Indexation des émotions dans les documents audiovisuels à partir de la modalité auditive

Xuân Hùng Lê

INSTITUT POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque

|__|__|__|__|__|__|__|__|__|__|

THESE EN COTUTELLE INTERNATIONALE

pour obtenir le grade de
DOCTEUR DE L'Institut Polytechnique de Grenoble
et
DE L'Institut Polytechnique de Hanoi

Spécialité : « Systèmes d'information »

préparée au **laboratoire LiG**
(Laboratoire d'Informatique de Grenoble)
dans le cadre de l'Ecole Doctorale
« Mathématiques, Sciences et Technologies de l'Information, Informatique »
et au **Centre de recherche international MICA**
(Multimédia Information, Communication and Application)

présentée et soutenue publiquement
par
LÊ Xuân Hùng
le 01 Juillet 2009

TITRE

INDEXATION DES ÉMOTIONS DANS LES DOCUMENTS AUDIOVISUELS À PARTIR DE LA MODALITÉ AUDITIVE

DIRECTEURS DE THÈSE : M. GEORGES QUÉNOT, M. ERIC CASTELLI

JURY

Mme Christine COLLET,	Présidente
M. Liming CHEN,	Rapporteur
Mme Laurence DEVILLERS,	Rapporteur
M. Georges QUÉNOT,	Directeur de thèse
M. Eric CASTELLI,	Directeur de thèse
M. Philippe JOLY,	Examinateur

Remerciement

Tout d'abord, je voudrais adresser tous mes remerciements, ainsi que toute ma gratitude, à mes directeur et co-directeur de thèse, Georges QUÉNOT et Eric CASTELLI, pour m'avoir accueilli dans leurs équipes respectives MRIM et MICA, et accompagné au cours de plus de cinq années de thèse avec leurs conseils, leurs aides, et leurs encouragements précieux.

Je tiens à remercier monsieur NGUYEN Trong Giang et madame PHAM Thi Ngoc-Yen, qui m'ont accueilli au Centre MICA, et qui m'ont beaucoup supporté du côté vietnamien pour cette thèse en cotutelle.

Je tiens à remercier monsieur Liming CHEN, madame Laurence DEVILLERS et monsieur Philippe JOLY, qui m'ont fait l'honneur d'être mes rapporteurs et mon examinateur, pour leur lecture attentive et pour toutes leurs remarques constructives sur le manuscrit. Je voudrais adresser un grand merci à madame Christine COLLET, d'avoir accepté d'être la Présidente du jury.

Je pense à ma famille qui m'a apporté un soutien important, non seulement à l'aspect sentimental, mais également par les encouragements dont j'avais besoin pour mener à bien ce travail.

Un grand merci à tous mes collègues français et vietnamiens du centre de recherche MICA, et de l'équipe MRIM pour leurs coopérations dans le travail, et leurs aides pour la correction de cette thèse.

Résumé

Cette thèse concerne la détection des émotions dans les énoncés audio multi-lingues. Une des applications envisagées est l'indexation des états émotionnels dans les documents audio-visuels en vue de leur recherche par le contenu.

Notre travail commence par l'étude de l'émotion et des modèles de représentation de celle-ci : modèles discrets, continus et hybride. Dans la suite des travaux, seul le modèle discret sera utilisé pour des raisons pratiques d'évaluation mais aussi parce qu'il est plus facilement utilisable dans les applications visées. Un état de l'art sur les différentes approches utilisées pour la reconnaissance des émotions est ensuite présenté. Le problème de la production de corpus annoté pour l'entraînement et l'évaluation des systèmes de reconnaissance de l'état émotionnel est également abordé et un panorama des corpus disponibles est effectué. Une des difficultés sur ce point est d'obtenir des corpus réalistes pour les applications envisagées. Afin d'obtenir des données plus spontanées et dans des langues plus variées, deux corpus ont été créés à partir de films cinématographiques, l'un en Anglais, l'autre en Vietnamien.

La suite des travaux se décompose en quatre parties : études et recherche des meilleurs paramètres pour représenter le signal acoustique pour la reconnaissance des émotions dans celui-ci, étude et recherche des meilleurs modèles et systèmes de classification pour ce même problème, expérimentation sur la reconnaissance des émotions inter-langues, et enfin production d'un corpus annoté en vietnamien et évaluation de la reconnaissance des émotions dans cette langue qui a la particularité d'être tonale. Dans les deux premières études, les cas mono-locuteur, multi-locuteur et indépendant du locuteur ont été considérés.

La recherche des meilleurs paramètres a été effectuée sur un ensemble large de paramètres locaux et globaux classiquement utilisés en traitement automatique de la parole ainsi que sur des dérivations de ceux-ci. Une approche basée sur la sélection séquentielle forcée avant a été utilisée pour le choix optimal des combinaisons de paramètres acoustiques. La même approche peut être utilisée sur des types de données différents bien que le résultat final dépende du type considéré. Parmi, les MFCC, LFCC, LPC, la fréquence fondamentale, l'intensité, le débit phonétique et d'autres coefficients extraits du domaine temporel, les paramètres de type MFCC ont donné les meilleurs résultats dans les cas considérés. Une approche de normalisation symbolique a permis d'améliorer les performances dans le cas indépendant du locuteur.

Pour la recherche du meilleur modèle et système de classification associé, une approche d'élimination successive selon des cas de complexité croissante (mono-locuteur, multi-locuteur et indépendant du locuteur) a été utilisée. Les modèle GMM, HMM, SVM et VQ (quantification vectorielle) on été étudiés. Le modèle GMM est celui qui donne les meilleurs résultats sur les données considérées.

Les expérimentations inter-langue (Allemand et Danois) ont montré que les méthodes développées fonctionnent bien d'une langue à une autre mais qu'une optimisation des paramètres spécifique pour chaque langue ou chaque type de données est nécessaire pour obtenir les meilleurs résultats. Ces langues sont toutefois des langues non tonales. Des essais avec le corpus créé en Vietnamien ont montré une beaucoup moins bonne généralisation dans ce cas. Cela peut être du au fait que le Vietnamien est une langue tonale mais cela peut aussi être dû à la différence entre les conditions de création des corpus : acté dans les premiers cas et plus spontané pour le Vietnamien.

Mots-clés : émotion, reconnaissance de l'émotion, indexation de l'émotion.

Abstract

This thesis concerns the detection of emotions in multi-lingual audio utterances. One application being considered is the indexing of emotional states in audio-visual documents for their search by contents.

Our work begins with the study of emotion and of its model representations: discrete, continuous and hybrid models. In the following of the work, only the discrete model will be used for practical reasons linked to evaluation but also because it is easier to use in the targeted applications. A state of the art on the different approaches used for emotion recognition is then presented. The problem of the production of annotated corpus for training and evaluation of emotional state recognition systems is also considered and an overview of the available corpus is given. One of the difficulties on this point is to obtain realistic corpus for the target applications. To obtain data more spontaneous and more diverse in languages, two corpora were created from motion pictures, one in English and one in Vietnamese.

The following work is divided into four parts: study and search for the best parameters to represent the acoustic signal for the emotion recognition, study and search for the best models and classification systems for the same problem, experiments on the recognition emotions across languages and, finally, production of an annotated Vietnamese corpus and assessment of emotion recognition in this language which has the specificity of being tonal. In the first two studies, mono-speaker, multi-speaker and speaker-independent cases were considered.

The search for the best parameters was performed on a broad set of global and local parameters traditionally used in automatic speech processing as well as derivations them. An approach based on the forward forced sequential selection was used for selecting optimal combinations of acoustic parameters. The same approach can be used on different data types, although the final result depends upon the type. Among the MFCC, LFCC, LPC, fundamental frequency, intensity, phonetic rate and other parameters from the time-domain, MFCC gave the best results in the considered cases. A symbolic normalization approach has helped to improve the performance in the speaker independent case.

For the search for the best models and associated classification systems, an approach by successive elimination within cases of increasing complexity (single-speaker, multi-speaker and speaker-independent) was used. The GMM, HMM, SVM and VQ (vector quantization) models have been studied. The GMM model is the one which led to the best results on the considered data.

Cross-language experiments (German and Danish) have shown that the developed methods work well from one language to another, but that a specific optimization of the parameters for each language and for each type of data is necessary for obtaining the best results. These languages are not tonal languages, however. Tests with the created Vietnamese corpus have shown a much less good generalization in this case. This may be due to the fact that the Vietnamese language is tonal but it may also be due to the difference between the conditions of creation of the corpora: action in the first case and more spontaneous for the Vietnamese.

Keywords: emotion, emotion recognition, emotion indexing.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1. **Introduction**

1.1. Contexte et problématiques

Actuellement, l'explosion de l'information ne se produit plus seulement en termes de quantité, de volume, mais aussi en termes de diversité des documents. Particulièrement, pendant ces dernières années, avec le développement des services de partage de l'image, de la musique ainsi que de la vidéo à la demande sur Internet et sur d'autres réseaux comme le téléphone portable, avec la capacité de participation à la production et à la diffusion d'œuvres personnelles, avec l'accélération énorme des capacités de stockage et de débit, ces bonds de développement nous confrontent à de réels besoins de gérer des archives des documents audio-visuels de taille conséquente. Pour traiter ces informations volumineuses, il nous faut donc des processus de traitement particulier qui visent à détecter et extraire automatiquement des éléments intéressants comme le thème, l'auteur des documents textuels, l'identification du locuteur, du présentateur dans les documents audiovisuels.

Dans la bande son et pour le signal de la parole, l'information linguistique (ce qui est dit) est devenue depuis longtemps un des éléments les plus étudiés parce qu'elle porte la plus grosse part du contenu des conversations. Cependant, aujourd'hui, beaucoup d'autres informations (extralinguistiques) contenues dans le signal de parole deviennent exploitables, y compris le genre et/ou l'identité du locuteur (qui parle ?) et son état émotionnel (comment le dit-il ?).

L'intérêt de ces informations est grandissant dans plusieurs domaines. Par exemple, en reconnaissance/synthèse automatique de la parole, des informations sur le locuteur et son état émotionnel peuvent aider à améliorer significativement le taux de la reconnaissance [Womack et al, 1996] [Bosch 2003] ou la qualité de la parole synthétique [Muray et Arnott, 2008]. Dans le domaine de l'interaction homme-machine, de la robotique, les systèmes deviennent plus naturels, plus amicaux, si une fonction de reconnaissance/génération de l'émotion y est intégrée. Pour la formation et l'éducation, la médecine ou la sécurité, les systèmes de surveillance

automatique seront plus efficaces s'ils peuvent détecter/identifier automatiquement les locuteurs et l'état émotionnel de ces locuteurs (patients, étudiants, clients, etc.). Dans notre travail de thèse, nous nous intéressons particulièrement à la détection de l'état émotionnel du locuteur.

Bien que l'émotion soit devenue un sujet d'intérêt depuis 1872 dans le cadre des études de Darwin sur l'expression et la transmission de l'émotion entre des êtres humains et entre des animaux, et que les études sur la composante audio de l'émotion aient aussi été démarrées depuis les années 1970, la reconnaissance automatique de l'émotion dans la parole ainsi que dans d'autres modalités n'attirent l'attention des chercheurs à une échelle importante que depuis ces dernières années.

La première question qui se pose pour toutes les recherches et aussi pour notre étude est de sélectionner les émotions. Effectivement, il existe encore beaucoup de discussions pour les deux problèmes principaux dans ce domaine : la nature et le nombre d'émotions d'une part, et l'universalité et l'innéité de l'émotion (à travers la culture, la langue, etc.) d'autre part.

Pour le premier problème, l'existence d'émotions primaires est largement reconnue depuis 1984 dans la plupart des recherches du psychologiste Paul Ekman [Ekman 1999], ainsi que par des dizaines d'autres travaux synthétisés par [Ortony et Turner 1990]. Cependant, la question « quels sont les émotions primaires ? » est toujours en discussion [Scherer 2003]. Le terme « big six » des émotions primaires semble recueillir le plus de suffrages dans les recherches du domaine, c'est la raison pour laquelle, nous suivrons aussi cette direction.

Le deuxième problème portant sur l'universalité et sur l'innéité de l'émotion attire également l'attention des chercheurs mais ce problème est aussi en discussion. Selon [Hager & Ekman 1983], une grande branche des recherches commencée par Darwin en 1982 cherche et réussit à certains niveaux à affirmer que l'universalité et l'innéité de l'émotion existent à certains degrés.

Ces deux problèmes principaux de la caractérisation de l'émotion influencent fortement les études dans le domaine, la nôtre comprise, non seulement pour le choix du corpus, de la langue et des émotions à étudier mais aussi pour l'estimation des possibilités d'application de notre système d'indexation automatique.

