(60) | 97.66 ± 2.92 | | |
| | GMM-UBM | V(39) | 93.75 ± 10.14 | 2.7752 | R |
| | | V(60) | 92.97 ± 7.48 | | |
| EXP II | GMM | V(39) | 71.65 ± 19.61 | 3.0410 | R |
| | | V(60) | 66.93 ± 14.58 | | |
| | GMM-UBM | V(39) | 73.23 ± 17.03 | 4.6580 | R |
| | | V(60) | 76.38 ± 18.02 | | |
| EXP III | GMM | V(39) | 17.18 ± 2.18 | 2.0639 | A |
| | | V(60) | 16.41 ± 4.57 | | |
| | GMM-UBM | V(39) | 17.87 ± 2.94 | 5.8807 | R |
| | | V(60) | 17.27 ± 3.82 | | |
| Dans la colonne décision A et R désignent respectivement acceptation et rejet (si t > 2.447) de l'hypothèse nulle. | | | | | |

### 4.7.3 Effet de la taille du nombre de mélanges de gaussiennes

Nous constatons, d'après les graphiques des **Figure 4.3** à **Figure 4.5**, que la tendance des performances des différents types de systèmes va dans le sens croissant quand nous augmentons le nombre de mélanges de gaussiennes (en partant de la valeur 2) jusqu'à ce qu'elles atteignent leurs limites de performance et recommencent à décroître par la suite. Ces systèmes performent moins avec un nombre de mélanges de gaussiennes réduit car de tels nombres ne sont pas suffisamment élevés pour modéliser toutes les variations fines du spectre du signal qui caractérisent les classes des émotions.

Par ailleurs, quand le nombre de mélanges de gaussiennes du modèle est très élevé, une quantité plus large de données est nécessaire pour l'apprentissage du modèle, tel que discuté

dans la section précédente, une condition qui fait défaut à notre corpus de données LDC. Par conséquent les valeurs médianes des nombres de mélanges de gaussiennes offrent un meilleur compromis qui tient compte de la taille des échantillons d'apprentissage et de la capacité de modélisation du spectre du signal de la parole. Nous constatons également que le maximum des performances est atteint avec un nombre de mélanges de gaussiennes plus réduit pour les expériences *EXP I* (**Figure 4.3**) et *EXP II* (**Figure 4.4**), où seulement deux catégories d'émotions participent, alors que ce nombre est plus élevé pour l'expérience *EXP III* (**Figure 4.5**) où quinze classes d'émotions sont modélisées.

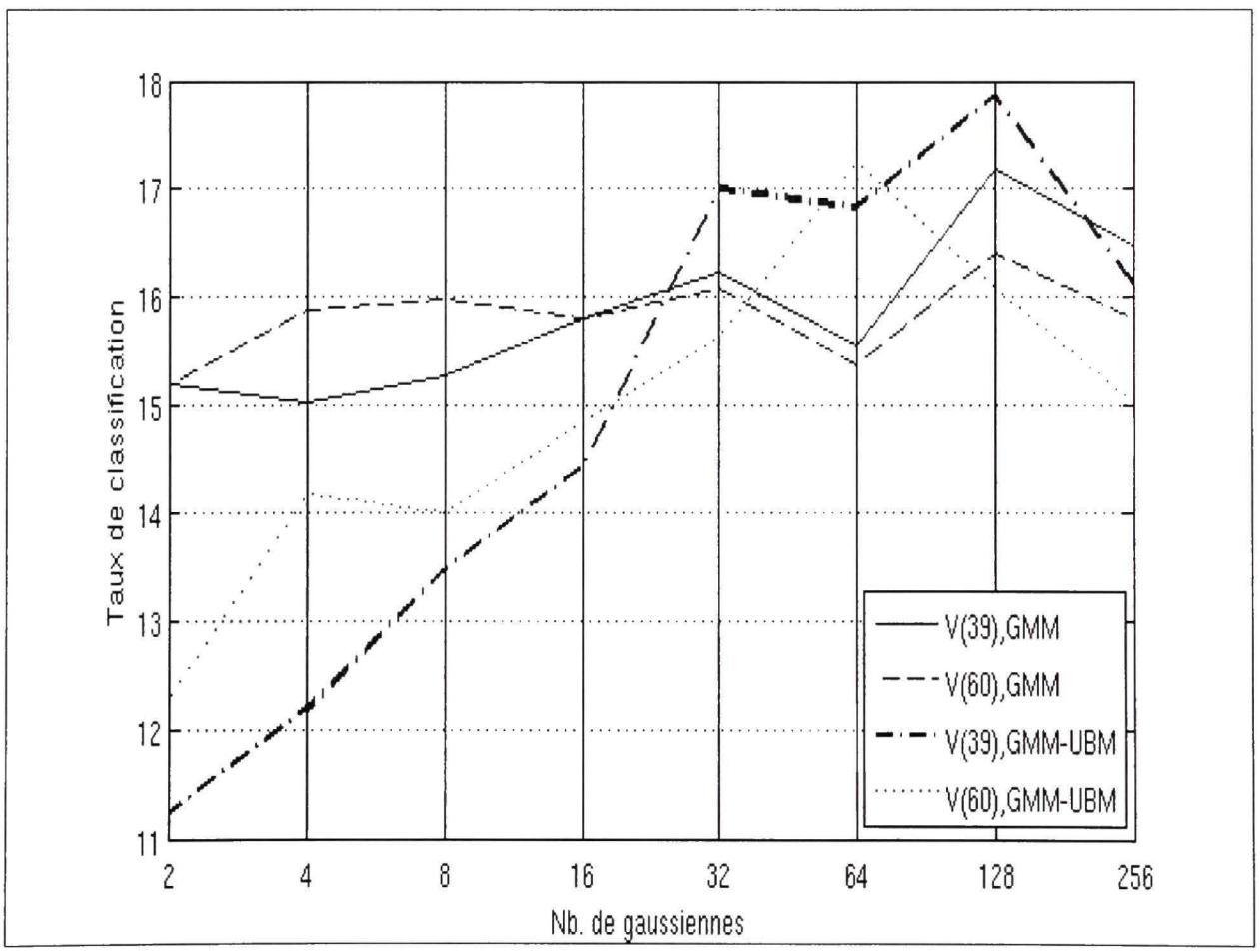


**Figure 4.5 Résultats de la reconnaissance automatique des 15 classes d'émotions (EXP III) en fonction du nombre de gaussiennes, du vecteur de traits et du type de classificateur.**

Le **Tableau 4.4** récapitule les meilleurs résultats obtenus pour chaque groupe d'expériences ainsi que les paramètres du système associé.

Tableau 4.4

Récapitulatif des résultats des trois groupes d'expériences avec les valeurs des paramètres correspondants aux meilleures performances

| **Expérience** | **Taux / f-mesure** | **Nb gauss.** | **Nb traits** | **UBM** |
|---|---|---|---|---|
| Neutre vs Colère | 97.66±2.92 / 0.98 | 16 | 60 | Non |
| Neutre vs Tristesse | 76.38±18.02 / 0.8 | 64 | 60 | Oui |
| 15 classes | 17.87±2.94 | 128 | 39 | Oui |

Le **Tableau 4.5** représente la matrice de confusion obtenue de la classification des quinze classes d'émotions *(EXP III)*, du système GMM-UBM-V(39) avec un mélange de 128 gaussiennes. Les valeurs des colonnes représentent les vraies classes d'émotions et les lignes représentent les classes reconnues par le classificateur. Les valeurs de cette matrice permettent de dégager deux remarques :

1. La première concerne la grande disparité qui existe dans les performances de classification des quinze classes d'émotions. En effet, certaines classes telles que la classe *neutre* et la classe *colère* sont relativement bien reconnues, avec un taux de 52% et 49% respectivement, contrairement à d'autres classes qui enregistrent de très faibles taux ne dépassant pas, par exemple, les 4% et 5% pour les classes *intérêt* et *joie* respectivement. Nous expliquons les très faibles taux de reconnaissance obtenus par certaines classes par la mauvaise représentativité des données associées à ces classes d'émotions utilisées dans les phases d'apprentissage et de test des modèles. Cette non-représentativité de ces échantillons de données est une conséquence de l'absence d'un consensus sur la manière d'exprimer ces classes d'émotions, ceci est

d'ailleurs bien mis en évidence par la classe intérêt, par exemple, où quatre des sept locuteurs enregistrent un taux de reconnaissance nul.

Tableau 4.5

Matrice de confusion des 15 classes d'émotions du système SMEG. Les valeurs des colonnes représentent les vraies classes et les lignes représentent les classes reconnues par le classificateur

| | **Anx** | **Enn** | **Fro** | **Mép** | **Dés** | **Dég** | **Exa** | **Joi** | **Col** | **Int** | **Neu** | **Pan** | **Fie** | **Tri** | **Hon** | **Moy.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Anxiété** | 13 | 5 | 5 | 3 | 15 | 7 | 2 | 3 | 0 | 6 | 3 | 3 | 1 | 6 | 13 | 15.29% |
| **Ennui** | 2 | 13 | 5 | 6 | 10 | 4 | 2 | 1 | 5 | 5 | 7 | 0 | 4 | 7 | 13 | 15.48% |
| **Col. froide** | 1 | 2 | 12 | 12 | 9 | 8 | 6 | 4 | 5 | 4 | 3 | 2 | 2 | 3 | 2 | 16.00% |
| **Mépris** | 5 | 5 | 3 | 14 | 12 | 6 | 3 | 3 | 10 | 4 | 2 | 8 | 4 | 0 | 6 | 16.47% |
| **Désespoir** | 10 | 8 | 3 | 2 | 14 | 7 | 3 | 6 | 0 | 4 | 1 | 3 | 2 | 3 | 12 | 17.95% |
| **Dégoût** | 3 | 6 | 11 | 1 | 14 | 9 | 2 | 8 | 5 | 1 | 7 | 6 | 5 | 1 | 4 | 10.84% |
| **Exaltation** | 0 | 1 | 5 | 3 | 6 | 4 | 5 | 6 | 14 | 3 | 4 | 12 | 6 | 3 | 1 | 6.85% |
| **Joie** | 1 | 3 | 8 | 6 | 8 | 5 | 8 | 4 | 6 | 2 | 5 | 7 | 9 | 3 | 9 | 4.76% |
| **Colère** | 1 | 0 | 8 | 4 | 3 | 2 | 5 | 2 | 37 | 0 | 1 | 12 | 1 | 0 | 0 | 48.68% |
| **Intérêt** | 1 | 13 | 2 | 5 | 7 | 4 | 0 | 2 | 5 | 3 | 10 | 6 | 3 | 9 | 14 | 3.57% |
| **Neutre** | 0 | 9 | 0 | 4 | 2 | 2 | 1 | 0 | 0 | 3 | 27 | 1 | 0 | 1 | 2 | 51.92% |
| **Panique** | 5 | 3 | 3 | 7 | 9 | 2 | 5 | 4 | 16 | 3 | 1 | 22 | 1 | 1 | 0 | 26.83% |
| **Fierté** | 1 | 5 | 4 | 9 | 6 | 5 | 4 | 6 | 5 | 6 | 7 | 0 | 9 | 3 | 6 | 11.84% |
| **Tristesse** | 5 | 8 | 2 | 2 | 11 | 4 | 3 | 0 | 0 | 6 | 5 | 1 | 3 | 10 | 15 | 13.33% |
| **Honte** | 6 | 10 | 3 | 6 | 6 | 4 | 0 | 3 | 0 | 3 | 7 | 0 | 2 | 6 | 16 | 22.22% |

Note : Les titres de colonnes représentent les trois premières lettres des noms des classes d'émotions affichés également, dans le même ordre, comme titres de lignes.

2. Une confusion caractérise certaines classes d'émotions, notamment les classes : *anxiété* avec *désespoir*, *ennui* avec *honte* et *colère* avec *panique*. Nous considérons

qu'il y a confusion entre deux classes A et B si la prédiction des données de la classe A entraine une classification erronée d'une quantité importante de ces données en classe B et vice-versa, c.-à-d. qu'une quantité importante de données de la classe B sont prédites incorrectement comme étant de la classe A.

### 4.7.4 Comparaison des résultats avec l'état de l'art

Le **Tableau 4.6** dresse une comparaison entre les performances du système que nous avons développé avec les systèmes des trois chercheurs déjà décrits dans la section 0 ayant utilisés le même corpus de données et le même protocole d'expérimentation, c'est à dire un test indépendant du locuteur pour mesurer les performances.

Tableau 4.6

Tableau comparatif de performance entre *SMEG* et les systèmes de l'état de l'art

| **Auteurs** | **Exp I** | **Exp II** | **Exp III** |
|---|---|---|---|
| Yacoub et al. | 94 % | 50 % | 9 % |
| Sethu et al. | 95 % | - | - |
| Huang et al. | 98 % | 69 % | 18 % |
| SMEG | 98 % | 76 % | 18 % |

Nous constatons que les taux de reconnaissance de notre système, *SMEG*, pour les expériences *EXP I* et *EXP III* sont similaires aux meilleures performances, qui ont été obtenues par Huang.

D'autre part, nous constatons que pour l'expérience II, nos résultats dépassent nettement les meilleures performances de l'état de l'art sur les données LDC, améliorant ainsi les taux de reconnaissance de 11%.

## 4.8 Conclusion

Dans ce chapitre, nous avons présenté la méthodologie suivie pour l'implémentation du système de RAE *SMEG*. Ce système a été conçu autour de l'*énoncé* comme unité d'analyse, les MFCC comme traits caractéristiques et les GMM comme modèle classificateur. Ce type de modèle a été expérimenté pour la première fois sur les données LDC. Nous avons cherché également à optimiser les performances de ce système en procédant à l'ajustement de trois de types de paramètres : le nombre de mélanges de gaussiennes, l'utilisation d'un modèle du monde et l'extension du nombre de coefficients MFCC au-delà des 13 premiers coefficients. Les performances du système obtenu ont dépassé les meilleurs résultats de l'état de l'art appliqué aux données LDC, pour l'expérience *EXP II* et nous avons reproduit les résultats des meilleures performances pour *EXP I* et *EXP III*. Par ailleurs, nous pensons que la précision des résultats obtenus peut faire objet d'autres améliorations dans les travaux futurs grâce à l'implémentation du rejet.

Dans le chapitre suivant, nous allons étudier les performances d'un système de RAE, basé sur une nouvelle unité d'analyse et une nouvelle méthode de calcul d'information à long terme dans notre domaine de recherche.

# CHAPITRE 5

# MÉTHODOLOGIE ET EXPÉRIMENTATION DU SYSTÈME SPPLG

## 5.1 Introduction

Dans la majorité des travaux réalisés jusqu'à présent, les auteurs se sont intéressés à la reconnaissance des émotions en considérant l'énoncé comme unité de base dans leurs modélisations. L'utilisation de cette unité s'adapte très bien au problème quand la charge de l'émotion exprimée est suffisamment répartie sur la majorité des phonèmes ou segments de l'énoncé et lorsque l'énoncé ne véhicule qu'une seule émotion. Cependant, parfois ces deux hypothèses ne sont pas vérifiées, notamment pour les énoncés issus des conversations réelles, où certains mots ou phonèmes sont émotionnellement plus saillants que le reste de l'énoncé. À titre d'exemple, un *oui* ou un *non* peut être émotionnellement plus marqué que d'autres mots d'un même énoncé. Partant de ce constat, d'autres unités d'analyse, tels que les mots, syllabes et phonèmes, ont été expérimentées par certains chercheurs, comme nous l'avons déjà vu dans la partie de la revue de littérature. Cependant, des contraintes et des difficultés importantes résultent de l'implémentation de systèmes basés sur ces unités. Parmi ces contraintes, la nécessité de munir le système RAE, non seulement d'un système de reconnaissance automatique de la parole, mais aussi d'un système pour l'alignement temporel entre la sortie du système de reconnaissance automatique de la parole et les phonèmes ou syllabes de l'énoncé. Aussi, toute erreur dans la sortie du système de reconnaissance automatique de la parole ou celui de l'alignement temporel, se répercute forcément sur les performances du système de reconnaissance de l'émotion.