Par ailleurs, la difficulté de construction / collection de corpus ne favorise pas non plus les recherches sur l'émotion. Ce sont les raisons pour lesquelles les recherches actuelles sont très variées au niveau des corpus, du nombre d'émotions, de la langue et des applications.

Comme mentionné ci-dessus, les travaux de cette thèse s'inscrivent dans ce contexte et s'intéressent plus particulièrement à un nouveau besoin des applications et de l'utilisateur : la recherche dans les documents audio-visuels par l'émotion ; ces documents peuvent être issus par exemple d'une émission de radio ou de télévision, des conversations dans un film, d'un clip ou d'une téléconférence. Il a été prouvé par plusieurs travaux de recherche [Ekman 1982], [Banse et al, 1996], [Burkhardt et al, 2000] que le signal audio en général et la parole en particulier véhiculent des informations fiables sur l'émotion bien que le signal de la parole ne soit pas le seul moyen pour l'émission des signes de l'émotion.

Comme la parole, l'état émotionnel est aussi une production physique et donc, par nature, il dépend fortement du locuteur. Il est aisé de remarquer que, pour pouvoir être appliqué dans un tel environnent multi-locuteurs, le système doit être le plus robuste possible ou le moins dépendant possible du locuteur. Cela peut-il être résolu par une sélection soigneuse de l'ensemble de paramètres ? Par l'utilisation de techniques de classification performantes ? Ou cela exige-t-il une normalisation adéquate ?

Cette problématique est aussi la problématique actuelle du domaine de la reconnaissance de l'émotion. Des études ont expérimenté la reconnaissance indépendante du locuteur mais les résultats obtenus ne sont pas satisfaisants. D'autres études se limitent au cas de la reconnaissance de l'émotion avec des locuteurs précis (mono-locuteur) avec des résultats obtenus qui peuvent être intéressants mais qui ne peuvent pas toujours être applicables. Dans cette thèse, nous proposons de travailler sur l'analyse, la sélection et la normalisation des paramètres, ainsi que sur les techniques de classification dans le but de trouver l'approche la plus adéquate pour le cas de la reconnaissance de l'émotion indépendante du locuteur, et ceci tout d'abord pour chaque langue disponible. Des expériences avec les langues croisées seront ensuite effectuées pour répondre d'une part à notre problème de la possibilité d'un système d'indexation automatique des émotions indépendante du locuteur et d'autre part à la question de l'universalité des émotions primaires.

1.2. Approche et contributions

Notre approche dans la reconnaissance de l'émotion se décompose en deux étapes.

La première étape a pour but d'analyser globalement des paramètres du signal de la parole pour avoir une première vue des comportements de ces paramètres lors de l'expression émotionnelle. Pour ce faire, nous proposons d'utiliser huit opérateurs que nous appelons « opérateurs globaux » qui nous permettent de caractériser à certains degrés la dynamique, la distribution et la variation des valeurs de chaque paramètre. Nous proposons également d'utiliser la normalisation par rapport au neutre pour réduire au maximum la dépendance des paramètres par rapport aux locuteurs et par rapport à la structure phonétique des énoncés. Bien qu'il y ait des résultats remarquables avec les informations globales, par exemple la possibilité de réaffirmation du rôle important de la prosodie en expression de l'émotion ou l'efficacité de l'intensité, nous n'allons pas plus loin avec ce type de paramètres car en utilisant seulement les opérateurs globaux, nous perdrons l'information instantanée. De plus, l'information globale dépend fortement du locuteur, ce qui ne favorise pas notre objectif de recherche. Cette étape nous donne simplement des suggestions pour la deuxième étape.

Dans la deuxième étape, en s'appuyant sur les paramètres instantanément mesurés tout au long du signal que nous appelons les « paramètres locaux », nous étudions trois cas en fonction de la tolérance des paramètres par rapport au locuteur : la reconnaissance mono-locuteur (le cas le plus tolérant), la reconnaissance multi-locuteur (un nombre fini de locuteurs connus) et la reconnaissance indépendante du locuteur (le cas le moins tolérant). En suivant ces trois cas, nous éliminons au fur à mesure des paramètres ainsi que des techniques de classification moins efficaces afin de trouver pour le dernier cas, celui de la reconnaissance indépendante du locuteur, l'ensemble des paramètres les plus adaptés. Pour sélectionner l'ensemble de paramètres efficaces, nous proposons d'utiliser la sélection forcée séquentielle en avant (SFSA) ; l'ajout d'un mécanisme pour forcer la continuité de cet algorithme en choisissant toujours le meilleur résultat obtenu nous assure d'avoir au moins la même performance obtenue par la méthode classique SSA (sélection séquentielle en avant, qui s'arrête s'il n'y a plus de meilleur résultat) et d'avoir un résultat supérieur à un seuil spécifié par l'utilisateur.

En travaillant avec les paramètres, nous proposons également une normalisation appelée la normalisation symbolique dont l'idée est d'unifier les intervalles différents des valeurs de paramètres en échelle unique en s'appuyant essentiellement sur la « signification sémantique » de ces intervalles. En comparaison avec les approches de normalisation par la moyenne et par l'écart-type, cette approche donne non seulement un meilleur équilibre de la dynamique, du décalage entre les paramètres et entre différents locuteurs, mais, comme elle prend en compte

aussi la signification de chaque paramètre, alors elle pourrait permettre une meilleure fusion entre des paramètres très différents, par exemple la fusion entre la fréquence fondamentale et les coefficients MFCCs.

A côté de l'étude sur les paramètres, les techniques de classification jouent aussi un rôle très important pour la performance des systèmes de reconnaissance. Dans la littérature, nous trouvons plusieurs approches utilisant plusieurs techniques, y compris des techniques assez simples comme l'utilisation d'arbre de décision dans [Yacoub et al, 2003], les k plus proches voisins dans [Dellaert et al, 1996], [Yacoub et al, 2003], [Yoon et al, 2007], mais aussi des techniques très complexes comme le mélange des modèles de Markov cachés, voir [Fernandez et al, 2003], les réseaux de neurones [Shi et al, 2003] [Fernandez et al, 2003] [Schüller et al, 2004] etc. Cependant, ces études ont été effectuées avec des corpus différents, avec des ensembles différents de paramètres et aussi pour des ensembles différents d'émotions, c'est la raison pour laquelle la comparaison de l'efficacité de ces techniques appliquées au domaine de la classification émotionnelle est tout-à-fait relative et n'a pas beaucoup de signification. Les études de [Batliner et al, 2006] sont les seules études que nous pouvons trouver dans la littérature qui ont effectué des comparaisons significatives entre différentes approches sur les mêmes données. Nous pouvons aussi trouver dans les études de [Batliner et al, 2006] des résultats qui montrent que la performance de la classification peut être améliorée par la fusion des meilleurs paramètres des différentes approches ou par la fusion des sorties de différents classificateurs.

Dans ce cadre de thèse, pour une étude systématique nous nous intéressons et distinguons les trois branches principales des techniques de reconnaissance/classification :

- le premier groupe se compose des techniques par lesquelles on cherche à modéliser de manière la plus efficace l'espace de paramètres pour chaque classe afin de la reconnaitre ; pour ce groupe de techniques, nous expérimentons deux modèles : le modèle de quantification vectorielle et le modèle de mélange de gaussiennes, qui peuvent être dits opposés en termes de complexité ;
- le deuxième groupe comprend des modèles qui cherchent à classifier des émotions en essayant de séparer explicitement l'espace de paramètres. Parmi les modèles que nous utilisons, les arbres de décision et les machines à vecteurs de support appartiennent à cette deuxième branche ;
- le troisième groupe est supérieur au premier groupe par sa capacité de capture des informations d'évolution temporelle. Parmi les techniques connues et utilisées dans le domaine du traitement de la parole comme la programmation dynamique (Dynamic Time Warping ou DTW en anglais), les réseaux de neurones (NN) et les modèles de Markov caché (HMM), nous avons choisi les modèles HMM pour notre étude en raison de leur proximité avec le modèle GMM ; cela favorise en premier lieu la comparaison entre ces deux groupes de techniques de classification et cela favorise en deuxième lieu l'estimation de l'existence et de l'utilité des informations sur l'évolution temporelle des paramètres.

Enfin, la plupart des rares études portant sur l'universalité de l'émotion comme celle de [Scherer et al, 2001], de Ekman 1982 [Scherer, 2003], et de [Haidt et al, 1999], n'a été effectuée que par des tests perceptifs sur des sujets humains. Nos travaux sur la reconnaissance inter-langue/interculturel à travers l'étude des performances d'un ordinateur s'avèrent aussi intéressants. Ils ont non seulement pour but de résoudre notre problème de reconnaissance indépendante du locuteur avec des langues/cultures précises mais aussi, à certains niveaux, ils nous permettent de contribuer à répondre au problème de l'universalité des émotions primaires à travers deux langues/cultures différentes.

1.3. Structure de la thèse

Après l'introduction, cette thèse se compose de trois grandes parties. Dans la première partie (chapitres 2 et 3) nous passons en revue les études théoriques sur l'émotion ainsi que les approches/résultats obtenus dans le domaine de reconnaissance de l'émotion. Dans la deuxième partie (chapitre 4) nous présentons les approches utilisées pour la construction des corpus. La troisième partie (chapitres 5, 6, 7 et 8) contient essentiellement notre contribution avec la présentation de nos approches et de nos expérimentations. Les résumés suivants détaillent un peu plus le contenu de chaque chapitre.

Dans le chapitre 2, premièrement, nous passons en revue les études psychologiques sur l'émotion afin de comprendre la nature de celle-ci et aussi afin de répondre à deux grandes questions avant de commencer nos études :

- la définition de l'émotion, le nombre d'émotions et les relations entre les émotions ;
- l'universalité de l'émotion (culturelle, langue, etc.).

Deuxièmement, les trois approches principales de modélisation des émotions sont considérées afin de choisir celle qui sera la plus adaptée à notre objectif :

- l'approche discrète qui repose sur l'existence d'un petit nombre d'émotions primaires discrètes et discriminantes entre elles, les autres émotions étant considérées comme des combinaisons de ces émotions primaires ;
- l'approche dimensionnelle considère l'état émotionnel comme un phénomène continu qui peut être représenté dans un espace dimensionnel ; les deux dimensions qui sont largement considérées sont la dimension « activité » (*arousal* en anglais) qui présente le dégrée d'activité (basse/haute) d'un état émotionnel et la dimension « valence » qui présente le degré de satisfaction (positif/négatif) ;
- et l'approche hybride de l'approche discrète et de l'approche dimensionnelle.

Dans le chapitre 3, nous passons au côté pratique décrit dans la littérature. Des approches pour la reconnaissance de l'émotion dans plusieurs modalités seront considérées, y compris le texte, l'image, la vidéo et, particulièrement le signal acoustique de la parole afin d'avoir, premièrement une vue globale des progrès du domaine, mais aussi deuxièmement pour évaluer si notre problème de reconnaissance de l'émotion en se basant seulement sur le signal acoustique est faisable et puis applicable. Les paramètres de la bande son, les techniques de classification ainsi que les résultats obtenus par les études sur la reconnaissance de l'émotion contenue dans la parole seront aussi synthétisés dans ce chapitre dans le but de chercher des prémisses de réponse pour les autres questions de notre problème : quels paramètres, quelles caractéristiques du signal de la parole sont discriminatives pour les émotions ? Quelle technique de classification est la plus adéquate ?

Dans le chapitre 4, nous parlons d'une partie très importante de tous les systèmes de reconnaissance, la construction/collection des corpus ; en effet la qualité des corpus utilisés décidera la qualité du système. Nous étudions les trois approches généralement utilisées pour construire des corpus de l'émotion, leurs avantages ainsi que leurs inconvénients/difficultés durant la construction et durant le travail de l'étude ; ces trois approches sont : la construction par la collection des conversations quotidiennes/spontanées, la construction par la simulation des échantillons de l'émotion en utilisant des acteurs/actrices, et la construction en extrayant des segments émotionnels à partir des films ou des documents télévisés. Dans ce chapitre, nous

décrivons précisément les corpus de la littérature que nous utilisons pour notre étude ainsi que les raisons pour lesquelles nous les avons choisis.

Le chapitre 5 commence nos expérimentations par l'analyse des paramètres. Nous vérifions la performance des paramètres en utilisant l'information statistique globale, pour éviter au maximum l'influence de la dépendance des paramètres par rapport aux locuteurs, la normalisation par rapport au neutre, et les coefficients relatifs sont proposés. Bien que des conclusions importantes puissent être retenues dans cette partie, nous focalisons notre étude dans la deuxième partie du chapitre sur les mesures instantanées des paramètres parce qu'elles sont plus spécifiques et elles sont aussi plus robustes pour la reconnaissance de l'émotion indépendante du locuteur.