Dans ce mémoire, nous proposons l'étude d'une nouvelle unité de modélisation pour la reconnaissance de l'émotion appelée *pseudosyllabe* (PS) qui permet de remédier aux inconvénients des unités précédentes. La *pseudosyllabe* est basée sur une segmentation

indépendante de tout système de reconnaissance de la parole, qui peut être réalisé dans des délais très courts, et est, par conséquent, bien adapté aux applications à temps réel.

La notion de *pseudosyllabe* a été déjà utilisée avec succès dans le domaine de l'identification du langage (Lin et Wang, 2005) et celui de la vérification du locuteur (Dehak, Dumouchel et Kenny, 2007) où elle a été utilisée dans le but de rechercher une information à long terme plus robuste. La méthode de segmentation des énoncés en *pseudosyllabes* que nous utiliserons dans notre système se base sur la même technique utilisée dans les deux derniers travaux cités. Nous combinerons également l'utilisation de la *pseudosyllabe* avec l'utilisation des coefficients de polynôme de Legendre pour le calcul de l'information à long terme, appliquée pour la première fois au domaine de la RAE.

## 5.2 Description du système

La Figure 5.1 présente l'architecture du système de RAE basé sur l'information prosodique comme traits caractéristiques, la *pseudosyllabe*, comme unité d'analyse, approximée par un *polynôme de Legendre* et utilisant les GMM comme modèle. Dans ce qui suit, ce système sera nommé **SPPLG** (**S**ystème-**P**rosodie-**P**seudosyllabe-polynôme-**L**egendre-**G**MM).

Le système, qui admet en entrée le signal de la parole, procède à l'extraction du contour de F0 et celui de l'énergie sur les régions voisées. Les contours de F0 et de l'énergie dont la taille minimale est supérieure à un seuil $T$ sont segmentés en *pseudosyllabes*. Les contours des *pseudosyllabes* sont par la suite approximés par l'utilisation d'une famille de polynômes orthogonaux appelée polynôme de Legendre (PL), déjà utilisée avec succès dans le domaine de la phonétique quantitative (Grabe, Kochanski et Coleman, 2003) et dans les travaux de Lin et Dehak (Lin et Wang, 2005; Dehak, Dumouchel et Kenny, 2007). La durée $T$ ms est imposée à la taille des segments dans le but d'avoir suffisamment de points pour l'interpolation des *pseudosyllabes* par des polynômes de Legendre. Les coefficients de ces PL, calculés pour chacun des contours des *pseudosyllabes* de F0 et de l'énergie, sont utilisés

comme vecteur de traits en combinaison avec la durée de la *pseudosyllabe*, le tremblotement et le vacillement. Le système *SPPLG* s'inscrit dans le diagramme de la revue de la littérature de la Figure 2.2, tel qu'illustré par la Figure 5.2. Dans ce qui suit, nous détaillerons chacune de ces opérations.

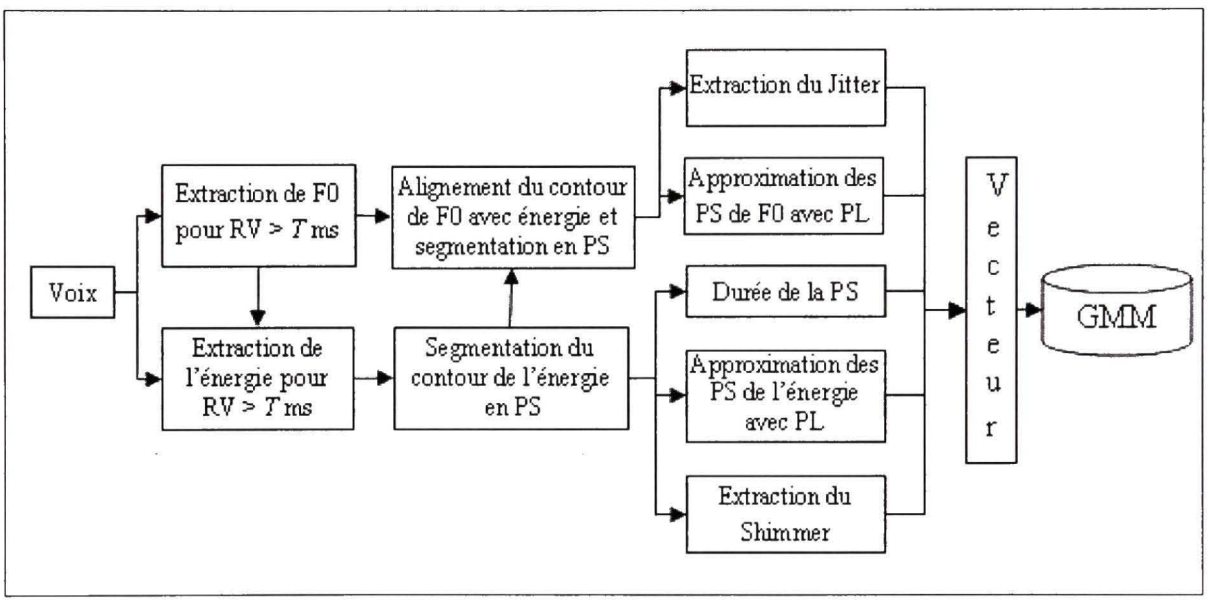


**Figure 5.1 Architecture du système de RAE *SPPLG*.**

## 5.3 Le corpus de données

Le corpus de données utilisé pour l'apprentissage et le test du système *SPPLG* est le corpus de données LDC, déjà décrit à la section 2.4.1.

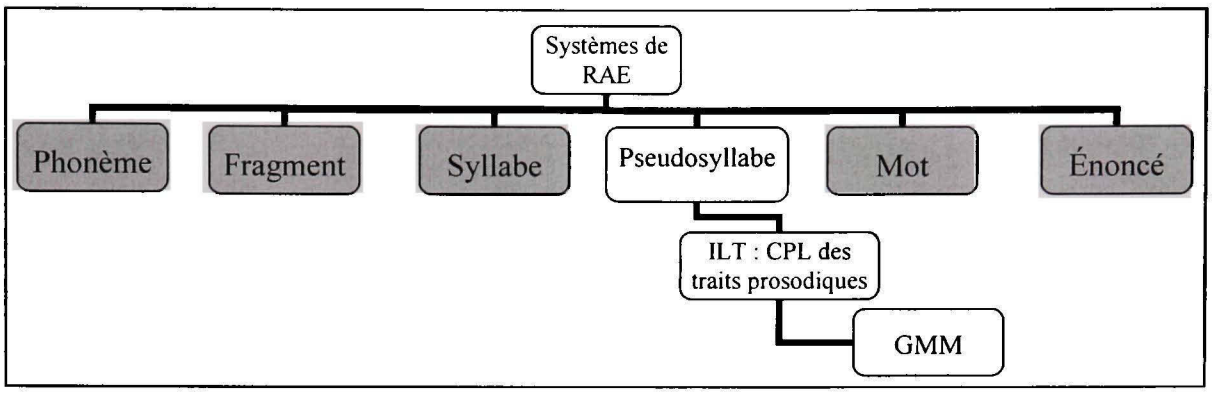


**Figure 5.2 Positionnement du système SPPLG par rapport à la revue de littérature.**

## 5.4 Extraction des caractéristiques

L'opération d'extraction des caractéristiques est réalisée en trois étapes : d'abord, les valeurs de la fréquence fondamentale et de l'énergie sont extraites au niveau de la trame, puis une segmentation de l'énoncé en unités de *pseudosyllabe* est effectuée, et enfin, l'information à long terme sur l'échelle d'une *pseudosyllabe* est calculée à travers une approximation par les coefficients du polynôme de Legendre.

### 5.4.1 Extraction des valeurs de F0 et de l'énergie

Nous utiliserons la fréquence fondamentale et l'énergie comme traits pour ce système. Les valeurs de F0 et de l'énergie sont extraites à partir du fichier audio pour chaque trame de 10 ms, en utilisant l'utilitaire *Praat* (Boersma et Weenink, 2008). Les valeurs de F0 sont calculées avec la méthode d'auto corrélation proposée par Boersma (Boersma et Weenink, 2008). Les valeurs des paramètres d'extraction de F0 sont affichées dans le Tableau 5.1

Notons que les valeurs de F0 sont indéfinies pour les régions non voisées car pour ces sons les cordes vocales n'entrent pas en vibration. La valeur de l'énergie est normalisée à l'échelle de l'énoncé en le soustrayant de la valeur maximale de tout l'énoncé (Dehak, Dumouchel et

Kenny, 2007) et la valeur de F0 en le divisant par la fréquence moyenne (Grabe, Kochanski et Coleman, 2003).

Tableau 5.1

Paramètres d'extraction de F0 avec *Praat*

| | |
|---|---|
| Taille de la fenêtre | 30 ms |
| Fréquence | 10 ms |
| Plafond de F0 (hommes) | 300 Hz |
| Plafond de F0 (femmes) | 500 Hz |
| Minimum de F0 (hommes) | 75 Hz |
| Minimum de F0 (femmes) | 100 Hz |
| Nombre max. de candidats | 5 |
| Seuil du silence | 0.03 |
| Seuil du voisement | 0.6 |
| Coût de l'octave | 0.01 |
| Coût du saut octave | 0.6 |
| Coût du voisée / non voisée | 0.14 |

La **Figure 5.3** et **Figure 5.4** représentent respectivement le contour de F0 et celui de l'énergie, de l'énoncé : « *One thousand six* », prononcé en simulant l'état émotionnel colère.

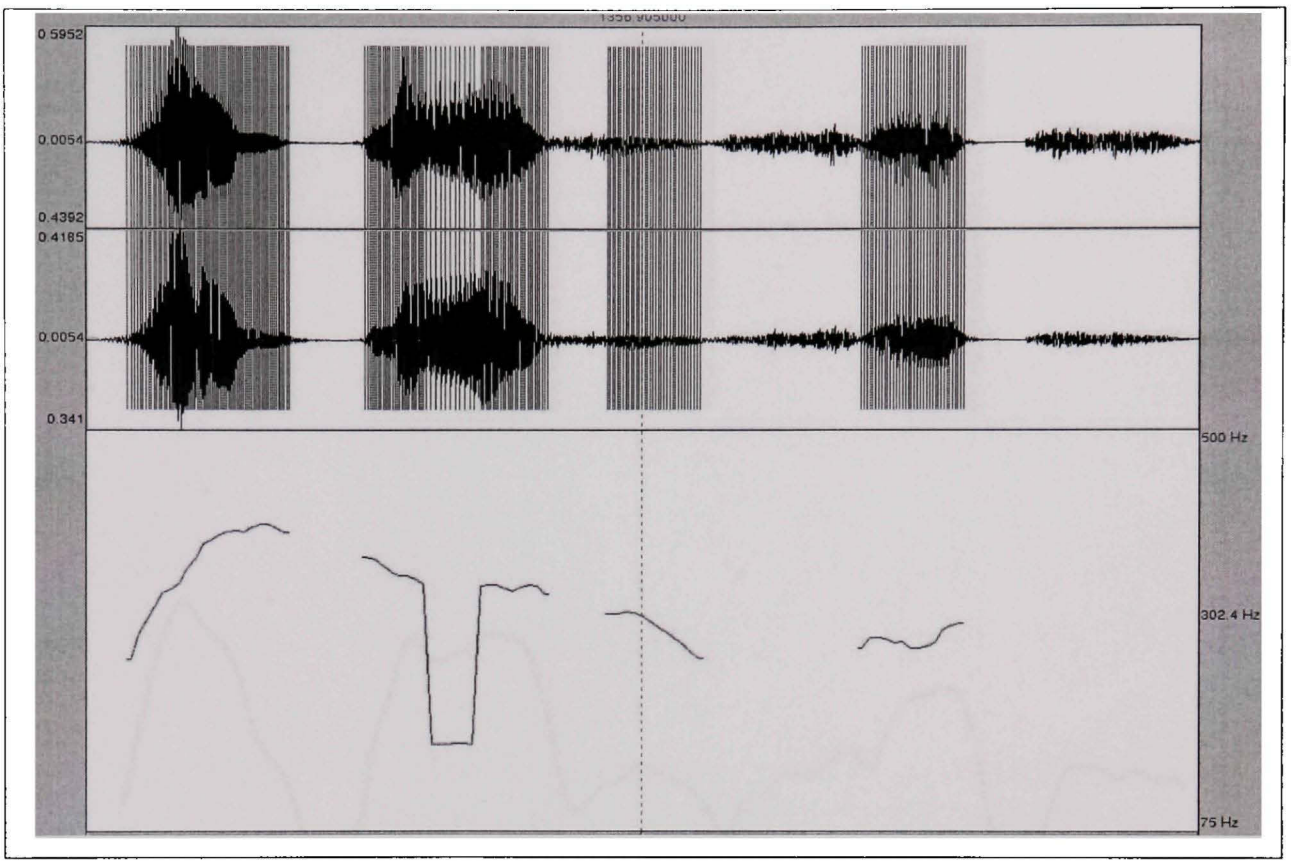


**Figure 5.3 Graphique représentant le contour de F0 (partie inférieure de l'image) de l'énoncé « One thousand six » prononcé en simulant l'état émotionnel colère forte, tracé avec l'outil Praat.**

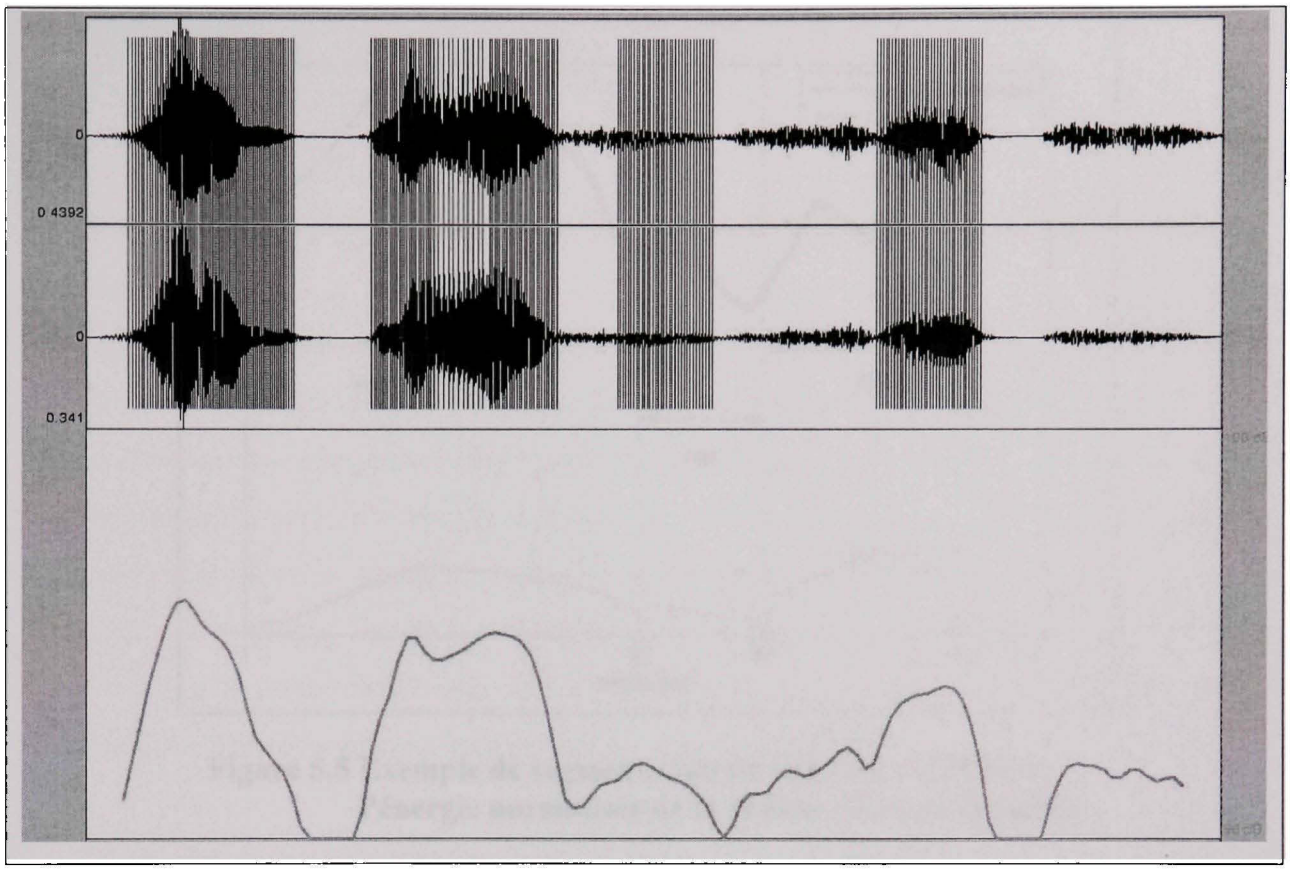


**Figure 5.4 Graphique représentant le contour de l'énergie (partie inférieure de l'image) de l'énoncé « One thousand six » prononcé en simulant l'état émotionnel colère forte, tracé avec l'outil Praat.**

### 5.4.2 Segmentation

Le but de cette étape est de segmenter les énoncés en des segments plus courts qui constitueront les *pseudosyllabes*. Après avoir réalisé un alignement du contour de F0 avec celui de l'énergie, nous procédons à une segmentation des deux courbes, guidée par le contour de l'énergie. Les points minimums locaux du contour de l'énergie détermineront les frontières des segments. Une *pseudosyllabe* peut correspondre à plusieurs syllabes ou à une partie d'une syllabe. La Figure 5.5 illustre l'opération de segmentation en *pseudosyllabes*.