Les paramètres locaux du signal de la parole auxquels nous nous intéressons se composent des paramètres de la prosodie, le nombre de passages par zéro, le nombre d'extrémités (points maxima, minima locaux) et les trois ensembles de coefficients connus : les MFCCs, les LFCCs et les LPCs. Leurs dérivées temporelles au premier et au second ordre sont aussi introduites pour améliorer la robustesse. Au total, nous aurons 273 paramètres des deux domaines temporel et fréquentiel. Pour sélectionner des paramètres et pour trouver les combinaisons de paramètres les plus performantes, nous proposons d'utiliser dans ce chapitre la sélection forcée séquentielle en avant (SFSA) qui est théoriquement plus efficace que sa voisine, la sélection séquentielle en avant (SSA), et de la comparer avec la méthode de l'analyse des composantes principales (ACP) très connue dans le domaine de l'exploration de données (data-mining en anglais). Ce chapitre présentera également le raffinement au fur à mesure des paramètres moins robustes en les considérant étape par étape dans les trois cas de reconnaissance : mono-locuteur, multi-locuteur et indépendante du locuteur. Cependant, pour que le processus de l'indexation automatique atteigne un niveau plus satisfaisant, une approche de normalisation symbolique est proposée et utilisée. La dernière partie de ce chapitre porte sur cette méthode de normalisation.

Dans le chapitre 6, nous effectuons nos études selon les trois axes principaux des techniques de classification mentionnés ci-dessus dans le but de :

- chercher la technique/le modèle la/le plus efficace pour notre problème de classification de l'émotion indépendante du locuteur ;
- de chercher la configuration la plus optimale de la technique choisie ;
- de vérifier si l'information de l'évolution temporelle des paramètres peut servir à discriminer les états émotionnels ;

Dans le dernier chapitre de la partie d'expérimentation, le chapitre 7, à côté du corpus DES (Danish Emotional Speech Database) qui a toujours été choisi pour notre étude de manière systématique en raison de l'équilibre de sa structure, nous effectuons également la validation de notre approche avec un autre corpus BES (Berlin Database of Emotional Speech) qui possède l'avantage de contenir un plus grand nombre de locuteurs. Les expérimentations de reconnaissance inter-langue, un cas plus général de la reconnaissance indépendante du locuteur, nous donnent une validation supplémentaire de notre approche, des conclusions à certains niveaux de l'universalité de l'émotion, et aussi des observations intéressantes sur les paramètres sur nos corpus.

Comme l'étude de l'émotion est un domaine encore nouveau et aussi l'émotion elle-même est un état difficile à obtenir, les corpus disponibles sont encore rares, particulièrement pour la parole spontanée. C'est la raison pour laquelle nous avons également effectué la construction de deux autres corpus, l'un en anglais et l'autre en vietnamien, en extrayant les segments émotionnels à partir de 10 films (en DVD). Par manque de temps, nous ne pouvons effectuer

que quelques expérimentations sur le corpus vietnamien. Les détails de ce corpus vietnamien et les résultats sont presentés dans le chapitre 8.

Le chapitre 9 terminera notre thèse par la conclusion des résultats obtenus, de la contribution de la thèse et des travaux en perspectives.

Chapitre 2. **Concepts et études de l'émotion**

Définir les émotions est une tâche difficile. Il est évident que nous pouvons tous exprimer et percevoir des émotions et que celles-ci constituent une force déterminant en partie au moins nos actions, mais il y a de grandes divergences d'opinion au sujet de la façon dont nous pouvons réaliser cette tâche et conceptualiser l'émotion. Cette section donne tout d'abord une vue globale des études théoriques sur l'émotion, ainsi que des approches pour modéliser et pour généraliser la représentation des états émotionnels, comme l'approche discrète, les approches dimensionnelles et l'approche de prototype.

2.1. L'émotion

[Frijda 1986] décrit l'émotion comme le changement dans un état de promptitude pour maintenir ou modifier des rapports avec l'environnement. De même, [Scherer et al, 1981] décrivent l'émotion comme interface de l'organisme vers le monde extérieur. [Darwin 1872] a indiqué dans « l'expression des émotions chez l'homme et les animaux » que celle-ci est largement commune et universelle. Il a présenté l'idée que les émotions sont inséparables des schémas d'actions sélectionnés par l'évolution en raison de leur valeur de survie [Cowie et al, 2001]. Ses exemples principaux d'expressions émotives étaient les mouvements faciaux et corporels chez l'homme et pour des animaux mais il décrit également quelques exemples d'expressions vocales.

Une expérience émotive ressentie se rapporte à la réaction interne humaine à un stimulus. Des actions extérieures qui montrent comment on se sent sont des expressions de l'émotion.

Les théoriciens considèrent que certaines propriétés fondamentales des émotions et de leur catégorisation sont universelles [Russell 1980]. Cependant, ils déclarent également que les facteurs innés et culturels influencent les expériences émotives.

Définie de façon plus pointue, l'émotion est un changement brusque en réponse à des stimuli particuliers qui dure pendant une période courte [Murray et Arnott 1993]. Un état d'esprit plus stable qui dure pendant une plus longue période est désigné sous le nom de l'humeur. Il existe également d'autres définitions qui sont différentes de cette connotation et qui se rapportent aussi à l'état mental humain comme l'impression, le sentiment, la personnalité… Définie de manière plus large, l'émotion inclut tous les termes apparus ci-dessus aux différents niveaux.

2.2. Approches discrètes

Les approches discrètes reposent sur l'existence d'un petit nombre d'émotions primaires discrètes, cela veut dire que ces émotions primaires sont supposées être discriminantes entre elles. Donc avec ces approches, les autres émotions sont considérées comme des mélanges des émotions primaires. Les émotions les plus communément considérées comme émotions primaires sont : la joie, la tristesse, la peur, le dégoût, la colère et la surprise. Cependant, le nombre d'émotions primaires varie de deux à dix-huit selon des théoriciens de l'émotion. Le Tableau 1 présente différentes listes d'émotions primaires proposées par les théoriciens ; cette synthèse est proposée par [Ortony et Turner 1990].

Les théoriciens ne s'accordent pas toujours au sujet de ce que sont les émotions. En effet, on peut constater que certaines listes d'émotions primaires contiennent des éléments qui n'existent pas dans d'autres listes. Les divergences d'opinion à propos du nombre d'émotions primaires correspondent à des divergences d'opinion à propos de leurs identités. Un exemple du désaccord est la surprise : bien qu'elle soit très souvent incluse dans les listes d'émotions primaires (Ekman et autres, 1982 ; Izard, 1971 ; Plutchik, 1980 ; Tomkins, 1984 ; (voir le Tableau 1)), ceux-ci remarquent aussi qu'il n'est pas évident qu'elle soit en elle-même une émotion. Une raison à cela est que, à la différence des émotions primaires indiscutables comme la peur ou la colère, la surprise n'a pas une présence certaine. Le désir est également un cas discutable bien que Descartes et quelques autres théoriciens (par exemple, Arnold 1960; Frijda 1986) aient inclus le désir dans leurs listes d'émotions primaires.

Un autre exemple incertain dans les listes d'émotions primaires est l'intérêt. Quelques théoriciens (e.g Frijda 1986 ; Izard 1971 ; Tomkins 1984 (voir le tableau 1)) le considèrent comme une émotion primaire contrairement à d'autres (Oatley et Johnson-Laid 1987 (voir le tableau 1); [Ortony et Turner 1990]). La raison de ceci est que l'intérêt relève plus d'une attitude que d'un état émotionnel [Ortony et Turner 1990].

Comme décrit ci-dessus, les approches discrètes de l'émotion essaient de conceptualiser les émotions à partir de quelques émotions primaires et de traiter chacune d'elles comme une émotion discrète. Cependant, il n'y a pas de consensus sur la définition des émotions primaires et ces approches ne fournissent pas de description claire pour des émotions non basiques. Néanmoins, la notion d'émotion primaire forme une base pour la conceptualisation de l'émotion [Iida 2002].

Référence	*Emotion primaire*	*Principe supposé sur lequel repose la sélection.*
Arnold (1960)	colère, aversion, courage, découragement, désir, désespoir, peur, haine, espoir, amour, tristesse	Relation aux tendances d'action
Ekman, Friesen, Ellsworth (1982)	colère, peur, dégoût, joie, tristesse, surprise	Expressions faciales universelles
Frijda (1986)	désir, bonheur, intérêt, surprise, étonnement, peine	État de préparation à l'action
Gray (1982)	fureur, terreur, anxiété, joie	Biologiquement câblé
Izard (1971)	colère, mépris, dégoût, détresse, peur, culpabilité, intérêt, joie, honte, surprise	Biologiquement câblé
James (1884)	peur, chagrin, amour, fureur	Participation corporelle
McDougall (1926)	colère, dégoût, exultation, peur, soumission, tendresse, étonnement	Relation aux instincts
Oatley and Johnson-Laid (1987)	colère, dégoût, anxiété, bonheur, tristesse	Pas de principe sous-jacent
Plutchik (1980)	acceptation, attente, joie, peur, colère, tristesse, surprise, dégoût	Relation aux processus biologiques adaptatifs
Tomkins (1984)	intérêt, détresse, colère, joie, mépris, peur, honte, surprise, dégoût	Niveau d'activité neuronale
Watson (1930)	peur, amour, fureur	Biologiquement câblé

Tableau 1 : Listes d'émotions primaires avec les principes justificateurs, synthétisées par [Ortony et Turner 1990].

2.3. Approches dimensionnelles

Les approches dimensionnelles considèrent les émotions comme un phénomène continu ou graduel. Les théoriciens essaient d'identifier les émotions en les plaçant dans un espace à plusieurs dimensions.

Des relations entre les émotions peuvent ainsi être capturées. Les approches dimensionnelles les plus rencontrées sont bidimensionnelles (par exemple [Schlosberg 1952] [Russel 1980]), tridimensionnelles [Daly 1983] et multidimensionnelles avec un nombre de dimensions supérieur à trois [Plutchik 1980]. La plupart des théoriciens de l'approche dimensionnelle incluent une dimension « valence »[1] (agréable/désagréable ou positif/négatif), une dimension

[1] Valence : le degré d'attraction ou d'aversion qu'un individu ressent envers un objet ou un événement spécifique - *The American Heritage® Dictionary of the English Language*

d'activité (haute/basse ou actif/passif « *arousal* »[2] en anglais) et peu avec une autre dimension d'intensité (forte/faible).

Modèle bidimensionnel de Schlosberg : [Schlosberg 1952] a proposé un modèle bidimensionnel par l'analyse des expressions faciales. Il a demandé à des sujets de classer les expressions faciales à partir de photos pour les six groupes d'émotions : 1) l'amour, le bonheur, la gaieté, 2) la surprise, 3) la peur, la souffrance, 4) la colère 5) le dégoût et 6) le mépris. Il a ensuite demandé aux sujets d'évaluer l'ensemble des photos selon une échelle monodimensionnelle sur 9 points en valence et pour l'attention. Il a constaté qu'il y avait une tendance à la confusion entre les catégories 6 et 1 ou 6 et 5, ce qui l'a amené à conclure que ces émotions peuvent être tracées dans un espace polaire plutôt que cartésien. Il a donc proposé un modèle à deux dimensions qui se compose de la valence sur l'axe vertical et l'attention (attention/rejet) sur l'axe horizontal, l'état neutre étant positionné au milieu. L'état neutre de l'esprit est difficile à définir et il est habituellement paraphrasé en tant qu'état « non émotionnel ». La Figure 1 présente le diagramme de ce modèle bidimensionnel. [Schlosberg 1954], en outre, a proposé un modèle tridimensionnel en ajoutant une dimension de l'activité (haute/basse).