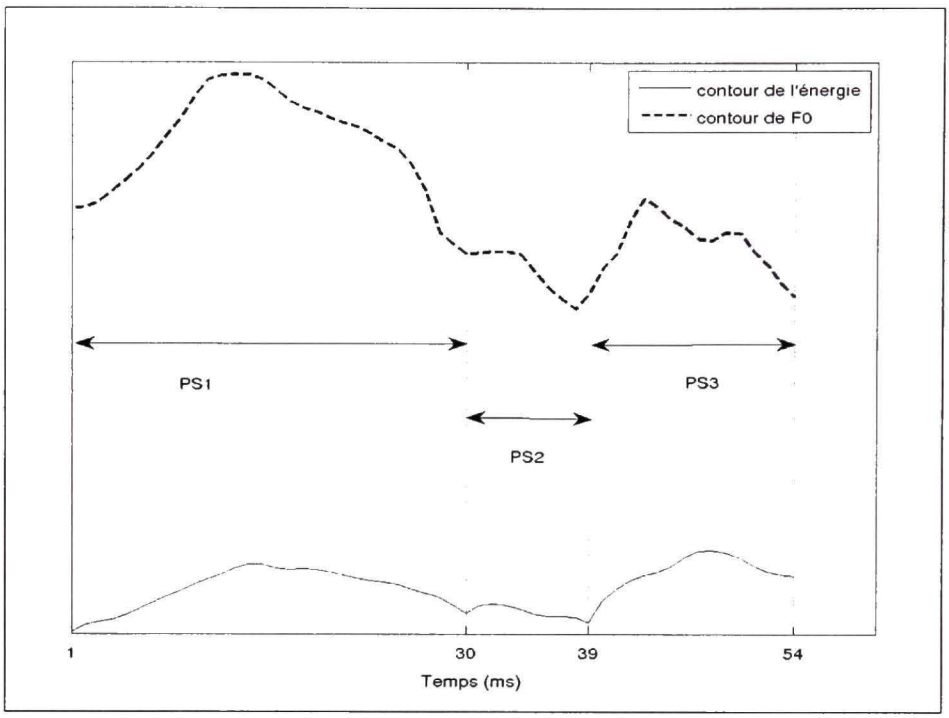


**Figure 5.5 Exemple de segmentation du contour de F0 et de l'énergie normalisée de la région voisée de la parole.**

### 5.4.3 Approximation

Tel que nous l'avons vu dans la partie revue de littérature, nous distinguons deux types d'information utilisés comme traits caractéristiques : l'information à court terme et l'information à long terme. Nous avons vu également que les paramètres statistiques de l'énergie et de F0 tels que le maximum, le minimum, l'étendue, la moyenne, l'écart-type sont utilisés comme fonctions pour obtenir l'information à long terme. Cependant, ces valeurs ne reflètent pas parfaitement les variations locales du contour de F0 et celui de l'énergie. C'est pourquoi nous utiliserons à la place une approximation du contour par la somme de polynômes de Legendre (PL).

L'approximation d'une fonction par une somme de polynômes de Legendre est définie comme suit :

$$f(t) = \sum_{i=0}^{M} a_i P_i(t) \qquad 5.1$$

où M représente le plus grand ordre du PL, $a_i$ le coefficient du PL d'ordre $i$ et $P_i$ le PL d'ordre $i$.

Les coefficients $a_0$,.., $a_M$, utilisés comme traits caractéristiques, offrent une meilleure représentation du contour du segment que les valeurs statistiques de la courbe qui, en fait, ne représentent qu'une partie de l'information véhiculée par les coefficients des PL. En effet, chaque coefficient modélise un aspect particulier de ce contour, par exemple, $a_0$ est interprété comme la moyenne du segment, $a_1$ représente la pente, $a_2$ donne des informations sur la courbure du segment, $a_3$ et $a_4$ modélisent les détails fins. Nous avons réalisé une normalisation de temps dans l'intervalle [−1, +1] afin que les coefficients soient comparables entre segments comme suggérée par Grabe (Grabe, Kochanski et Coleman, 2003).

La Figure 5.6 illustre l'approximation du contour de F0 en utilisant des polynômes de Legendre avec différents ordres.

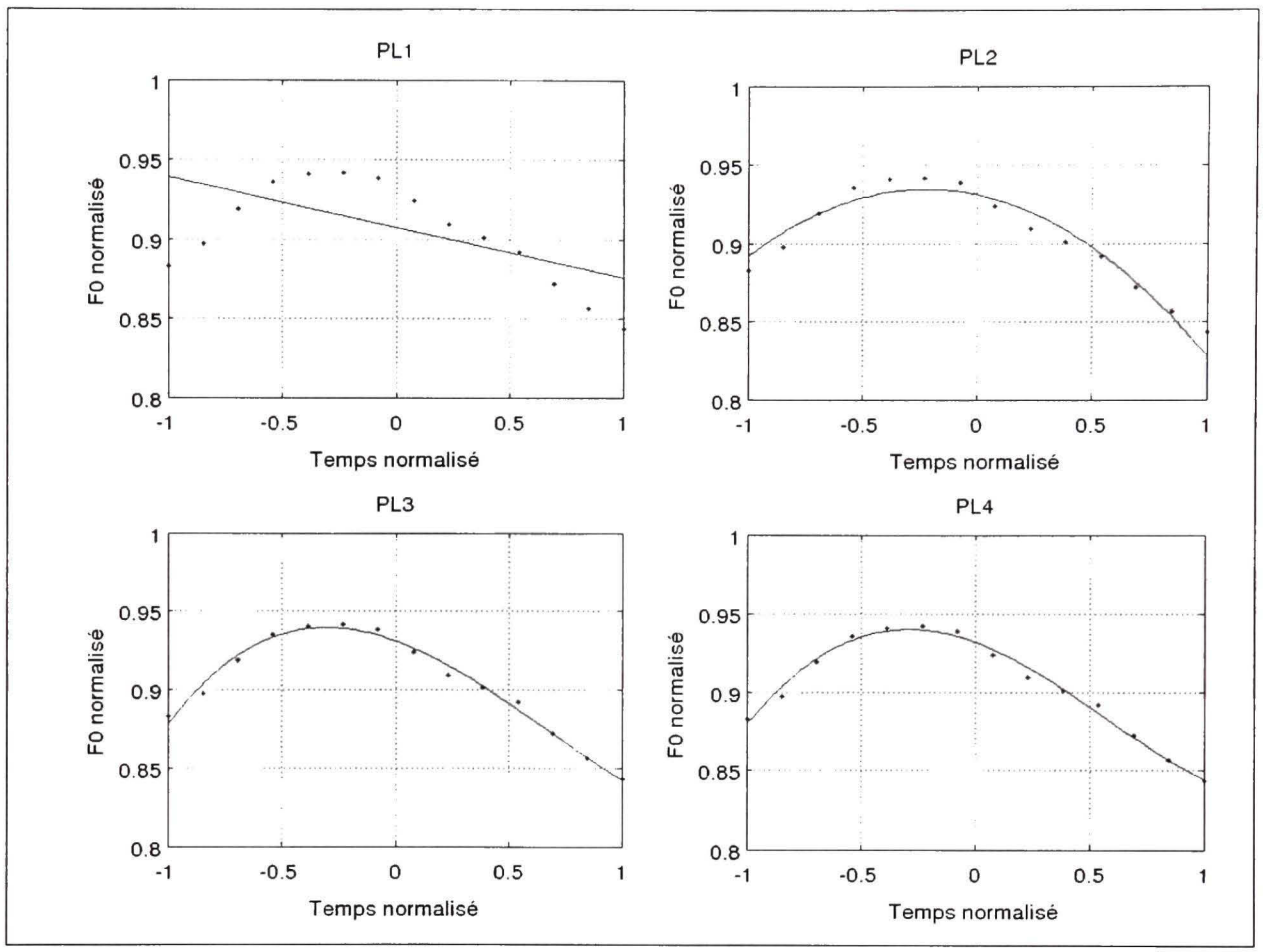


**Figure 5.6 Approximation du log de F0 par des polynômes de Legendre avec différents ordres.**

## 5.5 Apprentissage du modèle

Chaque classe d'émotion est modélisée par un GMM de quatre gaussiennes avec une matrice de covariance diagonale. Chaque GMM est entraîné directement à partir des vecteurs de traits extraits des *pseudosyllabes* des énoncés appartenant à la même classe d'émotion. Les paramètres de l'algorithme EM sont initialisés avec la méthode LBG décrite dans le paragraphe 3.5.

Pour la phase de test, la classe d'émotion à laquelle appartient un énoncé est déterminée selon l'équation 4.1, utilisée également pour le système *SMEG*, en calculant l'argument du maximum du logarithme de la vraisemblance des $m$ modèles de classes d'émotion de

l'énoncé. Cependant, la vraisemblance d'un énoncé est calculée différemment pour le système *SPPLG* et plusieurs stratégies d'évaluation de la vraisemblance d'un énoncé peuvent être considérées. Dans nos expériences, nous avons testé deux stratégies. Dans la première, le logarithme de la vraisemblance est calculé comme étant la somme du logarithme de la vraisemblance de chacune des *pseudosyllabes* constituant l'énoncé, tel que formulé par l'équation (5.2) :

$$\log p(\mathbf{X}|\lambda) = \sum_{n=1}^{p} \log p(\mathbf{S}_n | \lambda) \qquad (5.2)$$

où $\mathbf{S_n}$ représente le vecteur de traits de la n$^{ième}$ *pseudosyllabe* de l'énoncé à classifier segmenté en *p pseudosyllabes*.

Dans la deuxième stratégie, le $\log p(\mathbf{X}|\lambda)$ est considéré égal au maximum des logarithmes des vraisemblances des différentes *pseudosyllabes* constituant l'énoncé, tel que présenté par la formule (5.3) :

$$\log p(\mathbf{X}|\lambda) = \max_{n=1,..,p} \log p(\mathbf{S}_n | \lambda) \qquad (5.3)$$

Notons que c'est la stratégie de la somme du logarithme de la vraisemblance des *pseudosyllabes* qui a donné le meilleur taux de classification et ses résultats qui seront présentés ci-après.

## 5.6 Protocole d'expérimentation

Nous avons utilisé le protocole LUSO, méthode de validation croisée par locuteur, décrite dans le paragraphe 4.4, comme protocole d'expérimentation du système *SPPLG*.

## 5.7 Critères d'évaluation du système de RAE

Les critères *exactitude* et la *F-mesure* décrits à la section 4.5 sont utilisés comme mesure d'évaluation du système *SPPLG*.

## 5.8 Expérimentation et résultats

Dans cette deuxième partie de notre projet, nous avons implémenté le système *SPPLG*, où nous avons introduit deux nouveautés : l'utilisation de la *pseudosyllabe* comme unité d'analyse (versus *unité*) et les coefficients de polynôme de Legendre (CPL) (versus information à court terme) comme vecteurs de traits. Afin de mesurer l'impact de chacun de ces deux changements sur les performances du système de RAE, nous avons comparé le système *SPPLG* avec différents systèmes en introduisant pour chacun d'eux une seule variable à la fois. Les différents cas possibles de systèmes que nous pouvons obtenir sont donnés par le **Tableau 5.2** et décrits dans ce qui suit :

1. le système *+PS/+PL* : désigne le système *SPPLG*;
2. le système *+PS/-PL* : désigne un système basé sur la *pseudosyllabe* (PS) comme unité d'analyse et sur l'information prosodique à l'échelle de trame (information à court terme) comme traits caractéristiques.
3. le système *-PS/+PL* : désigne un système, basé sur l'*énoncé* comme unité d'analyse et sur les coefficients de polynôme de Legendre (CPL) comme traits caractéristiques. Notons que ce système ne sera pas implémenté, car l'approximation par des CPL doit se faire pour un segment de F0 qui s'étend sur tout l'énoncé, alors que ce dernier est composé, en plus des régions voisées, par des régions non voisées, c'est-à-dire des régions où les valeurs de F0 sont indéfinies.

4. le système *-PS/-PL* : désigne un système basé sur l'*énoncé* comme unité d'analyse et sur l'information à court terme de la prosodie comme traits caractéristiques.

Tableau 5.2

Types de systèmes de RAE en fonction de l'unité d'analyse et le type de traits caractéristiques

| | | **Type d'unité** | |
|---|---|---|---|
| | | **Pseudosyllabe** | **Énoncé** |
| **Type de traits** | **CPL de la prosodie** | +PS/+PL | -PS/+PL |
| | **ICT de la prosodie** | +PS/-PL | -PS/-PL |
| Le signe (+) signifie utilisé et le (-) signifie non utilisé. | | | |

Dans les prochaines sections, nous présenterons d'abord les résultats des expériences réalisées pour déterminer le vecteur caractéristique du système *SPPLG* ainsi que l'évaluation de ces attributs avec RELEF-F, suivie par une comparaison entre les différents systèmes.

### 5.8.1 Approximation du contour de l'énergie

Pour la courbe de l'énergie, nous avons modélisé chacun des segments des *pseudosyllabes*, dont la forme est une courbure, par les trois premiers CPL, c.-à-d. une fonction $f$ d'ordre $M$=2. Nous avons utilisé cette valeur, après avoir expérimenté des approximations avec des fonctions d'ordre M=1, 2, 3, 4, 5 et c'est avec la fonction d'ordre deux que nous avons obtenu le meilleur taux de classification.

L'évaluation de l'effet de chacune des deux variables (unité PS et les traits de type CPL) se fera à travers une comparaison de performance entre systèmes telle qu'illustrée par la Figure 5.7.

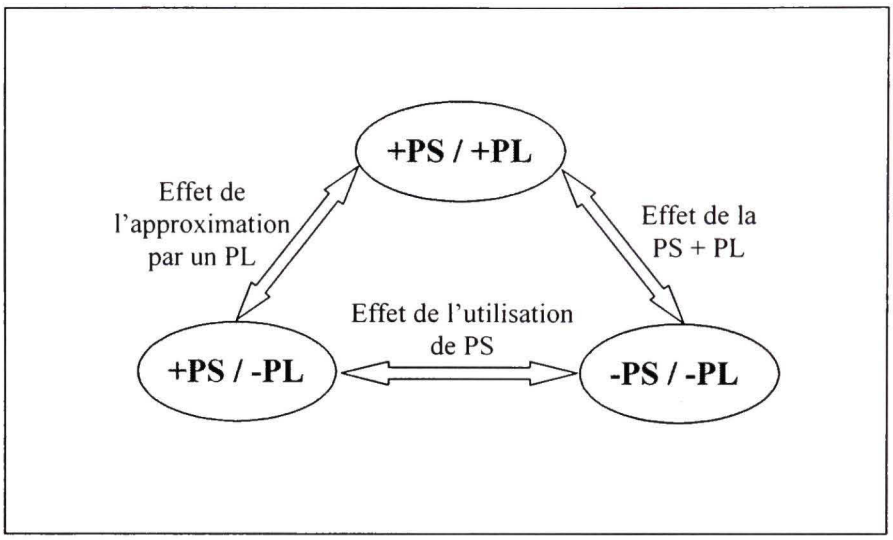


**Figure 5.7 Schéma de comparaison entre systèmes (illustrées par double flèches) pour l'évaluation de la pseudo-syllabe et les CPL.**

### 5.8.2 Approximation du contour de F0

Afin de déterminer le nombre de coefficients des polynômes de Legendre qui permet de mieux caractériser le contour des segments des *pseudosyllabes* de F0, nous avons réalisé plusieurs expériences. Dans chacune de ces expériences, nous avons utilisé les CPL d'un ordre $M$ allant de 1 à 8. Les CPL de F0 sont utilisés en combinaisons avec les trois CPL du contour de l'énergie.