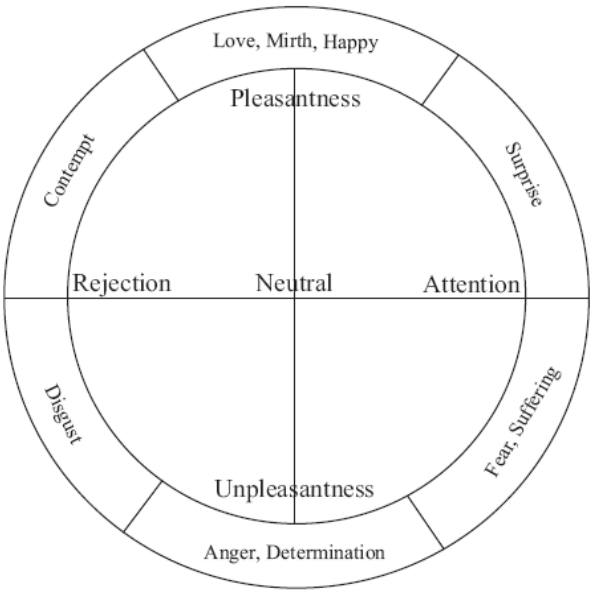

Figure 1 : Modèle bidimensionnel de [Schlosberg 1952]

Modèle circumplex de [Russel 1980] : le modèle circumplex de Russell est un modèle bidimensionnel polaire. Il est illustré par la Figure 2 avec 28 émotions déterminées expérimentalement. Russel a proposé ce modèle en analysant les résultats de la catégorisation des sujets et en les ordonnant sur un schéma polaire selon la graduation multidimensionnelle des états émotifs rapportés par ses expérimentateurs. L'axe horizontal du modèle est interprété comme la valence (agréable/désagréable) et l'axe vertical comme l'activité (haute/basse). Les étiquettes sont tracées par Russell sur ce circumplex selon le degré d'agrément et d'activité. Il

[2] Arousal : un état de réponse à la stimulation sensorielle ou à l'excitabilité. *Dorland's Medical Dictionary.* Les gens avec un niveau « arousal » haut répondent souvent aux stimuli sensoriels avec une réponse forte, fréquemment une réponse de bagarre/fuite/peur. Quand nous avons un niveau « arousal » bas, le système nerveux a une réaction diminuée à l'entrée sensorielle et donc ne réagit pas ou ne répond pas rapidement ou même ne répond pas du tout.

explique que le nombre de catégories est extensible et que n'importe quelle catégorie concernant l'émotion pourrait être ajoutée à ce modèle. De même que Schlosberg, Russell a considéré que le centre du cercle est un point neutre ou un niveau d'adaptation. La distance entre le point neutre et la position d'une émotion particulière représente l'intensité de cette émotion. Tandis que l'activité est mesurée comme la déviation de l'état physiologique normal d'une personne, l'intensité est considérée comme le degré auquel l'expérience émotionnelle produit un changement de l'état neutre. Il a aussi divisé l'espace circulaire en espaces plus étroits pour des observations plus fines. Une division des hémisphères nous donne des dimensions de satisfaction. Une division en quatre parties nous donne quatre quarts de cercle :

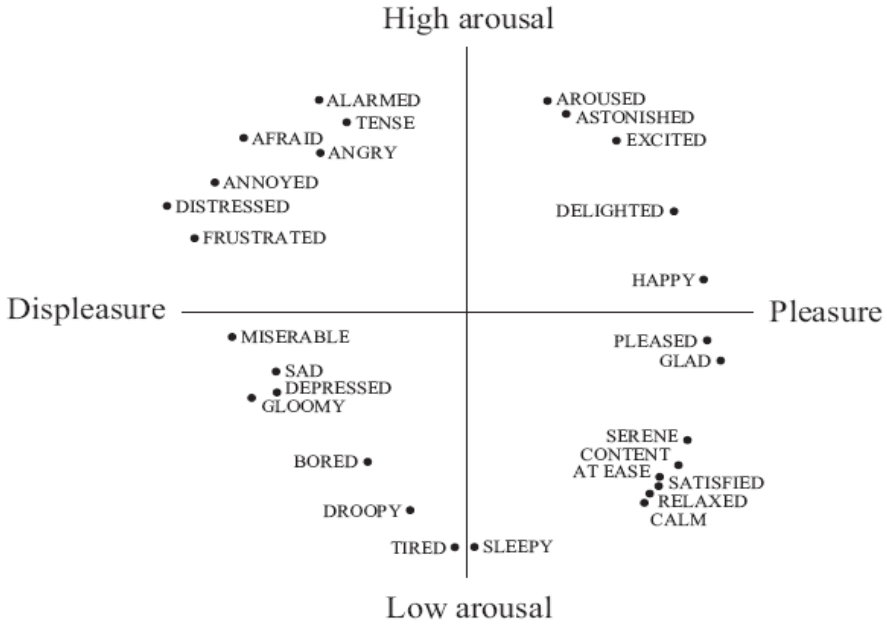

Figure 2 : Modèle circumplex de [Russel 1980]

1) agréable/haute activité (exultation) ;

2) désagréable/haute activité (détresse) ;

3) désagréable/base activité (dépression) et

4) agréable/base activité (calme).

Comme nous l'avons vu, le modèle de circumplex de Russell est simple et permet facilement de comprendre comment les émotions sont liées entre elles. Cependant, ce modèle a besoin d'être amélioré pour une analyse plus fine. Comme Shaver et d'autres (1987) le soulignent, il n'est pas évident de décider de classer comme émotions ou non certaines sensations comme la fatigue ou la somnolence.

Modèle du cône multidimensionnel de Plutchik : [Plutchik 1980] a élaboré sa théorie psycho-évolutionnaire sur l'approche discrète de l'émotion. Il a choisi la colère, l'attente, la joie, l'acceptation, la peur, la surprise, la tristesse et le dégoût comme émotions primaires pour son modèle. Il a présumé un arrangement circulaire des émotions primaires comme le présente la Figure 3 et ces émotions sont arrangées avec les autres émotions « relatives » dans un modèle de

cône tridimensionnel avec l'intensité, la polarité, et les dimensions de similitude. La Figure 4 présente son diagramme de ce modèle sous une forme géométrique proche d'un épi de maïs. Dans les deux figures (Figure 3 et la Figure 4) la polarité est montrée par des émotions opposées autour du point neutre, par exemple : la joie face à la tristesse.

Plutchik considère qu'une émotion complexe peut être expliquée comme un mélange d'émotions primaires et que n'importe quelle paire adjacente d'émotions primaires peut être combinée pour produire une émotion complexe. Il les appelle des dyades primaires et selon lui, « des mélanges de deux émotions primaires séparées par une seule émotion peuvent s'appeler des dyades secondaires et, plus loin, des mélanges de deux émotions primaires séparées par deux autres émotions peuvent être appelées des dyades tertiaires ». Par exemple, pour l'émotion primaire : l'amour = la joie + l'acceptation, pour l'émotion secondaire : la culpabilité = la joie + la peur et, pour l'émotion tertiaire : le délice = la joie + surprise.

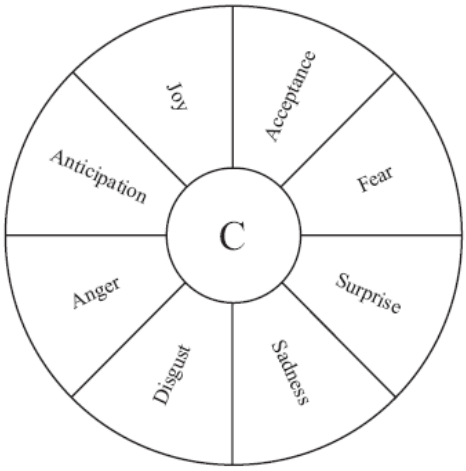

Figure 3 : Les émotions primaires de [Plutchik 1980]

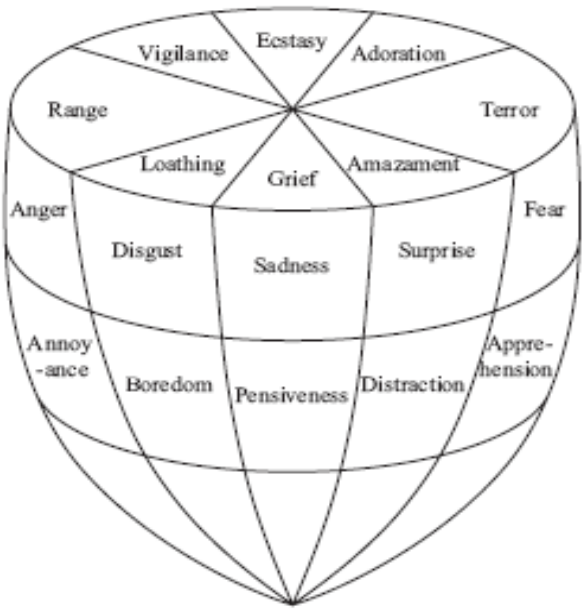

Figure 4 : Modèle du cône multidimensionnel [Plutchik 1980]

Bien que le modèle de Plutchik soit bien structuré, quelques critiques sont apparues, semblables à celles émises pour l'approche discrète d'émotions. Comme abordé dans la section 2.2, des émotions « non-valences », telles que la surprise, qui sont incertaines à s'appeler une émotion, sont introduites dans le modèle de cône. De plus, pour les oppositions polaires définies par Plutchik telles que l'attente (l'espoir) et la surprise, il est discutable de les appeler des émotions opposées.

2.4. Approche hybride

L'approche hybride est un compromis entre l'approche discrète et les approches dimensionnelles. L'étude par [Shaver et al, 1987] est un bon exemple de cette approche. Les auteurs ont conçu une analyse de groupe hiérarchique et construit un modèle de trois-couches pour conceptualiser les émotions avec une catégorisation sémantique par 112 sujets et par 135 mots de l'émotion.

La couche la plus abstraite comporte seulement les deux catégories : « valence » positive et « valence » négative. Les catégories de la couche du milieu sont les catégories d'émotions primaires : la joie, l'amour, la colère, la tristesse et la peur. Ce sont des équivalents aux émotions primaires définies dans l'approche de l'émotion discrète. La plus basse couche se compose d'émotions non-basiques et concrètes (par exemple : l'adoration, la tendresse pour l'amour ; l'enthousiasme, le zèle pour la joie; l'agitation, la gêne pour la colère, etc). On a conduit plus loin une analyse multidimensionnelle et tracé un diagramme avec deux dimensions orthogonales qui ressemble au modèle de circumplex de Russell. L'axe vertical du diagramme peut être considéré comme la dimension de « valence » (positive – négative) et l'axe horizontal peut être considéré comme la dimension d'activité.

La discussion est la même pour cette approche : est-ce que toutes les émotions de bas niveau sont des émotions primaires ? En outre, les données pour construire l'analyse de groupe hiérarchique sont-elles suffisantes ?

2.5. Définition des émotions utilisées dans cette thèse

Nous avons passé en revue dans cette partie les théories conceptuelles des émotions comprenant les approches discrètes, les approches dimensionnelles et l'approche hybride. Comme nous avons pu l'observer, caractériser et organiser les émotions sont des tâches difficiles. Il y a des divergences d'opinions parmi les spécialistes sur le nombre d'émotions primaires et sur la façon dont nous pouvons conceptualiser l'émotion.

Comme nous l'avons vu, il y a fondamentalement deux façons d'interpréter l'émotion. La première consiste à capturer l'émotion comme un changement discret et brusque dans l'état mental. La seconde préfère considérer l'émotion comme un changement progressif de l'état mental. Les approches discrètes de l'émotion prennent en compte la présence ou l'absence d'émotions précises seulement, tandis que les approches dimensionnelles et l'approche hybride utilisent l'intensité et le positionnement dans un espace continu.

Dans cette thèse, nous avons choisi de travailler dans le cadre d'une approche discrète : nous considèrerons un petit nombre d'émotions précises et nous chercherons à détecter leur absence ou leur présence seulement. Nous ne chercherons pas à placer l'état émotif du locuteur dans un espace continu ni non plus à évaluer l'éventuelle intensité de l'émotion ressentie. Ceci ne correspond pas à une préférence d'un type d'approche par rapport à un autre, mais plutôt à des considérations pragmatiques liées d'une part à la nécessité de produire des corpus annotés pour

l'entraînement et l'évaluation des systèmes et d'autre part aux besoins que nous souhaitons couvrir dans le cadre des applications visées.

Produire un corpus annoté et définir une émotion recherchée dans un document serait très difficile dans le cadre d'une approche continue. En ce qui concerne le nombre d'émotions discrètes, nous nous sommes restreints à un petit nombre (huit) pour des raisons pratiques et aussi pour rester dans une gamme qui fait largement consensus. Le constat sous-jacent de notre position s'exprime par le fait que les expressions émotives vocales, qui sont le thème de cette thèse, sont essentiellement l'interprétation (le résultat) des processus mentaux internes et, même si on peut classer les émotions ressenties dans un espace continu, nous avons tout de même naturellement tendance à les classer de manière discrète et limitée lorsque nous en parlons.

Un nombre très important d'émotions ou d'états mentaux liés à l'émotion peuvent être considérés (voir le Tableau 2). Comme mentionné précédemment, nous avons retenu les huit emotions : la colère, la joie, la peur, la tristesse, la surprise, le dégoût, l'ennui et le neutre pour notre recherche sur leur reconnaissance. Parmi ces huit émotions la colère, la joie, la peur, la tristesse, la surprise, le dégoût sont les émotions primaires répandues et consensuelles dans plusieurs d'études (voir le Tableau 1), l'ennui et le neutre sont choisis en raison pratique du corpus. Celles-ci, naturellement ne couvrent pas toutes les émotions mais comme dans d'autres recherches considérées précédemment, nous les avons considérées comme des émotions essentielles de la vie quotidienne.