Tableau 5.3

Résultats de *EXP I* (*neutre* vs *colère*) de *SPPLG*

| **Traits** | **Neutre vs Colère** |
|---|---|
| F0 (3) + E (3) | 85.60 ± 11.46 / 0.88 |
| F0 (3) + E (3) + T | **90.40 ± 9.16 / 0.92** |
| F0 (3) + E (3) + Vacil. | 61.98 ± 11.06 / 0.85 |
| F0 (3) + E (3) + Tremb. | 81.97 ± 11.38 / 0.83 |
| F0 (3) + E (3) + T + Tremb. + Vacil. | 82.40 ± 10.85 / 0.84 |

Les résultats obtenus montrent que l'utilisation des trois premiers CPL offre une meilleure modélisation pour la classification des émotions *neutre* vs *colère*, alors que pour les expériences *EXP II* (neutre vs tristesse) et *EXP III* (15 classes d'émotions), c'est avec les quatre premiers CPL que les meilleurs taux de classification sont atteints.

Nous présentons dans les tableaux 5.3 à 5.5 les résultats les plus importants obtenus pour chaque groupe d'expériences en fonction des traits utilisés. Dans ces tableaux, la lettre E désigne l'énergie, F0 désigne la fréquence fondamentale et le nombre entre parenthèses qui suit F0 et E représente le nombre de CPL utilisés (il est égal à l'ordre du PL plus 1), *T*, *Tremb* et *Vacil* désignent respectivement la durée de la *pseudosyllabe*, le tremblotement et le vacillement.

Tableau 5.4

Résultats de *EXP II* (*neutre* vs *tristesse)* de *SPPLG*

| **Traits** | **Neutre vs Tristesse** |
|---|---|
| F0 (4) + E (3) | **69.60 ± 13.81 / 0.71** |
| F0 (4) + E (3) + T | 68.80 ± 13.60 / 0.71 |
| F0 (4) + E (3) + Vacil. | 56.80 ±16.72 / 0.60 |
| F0 (4) + E (3) + Tremb. | 56.80 ±15.97 / 0.58 |
| F0 (4) + E (3) + T + Tremb. + Vacil. | 63.20 ± 18.81 / 0.67 |

### 5.8.3 Pertinence du trait durée de la pseudosyllabe

D'après les résultats montrés au **Tableau 5.3**, nous constatons que la durée de la *pseudosyllab*e est un trait important dans la discrimination de la classe *colère* versus *neutre* et que le taux de classification passe de 85.6 % à 90.4 % avec l'ajout de ce trait, contrairement aux expériences *EXP II* et *EXP III*, où nous enregistrons une légère baisse de performance. Afin d'expliquer la raison de cette différence dans l'importance de la durée de la

*pseudosyllabe*, nous avons mesuré la durée moyenne de la *pseudosyllabe* pour chaque catégorie d'émotion. Nous constatons d'après les valeurs du **Tableau 5.7**, que la classe colère possède la plus petite valeur de durée et la classe neutre la plus grande valeur, ce qui représente un bon critère de discrimination entre les deux classes d'émotions.

Tableau 5.5

Résultats de *EXP III* (15 classes d'émotions) de *SPPLG*

| **Traits** | **15 classes** |
|---|---|
| F0 (4) + E (3) | **13.62 ± 2.49** |
| F0 (4) + E (3) + T | 13.45 ± 4.64 |
| F0 (4) + E (3) + Vacil. | 11.53 ± 3.24 |
| F0 (4) + E (3) + Tremb. | 9.69 ± 2.39 |
| F0 (4) + E (3) + T + Tremb. + Vacil. | 11.62 ± 2.28 |

Tableau 5.6

Récapitulatif des résultats du système *SPPLG*
pour les trois groupes expériences

| **Expériences** | **Taux / f-score** | **Traits** |
|---|---|---|
| Neutre vs Colère | 90.40 ± 9.16 / 0.92 | F0(3) + E(3) + T |
| Neutre vs Tristesse | 69.60 ± 13.81 / 0.71 | F0(4) + E(3) |
| 15 classes | 13.62 ± 2.49 | F0(4) + E(3) |

Tableau 5.7

Durée moyenne de la pseudo-syllabe par classe d'émotion

| **Émotions** | **Durée (ms)** |
|---|---|
| Anxiété | 90 |
| Ennui | 99 |
| Mépris | 94 |
| Col. froide | 103 |
| Désespoir | 91 |
| Dégoût | 98 |
| Exaltation | 88 |
| Joie | 90 |
| Col. Forte | 81 |
| Intérêt | 90 |
| Neutre | 100 |
| Panique | 84 |
| Fierté | 90 |
| Tristesse | 87 |
| Honte | 92 |
| **Moyenne** | 92 |
| **Écart type** | 6.1 |

- **Matrice de confusion**

Le **Tableau 5.8** représente la matrice de confusion de la classification des 15 classes d'émotions du système *SPPLG*. Cette matrice met clairement en évidence une confusion qui caractérise un certain nombre de classes d'émotions, particulièrement les classes : ennui avec anxiété, dégoût avec colère froide, exaltation avec colère forte, panique avec exaltation. Nous constatons également, à l'instar de la matrice de confusion du système SMEG, qu'il existe des classes qui sont relativement bien reconnues telles que les classes neutre, intérêt, colère et exaltation et d'autres très mal reconnues comme la classe désespoir, fierté et honte.

### 5.8.4 Comparaison des matrices de confusion du système *SMEG* et *SPPLG*

Une comparaison entre les matrices de confusion des deux systèmes *SMEG* et *SPPLG*, permet de relever deux remarques intéressantes :

1. Les confusions qui existent entre les paires de classes d'émotions ne sont pas identiques pour les deux systèmes.
2. Chacun des deux systèmes est plus habile à reconnaître certaines classes d'émotions que d'autres. Le point intéressant est que les classes des émotions que le premier système reconnaît faiblement, le deuxième système le reconnaît avec plus de précision et vice-versa. Ceci est vrai par exemple pour les classes d'émotions désespoir, exaltation, colère forte, intérêt, panique, fierté et honte.

Par conséquent, nous pouvons déduire que chacun des deux systèmes joue un rôle complémentaire par rapport à l'autre, et un système de RAE basé sur la combinaison des deux classificateurs (systèmes) devrait permettre d'améliorer sensiblement les taux de classification.

### 5.8.5 Évaluation des attributs avec la méthode *RELIEF-F*

Dans nos expériences précédentes, nous avons constitué un ensemble de traits caractéristiques, composé des trois premiers CPL de F0 pour *EXP I* et les quatre premiers CPL de F0 pour *EXP II* et *III*, les trois premiers CPL de l'énergie, la durée de la *pseudosyllabe* le tremblotement et le vacillement. Nous avons appliqué par la suite la méthode de sélection des caractéristiques *RELIEF-F* pour évaluer les attributs de cet ensemble en vue de sélectionner le sous-ensemble de traits optimal permettant d'améliorer les résultats de la classification. Étant donné que la méthode *RELIEF-F* dépend des données utilisées, nous avons procédé à l'évaluation des attributs séparément pour chaque type d'expérience. Nous avons donc utilisé dans la méthode *RELIERF-F* les données de la classe *neutre* et *colère* pour *EXP I*, *neutre* et *tristesse* pour *EXP II* et les données des quinze classes pour *EXP III*. Les résultats de l'évaluation des attributs sont donnés dans le **Tableau 5.9**.

D'après ces résultats, nous constatons que l'ordre d'importance des traits diffère en fonction du type d'expérience, c'est-à-dire selon les classes de données utilisées.

Tableau 5.8

Matrice de confusion pour les 15 classes d'émotions du système *SPPLG*

| | **Anx** | **Enn** | **Fro** | **Mép** | **Dés** | **Dég** | **Exa** | **Joi** | **For** | **Int** | **Neu** | **Pan** | **Fier** | **Tri** | **Hon** | **Moy.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Anxiété** | 12 | 7 | 2 | 7 | 2 | 2 | 4 | 2 | 1 | 5 | 14 | 6 | 5 | 9 | 4 | 14.63% |
| **Ennui** | 14 | 10 | 5 | 5 | 3 | 3 | 5 | 2 | 1 | 1 | 14 | 6 | 2 | 5 | 7 | 12.05% |
| **Col. froide** | 4 | 1 | 7 | 8 | 6 | 6 | 7 | 5 | 8 | 4 | 4 | 4 | 5 | 3 | 2 | 9.46% |
| **Mépris** | 8 | 4 | 2 | 10 | 3 | 5 | 2 | 8 | 3 | 6 | 11 | 5 | 2 | 6 | 8 | 12.05% |
| **Désespoir** | 7 | 6 | 5 | 4 | 2 | 7 | 7 | 5 | 6 | 3 | 6 | 5 | 2 | 8 | 4 | 2.60% |
| **Dégoût** | 6 | 4 | 11 | 6 | 3 | 8 | 4 | 3 | 9 | 2 | 6 | 5 | 2 | 9 | 5 | 9.64% |
| **Exaltation** | 3 | 6 | 3 | 0 | 0 | 4 | 11 | 5 | 15 | 7 | 2 | 13 | 3 | 1 | 0 | 15.07% |
| **Joie** | 8 | 6 | 3 | 8 | 6 | 5 | 5 | 5 | 3 | 6 | 2 | 2 | 10 | 8 | 7 | 5.95% |
| **Col. Forte** | 3 | 2 | 6 | 0 | 0 | 7 | 15 | 1 | 13 | 1 | 1 | 13 | 2 | 3 | 6 | 17.81% |
| **Intérêt** | 9 | 5 | 3 | 3 | 3 | 1 | 2 | 10 | 4 | 21 | 4 | 6 | 5 | 3 | 4 | 25.30% |
| **Neutre** | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 0 | 0 | 3 | 30 | 3 | 2 | 2 | 4 | 57.69% |
| **Panique** | 4 | 5 | 8 | 5 | 3 | 4 | 15 | 5 | 11 | 1 | 2 | 12 | 1 | 2 | 1 | 15.19% |
| **Fierté** | 9 | 6 | 1 | 12 | 4 | 5 | 4 | 7 | 3 | 5 | 5 | 3 | 3 | 5 | 4 | 3.95% |
| **Tristesse** | 12 | 5 | 2 | 5 | 4 | 1 | 1 | 3 | 1 | 4 | 14 | 3 | 7 | 7 | 4 | 9.59% |
| **Honte** | 14 | 6 | 1 | 2 | 6 | 2 | 3 | 0 | 2 | 5 | 10 | 3 | 5 | 6 | 5 | 7.14% |

Note : Les titres de colonnes représentent les trois premières lettres des noms des classes d'émotions affichés également, dans le même ordre, comme titres de lignes.

Tableau 5.9

Résultats de l'ordonnancement des traits avec *RELIEF-F* en fonction du type d'expérience

| 15 émotions | Neutre & colère | Neutre & tristesse |
|---|---|---|
| F0:C3 | F0:C3 | Tremblotement |
| Tremblotement | E:C0 | E:C1 |
| T | F0:C2 | T |
| F0:C1 | E:C2 | F0:C2 |
| E:C1 | Vacillement | E:C2 |
| E:C2 | E:C1 | F0:C0 |
| F0:C2 | F0:C0 | F0:C3 |
| E:C0 | F0:C1 | F0:C1 |
| F0:C0 | T | E:C0 |
| Vacillement | Tremblotement | Vacillement |
| E:Ci et F0:Ci désignent le $i^{ième}$ CPL pour la courbe de l'énergie et de F0 respectivement. | | |

Nous avons utilisé le classement des attributs dans nos expériences pour la classification des émotions, en éliminant à chaque itération d'expérience le trait le moins pertinent. Les taux de reconnaissance obtenus en utilisant le classement de *RELIEF-F* sont plus faibles que ceux affichés dans le Tableau 5.6. En conclusion, l'utilisation de la méthode *RELIEF-F* ne nous a pas permis de retrouver le sous-ensemble optimal de traits pour notre système et que l'utilisation d'une autre méthode de sélection de traits telle *Forward Selection* dans nos travaux futurs serait peut-être plus adaptée pour notre corpus de données.

### 5.8.6 Évaluation des traits en fonction du locuteur

Dans cette section, nous présentons les résultats du classement des traits caractéristiques selon leurs capacités de discrimination entre les classes d'émotions, en fonction du locuteur. Nous avons évalué les différents traits caractéristiques en utilisant l'algorithme *RELIEF-F*.

Tableau 5.10

Résultats de l'ordonnancement des traits par locuteur avec RELIEF-F en utilisant les données des 15 classes d'émotions

| | **Locuteurs** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **CC** | **CL** | **GG** | **JG** | **MF** | **MK** | **MM** | **Tous** |
| **Traits ordonnés** | F0:C0 | F0:C3 | E:C0 | F0:C3 | F0:C0 | F0:C2 | E:C1 | F0:C3 |
| | E:C0 | E:C0 | E:C1 | E:C2 | T | Tremb. | E:C0 | Tremb. |
| | F0:C1 | F0:C2 | T | T | E:C0 | E:C2 | F0:C2 | T |
| | Vacil. | E:C2 | F0:C0 | Vacil. | F0:C2 | F0:C3 | F0:C0 | F0:C1 |
| | F0:C2 | F0:C1 | F0:C3 | F0:C2 | E:C2 | F0:C1 | F0:C1 | E:C1 |
| | F0:C3 | F0:C0 | F0:C1 | F0:C1 | Tremb. | F0:C0 | Vacil. | E:C2 |
| | T | Vacil. | F0:C2 | E:C1 | F0:C1 | E:C0 | T | F0:C2 |
| | E:C1 | Tremb. | Tremb. | F0:C0 | F0:C3 | E:C1 | F0:C3 | E:C0 |
| | Tremb. | T | E:C2 | E:C0 | E:C1 | T | E:C2 | F0:C0 |
| | E:C2 | E:C1 | Vacil. | Tremb. | Vacil. | Vacil. | Tremb. | Vacil. |

E:Ci et F0:Ci désignent le $i^{ième}$ CPL pour la courbe de l'énergie et de F0 respectivement.

Une remarque importante que nous relevons des résultats présentés dans le **Tableau 5.10** est la grande dissimilitude qui existe entre les différents locuteurs dans l'ordonnancement de leurs traits caractéristiques. À titre d'exemple, la durée de la *pseudosyllabe* qui joue un rôle important (classée en deuxième position) dans la discrimination des classes pour le locuteur *MF* n'est que d'une faible influence (classée en avant-dernière position) pour les locuteurs *CL* et *MK*. Ceci s'explique par le fait que chacun des sept locuteurs met l'accent sur la fréquence fondamentale, l'intensité et le rythme de la parole dans un ordre et des proportions différentes par rapport aux autres locuteurs pour distinguer les catégories des émotions au cours de leurs simulations. Ceci illustre l'un des aspects des incertitudes liées à l'universalité du mode d'expression des émotions sur le plan acoustique.

Une conséquence directe de différence entre l'ordre d'importance des traits des locuteurs, se manifeste au cours de l'application des méthodes de sélection des traits caractéristiques.

Ainsi à chaque élimination d'un trait ou d'un sous-ensemble de traits, nous enregistrons un gain en performance pour certains locuteurs et une baisse pour d'autres. Ceci représente un des facteurs qui pourrait expliquer les résultats mitigés obtenus par l'application de la méthode de sélection de traits *RELIEF-F*.

### 5.8.7 Effet de l'utilisation de l'unité pseudosyllabe versus énoncé

Notre objectif dans cette section est de mesurer l'impact que pourrait avoir l'utilisation de l'unité *pseudosyllabe* versus *énoncé* sur les performances d'un système de RAE. C'est pourquoi nous avons développé deux autres systèmes, *-PS/-PL* et *+PS/-PL*, que nous avons décrits dans la section 5.8. Nous rappelons que ces deux systèmes utilisent les valeurs de F0 et de l'énergie à l'échelle de trame comme traits caractéristiques à court terme au niveau de l'*énoncé* pour le système *-PS/-PL* et de la *pseudosyllabe* pour le système *+PS/-PL*

Tableau 5.11

Comparaison de performance entre systèmes utilisant l'unité *pseudosyllabe* versus *énoncé*

| **Système** | **Neutre vs Colère** | **Neutre vs Tristesse** | **15 émotions** |
|---|---|---|---|
| *+PS/-PL* | 76.00 ± 19.19 / 0.81 | **38.89 ± 20.18 / 0.48** | **10.56 ± 3.60** |
| *-PS/-PL* | **80.80 ± 16.18 / 0.85** | 36.51 ± 19.45 / 0.46 | 9.38 ± 2.81 |
| t (CVP t-Test) | **R** (3.1744) | **A** (1.3849) | **R** (2,7713) |
| À la 4 ième ligne, R et A désignent respectivement l'acceptation ou le rejet de l'hypothèse nulle du K-Fold CrossValidation Paired t Test et le nombre entre parenthèse représente la valeur calculée *t*. | | | |

D'après les résultats du **Tableau 5.11**, nous constatons que l'utilisation de la *pseudosyllabe* comme unité d'analyse a permis d'améliorer les taux de classification avec un gain relatif de 7% pour l'expérience *EXP II* et de 13% pour *EXP III*. Par contre, nous enregistrons gain relatif de l'ordre de 6% au profit de l'unité *énoncé* pour l'expérience *EXP I*. D'après les résultats du CVP t-Test, nous constatons que la différence de performances entre les deux

types de systèmes *+PS/-PL* et *-PS/-PL* est significative pour les expériences *EXP I* et *EXP III* et non significatives pour le *EXP II*.

### 5.8.8 Effet de l'approximation par un polynôme de Legendre

Notre objectif dans cette section est de mesurer l'effet de l'approximation du contour de l'énergie et celui de la fréquence fondamentale par des polynômes de Legendre et l'utilisation de leurs coefficients comme traits caractéristiques sur les performances du système de reconnaissance des émotions. À cette fin, nous avons comparé les performances du système +PS/+PL (*SPPLG*), qui utilise l'unité *pseudosyllabe* et les CPL, avec le système +PS/-PL basé également sur l'unité *pseudosyllabe* mais qui utilise l'information à court terme de la prosodie comme traits.

Tableau 5.12

Comparaison de performances entre les deux systèmes +PS/+PL et +PS/-PL pour la reconnaissance de l'émotion neutre vs colère

| **Traits** | **F0 + E** | **F0 + E + T** |
|---|---|---|
| +PS/+PL | **85.60 ± 11.46 / 0.88** | **90.40 ± 9.16 / 0.92** |
| +PS/-PL | 76.00 ± 19.19 / 0.81 | 78.40 ± 19.67 / 0.83 |
| t (CVP t-Test) | **R** (4.8862) | **R** (4.6650) |
| À la 4 ième ligne, R et A désignent respectivement l'acceptation ou le rejet de l'hypothèse nulle du K-Fold CrossValidation Paired t Test et le nombre entre parenthèse représente la valeur calculée *t*. | | |

Les résultats des **Tableau 5.12** à **Tableau 5.14** montrent que l'utilisation des coefficients du PL comme traits caractéristiques à long terme apporte une amélioration significativement importante aux performances du système de RAE. Cette amélioration est de l'ordre de **15%**

pour l'expérience *EXP I*, **69%** pour *EXP II* et **29%** pour *EXP III* quand F0, l'énergie et la durée de la *pseudosyllabe* sont utilisées.

Tableau 5.13

Comparaison de performances entre les deux systèmes +PS/+PL et +PS/-PL pour la reconnaissance de l'émotion neutre vs tristesse

| **Traits** | **F0 + E** | **F0 + E + T** |
|---|---|---|
| +PS/+PL | **69.60 ± 13.81 / 0.71** | **68.8 ± 13.60 / 0.71** |
| +PS/-PL | 38.89 ± 20.18 / 0.48 | 41.27 ± 13.19 / 0.53 |
| t (CVP t-Test) | **R** (4.1155) | **R** (4,6650) |
| À la 4 ième ligne, R et A désignent respectivement l'acceptation ou le rejet de l'hypothèse nulle du K-Fold CrossValidation Paired t Test et le nombre entre parenthèse représente la valeur calculée *t*. | | |

Tableau 5.14

Comparaison de performances entre les deux systèmes +PS/+PL et +PS/-PL pour la reconnaissance des 15 classes d'émotions

| **Traits** | **F0 + E** | **F0 + E + T** |
|---|---|---|
| +PS/+PL | **13.62 ± 2.49** | **13.45 ± 4.64** |
| +PS/-PL | 10.56 ± 3.60 | 9.16 ± 5.09 |
| t (CVP t-Test) | **R** (5.4780) | **R** (3.4559) |
| À la 4 ième ligne, R et A désignent respectivement l'acceptation ou le rejet de l'hypothèse nulle du K-Fold CrossValidation Paired t Test et le nombre entre parenthèse représente la valeur calculée *t*. | | |

### 5.8.9 Effet de l'utilisation combinée de l'unité PS et de l'approximation par un PL

Finalement, nous avons comparé les performances du système *-PS/-PL* avec celle du système *+PS/+PL* afin de mesurer l'impact de l'utilisation combinée de la *pseudosyllabe* comme unité d'analyse et des CPL comme traits. Le **Tableau 5.15** montre les résultats de ces expériences, où seules F0 et l'énergie sont utilisés comme traits caractéristiques.

Tableau 5.15

Comparaison de résultats entre les systèmes *-PS/-PL et +PS/+PL* utilisant F0 et l'énergie comme traits

| **Système** | **Neutre vs. Colère** | **Neutre vs. Tristesse** | **15 classes d'émotions** |
|---|---|---|---|
| *+PS/+PL* | **85.60 ± 11.46 / 0.88** | **69.60 ± 13.81 / 0.71** | **13.62 ± 2.49** |
| *-PS/-PL* | 80.80 ± 16.18 / 0.85 | 36.51 ± 19.45 / 0.46 | 9.38 ± 2.81 |
| t (CVP t-Test) | **R** (4.5365) | **R** (4.0839) | **R** (4.0176) |
| À la 4 ième ligne, R et A désignent respectivement l'acceptation ou le rejet de l'hypothèse nulle du K-Fold CrossValidation Paired t Test et le nombre entre parenthèse représente la valeur calculée *t*. | | | |

D'après ces résultats, nous constatons que le système proposé *+PS/+PL* basé sur la *pseudosyllabe* et les coefficients du polynôme de Legendre apporte un gain relatif important en performance, qui est de l'ordre de **91%** dans l'expérience *EXP II*, **41%** dans *EXP III* et **6%** dans l'expérience *EXP I*.

Nous pouvons donc conclure que l'unité *pseudosyllabe* combinée avec une approximation par un polynôme de Legendre permet d'obtenir un système de reconnaissance de l'émotion beaucoup plus performant en comparaison avec un système utilisant l'unité *énoncé* et l'information à court terme lorsque ces systèmes sont entrainés et testés avec le corpus LDC.

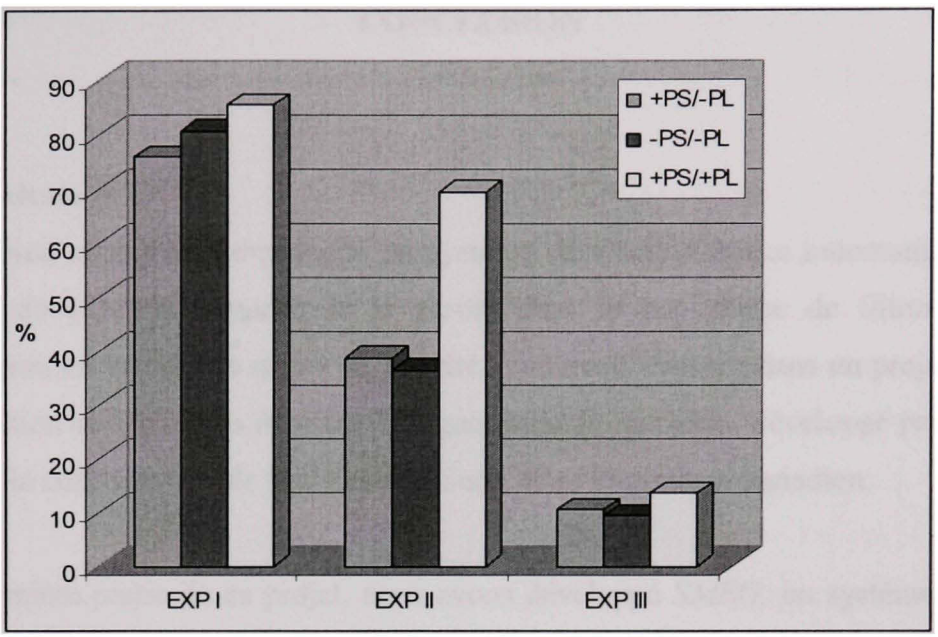


**Figure 5.8 Comparaison de performance entre les systèmes +PS/+PL, +PS/-PL et –PS/-PL pour les trois types d'expériences.**

## 5.9 Conclusion

Dans ce chapitre, nous avons présenté la méthodologie suivie pour l'implémentation du système de RAE appelé *SPPLG*. Dans ce système, nous avons expérimenté une nouvelle unité de modélisation pour les systèmes de RAE, appelée *pseudosyllabe*, obtenue par la segmentation des contours de l'énergie et de F0, suivant les points minimums du contour de l'énergie. Les différents segments sont approximés par des sommes de polynômes de Legendre, et ses coefficients sont utilisés comme traits caractéristiques. Nous avons montré, qu'un tel système appliqué aux données LDC performe mieux qu'un système basé sur l'énoncé comme unité de modélisation et l'information à court terme comme traits caractéristiques. Il est possible d'apporter d'autres améliorations aux performances de ce système à travers l'enrichissement du vecteur de traits par d'autres informations et le choix d'un autre type de classificateurs que nous détaillerons un peu plus dans la section travaux futurs du chapitre suivant.

# CONCLUSION

## 6.1 Conclusions

Dans ce projet, nous avons développé un système de reconnaissance automatique de l'état émotionnel du locuteur à partir de la parole dans le but ultime de filtrer les appels problématiques en vue de les traiter en priorité. Ce travail s'inscrit dans un projet plus vaste appelé «Gestion des émotions dans les dialogues humain-machine» développé par l'ÉTS et le CRIM avec la collaboration de Bell Canada Corp. et de Patrimoine canadien.

Dans la première partie de ce projet, nous avons développé *SMEG,* un système bâti autour des traits de type MFCC extraits au niveau d'une trame, analysés à l'échelle d'un *énoncé* et modélisé par un mélange de gaussiennes. Notre objectif était de reproduire avec ce système, les meilleurs résultats de l'état de l'art et éventuellement améliorer les performances. Ceci a été entrepris à travers l'optimisation de certains paramètres de ce système, sur le plan de l'extraction des caractéristiques, celui de la représentativité des données et celui du paramétrage du classificateur. Nous avons amélioré les résultats de l'état de l'art, lorsque les systèmes sont entrainés et testés avec le corpus LDC, avec un gain relatif de 11% pour la reconnaissance de l'émotion *neutre vs tristesse* alors que nous avons reproduit les meilleures performances pour l'expérience *neutre* vs *colère* et pour l'expérience avec 15 classes d'émotions.

Les résultats des expériences réalisées ont montré qu'il pouvait exister dans les coefficients cepstraux supérieurs – au-delà des treize premiers, une information utile pour la RAE. Nous concluons aussi que l'utilisation d'un modèle du monde (UBM) permet d'améliorer l'apprentissage des modèles de certaines classes d'émotions quand les données d'apprentissage ne sont pas suffisantes.

Par ailleurs, nous avons étudié et testé, pour la première fois dans le domaine de la RAE, un système, nommé *SPPLG,* basé sur la *pseudosyllabe* comme unité d'analyse et les coefficients

du *polynôme de Legendre* (CPL), résultants de l'approximation des courbes de l'énergie et de la fréquence fondamentale, comme traits caractéristiques à long terme. Les résultats obtenus de la comparaison entre ce système et un système utilisant l'information prosodique à l'échelle de trame comme traits à court terme, analysé au niveau de l'*énoncé* sont très prometteurs. Les taux de reconnaissance obtenus avec un système basé sur la *pseudosyllabe* et les coefficients de *polynômes de Legendre* sont nettement supérieurs à ceux d'un système basé sur l'*énoncé* et l'information à court terme. Le gain relatif réalisé est de l'ordre de 6% pour la reconnaissance des émotions *neutre* vs *colère*, tandis que ce gain est de l'ordre 91% pour *neutre* vs *tristesse*. Enfin, nous avons obtenu une amélioration de l'ordre de 41% dans la détection de 15 classes d'émotions. Ces résultats sont obtenus avec les données LDC, un corpus d'émotions de type simulées. Il serait opportun de consolider ces conclusions en testant ces systèmes avec un corpus d'émotions vécues, issu d'un centre d'appels ou des extraits de bandes sonores par exemple.

## 6.2 Travaux futurs

Le système de RAE que nous avons développé est destiné à une utilisation dans le cadre d'un centre d'appels, où des centaines de milliers d'appels anglophones et francophones sont reçus par jour au Canada. De ce fait, tout gain en performance du système aura un impact considérable sur la qualité du service à la clientèle du centre d'appels. C'est pourquoi nous envisageons de réaliser, comme travaux futurs, ce qui suit :

1. Évaluer les performances des deux systèmes *SPPLG* et *SMEG* sur un corpus de données représentants des émotions réelles (non simulées) issues des conversations des clients d'un centre d'appels automatisé.
2. Optimiser la précision du système *SPPLG*, que nous avons implémenté dans ce projet dans le but d'évaluer l'efficacité de l'utilisation des coefficients de polynôme de Legendre extraits sur l'échelle d'une *pseudosyllabe* sans souci d'optimisation. À l'issue des résultats prometteurs obtenus, nous pensons améliorer les performances du système à travers :

a. L'utilisation de la méthode MAP et l'expérimentation de différents nombres de mélanges de gaussiennes à l'instar de la procédure suivie avec le système *SMEG.*
b. Utiliser, en plus de fréquence fondamentale et de l'énergie, les valeurs des premiers et seconds formants ainsi que les valeurs de leurs bandes passantes comme traits. Nous pouvons également envisager la combinaison de l'information prosodique avec l'information spectrale.
c. Combiner les notions de *pseudosyllabe* et de l'approximation avec des CPL du système *SPPLG* avec l'utilisation des MFCC utilisées dans le système *SMEG.*
d. Récupérer l'information temporale des traits prosodiques pour capter les caractéristiques d'une classe d'émotion spécifique. Ceci peut se réaliser à travers une modélisation du système *SPPLG* Avec les HMM.

3. Construire un système de RAE, basé sur la combinaison des classificateurs des deux systèmes *SPPLG* et *SMEG.* Tel que nous l'avons mentionné au chapitre précédent, ce choix est essentiellement motivé par le fait que les deux systèmes possèdent des matrices de confusion différentes, et qui se complémentent par rapport aux classes d'émotions qu'ils reconnaissent le mieux.

# ANNEXE I

## PARAMETRES D'EXTRACTION DES MFCC

Le code suivant représente le fichier de configuration des paramètres que nous avons utilisé pour l'extraction des coefficients MFCC avec l'outil *HTK*.

```
# Coding parameters
BYTEORDER     = VAX
SOURCEKIND = WAVEFORM
SOURCEFORMAT = NOHEAD
SOURCERATE   = 453.51473922902494331776956926205457421 02
TARGETKIND = MFCC_0_D_A                    # Identifier of the coefficients to use
TARGETFORMAT   = HTK
# Unit = 0.1 micro-second
WINDOWSIZE = 250000.0                      # = 25 ms = length of a time frame
TARGETRATE = 100000.0                      # = 10 ms = frame periodicity
NUMCEPS = 12                               # Number of MFCC coeffs (here from c1 to c12)
USEHAMMING = T                             # Use of Hamming function for windowing frames
PREEMCOEF = 0.97                           # Pre-emphasis coefficient
NUMCHANS = 26                              # Number of filterbank channels
CEPLIFTER = 22                             # Length of cepstral liftering
SAVEWITHCRC = FALSE
ENORMALISE = T
ZMEANSOURCE    = TRUE

# The End
```

# ANNEXE II

## Résultats détaillés des expériences du système SMEG

Dans ce qui suit, nous présentons les résultats des expériences de RAE du système SMEG, dont les résultats ont été présentés sous forme de graphiques dans le chapitre CHAPITRE 4.