Il est difficile de définir un état neutre mais, comme d'autres producteurs de corpus annotés que nous utilisons dans le cadre de cette thèse, nous considérons un état neutre comme l'état intentionnellement non-émotionnel dans l'interprétation des acteurs ou des actrices (voir le Chapitre 4).

2.6. Corrélation entre l'aspect acoustique et les émotions

Nous faisons l'hypothèse que la réalisation des systèmes de synthèse, ainsi que la reconnaissance des émotions contenues dans la parole, sont possibles si et seulement si il existe des corrélations fiables entre les émotions et les caractéristiques acoustiques du signal. Un certain nombre de chercheurs ont déjà étudié cette question comme [Banse et al, 1996] [Burkhardt et al, 2000] et leurs résultats s'accordent avec des corrélations venant des contraintes physiologiques pour des classes d'émotions primaires.

En effet, certains des états émotifs sont souvent corrélés avec des états physiologiques particuliers [Picard 1997] qui ont à leur tour des effets assez « mécaniques » et prévisibles sur la parole, particulièrement sur la fréquence fondamentale, le débit et la qualité de la parole. Par exemple, quand on est dans un état de colère, de peur ou de joie, le système nerveux sympathique est agité, le rythme cardiaque et la pression du sang augmentent, la bouche devient sèche, il y a aussi des tremblements occasionnels des muscles. La voix est donc forte, rapide et parlée avec une forte énergie de haute fréquence, la moyenne et la variation de la fréquence fondamentale sont également plus importantes [Breazeal 2000]. Au contraire, quand on est fatigué, ennuyé ou triste, le système nerveux parasympathique produit une diminution de la fréquence cardiaque, une diminution de la pression du sang et l'augmentation de salivation, ce qui a pour conséquence une voix lente, à basse intonation et avec peu d'énergie à haute fréquence [Breazeal 2000].

abaissé	déception	Faible	panique
abandonné	découragement	fatigué	paralysé
abhorrer	défensif	fébrilité	paresse
absent	dégoût	fermé	passion
admiration	délaissé	fierté	patient
adorer	délectation	figé	peine
affection (affectueux)	dépendant	flottement	perdu
affolement	déprimé	fort	persécuté
agitation	dérangé	fou	perturbé
agité	désespoir	frayeur	pessimiste
agrément	désir	frustré	peur
agressif	désœuvrement	fureur	peur-panique
aimable	détester	gelé	pitié
aimé	de trop	gêne	plaisir
ambivalent	dévalorisé	gentil	positif
amertume	différent	gratitude	proche
amitié	diminué	haine	rage
amour	distant	harmonieux	raisonnable
angoisse	douleur	heureux	rancune
anxiété	doute	honte	ravissement
apathie (apathique)	doux	hostilité	reconnaissance
appréhension	drôle	humiliation	refusé
arrogant	écarté	impatience	regret
attachement(attaché)	écœuré	impuissant	rejeté
attendrissement	effroi	impulsif	reposé
béatitude	égoïsme	incompris	repoussé
bégaiement	éjecté	indifférent	réservé
bizarre	éloigné	inférieur	ressentiment
blessé	émerveillé	inquiétude	retiré
bloqué	embarras	insécurité	révolté
bonheur	emprisonné	intimidé	ridicule
boule dans la gorge	enchanté	irrité	rougissement
brisé	énervement (énervé)	jalousie-amoureuse	satisfaction
cafard	engourdissement	jalousie-envie	sensuel
calme	ennui	joie	sérénité (serein)
chagrin	enragé	jouissance	solitude (seul)
chérir	enthousiaste	jugé	soupir
choqué (colère)	envahi	loin	stressé
coincé	envahissant	malaise (mal à l'aise)	supérieur
colère	envie	mal de tête	surpris
compassion	épouvante	manipulation	sympathie
confiance	estime	manipulé	tendresse
confusion	étouffement	migraine de tension	tensions diverses
considération	étourdissement	mécontentement	terreur
contentement	euphorie	méfiance	tics
crainte	évalué	mélancolie	timide
crise de panique	évanouissement	mépris	trac
culpabilité	exaspération	mort	trahi
	excitation	nausées	transpiration excessive
	exclu	négatif	tremblement
	exécrer	nervosité	tristesse
	extase	non-désiré	vanité
	extrémités froides	nostalgie	victime
		optimiste	vide
		ouvert	violence
			vivant
			volupté

Tableau 2 : Liste d'expériences émotives [Larivey 2002]

Ceci peut être considéré comme un signe pour confirmer l'universalité des émotions primaires par le fait que leurs effets physiologiques sont plutôt universels ; cela veut dire aussi qu'il y a des tendances communes dans la corrélation entre l'acoustique et les émotions à travers des cultures différentes. A côté des études de l'universalité des émotions primaires en expressions faciales de [Ekman et al, 1971], plusieurs autres études comme celles d'[Abelin et al, 2000] et [Tickle 2000] ont réalisé des expérimentations sur le signal acoustique de la parole. Dans ces expérimentations, les auteurs ont demandé aux évaluateurs d'identifier l'émotion (la joie, la tristesse, la colère, la peur ou le calme) en se basant uniquement sur l'information acoustique (les expressions n'avaient pas de signification, donc il n'y avait pas d'informations sémantiques), les évaluateurs américains ont du identifier l'émotion de personnes soit américaines soit japonaises et, vice versa, les évaluateurs japonais ont dû décider quelles émotions des locuteurs japonais ou américains essayaient d'exprimer. A partir de ces expérimentations, deux résultats essentiels ont été produits :

- il y a seulement peu de différence de performance entre la détection des émotions exprimées dans la même langue ou dans une autre langue ; cela est vrai pour des japonais aussi bien que pour des américains ;
- la qualité de la reconnaissance des émotions chez humain est loin d'être parfaite ; le meilleur taux obtenu n'a été que de 60% seulement [Tickle 2000].

Les études précédentes sur l'universalité des émotions primaires et ce premier résultat signifient que des résultats obtenus dans le domaine de la reconnaissance des émotions sur le signal acoustique pourraient être transférables d'une langue à une autre langue. Cela donne une possibilité de portabilité des résultats obtenus entre corpus construits dans différentes langues. Cependant, il faudra vérifier les résultats ainsi qu'adapter le système aux langues spécifiques car, premièrement les résultats obtenus par l'expérimentation avec quelques paires de langues différentes ne sont pas généraux, et puis parce que beaucoup d'autres facteurs qui peuvent causer des variations doivent être étudiés, par exemple les tons d'une langue tonale.

Chapitre 3. **Travaux sur la reconnaissance de l'émotion**

A côté des études théoriques sur l'émotion, depuis ces dernières années on commence à s'intéresser à ce que l'étude de l'émotion peut apporter pour l'amélioration des performances des systèmes de reconnaissance automatique de la parole, pour l'amélioration du niveau naturel de la parole synthétisée, ainsi que dans la conception des systèmes d'interaction homme machine, etc. Effectivement, les études de l'émotion dans toutes les modalités suivent également ces deux orientations : la synthèse de l'émotion et la reconnaissance de l'émotion. C'est pourquoi, dans ce chapitre nous présenterons tout d'abord le principe général d'un système de reconnaissance de l'émotion. Puis les aspects émotifs exploités dans l'image et la vidéo seront étudiés, ainsi que l'étude de l`émotion dans le texte qui est aussi une branche proche de notre travail car les transcriptions des systèmes de reconnaissance automatique de la parole sont aussi du texte et traitées comme tel. Enfin, nous terminerons le chapitre avec les approches et les résultats obtenus ces dernières années sur la reconnaissance de l'émotion dans la bande son, ce qui est aussi l'objectif de cette thèse.

3.1. Formalisation du problème

3.1.1. Étiqueter des énoncés avec des émotions

Le problème que nous considérons est celui de l'attribution d'étiquettes d'émotion à des segments de paroles ou « énoncés ».

Deux cas se présentent selon que les émotions à reconnaître sont considérées comme exclusives ou non. Dans le premier cas, une et une seule étiquette doit être associée à chaque énoncé : on

doit faire un choix unique parmi tous les états émotionnels possibles. Dans le second cas, on décide indépendamment pour chaque étiquette si elle doit ou non être associée à l'énoncé : on fait alors un choix pour chaque état émotionnel.

On peut également attribuer pour les étiquettes soit une valeur binaire, l'émotion est alors soit présente soit absente, soit un score qui indique un degré et/ou une probabilité de présence pour chacune des émotions possibles.

Le cas le plus fréquemment rencontré dans la littérature est celui où l'on attribue de manière binaire des états émotionnels exclusifs à des énoncés. Cela n'est pas toujours réaliste car on peut trouver en pratique des cas où plusieurs émotions sont présentes simultanément, par exemple la surprise et la colère. Dans le contexte de la recherche d'information qui est le nôtre, disposer du degré de probabilité et/ou d'intensité est très utile pour ordonner les résultats de la requête par ordre de pertinence.

3.1.2. Segmenter un énoncé selon les émotions.

Dans le contexte de l'indexation des documents audio et vidéo, l'étiquetage par des émotions n'a généralement pas de sens au niveau du document complet. Il peut exister ou non des segmentations à un niveau de détail plus fin comme les tours de parole mais ces unités ne sont pas forcément homogènes non plus en ce qui concerne leur contenu émotionnel. Il se pose alors le problème de la segmentation des énoncés selon les états émotionnels présents. Il ne suffit plus de détecter la présence d'une émotion, il faut en plus déterminer où son expression commence et où elle finit. Les problèmes de l'étiquetage et de la segmentation sont interdépendants.

Nous reviendrons sur le problème de l'annotation et de la segmentation dans le chapitre portant sur les corpus ainsi que dans le chapitre sur les expérimentations avec des données réelles.

3.2. Systèmes de reconnaissance de l'émotion

3.2.1. Apprentissage supervisé

Pratiquement toutes les méthodes actuelles appliquées sur la reconnaissance de l'émotion fonctionnent selon le principe général de l'apprentissage supervisé. Un système basé sur ce principe est articulé autour d'un modèle qui est supposé représenter la relation entre des échantillons et les classes à reconnaître. Une première partie du système, le module d'entraînement ou d'apprentissage, a pour objet la production du modèle à partir d'un ensemble d'échantillons annotés. Une seconde partie, le module de prédiction ou de reconnaissance, a pour objet d'attribuer des étiquettes aux nouveaux échantillons à partir du modèle. Selon le type d'approche, la prédiction peut être une classe unique parmi plusieurs classes exclusives, une ou plusieurs classes parmi plusieurs classes indépendantes (et non exclusives), ou des scores ou des probabilités pour chacune des classes existantes, qu'elles soient ou non exclusives. Si les classes sont exclusives, il y a un seul classifieur global, dans le cas contraire, il y a autant de classifieurs indépendants les uns des autres que de classes à reconnaître.

La Figure 5 montre l'architecture générale d'un système de classification par apprentissage supervisé. Les modules d'entraînement et de prédiction sont des programmes informatiques et ils fonctionnent sans intervention humaine. Le module d'entraînement a besoin d'échantillons annotés pour produire le modèle utilisé par le module de prédiction. Les échantillons proviennent normalement du monde réel au travers d'un processus d'acquisition (non montré). Leur annotation doit être faite par un opérateur humain. Celui-ci effectue des jugements qui consistent en l'attribution de classes aux échantillons. Comme pour le module de prédiction, ces

jugements peuvent correspondre à des classes exclusives ou non et ils peuvent être faits de manière binaire ou non. L'objectif du système de classification est de produire des prédictions qui soient aussi conformes que possible au jugement de l'annotateur humain. Les évaluations de performances sont d'ailleurs faites selon ce principe comme cela est illustré dans la Figure 6 :

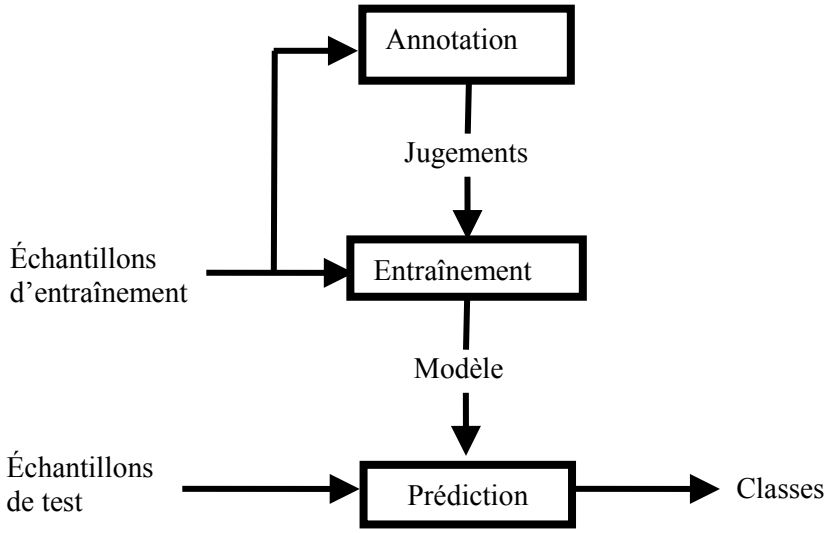

Figure 5 : Apprentissage supervisé

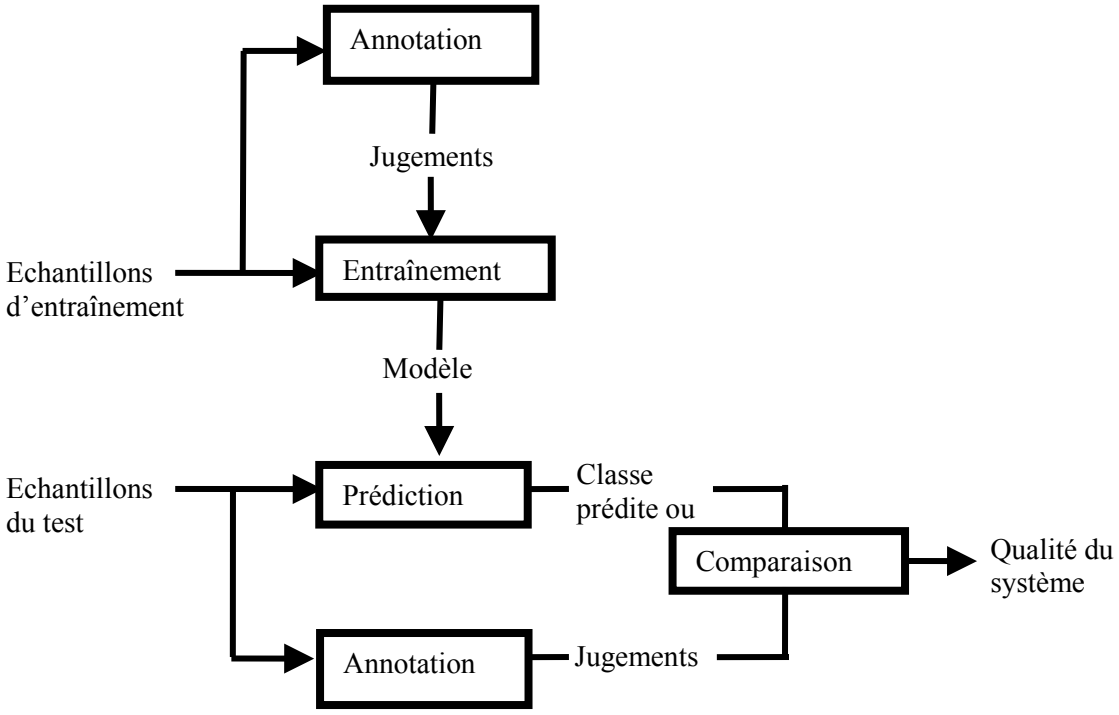

Figure 6 : Évaluation des systèmes de classification

3.2.2. Prétraitements

Les échantillons ne sont en général pas envoyés directement aux modules d'entraînement et de prédiction. Au départ, ils ne se trouvent même pas forcément sous forme informatique ou numérique. Une chaîne de prétraitements est alors mise en place. Elle peut ou non contenir une étape d'acquisition et elle vise à extraire des informations caractéristiques plus adaptées pour les processus d'entraînement et de prédiction. Un module supplémentaire est inséré dans le système en amont des deux autres modules de traitement. Les prétraitements sont pratiquement toujours les mêmes pour les deux types de modules. Ils visent en général à réduire le volume de données associées aux échantillons et à extraire des « grandeurs » plus invariantes (par rapport aux classes recherchées) que les données de départ. Dans la plupart des cas, les échantillons sont finalement représentés par des vecteurs de nombres réels de dimension fixe.

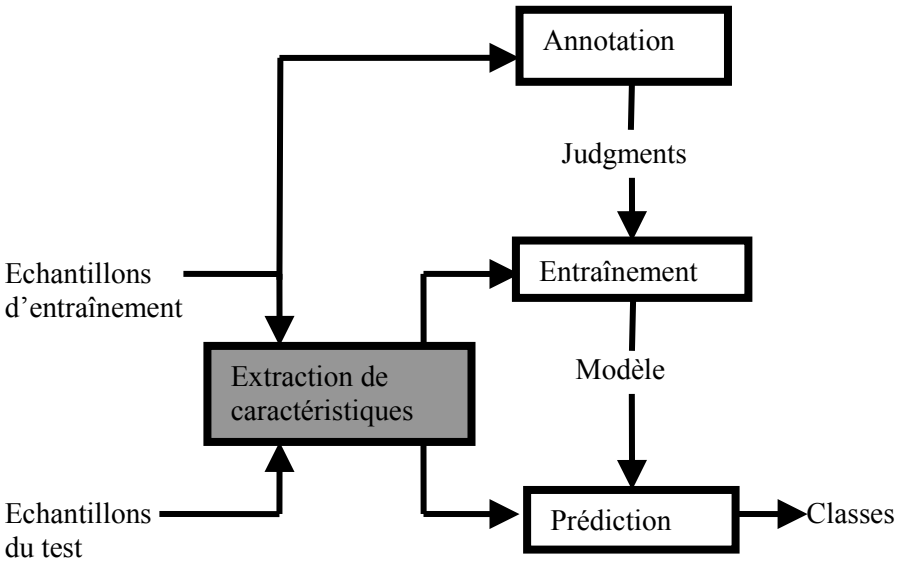

Figure 7 : Système de classification avec prétraitements par le processus d'extraction de caractéristiques

3.3. Approches dans la reconnaissance des émotions

Comme cela a été remarqué par [Salovey 1990] et [Goleman 1995], l'aptitude émotionnelle est une part essentielle de l'aptitude appelée « l'intelligence ». L'émotion contrôle et modifie presque tous les modes de communication humaine : l'expression faciale, les gestes, la posture, la tonalité de la voix, le choix des mots, la respiration, le rythme du cœur, la pression du sang, la température, l'humidité de la peau, etc. L'émotion peut aussi changer de manière significative le message transféré : parfois ce n'est pas ce qui a été dit qui est le plus important mais comment cela a été dit. Le visage tend à être la forme la plus visible de la communication émotionnelle. Il est également l'un des moyens les plus importants de communication expressive. Il est appelé « l'organe de l'émotion », il est peut-être le canal le plus puissant de communication non-verbale [Archer 1996].

Comme cela a été remarqué par [Picard 1997], l'identification des émotions est plus précise si elle se fait à partir de plusieurs modalités. Une combinaison des caractéristiques visuelles et acoustiques de bas niveau et d'un raisonnement de haut niveau à partir du langage naturel pourrait produire la meilleure inférence de l'émotion. Cependant, la réalisation d'analyseurs de l'état émotionnel humain robustes, multimodaux, adaptatifs et sensibles au contexte est loin d'être une réalité car :

- malgré son rôle important, l'émotion n'est devenue un sujet de recherche vraiment actif que ces dernières années ; ce domaine fait appel à des méthodes pour chaque modalité, pour chaque domaine, comme dans le traitement de l'audio, de la vidéo, de l'image etc., et pour l'intégration de ces modalités ;
- malgré l'ancienneté des études théoriques sur l'émotion, la compréhension des relations entre le comportement individuel et le fonctionnement du système sensoriel humain reste insuffisante.

Des essais de combinaisons de plusieurs modalités ont été conduits comme la combinaison de l'image et du son ou du texte et du son (voir la section sur la Multimodalité 3.3.4).

En parallèle avec les recherches sur la multimodalité, le développement et l'amélioration de la performance des systèmes unimodaux sont toujours nécessaires et importants. Les approches unimodales sont toujours des sujets de recherche actifs et le travail de cette thèse se place dans le cadre de l'une d'entre elles : l'identification des états émotionnels à partir du signal audio.

Comme le visage, la parole est un composant indispensable pour l'expression émotionnelle dans les conversations quotidiennes. L'information la plus importante est le contenu sémantique qui est contextuellement et directement ou indirectement transférée entre les personnes. Jusqu'à aujourd'hui, malgré des progrès énormes de la technologie dans tous les domaines, l'être humain reste toujours le seul capable de capturer parfaitement et de réagir naturellement à des informations de si haut niveau parce que l'être humain possède une capacité de fusionner toutes les sources d'informations y compris les deux types d'informations portés par le signal de la parole : l'information linguistique et l'information extralinguistique. Pour l'ordinateur, jusqu'à maintenant, le contenu sémantique n'est essentiellement obtenu qu'en analysant l'aspect linguistique par l'analyse textuelles des transcriptions. Cependant, comme précédemment mentionné, si nous ne travaillons qu'avec l'aspect linguistique nous perdons un aspect important de l'information qui contribue fortement à la signification de ce qui est dit (le contenu sémantique), c'est l'aspect « acoustique » de la parole. Cet aspect correspond à la façon dont le message a été dit.

La section 3.3.1 présente un résumé des études sur la reconnaissance de l'émotion dans les autres médias comme l'image ou la vidéo (considérée comme des séquences d'images). La section 3.3.2 présente un résumé des études sur la reconnaissance de l'émotion à partir du traitement lexical et langagier de la parole. La section 3.3.3 passe en revue les études actuelles sur la reconnaissance de l'émotion dans le média audio au niveau signal (c'est-à-dire hors aspects linguistiques). La section 3.3.4 présente quelques travaux récents utilisant une approche multimodale.

3.3.1. Détection des émotions dans les images et les vidéos

Parmi les modalités de l'expression émotionnelle mentionnées, l'expression faciale est l'élément le plus fréquemment capturé à partir des images ou des séquences d'images. Les informations utilisées sur le visage peuvent être le mouvement de la bouche, des lèvres [Söderström et al.

2004], des yeux [Busso et al, 2004], etc. L'obtention d'informations plus complexes provenant des gestes ou de la posture reste encore un défi pour la reconnaissance des émotions.

L'impulsion majeure pour l'analyse automatique des expressions faciales vient du rôle significatif du visage dans nos vies émotives et sociales. Le visage fournit des signaux conversationnels et interactifs qui clarifient notre centre d'attention et règlent nos interactions avec l'environnement et les personnes qui nous entourent [Russell et al, 1997]. De plus, les états et mouvements faciaux sont nos moyens directs et naturellement prépondérants pour exprimer les émotions [Keltner et al, 1999], [Russell et al, 1997]. Par conséquent, pratiquement toutes les études actuelles sur la reconnaissance visuelle de l'émotion travaillent sur les visages. C'est pourquoi, nous présentons dans cette partie des études portant sur l'analyse des expressions faciales pour la reconnaissance des émotions.

Depuis le début des années 70, Paul Ekman et ses collègues ont effectué des études extensives sur l'expression faciale humaine [Ekman 1994]. Ils ont étudié les expressions faciales dans différentes cultures, y compris des cultures pré-littéraires, et ils ont trouvé beaucoup de similarités dans l'expression et l'identification des émotions par le visage. Ces observations leur ont permis de conclure à l'universalité de quelques expressions faciales. On les appelle les « expressions faciales universelles ». Ce sont le bonheur, la tristesse, la colère, la peur, la surprise et le dégoût. [Ekman et al, 1986] et [Matsumoto 1998] ont aussi rapporté la découverte d'une septième expression faciale universelle : le mépris, mais ceci est encore en discussion car dans quelques travaux plus récents comme [Haidt et al, 1999], les sujets ne parviennent pas à reconnaître le mépris à partir des expressions faciales. Ceci suggère que le mépris ne devrait pas être considéré comme une émotion primaire.

Malgré la similitude importante dans les expressions faciales interculturelles, des différences ont aussi été observées. Par exemple, les japonais et les américains ont produit des expressions faciales semblables tout en regardant le même film stimulus. Mais, en présence d'autorités, les japonais étaient peu disposés à montrer leurs émotions. C'est la raison pour laquelle ils ont conclu que les expressions faciales étaient régies par « des règles » dans différents contextes sociaux. Enfin, l'observation des enfants sourds et aveugles de naissance exprimant les mêmes expressions émotionnelles permet à Ekman de conclure l'innéité des expressions faciales parce qu'il n'y a aucun moyen pour que ces enfants puissent apprendre ces comportements à travers leurs sens. La même conclusion a été retenue par Segerstrale et Molnar 1997 en observant l'expression de peur des enfants de moins de 6 mois pour des visages « méchants » ; comme ces enfants sont trop petits pour pouvoir apprendre quels visages sont « méchants », il semble possible de conclure que leurs réactions sont innées.