### Tableau II.1

Résultats de la reconnaissance automatique des émotions *neutre* et *colère* en fonction du nombre de gaussiennes, du vecteur de traits et le type de classificateur

| | | Accuracy ± écart-type / F-score$_{colère}$ | | | |
|---|---|---|---|---|---|
| | | **GMM** | | **GMM-UBM** | |
| | | **39 traits** | **60 traits** | **39 traits** | **60 traits** |
| **Nb. Gaussiennes** | **256** | 95,31±5.95 / 0.96 | 85,94±12.4 / 0.89 | 78,91±13.16 / 0.8 | 72,66±19.34 / 0.74 |
| | **128** | 96,09±2.67 / 0.97 | 89,84±11.1 / 92 | 91,41±15.2 / 0.93 | 75,78±16.4 / 0.77 |
| | **64** | 96,88±2.89 / 0.97 | 95,31±5.95 / 0.96 | 92,97±12.92 / 0.94 | 82,81±19.73 / 0.84 |
| | **32** | 96,09±6.49 / 0.97 | 96,88±2.89 / 0.97 | **93,75±10.14 / 0.95** | 89,84±16.79 / 0.91 |
| | **16** | **96,88±4.52 / 0.97** | **97,66±2.92 / 0.98** | 90,62±10.21 / 0.92 | 92,19±12.57 / 0.93 |
| | **8** | 92,97±11.08 / 0.94 | 96,09±4.99 / 0.97 | 85,94±16.04 / 0.88 | 92,19±9.39 / 0.94 |
| | **4** | 90,62±11.16 / 0.92 | 95,31±6.75 / 0.96 | 85,94±11.67 / 0.87 | **92,97±7.48 / 0.94** |
| | **2** | 87,50±13.99 / 0.88 | 89,84±10.76 / 0.91 | 84,38±15.3 / 0.86 | 82,03±15.28 / 0.84 |

**Tableau II.2**

Résultats de la reconnaissance automatique des émotions *neutre* et *tristesse* en fonction du nombre de gaussiennes, du vecteur de traits et le type de classificateur

| | | Accuracy ± écart-type / F-score$_{tristesse}$ | | | |
|---|---|---|---|---|---|
| | | **GMM** | | **GMM-UBM** | |
| | | **39 traits** | **60 traits** | **39 traits** | **60 traits** |
| **Nb. Gaussiennes** | **256** | 64,57±13.71 / 0.76 | 58,27±9.94 / 0.73 | 67,72±13.37 / 0.67 | 64,57±12.26 / 0.66 |
| | **128** | 59,06±15.56 / 0.71 | 58,27±11.3 / 0.72 | 71,65±9.34 / 0.74 | 72,44±13.57 / 0.74 |
| | **64** | 64,57±15.5 / 0.75 | 59,06±17.3 / 0.72 | 71,65±12.69 / 0.74 | **76,38±18.02 / 0.8** |
| | **32** | 66,93±21.75 / 0.76 | 59,84±17.25 / 0.73 | **73,23±17.03 / 0.77** | 70,87±18.27 / 0.73 |
| | **16** | 66,14±16.53 / 0.76 | 63,78±15.27 / 0.75 | 70,08±20.23 / 0.72 | 67,72±22.6 / 0.72 |
| | **8** | 69,27±14.69 / 0.77 | 62,99±19.11 / 0.73 | 69,29±22.87 / 0.73 | 68,50±24.89 / 0.73 |
| | **4** | **71,65±19.61 / 0.76** | 65,35±19.26 / 0.69 | 70,08±24.42 / 0.72 | 68,50±24.65 / 0.71 |
| | **2** | 62,20±15.87 / 0.68 | **66,93±14.58 / 0.7** | 68,50±20.91 / 0.71 | 65,35±21.81 / 0.69 |

**Tableau II.3**

Résultats de la reconnaissance automatique des 15 classes d'émotions en fonction du nombre de gaussiennes, du vecteur de traits et le type de classificateur

| | | **GMM** | | **GMM-UBM** | |
|---|---|---|---|---|---|
| | | **39 traits** | **60 traits** | **39 traits** | **60 traits** |
| **Nb. Gaussiennes** | **256** | 16,49±4.36 | 15,81±3.06 | 16,15±3.60 | 15,03±3.78 |
| | **128** | **17,18±2.18** | **16,41±4.57** | **17,87±2,94** | 16,07±3.67 |
| | **64** | 15,55±3.52 | 15,38±3.25 | 16,84±2.73 | **17,27±3.82** |
| | **32** | 16,24±4.24 | 16,07±3.35 | 17,01±3.04 | 15,64±3.28 |
| | **16** | 15,81±4.5 | 15,81±3.68 | 14,43±2.72 | 14,86±2.61 |
| | **8** | 15,29±6.04 | 15,98±3.14 | 13,49±2.50 | 14,00±3.34 |
| | **4** | 15,03±4.72 | 15,89±3.38 | 12,20±3.96 | 14,18±3.74 |
| | **2** | 15,21±5.15 | 15,21±3.65 | 11,25±4.58 | 12,29±2.95 |

## LISTE DE RÉFÉRENCES


Aiyer, Anuradha K. 2001. « Robust image compression using Gauss mixture models ». California, Stanford University, 152 pages.

Alvarez, A., I. Cearreta, J. M. Lopez, A. Arruti, E. Lazkano, B. Sierra et N. Garay. 2007. « A comparison using different speech parameters in the automatic emotion recognition using feature subset selection based on evolutionary algorithms ». In *Proceedings 10th International Conference, TSD 2007*. p. 423-430. Pilsen, Czech Republic: The Institution of Engineering and Technology.

Arnfield, Simon, Peter Roach, Jane Setter, Peter Greasley et Dave Horton. 1995. « Emotional Stress and Speech Tempo Variation ». In *Proceedings of ESCA-NATO Tutorial and Research Workshop on Speech Under Stress*. p. 13-15. Lisbon.

Banse, R., et K. R. Scherer. 1996. « Acoustic Profiles in Vocal Emotion Expression ». *Journal of Personality and Social Psychology*, p. 614–636.

Barra, R., J. Macias-Guarasa, J. M. Montero, C. Rincon, F. Fernandez et R. Cordoba. 2007. « In search of primary rubrics for language independent emotional speech identification ». In *International Symposium on Intelligent Signal Processing*. p. 923-928. Alcala de Henares, Spain: The Institution of Engineering and Technology.

Batliner, Anton, Jan Buckow, Richard Huber, Volker Warnke, Elmar Noth et Heinrich Niemann. 1999. « Prosodic Feature Evaluation: Brute Force or Well Designed ». In *Proc. 14th Int. Congress of Phonetic Sciences*. p. 2315-2318. <http://www5.informatik.uni-erlangen.de/literature/ps-dir/1999/Batliner99:PFE.ps.gz>.

Batliner, Anton, et Richard Huber. 2007. « Speaker characteristics and emotion classification ». *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4343 NAI, p. 138-151.

Baum, Leonard E., et Ted Petrie. 1966. « Statistical Inference for Probabilistic Functions of Finite State Markov Chains ». *Annals of Mathematical Statistics,* vol. 37, p. 1554-1563.

Beritelli, Francesco, Salvatore Casale, Alessandra Russo, Salvatore Serrano et Donato Ettorre. 2007. « Speech emotion recognition using MFCCs extracted from a mobile terminal based on ETSI front end ». In *International Conference on Signal Processing Proceedings, ICSP*. Vol. 2, p. 4129151. Guilin, China: Institute of Electrical and Electronics Engineers Inc., Piscataway, NJ 08855-1331, United States. <http://dx.doi.org/10.1109/ICOSP.2006.345670>.

Bhatti, M. W., Wang Yongjin et Guan Ling. 2004. « A neural network approach for human emotion recognition in speech ». In *International Symposium on Circuits and Systems (IEEE Cat. No.04CH37512)*. Vol. 2, p. 181-184. Vancouver, BC, Canada: IEEE. <http://dx.doi.org/10.1109/ISCAS.2004.1329238 >.

Boersma, Paul, et David Weenink. 2008. « Praat: doing phonetics by computer (Version 5.0.32) ». In. <http://www.praat.org/>. Consulté le 21/08/2008.

Boite, René, Hervé Bourlard, Thierry Dutoit, Joël Hancq et Henri Leich. 2000. *Traitement de la parole*. Lausanne: Presses polytechniques et universitaires Romandes.

Bronakowski, L., K. Slot, J. Cichosz et Kim Jongman. 2007. « Application of Poincare map-based description of vowel pronunciation variability for emotion assessment in speech signal ». In *International Symposium on Information Technology Convergence - ISITC '07*. p. 175-178. Joenju, South Korea: The Institution of Engineering and Technology.

Buscicchio, Cosimo A., Przemyslaw Gorecki et Laura Caponetti. 2006. « Speech emotion recognition using spiking neural networks ». In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 4203 NAI, p. 38-46. Bari, Italy: Springer Verlag, Heidelberg, D-69121, Germany.

Cao, Wenming, et Tiancheng He. 2006. « Mandarin speech emotion recognition based on high dimensional geometry theory ». *Chinese Journal of Electronics,* vol. 15, n° 4 A, p. 818-821.

Casale, Salvatore, Alessandra Russo et Salvatore Serrano. 2007. « Multistyle classification of speech under stress using feature subset selection based on genetic algorithms ». *Speech Communication,* vol. 49, n° 10-11, p. 801-810.

Chen, Chun, Mingyu You, Mingli Song, Jiajun Bu et Jia Liu. 2006. « An enhanced speech emotion recognition system based on discourse information ». In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 3991 NCS - I, p. 449-456. Reading, United Kingdom: Springer Verlag, Heidelberg, D-69121, Germany.

Chi-Yueh, Lin, et Wang Hsiao-Chuan. 2006. « Language identification using pitch contour information in the ergodic Markov model ». In *International Conference on Acoustics, Speech, and Signal Processing (IEEE Cat. No. 06CH37812C)*. p. 193-196. Toulouse, France: The Institution of Engineering and Technology.

Chuang, Ze-Jing, et Chung-Hsien Wu. 2004. « Emotion recognition using acoustic features and textual content ». In *International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*. Vol. 1, p. 53-56. Taipei, Taiwan: IEEE.

Ciota, Zygmunt. 2007. « Feature extraction of spoken dialogs for emotion detection ». In *International Conference on Signal Processing Proceedings, ICSP.* Vol. 1, p. 4128934. Guilin, China: Institute of Electrical and Electronics Engineers Inc., Piscataway, NJ 08855-1331, United States. <http://dx.doi.org/10.1109/ICOSP.2006.345519>.

Clavel, Chloe, Ioana Vasilescu, Gael Richard et Laurence Devillers. 2006. « De la construction du corpus émotionnel au système de détection le point de vue applicatif de la surveillance dans les lieux publics ». *Revue d'Intelligence Artificielle,* vol. 20, n° 4-5, p. 529-551.

D'Mello, Sidney K., Scotty D. Craig, Amy Witherspoon, Bethany McDaniel et Arthur Graesser. 2008. « Automatic detection of learner's affect from conversational cues ». *User Modelling and User-Adapted Interaction,* vol. 18, n° 1-2, p. 45-80.

Dash, M., et H. Liu. 1997. « Feature selection for classification ». *Intelligent Data Analysis,* vol. 1, n° 3.

Dehak, N., P. Dumouchel et P. Kenny. 2007. « Modeling prosodic features with joint factor analysis for speaker verification ». *IEEE Transactions on Audio, Speech and Language Processing,* vol. 15, n° 7, p. 2095-103.

Dempster, A. P., N. M. Laird et D. B. Rubin. 1977. « Maximum likelihood from incomplete data via the em algorithm ». *Journal of the Royal Statistical Society. Series B,* vol. 39, p. 1-38.

Devillers, Laurence, et Laurence Vidrascu. 2007. « Real-life emotion recognition in speech ». *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* vol. 4441 NAI, p. 34-42.

Do, M.N. 2000. *An Automatic Speaker Recognition System.* Coll. « DSP Mini-Project ». Lausanne, Switzerland.

El Ayadi, M. M. H., M. S. Kamel et F. Karray. 2007. « Speech emotion recognition using Gaussian mixture vector autoregressive models ». In *International Conference on Acoustics, Speech, and Signal Processing (IEEE Cat. No. 07CH37846).* p. 957-960. Honolulu, HI, USA: The Institution of Engineering and Technology.

Forbes-Riley, Kate, et Diane J. Litman. 2004. « Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources ». In *Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics.* <http://acl.ldc.upenn.edu/hlt-naacl2004/main/pdf/55_Paper.pdf>.

Fotinea, Stavroula-Evita, Stelios Bakamidis, Theologos Athanaselis, Ioannis Dologlou, George Carayannis, Roddy Cowie, E. Douglas-Cowie, N. Fragopanagos et John G. Taylor. 2003 «Emotion in Speech: Towards an Integration of Linguistic, Paralinguistic, and Psychological Analysis ». In *Joint International Conference ICANN/ICONIP 2003 Istanbul, Turkey, June 26-29, 2003 Proceedings*. Vol. 2714, p. 1125 - 1132. <http://www.springerlink.com/(u0n0bj55qvs1vwvrjh1s3l45)/app/home/contribution.asp?referrer=parent&backto=issue,134,140;journal,1244,3844;linkingpublicationresults,1:105633,1>.

Fujie, Shinya, Daizo Yagi, Hideaki Kikuchi et Tetsunori Kobayashi. 2004a. « Prosody based Attitude Recognition with Feature Selection and Its Application to Spoken Dialog System as Para-Linguistic Information ». In *International Conference on Spoken Language Processing*. Jeju Island, Korea. <http://www.pcl.cs.waseda.ac.jp/publications/data/att/57d5.text.pdf>.

Fujie, Shinya, Daizo Yagi, Yosuke Matsusaka, Hideaki Kikuchi et Tetsunori Kobayashi. 2004b. « Spoken Dialogue System Using Prosody as Para-Linguistic Information ». In *ISCA, Speech Prosody 2004*. Nara, Japan. <http://www.isca-speech.org/archive/sp2004/sp04_387.pdf>.

Galarneau, A., P. Tremblay et P. Martin. « Dictionnaire de la parole ». In. <http://www.lli.ulaval.ca/labo2256/lexique/dico.html>.

Gao, Hui, Shanguang Chen et Guangchuan Su. 2007. « Emotion classification of mandarin speech based on TEO nonlinear features ». In *Proceedings - SNPD 2007: Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*. Vol. 3, p. 394-398. Qingdao, China: Institute of Electrical and Electronics Engineers Computer Society, Piscataway, NJ 08855-1331, United States. <http://dx.doi.org/10.1109/SNPD.2007.267>.

Giripunje, S., et N. Bawane. 2007. « ANFIS based emotions recognision in speech ». In *Knowledge-Based Intelligent Information and Engineering Systems: KES 2007 - WIRN 2007. Proceedings 11th International Conference, KES 2007. XVII Italian Workshop on Neural Networks. (Lecture Notes in Artificial Intelligence vol. 4692)*. Vol. 1, p. 77-84. Vietri sul Mare, Italy: The Institution of Engineering and Technology.

Giripunje, Shubhangi, et Ashish Panat. 2004. « Speech recognition for emotions with neural network: a design approach ». In *Knowledge-Based Intelligent Information and Engineering Systems. 8th International Conference, KES 2004. Proceedings (Lecture Notes in Artificial Intelligence Vol.3214)*. p. 640-645. Wellington, New Zealand: Springer-Verlag.

Gish, Herbert, et Michael Schmidt. 1994. « Text-independent speaker identification ». *IEEE Signal Processing Magazine,* vol. 11, n° 4, p. 18-32.

Gorodnitsky, Irina, et Claudia Lainscsek. 2004. « Machine Emotional Intelligence: A Novel Method for Spoken Affect Analysis ». In *Intern. Conf. on Development and Learning ICDL 2004.* <http://www.lis.inpg.fr/pages_perso/hammal/hammal_eusipco_soumis05.pdf>.

Grabe, Esther, Greg Kochanski et John Coleman. 2003. « Quantitative Modelling of Intonational Variation ». In *Proc. Speech Analysis and Recognition in Technology, Linguistics and Medicine.*

Grimm, Michael, et Kristian Kroschel. 2005. « Rule-based emotion classification using acoustic features ». In *Proc. 3rd Internat. Conf. on Telemedicine and Multimedia Communication.* Kajetany, Poland.

Grimm, Michael, Kristian Kroschel, Emily Mower et Shrikanth Narayanan. 2007. « Primitives-based evaluation and estimation of emotions in speech ». *Speech Communication,* vol. 49, n° 10-11, p. 787-800.

Grimm, Michael, Kristian Kroschel et Shrikanth Narayanan. 2007. « Support vector regression for automatic recognition of spontaneous emotions in speech ». In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings.* Vol. 4, p. 1085-1088. Honolulu, HI, United States: Institute of Electrical and Electronics Engineers Inc., Piscataway, NJ 08855-1331, United States. <http://dx.doi.org/10.1109/ICASSP.2007.367262>.

Hammal, Z., B. Bozkurt, L. Couvreur, D. Unay, A. Caplier et T. Dutoit. 2005. « Quiet versus agitated: vocal classification system ». In *EUSIPCO 2005* Antalya (Turkey).

Hirschberg, Julia, Stefan Benus, Jason M. Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, Bryan Pellom, Elizabeth Shriberg et Andreas Stolcke. 2005. « Distinguishing deceptive from non-deceptive speech ». In *9th European Conference on Speech Communication and Technology.* p. 1833-1836. Lisbon, Portugal: International Speech and Communication Association, Baixas, 66390, France.

Hoque, Mohammed E., Mohammed Yeasin et Max M. Louwerse. 2006. « Robust recognition of emotion from speech ». In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* Vol. 4133 NAI, p. 42-53. Marina Del Rey, CA, United States: Springer Verlag, Heidelberg, D-69121, Germany.

Hozjan, Vladimir, et Zdravko Kacic. 2003. « Context-independent multilingual emotion recognition from speech signals ». *International Journal of Speech Technology,* vol. 6, n° 3, p. 311-320.

Hu, Hao, Ming-Xing Xu et Wei Wu. 2007. « GMM supervector based SVM with spectral features for speech emotion recognition ». In *International Conference on Acoustics, Speech, and Signal Processing (IEEE Cat. No. 07CH37846)*. p. 413-416. Honolulu, HI, USA: The Institution of Engineering and Technology.

Huang, Huiqin, Qi Luo et Aiqin Zhu. 2008. « Speech emotion recognition in web based service ». In *International Conference on Communications, Circuits and Systems 2007*. p. 804-806. Kokura, Japan: Institute of Electrical and Electronics Engineers Computer Society, Piscataway, NJ 08855-1331, United States.

Huang, Rongqing, et Changxue Ma. 2006. « Toward a speaker-independent real-time affect detection system ». In *18th International Conference on Pattern Recognition* p. 4 pp. Hong Kong, China: The Institution of Engineering and Technology.

Huang, Xuedong, Alex Acero et Hsiao-wuen Hon. 2001. *Spoken language processing : a guide to Theory, Algorithm, and System Development*. United states of America: Prentice Hall PTR.

HUMAINE. 2009. *Research on Emotions and Human-Machine Interaction.* <http://emotion-research.net/>. Consulté le 15 janvier 2009.

Hung, LE Xuan, Georges Quénot et Eric Castelli. 2004. « Speaker-Dependent Emotion Recognition For Audio Document Indexing ». In *International Conference on Electronics, Information, And Communications (ICEIC'04)*. Hanoi, Vietnam.

Hyun, Kyung-Hak, Eun-Ho Kim et Yoon-Keun Kwak. 2006. « Robust speech emotion recognition using log frequency power ratio ». In *SICE-ICASE International Joint Conference*. p. 2586-2589. Busan, South Korea: Institute of Electrical and Electronics Engineers Computer Society, Piscataway, NJ 08855-1331, United States. <http://dx.doi.org/10.1109/SICE.2006.314794>.

Hyun, Kyung Hak, Eun Ho Kim et Yoon Keun Kwak. 2007. « Emotional feature extraction based on phoneme information for speech emotion recognition ». In *16th IEEE International Conference on Robot and Human Interactive Communication*. p. 802-806. Jeju, South Korea: The Institution of Engineering and Technology.

Inanoglu, Zeynep, et Ron Caneel. 2005. « Emotive alert: HMM-based emotion detection in voicemail messages ». In *International Conference on Intelligent User Interfaces, Proceedings IUI*. p. 251-253. San Diego, CA, United States: Association for Computing Machinery, New York, NY 10036-5701, United States.

Jia, Rong, Y. P. P. Chen, M. Chowdhury et Li Gang. 2007. « Acoustic features extraction for emotion recognition ». In *International Conference on Computer and Information Science*. p. 404-409. Melbourne, Qld., Australia: The Institution of Engineering and Technology.

Jiang, Dan-Ning, et Lian-Hong Cai. 2004. « Speech emotion classification with the combination of statistic features and temporal features ». In *International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*. Vol.3, p. 1967-1970. Taipei, Taiwan: IEEE.

Jones, Christian Martyn, et Ing-Marie Jonsson. 2007. « Performance analysis of acoustic emotion recognition for in-car conversational interfaces ». In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, PART 2. Vol. 4555 NCS, p. 411-420. Beijing, China: Springer Verlag, Heidelberg, D-69121, Germany.

Kim, Eun Ho, Kyung Hak Hyun, Soo Hyun Kim et Yoon Keun Kwak. 2007a. « Emotion interactive robot focus on speaker independently emotion recognition ». In *Proceedings of the 2007 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM2007)*. p. 313-318. Zurich, Switzerland: The Institution of Engineering and Technology.

Kim, Eun Ho, Kyung Hak Hyun, Soo Hyun Kim et Yoon Keun Kwak. 2007b. « Speech emotion recognition using eigen-FFT in clean and noisy environments ». In *16th IEEE International Conference on Robot and Human Interactive Communication*. p. 689-694. Jeju, South Korea: The Institution of Engineering and Technology.

Kira, K., et L. A. Rendell. 1992. « The feature selection problem: traditional methods and a new algorithm ». In *Proceedings Tenth National Conference on Artificial Intelligence*. p. 129-134. San Jose, CA, USA: AAAI Press.

Klasmeyer, Gudrun. 2000. « An automatic description tool for time-contours and long-term average voice features in large emotional speech databases ». In *SpeechEmotion*. p. 66-71. Newcastle, Northern Ireland, UK.

Kononenko, I. 1994. « Estimating attributes: analysis and extensions of RELIEF ». In *Proceedings of the European conference on machine learning on Machine Learning*. p. 171-182. Coll. « Machine Learning: ECML-94. European Conference on Machine Learning. Proceedings ». Catania, Italy: Springer-Verlag.

Kostoulas, T., T. Ganchev, I. Mporas et N. Fakotakis. 2008. « Detection of negative emotional states in real-world scenario ». In *19th IEEE International Conference on Tools with Artificial Intelligence*. p. 502-509. Patras, Greece: The Institution of Engineering and Technology.

Kostov, V., et S. Fukuda. 2000. « Emotion in user interface, voice interaction system ». In *SMC 2000 Conference Proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics. `Cybernetics Evolving to Systems, Humans, Organizations, and their Complex Interactions' (Cat. No.00CH37166)*. Vol. 2, p. 798-803. Nashville, TN, USA: IEEE. <http://dx.doi.org/10.1109/ICSMC.2000.885947 >.

Kuncheva, L. I. 2004. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley.

Kurematsu, M., J. Hakura et H. Fujita. 2007. « A framework of a speech communication system with emotion processing ». *WSEAS Transactions on Systems and Control,* vol. 2, n° 3, p. 283-288.

Kwon, Oh-Wook, Kwokleung Chan, Jiucang Hao et Te-Won Lee. 2003. « Emotion Recognition by Speech Signals ». In *EUROSPEECH 2003* Geneva, Switzerland.

Langley. 1994. « Selection of relevant features in machine learning ». In *Proceedings of the AAAI Fall Symposium on Relevance*. New Orleans: AAAI Press.

Lee, Ch Min, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee et Shrikanth Narayanan. 2004. « Emotion Recognition based on Phoneme Classes ». In *ICSLP*. Korea. <http://sail.usc.edu/publications/icslp04_chulminlee.pdf>.

Lee, Chul Min. 2004. « Recognizing emotions from spoken dialogs: a signal processing approach ». University of southern California.

Lee, Chul Min, S. S. Narayanan et R. Pieraccini. 2002. « Classifying emotions in human-machine spoken dialogs ». In *Proceedings 2002 IEEE International Conference on Multimedia and Expo (Cat. No.02TH8604)*. Vol. 1, p. 737-740. Lausanne, Switzerland: IEEE. <http://dx.doi.org/10.1109/ICME.2002.1035887 >.

Lee, Chul Min, et Shrikanth Narayanan. 2003. « Emotion Recognition Using a Data-Driven Fuzzy Inference System ». In *Eurospeech*. Geneva.

Lee, Tae-Seung, Mikyoung Park et Tae-Soo Kim. 2006. « Toward more reliable emotion recognition of vocal sentences by emphasizing information of Korean ending boundary tones ». In *Proceedings of SPIE - The International Society for Optical Engineering*. Vol. 6105, p. 304-313. San Jose, CA, United States: International Society for Optical Engineering, Bellingham WA, WA 98227-0010, United States.

Li, Wu, Yanhui Zhang et Yingzi Fu. 2007. « Speech emotion recognition in E-learning system based on affective computing ». In *Proceedings - Third International Conference on Natural Computation, ICNC 2007*. Vol. 5, p. 809-813. Haikou,

Hainan, China: Institute of Electrical and Electronics Engineers Computer Society, Piscataway, NJ 08855-1331, United States. <http://dx.doi.org/10.1109/ICNC.2007.677>.

Li, Xi, Jidong Tao, Michael T. Johnson, Joseph Soltis, Anne Savage, Kirsten M. Leong et John D. Newman. 2007. « Stress and emotion classification using jitter and shimmer features ». In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. 4, p. 1081-1084. Honolulu, HI, United States: Institute of Electrical and Electronics Engineers Inc., Piscataway, NJ 08855-1331, United States. <http://dx.doi.org/10.1109/ICASSP.2007.367261>.

Lin, Chi-Yueh, et Hsiao-Chuan Wang. 2005. « Language identification using pitch contour information ». In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Vol. I, p. 601-604. Philadelphia, PA, United States: Institute of Electrical and Electronics Engineers Inc. <http://dx.doi.org/10.1109/ICASSP.2005.1415185>.

Lin, Yi-Lin, et Gang Wei. 2005. « Speech emotion recognition based on HMM and SVM ». In *Proceedings of 2005 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 05EX1059)*. Vol. 8, p. 4898-4901. Guangzhou, China: IEEE.

Linde, Yoseph, Andres Buzo et Robert M. Gray. 1980. « Algorithm for vector quantizer design ». *IEEE Transactions on Communications,* vol. CM-28, n° 1, p. 84-95.

Lindemann, Andreas, Christian L. Dunis et Paulo Lisboa. 2004. « Probability distributions, trading strategies and leverage: An application of Gaussian mixture models ». *Journal of Forecasting,* vol. 23, n° 8, p. 559-585.

Litman, Diane J., et Kate Forbes-Riley. 2006. « Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors ». *Speech Communication,* vol. 48, n° 5, p. 559-590.

Liu, Jia, Chun Chen, Jiajun Bu, Mingyu You et Jianhua Tao. 2007a. « Speech emotion recognition based on a fusion of all-class and pairwise-class feature selection ». In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 4487 NCS, p. 168-175. Beijing, China: Springer Verlag, Heidelberg, D-69121, Germany.

Liu, Jia, Chun Chen, Jiajun Bu, Mingyu You et Jianhua Tao. 2007b. « Speech emotion recognition using an enhanced co-training algorithm ». In *International Conference on Multimedia & Expo*. p. 999-1002. Beijing, China: The Institution of Engineering and Technology.

Lugger, M., et Yang Bin. 2007. « The relevance of voice quality features in speaker independent emotion recognition ». In *IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE Cat. No. 07CH37846)*. p. 17-20. Honolulu, HI, USA: The Institution of Engineering and Technology.

Maganti, H. K., S. Scherer et G. Palm. 2007. « A novel feature for emotion recognition in voice based applications ». In *Affective Computing and Intelligent Interaction. Proceedings Second International Conference, ACII 2007. (Lecture Notes in Computer Science vol. 4738)*. p. 710-711. Lisbon, Portugal: The Institution of Engineering and Technology.

McLachlan, G. 1988. *Mixture Models*. New York: Marcel Dekker.

Mingmin, Gong, et Luo Qi. 2007. « Speech emotion recognition in Web based education ». In *Proceedings of the 2007 IEEE International Conference on Grey Systems and Intelligent Services*. p. 1082-1086. Nanjing, China: The Institution of Engineering and Technology.

Minker, W., J. Pittermann, A. Pittermann, P. M. Strauss et D. Buhler. 2006. « Next-generation human-computer interfaces - towards intelligent, adaptive and proactive spoken language dialogue systems ». In *2nd IET International Conference on Intelligent Environments IE 06*. Vol. 1, p. 7. Athens, Greece: The Institution of Engineering and Technology.

Morrison, D., et L. C. De Silva. 2007. « Voting ensembles for spoken affect classification ». *Journal of Network and Computer Applications,* vol. 30, n° 4, p. 1356-1365.

Morrison, D., Wang Ruili, L. C. De Silva et W. L. Xu. 2005. « Real-time spoken affect classification and its application in call-centres ». In *Proceedings. Third International Conference on Information Technology and Applications*. Vol. 1, p. 483-487. Sydney, NSW, Australia: IEEE Comput. Soc.

Nakatsu, Ryohei, Joy Nicholson et Naoko Tosa. 1999 «Emotion recognition and its application to computer agents with spontaneous interactive capabilities ». In *International Multimedia Conference, Proceedings of the seventh ACM international conference on Multimedia*. p. 343 - 351 Orlando, Florida, United States ACM Press New York, NY, USA. <http://portal.acm.org/citation.cfm?id=319463.319641&dl=GUIDE&dl=ACM&type=series&idx=SERIES316&part=Proceedings&WantType=Proceedings&title=International%20Multimedia%20Conference>.

Neiberg, Daniel, Kjell Elenieus et Kornel Laskowski. 2006. « Emotion Recognition in Spontaneous Speech Using GMMs ». In *International Conference on Spoken Language Processing, Interspeech 2006—ICSLP*. Pittsburgh.

Noda, Tetsuya, Yoshikazu Yano, Shinji Doki et Shigeru Okuma. 2007. « Adaptive emotion recognition in speech by feature selection based on KL-divergence ». In *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*. Vol. 3, p. 1921-1926. Taipei, Taiwan: Institute of Electrical and Electronics Engineers Inc., New York, NY 10016-5997, United States. <http://dx.doi.org/10.1109/ICSMC.2006.385011>.

Nogueiras, Albino, Asunción Moreno, Antonio Bonafonte et José B. Mariño. 2001. « Speech Emotion Recognition Using Hidden Markov Models ». In *Eurospeech*. Aalborg, Denmark.

Nwe, Tin Lay, Say Wei Foo et Liyanage C. De Silva. 2003. « Speech emotion recognition using hidden Markov models ». *Speech Communication,* vol. 41, n° 4, p. 603-623.

O'Shaughnessy, Douglas. 2000. « Speech communications human and machine ». In, 2. New York: The Institute of Electrical and Electronics Engineers, Inc.

Oudeyer, Pierre-yves. 2002. « Novel useful features and algorithms for the recognition of emotions in speech ». In *In Proceedings of the 1st International Conference on Speech Prosody*. p. 547-550. Aix-en-Provence. <http://www.csl.sony.fr/~py/prosodyRec.ps>.