[Ekman et al, 1978] ont développé un système de codage des actions faciales FACS (*Facial Action Coding System* en anglais)[3] où les mouvements du visage sont décrits par un ensemble d'unités d'actions appelées AU. Physiquement, chaque AU correspond à certaines bases musculaires physiques et les expressions faciales peuvent être décrites comme une combinaison des AU. Ce système de codage des expressions faciales a été fait manuellement en suivant un ensemble de règles. Les entrées sont des images d'expressions faciales, ces images ont été prises aux moments où les expressions atteignaient un degré maximum. Ce processus est très laborieux et coûteux en temps.

Le travail d'Ekman a inspiré plusieurs chercheurs et a fourni un outil pour reconnaître des expressions faciales par traitement de l'image et de la vidéo. En extrayant des caractéristiques faciales et en mesurant les quantités de mouvements faciaux, ils ont essayé de classer différentes

[3] http://face-and-emotion.com/dataface/facs/new_version.jsp

expressions faciales. On peut trouver des travaux récents d'analyse ou de reconnaissance des expressions faciales qui utilisent ce codage dans [Kanade et al, 2000] [Pantic et al, 2005].

La reconnaissance des expressions faciales par ordinateur n'a commencé que dans les années 90. [Mase 1991] a employé le flot optique appelé OF (*Optical Flow* en anglais) pour identifier les expressions faciales. Il était l'un des premiers à utiliser les techniques de traitement d'image dans ce domaine.

[Lanitis et al, 1995] ont employé un modèle de l'apparence faciale utilisant des formes flexibles pour le codage d'image, l'identification des sujets, l'identification de la pose, l'identification du genre et la reconnaissance de l'expression faciale. [Black et al, 1995] ont employé des modèles paramétrés locaux des mouvements de l'image pour récupérer des mouvements souples. Après récupération, ces paramètres ont été utilisés dans un classifieur basé sur des règles afin de reconnaitre les six émotions primaires (actées) de l'expression faciale (la surprise, la tristesse, la colère, la joie, le dégoût, et la peur) contenues dans un corpus de 70 séquences d'images de 40 sujets, ce qui donne 128 expressions émotionnelles. Ils ont obtenu avec cette méthode un taux de reconnaissance de 92 %.

Les modèles de Markov cachés (HMM) ont aussi souvent été employés pour la modélisation et la reconnaissance visuelle des expressions. Le système construit par [Otsuka et al, 1997] peut reconnaître l'une des six émotions primaires (la colère, le dégoût, la peur, la tristesse, la joie, et la surprise) presque en temps réel (à une fréquence d'image de 10Hz) sur un corpus contenant 240 séquences d'images des expressions émotionnelles actées de 4 sujets (3 hommes et une femme). Les auteurs ont tout d'abord calculé les vecteurs de la vitesse en utilisant un algorithme de flot optique. Les coefficients obtenus par la transformée de Fourier bidimensionnelle autour de la région des yeux et de la bouche ont été employés comme vecteurs de caractéristiques dans un modèle de Markov caché afin de classifier les expressions de plusieurs sujets. Un mélange de densités a été employé pour s'adapter à la variation des expressions faciales entre les sujets. Les auteurs ont constaté expérimentalement que le mélange de densités était un facteur d'efficacité important : le taux de reconnaissance s'améliore avec l'augmentation du nombre de composantes des mélanges, (nous avons retiré la même conclusion par notre expérimentation présentée à la section 6.3 du Chapitre 6). Ces auteurs ont obtenu un taux de reconnaissance de 93 %. Une approche semblable en utilisant des caractéristiques différentes a également été suivie par [Lien 1998].

En 2003, [Cohen et al, 2003] décrivent deux méthodes de classification des expressions faciales : statique et dynamique (sur un corpus de 6 émotions actées : la joie, la surprise, la colère, le dégoût, la tristesse et la peur de 5 sujets avec 180 séquences d'images). Dans le cas statique, les images des séquences vidéo sont classifiées par un réseau bayésien. Dans le cas dynamique, elles sont classifiées en utilisant un HMM multi-niveau capable de capturer l'information temporelle. Ce dernier permet la classification des émotions mais aussi la segmentation de longues séquences selon l'émotion présente. Ils ont constaté expérimentalement que l'approche statique était plus facile à utiliser et à entraîner que l'approche dynamique mais qu'appliquée sur des séquences de vidéo, elle pouvait devenir peu fiable car l'image traitée ne correspondait pas toujours à l'état d'expression maximale. L'approche statique s'est montrée meilleure dans le cas de la reconnaissance avec plusieurs sujets car elle semblait être moins sensible aux variations interpersonnelles et aux variations temporelles.

Auteur(s)	*Traitement*	*Corpus*	*Classification et Taux de classification*
[Black et al, 1995]	Modèle paramétré avec l'approche statique	+ 6 émotions primaires actées : surprise, tristesse, colère, joie, dégoût, peur + 40 sujets + 128 échantillons contenus dans 70 séquences d'images	Basé sur des règles : 92 %
[Yacoob et al, 1996]	Flot optique avec l'approche statique	+ 6 émotions primaires actées : surprise, tristesse, colère, joie, dégoût, peur + 32 sujets + 116 échantillons contenus dans 40 séquences d'images	Basé sur des règles : 95 %
[Rosenblum et al, 1996]	Flot optique avec l'approche dynamque	+ 2 émotions actées : joie (souris) et surprise + 32 sujets	Réseaux neuronaux : 88 %
[Essa et al, 1997]	Flot optique avec l'approche dynamique	+ 5 états actées : joie (souris) surprise, colère, dégoût, tristesse et « raise-brow ». + 8 sujets + 52 séquences d'images.	Basé sur la distance : 98 %
[Otsuka et al, 1997]	FFT 2D du flot optique avec l'approche dynamique	+ 6 émotions actées : joie, surprise, colère, dégoût, tristesse, peur + 4 sujets + 240 séquences d'images	HMM : 93 %
[Lyons et al, 1998]	Filtres ondelettes de Gabor avec l'approche statique	+ 6 émotions actées : joie, surprise, colère, dégoût, tristesse, peur + 10 sujets + 219 images en expression faciale.	Basé sur la corrélation avec l'évaluation humaine, le rang de corrélation : 56.8 %
[Kanade et al, 2000]	Base de données avec le codage FACS avec l'approche dynamque	+ 6 émotions actées : joie, surprise, colère, dégoût, tristesse, peur + 182 sujets + 1917 séquences d'images	Pas de resultats
[Cohen et al, 2003]	Unité des Mouvements (MU) avec l'approche dynamque	+ 6 émotions actées : joie, surprise, colère, dégoût, tristesse, peur + 5 sujets + 180 séquences d'images	Réseau bayésien et HMM : 82 % (dépendante du sujet) et 66 % (indépendante du sujet)
[Pantic et al, 2005][4]	Base de données avec le codage FACS avec l'approche statique et dynamque	+ Plusieurs émotions actées. + 19 sujets + 1500 échantillons : images et séquence d'images.	Pas de resultats
[O'Toole et al, 2005]	Base de données avec l'approche statique et dynamque	+ Plusieurs émotions actées. + 284 sujets	Pas de resultats

Tableau 3 : Résumé de quelques études sur l'émotion en expression faciale dans l'image et dans la vidéo

[4] http://www.mmifacedb.com/

En conclusion, ces méthodes sont semblables dans leur principe à celles de tous les systèmes de reconnaissance. En effet, elles extraient d'abord des paramètres à partir du signal brut (les images), puis ces paramètres sont traités dans un système de classification dont la sortie est l'une des émotions à rechercher ou un score pour chacune des émotions. Elles diffèrent principalement par les caractéristiques utilisées ou dans la phase de traitement de l'image ou de la vidéo. On distingue généralement deux classes de traitement de l'image ou de la vidéo. La première est appelée : « *feature-based* » : le principe est de détecter et de suivre des caractéristiques spécifiques comme les coins de la bouche, des sourcils, etc. La deuxième est appelée « *region-based* » : les mouvements faciaux sont mesurés dans certaines régions du visage comme la région des yeux, la région des sourcils ou la région de la bouche. Plusieurs algorithmes de classification ont été utilisés pour reconnaître des émotions dans les images, les séquences d'images ou les vidéos. Et enfin, toutes les études portant sur la détection des émotions en se basant sur l'image et sur des sequences d'images que nous avons trouvées dans la literature utilisent les corpus non-naturels dont les émotions sont actées par les acteurs/actrices. Le Tableau 3 donne une vue globale de quelques algorithmes appliqués dans la reconnaissance des expressions faciales. Ces algorithmes fonctionnent généralement bien en comparaison relative avec la reconnaissance humaine qui est d'environ 87% comme rapporté par [Bassili 1979].

Le dernier aspect intéressant est la confusion entre les six expressions faciales des émotions primaires : la colère et le dégoût sont souvent confondus par les êtres humains, de même que la peur et la surprise. Ces confusions sont expliquées par la similitude des actions faciales [Ekman et al, 1978]. Ce problème se rencontre également dans la reconnaissance assistée par ordinateur [Black et al, 1995] [Cohen et al, 2003] [Yacoob et al, 1996] et d'autres.

3.3.2. Détection des émotions à partir du texte

Traditionnellement, les études sur l'identification des émotions à partir du texte se sont concentrées sur la découverte et l'utilisation de « mots-clés émotifs », c'est-à-dire sur la recherche de mots spécifiques qui indiquent l'état émotif du locuteur. L'utilisation de mots-clés émotifs est l'approche la plus directe et la plus souvent utilisée jusqu'à présent dans la littérature pour la reconnaissance de l'état émotionnel dans le texte.

[Yanaru 1995] a suivi des locuteurs tandis qu'ils parlaient dans un contexte naturel en utilisant des mots-clés émotifs. [Subasic et al, 2001] ont construit un groupe de mots émotifs en marquant manuellement le degré d'émotion pour chacun d'entre eux. [Boucouvalas et al, 2002] ont essayé d'extraire l'état émotionnel en temps réel dans des chats sur internet en appliquant un analyseur (*parser* en anglais) pour identifier des objets associés avec des mots-clés émotifs et avec des règles grammaticales. [Devillers et al, 2002], [Devillers et al, 2003] ont recherché les états émotionnels en calculant la probabilité conditionnelle entre les mots-clés émotifs et les états émotionnels. [Devillers et al, 2004] ont combiné plusieurs niveaux (voir la partie multimodale) : l'information lexicale avec le modèle markovien unigramme et l'information prosodique comme la fréquence fondamentale, le débit, etc.

[Tao et al, 2004] supposent aussi que le contenu émotionnel dans le texte d'une phrase est essentiellement porté par le type des mots. Ils ont donc classé les mots en deux groupes : les mots de contenu (*content words* en anglais) et les mots fonctionnels de l'émotion - EFW (*Emotion Functional Word* en anglais). Pour chaque EFW, des degrés d'association avec les émotions primaires sont définis manuellement. Pour chaque phrase d'entrée, ces valeurs sont combinées par un réseau de neurones pour produire une décision finale. Plus récemment,

[Zhang et al, 2005] ont utilisé un thésaurus semi-automatique pour améliorer la reconnaissance de la catégorie et la valeur émotionnelle de texte d'entrée.

En général, tous les systèmes basés sur des mots présentent les problèmes suivants :

- l'ambiguïté liée à la polysémie des mots-clés émotifs ;
- l'incapacité à reconnaître l'émotion dans le cas où il n'y a pas de mots-clés émotifs ;
- le manque d'informations sémantiques et syntaxiques qui affectent fortement le sens des mots.

Certains chercheurs comme [Dijkstra et al, 1994] sont allés plus loin dans l'exploitation de l'information textuelle en utilisant d'autres indices, de plus haut niveau, extraits à partir du texte comme : l'intention pragmatique, la structure des paragraphes, le degré d'imagination, l'activité et la plausibilité du contenu pour estimer et chercher la relation entre l'émotion réelle des lecteurs et l'émotion fictive des personnages dans des histoires. [Lee et al, 2007] ont prouvé également que des améliorations de la détection des états émotionnels à partir du texte sont possibles en combinant les trois types d'informations : linguistique, pragmatique et mots-clés.

[Wu et al, 2006] ont introduit une « ontologie lexicale universelle » en combinaison avec des règles de génération de l'émotion EGR (*emotion generation rules* en anglais) qui sont définies manuellement en se basant sur la psychologie. Avec les EGR, des règles d'association de l'émotion EAR (*emotion association rules* en anglais) sont automatiquement dérivées pour chaque émotion par un algorithme d'apprentissage sur un corpus annoté. L'état émotionnel des phrases d'entrée est estimé par la similitude de leurs EAR avec les EAR de chaque état émotionnel en utilisant un modèle de mélange séparable SMM (*separable mixture model* en anglais) (voir [Wu et al, 2006]). Enfin, les auteurs ont expérimentalement constaté les apports de l'application de la psychologie et du modèle de mélange séparable.