Panat, A. R., et V. T. Ingole. 2008. « Affective state analysis of speech for speaker verification: Experimental study, design and development ». In *Proceedings - International Conference on Computational Intelligence and Multimedia Applications, ICCIMA 2007*. Vol. 1, p. 255-261. Sivakasi, Tamil Nadu, India: Institute of Electrical and Electronics Engineers Computer Society, Piscataway, NJ 08855-1331, United States. <http://dx.doi.org/10.1109/ICCIMA.2007.59>.

Pao, Tsang-Long, Yu-Te Chen, Jun-Heng Yeh et Pei-Jia Li. 2006. « Mandarin emotional speech recognition based on SVM and NN ». In *Proceedings - International Conference on Pattern Recognition*. Vol. 1, p. 1096-1099. Hong Kong, China: Institute of Electrical and Electronics Engineers Inc., Piscataway, NJ 08855-1331, United States. <http://dx.doi.org/10.1109/ICPR.2006.780>.

Pao, Tsang-Long, Yu-Te Chen, Jun-Heng Yeh et Wen-Yuan Liao. 2005. « Detecting Emotions in Mandarin Speech ». *Computational Linguistics and Chinese Language Processing,* vol. 10, p. 347-362.

Pao, Tsang-Long, C. S. Chien, Yu-Te Chen, Jun-Heng Yeh, Yun-Maw Cheng et Wen-Yuan Liao. 2008a. « Combination of multiple classifiers for improving emotion recognition in mandarin speech ». In *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. Vol. 1, p. 33-36. Kaohsiung, Taiwan: The Institution of Engineering and Technology.

Pao, Tsang-Long, Wen-Yuan Liao, Yu-Te Chen, Jun-Heng Yeh, Yun-Maw Cheng et C. S. Chien. 2008b. « Comparison of several classifiers for emotion recognition from noisy mandarin speech ». In *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. Vol. 1, p. 21-24. Kaohsiung, Taiwan: The Institution of Engineering and Technology.

Pelletier, Charles. 2006. « Classification des sons respiratoires en vue d'une détection automatique des sibilants ». Mémoire de maîtrise Chicoutimi : Université du Québec à Chicoutimi ; Rimouski : Université du Québec à Rimouski, 2006. <http://theses.uqac.ca/resume_these.php?idnotice=24968894

Petrushin, Valery A. 2000. « Emotion recognition in speech signal: experimental study, development and application ». In *International Conference on Spoken Language Processing (ICSLP 2000)*. <http://www.accenture.com/NR/rdonlyres/288F697E-EC02-4E84-9016-88ADDDD0CD40/0/EmotionRecognitionICSLP20002.pdf>.

Pirker, H. 2007. « Mixed feelings about using phoneme-level models in emotion recognition ». In *Affective Computing and Intelligent Interaction. Proceedings Second International Conference, ACII 2007. (Lecture Notes in Computer Science vol. 4738)*. p. 772-773. Lisbon, Portugal: The Institution of Engineering and Technology.

Pittermann, Angela, et Johannes Pittermann. 2007a. « Getting bored with HTK? Using HMMs for emotion recognition from speech signals ». In *International Conference on Signal Processing Proceedings, ICSP*. Vol. 1, p. 4128935. Guilin, China: Institute of Electrical and Electronics Engineers Inc., Piscataway, NJ 08855-1331, United States. <http://dx.doi.org/10.1109/ICOSP.2006.345520>.

Pittermann, Johannes, et Angela Pittermann. 2007b. « A post-processing approach to improve emotion recognition rates ». In *International Conference on Signal Processing Proceedings, ICSP*. Vol. 1, p. 4128936. Guilin, China: Institute of Electrical and Electronics Engineers Inc., Piscataway, NJ 08855-1331, United States. <http://dx.doi.org/10.1109/ICOSP.2006.345521>.

Prasad, J. R., R. S. Prasad et U. V. Kulkarni. 2007. « Analysis of feature extraction methods for emotions recognition using pattern recognition approach ». In *International MultiConference of Engineers and Computer Scientists*. Vol. 1, p. 514-519. Kowloon, China: The Institution of Engineering and Technology.

Rahurkar, Mandar A., et John H.L Hansen. 2003. « Frequency Distribution Based Weighted Sub-Band Approach for Classification Of Emotional/Stressful Content in Speech ». In *EUROSPEECH-2003 / INTERSPEECH-2003*. Switzerland: Robust Speech Processing Group Center for Spoken Language Research University of Colorado Boulder, Campus Box 594. <http://cslr.colorado.edu/rspl/PUBLICATIONS/PDFs/CP-Euro03-TEO-Wcover.PDF>.

Ramamohan, S., et S. Dandapat. 2006. « Sinusoidal model-based analysis and classification of stressed speech ». *IEEE Transactions on Audio, Speech and Language Processing,* vol. 14, n° 3, p. 737-746.

Razak, A. A., M. I. Z. Abidin et R. Komiya. 2003. « A preliminary speech analysis for recognizing emotion ». In *Student Conference on Research and Development SCOReD 2003. Proceedings (IEEE Cat. No.03EX752)*. p. 49-54. Putrajaya, Malaysia: IEEE.

Razak, A. A., M. H. M. Yusof et R. Komiya. 2004. « Towards automatic recognition of emotion in speech ». In *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No.03EX795)*. p. 548-551. Darmstadt, Germany: IEEE.

Resch, B. *Mixtures of Gaussians: A tutorial for the course computational intelligence.* <http://www.igi.tugraz.at/lehre/CI>. Consulté le 9 mai 2008.

Reynolds, D. A., T. F. Quatieri et R. B. Dunn. 2000. « Speaker verification using adapted Gaussian mixture models ». In *Digit. Signal Process.*, 1-3.Vol. 10, p. 19-41. Gaithersburg, MD, USA: Academic Press. <http://dx.doi.org/10.1006/dspr.1999.0361>.

Reynolds, D. A., et R. C. Rose. 1995. « Robust text-independent speaker identification using Gaussian mixture speaker models ». *IEEE Transactions on Speech and Audio Processing,* vol. 3, n° 1, p. 72-83.

Rudra, T., M. Kavakli et D. Tien. 2007. « Emotion detection from male speech in computer games ». In *TENCON 2007 - 2007 IEEE Region 10 Conference*. p. 32-35. Taipei, Taiwan: The Institution of Engineering and Technology.

Ruili, Wang, D. Morrison et L. C. De Silva. 2007. « Ensemble methods for spoken emotion recognition in call-centres ». *Speech Communication,* vol. 49, n° 2, p. 98-112.

Scherer, K.R. 2000. « Psychological models of emotion ». In *The Neuropsychology of Emotion* sous la dir. de Borod, Joan C. p. 137–162. New York: Oxford University Press.

Scherer, Klaus R. 1996. « Adding the affective dimension: a new look in speech analysis and synthesis ». In *In ICSLP*. Philadelphia: 4th International Conference on Spoken Language Processing. <http://www.asel.udel.edu/icslp/cdrom/vol3/1014/a1014.pdf >.

Scherer, Klaus R. 2003. « Vocal communication of emotion: A review of research paradigms». *Speech Communication,* vol. 40, n° 1-2, p. 227-256.

Schuller, B. 2002. « Towards intuitive speech interaction by the integration of emotional aspects ». In *IEEE International Conference on Systems, Man and Cybernetics. Conference Proceedings (Cat. No.02CH37349)*. Vol. 6, p. 6. Yasmine Hammamet, Tunisia: IEEE. <http://dx.doi.org/10.1109/ICSMC.2002.1175635 >.

Schuller, B., M. Lang et G. Rigoll. 2002. « Automatic emotion recognition by the speech signal ». In *6th World Multiconference on Systemics, Cybernetics and Informatics. Proceedings*. Vol. 9, p. 367-72. Orlando, FL, USA: Int. Inst. Inf. & Syst.

Schuller, B., D. Seppi, A. Batliner, A. Maier et S. Steidl. 2007a. « Towards more reality in the recognition of emotional ». In *IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE Cat. No. 07CH37846)*. p. 941-944. Honolulu, HI, USA: The Institution of Engineering and Technology.

Schuller, B., B. Vlasenko, R. Minguez, G. Rigoll et A. Wendemuth. 2007b. « Comparing one and two-stage acoustic modeling in the recognition of emotion in speech ». In *IEEE Workshop on Automatic Speech Recognition and Understanding*. p. 596-600. Kyoto, Japan: The Institution of Engineering and Technology.

Schuller, Bjorn, Ronald Muller, Manfred Lang et Gerhard Rigoll. 2005. « Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles ». In *9th European Conference on Speech Communication and Technology*. p. 805-808. Lisbon, Portugal: International Speech and Communication Association, Baixas, 66390, France.

Seppänen, Tapio, Eero Väyrynen et Juhani Toivanen. 2003. « Prosody-based classification of emotions in spoken Finnish ». In *EUROSPEECH*. Geneva, Switzerland. <http://www.mediateam.oulu.fi/publications/pdf/438.pdf>.

Sethu, V., E. Ambikairajah et J. Epps. 2007. « Speaker normalisation for speech-based emotion detection ». In *15th International Conference on Digital Signal Processing*. p. 611-614. Cardiff, UK: The Institution of Engineering and Technology.

Sigmund, M., et T. Dostal. 2007. « Detection of psychological stress by analysing of glottal pulse waveform ». In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications*. p. 526-529. Innsbruck, Austria: The Institution of Engineering and Technology.

Sim, Kwee-Bo, In-Hun Jang et Chang-Hyun Park. 2007. « The development of interactive feature selection and GA feature selection method for emotion recognition ». In *Knowledge-Based Intelligent Information and Engineering Systems: KES 2007-WIRN 2007. 11th International Conference, KES 2007*. p. 73-81. Vietri sul Mare, Italy: The Institution of Engineering and Technology.

Tabatabaei, T. S., S. Krishnan et Guergachi Aziz. 2007. « Emotion recognition using novel speech signal features ». In *IEEE International Symposium on Circuits and Systems (IEEE Cat. No.07CH37868)*. p. 4. New Orleans, LA, USA: The Institution of Engineering and Technology.

Ten Bosch, Louis. 2003. « Emotions, speech and the ASR framework ». *Speech Communication,* vol. 40, n° 1-2, p. 213-225.

Tickle, A. 2000. « English and Japanese Speaker's Emotion Vocalizations and Recognition: A Comparison Highlighting Vowel Quality ». In *ISCA Workshop on Speech and Emotion* Belfast.

Tomas, B., M. Maletic et Z. Raguz. 2007. « Determination and evaluation pitch harmonics parameters with emotions classification ». In *15th International Conference on Software, Telecommunications & Computer Networks*. p. 375-379. Split-Dubrovnik, Croatia: The Institution of Engineering and Technology.

Verhelst, W., et M. Shami. 2007. « An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech ». *Speech Communication,* vol. 49, n° 3, p. 201-12.

Ververidis, D., et C. Kotropoulos. 2005. « Emotional speech classification using Gaussian mixture models ». In *IEEE International Symposium on Circuits and Systems (ISCAS) (IEEE Cat. No. 05CH37618)*. Vol. 3, p. 2871-2874. Kobe, Japan: IEEE.

Ververidis, D., et C. Kotropoulos. 2007. « Accurate estimate of the cross-validated prediction error variance in Bayes classifiers ». In *IEEE Workshop on Machine Learning for Signal Processing*. p. 354-359. Thessaloniki, Greece: The Institution of Engineering and Technology.

Vlasenko, B., B. Schuller, A. Wendemuth et G. Rigoll. 2007. « Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing ». In *Affective Computing and Intelligent Interaction. Proceedings Second International Conference, ACII 2007. (Lecture Notes in Computer Science vol. 4738)*. p. 139-147. Lisbon, Portugal: The Institution of Engineering and Technology.

Vogt, T., et E. Andre. 2005. « Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition ». In *IEEE International Conference on Multimedia and Expo*. p. 4 pp. Amsterdam, Netherlands: The Institution of Engineering and Technology.

Wagner, J., T. Vogt et E. Andre. 2007. « A systematic comparison of different HMM designs for emotion recognition from acted and spontaneous speech ». In *Affective Computing and Intelligent Interaction. Proceedings Second International Conference, ACII*

*2007. (Lecture Notes in Computer Science vol. 4738)*. p. 114-125. Lisbon, Portugal: Springer-Verlag.

Wahab, A., Quek Chai et S. De. 2007. « Speech emotion recognition using auditory cortex ». In *IEEE Congress on Evolutionary Computation*. p. 2658-2664. Singapore, Singapore: The Institution of Engineering and Technology.

Walker, Marilyn A., Diane J. Litman, Ace A. Kamm et Alicia Abella. 1997. « PARADISE: A Framework for Evaluating Spoken Dialogue Agents ». In *In Proc. 35th Annual Meeting of the Association for Computational Linguistics and 8th Conf. of the European Chapter of the Association for Computational Linguistics*. p. 271-280.

Xiao, Zhongzhe, E. Dellandrea, Weibei Dou et Liming Chen. 2007. « Automatic hierarchical classification of emotional speech ». In *Ninth IEEE International Symposium on Multimedia 2007 - Workshops*. p. 291-296. Beijing, China: The Institution of Engineering and Technology. <http://dx.doi.org/10.1109/ISM.Workshops.2007.56>.

Xie, Bo, Ling Chen, Gen-Cai Chen et Chun Chen. 2007. « Feature selection for emotion recognition of mandarin speech ». *Zhejiang Daxue Xuebao (Gongxue Ban)/Journal of Zhejiang University (Engineering Science)*, vol. 41, n° 11, p. 1816-1822.

Yacoub, Sherif, Steve Simske, Xiaofan Lin et John Burns. 2003. « Recognition of Emotions in Interactive Voice Response Systems ». In *EUROSPEECH*. Geneva, Switzerland.

Yanushevskaya, I., M. Tooher, C. Gobi et Chasaide Ailbhe Ni. 2007. « Time- and amplitude-based voice source correlates of emotional portrayals ». In *Affective Computing and Intelligent Interaction. Proceedings Second International Conference, ACII 2007. (Lecture Notes in Computer Science vol. 4738)*. p. 159-170. Lisbon, Portugal: Springer-Verlag.

Yoon, Won-Joong, et Kyu-Sik Park. 2007. « A study of emotion recognition and its applications ». In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 4617 NAI, p. 455-462. Kitakyushu, Japan: Springer Verlag, Heidelberg, D-69121, Germany.

You, Mingyu, Chun Chen, Jiajun Bu, Jia Liu et Jianhua Tao. 2006. « Emotion recognition from noisy speech ». In *IEEE International Conference on Multimedia and Expo, ICME 2006 - Proceedings*. Vol. 2006, p. 1653-1656. Toronto, ON, Canada: Institute of Electrical and Electronics Engineers Computer Society, Piscataway, NJ 08855-1331, United States. <http://dx.doi.org/10.1109/ICME.2006.262865>.

Young, S., P. Woodland et W. Byrne. 1993. « HTK: Hidden Markov Model Toolkit V1.5 ». In *Cambridge University Engineering Department Speech Group and Entropic Research Laboratories Inc.* <http://htk.eng.cam.ac.uk/>. Consulté le 21/08/2008.

Yu, Chen, Paul M. Aoki et Allison Woodruff. 2004. « Detecting user engagement in everyday conversations ». In *In Proc. 8th International Conference on Spoken Language Processing (ICSLP 2004)*. p. 1329-1332. Jeju Island; Korea.

Yu, Dong-Mei, et Jian-An Fang. 2008. « Research on a methodology to model speech emotion ». In *Proceedings of the 2007 International Conference on Wavelet Analysis and Pattern Recognition*. p. 825-830. Beijing, China: The Institution of Engineering and Technology.

Yu, Feng, Eric Chang, Yingqing Xu et Heung-Yeung Shum. 2001. « Emotion Detection from Speech to Enrich Multimedia Content ». *Advances in Multimedia Information Processing - PCM 2001: Second IEEE Pacific Rim Conference on Multimedia Bejing, China, Proceedings,* vol. 2195 / 2001 p. 550.

Zervas, Panagiotis, Iosif Mporas, Nikos Fakotakis et George Kokkinakis. 2007. « Evaluating intonational features for emotion recognition from speech ». *International Journal on Artificial Intelligence Tools,* vol. 16, n° 6, p. 1001-1014.

Zhu, Aiqin, et Qi Luo. 2007. « Study on speech emotion recognition system in E-learning ». In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, PART 3. Vol. 4552 NCS, p. 544-552. Beijing, China: Springer Verlag, Heidelberg, D-69121, Germany.