Un exemple d'une règle EAR pour la phrase : *i have got less than 100 dollars*

[NEGATIVE] + [NUM] → UNHAPPY où [NEGATIVE] est le libellé sémantique correspondant avec « *less* than » qui est déterminé en se basant sur les EGR, [NUM] est l'attribut correspondant avec « 100 ».

En conclusion, la reconnaissance de l'émotion à partir du texte est encore un sujet très ouvert. L'extraction et l'utilisation d'informations sémantiques de plus haut niveau, de connaissances psychologiques, de connaissances contextuelles et la prise en compte des progrès dans le traitement du langage naturel comme les réseaux sémantiques [Woods 1970] pour enrichir, pour symboliser et pour désambiguïser [Chan et al, 1998] sont encore des sujets de recherches.

3.3.3. Détection des émotions dans le signal acoustique

Le deuxième composant de base et indispensable de la communication orale est le contenu non-verbal. Celui-ci correspond à la façon dont le message a été prononcé et il porte divers types d'informations. Si nous ne considérons que la partie verbale comme dans la section précédente, nous pouvons manquer des aspects importants de la communication et même mal comprendre le message. Contrairement à la reconnaissance de la parole qui a connu des avancées significatives dans ces dernières décennies, la reconnaissance des émotions contenues dans la parole n'a fait l'objet de recherches importantes que ces dernières années [Bosh 2000].

Dans cette partie, nous passerons en revue les principales études ainsi que leurs résultats portant sur la reconnaissance des émotions. Celles-ci utilisent le même genre de caractéristiques que celles qui sont les plus fréquemment utilisées dans le domaine de la reconnaissance de la parole. Elles utilisent des techniques de classification comme les modèles de Markov cachés (HMM),

les réseaux neuronaux artificiels (NN), l'analyse discriminante linéaire (LDA), les K plus proches voisins (KNN) et les machines à vecteurs de support (SVM).

Comme précédemment présenté, dans la littérature les techniques ou les modèles utilisés pour la reconnaissance en général, et de l'émotion en particulier, peuvent être divisés en deux catégories :

- les techniques « statiques » qui capturent des caractéristiques statistiques du signal comme : les K plus proche voisins (KNN), la quantification vectorielle (VQ), le modèle de mélange des gaussiennes (GMM) ;
- les techniques « dynamiques » qui capturent aussi des caractéristiques temporelles du signal comme : les réseaux neuronaux (NN), les modèles de Markov cachés (HMM), la programmation dynamique DTW (*Dynamic Time Warping* en anglais) etc.

Les modèles statiques peuvent utiliser ou non la fonction de densité de probabilité (*pdf*) (*probability density function* en anglais).

Etant donné $y = (y_1, y_2, \ldots, y_D)$ le vecteur contenant des caractéristiques extraites d'un échantillon e de la collection où D est le nombre des caractéristiques.

Selon la classification de Bayes, un échantillon e est assigné à une classe Ω_{c*} si :

$$c^* = \arg\max_{c=1}^{C} \{P(y_e \mid \Omega_c) P(\Omega_c)\} \tag{3.1}$$

Où C est le nombre de classes, $P(y|\Omega_c)$ est la densité de probabilité de y étant donné la classe Ω_c et $P(\Omega_c)$ est la probabilité à priori de la classe Ω_c (par exemple un état émotionnel). $P(\Omega_c)$ représente donc la connaissance que nous avons sur la classe d'un échantillon avant la mesure du vecteur de cet échantillon.

Il y a plusieurs méthodes pour modéliser $P(y|\Omega_c)$. Par exemple le modèle de gaussienne ou le modèle de mélange des gaussiennes (GMM) supposent que les vecteurs y du signal d'entrée appartenant à Ω_c sont distribués selon une distribution gaussienne. L'expression (3.2) montre la fonction gaussienne multi-variable simple :

$$P(y \mid \Omega_c) = g(y, \mu_c, \Sigma_c) = \frac{e^{\left[-\frac{1}{2}(y-\mu_c)^T \Sigma_c^{-1} (y-\mu_c)\right]}}{(2\pi)^{D/2} \mid \det(\Sigma_c) \mid^{1/2}} \tag{3.2}$$

où μ_c et Σ_c sont respectivement le vecteur moyen et la matrice de covariance, **det** est le déterminant de cette matrice.

Dans le cas où la distribution des données est inconnue, la méthode dite des fenêtres de Parzen permet d'estimer la fonction *pdf* d'une variable aléatoire :

Pour chaque classe Ω_c, il y a un ensemble d'échantillons associés $T_c = \{ e_c ; 1 \le c \le C \}$. Étant donnés ces échantillons, l'estimation par des fenêtres de Parzen permet d'extrapoler des données pour d'autres valeurs $P(y_c|\Omega_c)$ de y_c. L'idée de cette méthode est : il est certain que $P(y_{c\xi}|\Omega_c) \neq 0$ au point $y_{c\xi}$ où $y_{c\xi}$ est le vecteur des caractéristiques d'un échantillon e_c associé à la classe Ω_c. Puisque la fonction densité de probabilité *pdf* de la classe sortie est continue, on s'attend à ce que $P(y_c|\Omega_c)$ dans le voisinage de $y_{c\xi}$ ne soit pas nul. Plus on s'éloigne du $y_{c\xi}$, moins nous pouvons prédire $P(y_c|\Omega_c)$. Donc, la connaissance de $P(y_c|\Omega_c)$ peut être obtenue en ne représentant

le y_c par une fonction appelée le noyau (*kernel* en anglais). La fonction noyau $h(.)$ peut être n'importe quelle fonction de $R+ \rightarrow R+$ qui donne le maximum au point y_c et qui augmente monotoniquement quand y_c s'approche $y_{c\xi}$. Etant donnée $d(y_c, y_{c\xi})$ être la distance entre y_c et $y_{c\xi}$ (Euclidien, de Mahalanobis, etc.), la fonction *pdf* de la classe de sortie Ω_c est estimée par :

$$P(y_c \mid \Omega_c) = \frac{1}{N_c} \sum_{y_{c\xi} \in \Omega_c} h(d(y_c, y_{c\xi})) \tag{3.3}$$

[Ververidis et al, 2004-2] ont obtenu 53 % de taux de classification correcte en travaillant avec 5 états émotionnels actés (la joie, la tristesse, la surprise, la colère et le neutre) du corpus DES de 4 locuteurs (2 hommes et 2 femmes) (voir le Chapitre 4), par la classification de Bayes et une fonction *pdf* estimée par des fenêtres de Parzen.

Le Tableau 4 présente quelques études dans cette branche.

Classificateurs		***Auteurs***
Avec la fonction pdf	Classificateur bayésien utilisant la fonction pdf gaussienne	[Dellaert et al, 1996]
	Classificateur bayésien utilisant la fonction pdf gaussienne avec l'analyse discriminante linéaire	[France et al, 2000], [Lee et al, 2005]
	Classificateur bayésien utilisant la fonction pdf gaussienne estimée par des fenêtres de Parzen	[Dellaert et al, 1996], [Ververidis et al, 2004-2]
	Classificateur bayésien utilisant la fonction pdf de mélange des Gaussiennes	[Slaney et al, 1998] [Schüller et al, 2004] [Jiang et al, 2004-2] [Ververidis et al, 2005-1]
Sans fonction pdf	K plus proches voisins	[Dellaert et al, 1996] [Petrushin 1999] [Picard et al, 2001]
	Machine à vecteurs de support	[McGilloway et al, 2000] [Fernandez et al, 2003] [Kwon et al, 2003]
	Réseaux neuronaux artificiels	[Petrushin 1999] [Tato 2002] [Shi et al, 2003] [Fernandez et al, 2003] [Schüller et al, 2004]

Tableau 4 : Classificateurs pour la reconnaissance des émotions.

Les caractéristiques utilisées avec ces techniques peuvent être des caractéristiques de la prosodie, des MFCCs, des coefficients spectraux, etc. (Voir la section 5.1 pour le détail).

[Dellaert et al, 1996] ont obtenu ~70 % pour le taux de reconnaissance correcte entre les 4 états émotionnels (joie, tristesse, colère, peur) d'un corpus de 1000 énoncés actés par 5 locuteurs en utilisant la technique des K plus proches voisins (KNN). Les inconvénients de cette technique sont la difficulté de la recherche des méthodes systématiques permettant de choisir le nombre optimal de voisins les plus proches, ainsi que la difficulté dans la détermination de la meilleure

mesure de distance appliquée dans ce modèle. En utilisant la fonction *pdf* et la formule de Bayes sur les statistiques de l'énergie et du contour de la fréquence fondamentale, ces auteurs ont aussi obtenu le taux de reconnaissance ~56 %.

Avec le modèle de mélange des gaussiennes GMM, on suppose que les vecteurs de caractéristiques y d'un état émotionnel Ω_c sont distribués en faisceaux, et les vecteurs dans chaque faisceau suivent une gaussienne. ~79 % est alors le taux de classification correcte obtenu pour les 3 états affectifs (l'approbation, l'attention, la prohibition) de la parole en utilisant le modèle de mélange de gaussiennes multidimensionnel par [Slaney et al, 1998] pour modéliser la fonction de densité de probabilité $P(y|\Omega_c)$ du contour de F_0, de l'énergie et des formants. Ces auteurs ont travaillé sur un corpus de 500 énoncés obtenus par l'enregistrement de la parole de 12 parents (6 pères et 6 mères) parlée à leurs enfants dans la salle d'enregistrement. En comparaison avec les corpus construit en utilisant des acteurs ou actrices, à notre avis, ce type de corpus est plus naturel car des expressions émotionnelles sont obtenues dans le contexte et elles ne sont pas donc trop exagerée. Par contre, les trois états étudiés (l'approbation, l'attention, la prohibition) ne sont pas très repandus dans les autres édudes et c'est une des difficultés pour les mettre en comparaison.

En conclusion, l'avantage du modèle GMM est qu'il pourrait modéliser plusieurs locuteurs par des faisceaux. Mais l'inconvénient de ce modèle est que l'algorithme EM ne converge qu'à un optimum local.

Avec la classification discriminante linéaire [Lee et al, 2005] ont obtenu un taux de reconnaissance correcte remarquable de 93 % en travaillant avec les 2 classes émotives (neutre/non-neutre) et en utilisant des statistiques du contour de la F_0 et d'énergie. Effectivement, les données que ces auteurs ont utilisées se composent de 7200 quasi-réel énoncés qui sont obtenus en enregsitrement les appels des utilisateurs à un centre d'appel. Cependant, selon [Fukunaga, 1990], la technique de l'analyse discriminante linéaire LDA possède son inconvénient : si la fonction *pdf* de chaque émotion dans l'espace Y n'est pas une gaussienne ou sysmétrique, LDA ne parvient pas à trouver les directions discriminantes.

La machine à vecteurs de support sépare les états émotionnels avec une marge maximale. L'avantage de cette technique de classification est la capacité d'être étendue à la séparation non-linéaire par quelques techniques de la fonction noyau. [Fernandez et al, 2003] ont pu atteindre 61.2 % pour le taux de reconnaissance correcte dépendante du locuteur, 51.2 % pour le taux de reconnaissance correct indépendante du locuteur, en utilisant cette technique sur les 4 états affectifs (FF, FS, SF, SS) des énoncés. Ces 4 états sont definis en se basant sur le contexte où les conducteurs se trouvaient : la vitesse de conduite rapide ou lente, la vitesse de réponse rapide ou lente, par exemple, FF correspond avec l'état du conducteur qui conduit en haute vitesse et qui doit répondre en même temps aux questions apparues en haute fréquence, SS correspond avec l'état du conducteur qui conduit en basse vitesse et qui doit répondre aux questions apparues en basse fréquence. Les caractéristiques que ces auteurs ont utilisées sont 20 coefficients d'énergies des bandes fréquentielles extraits en appliquant l'opérateur TEO (Teager Energy Operator) sur les sous-bandes (voir [Fernandez et al, 2003]). Comme les études de [Slaney et al, 1998] (voir ci-dessus), le corpus utilisé dont les 4 etats affectifs se trouvent dans un contexte assez particulier et donc, la comparaison avec les autres approches est presque impossible.

Récemment, [Vidrascu & Devillers, 2007] ont choisi ce modèle pour la comparaison de l'efficacité de plusieurs types de paramètres paralinguistiques appliqués sur une base des données audio réelles en français avec 5 émotions : la peur, la colère, la tristesse, le neutre et la tension (relief en anglais). Selon les auteurs, les 25 meilleurs paramètres choisis avec la

connaissance dérivée de la transcription orthographique donne le meilleur taux de reconnaissance : 56%.

Comme les SVM, des classificateurs basés sur les réseaux neuronaux artificiels (ANN) ont également été utilisés dans la classification des émotions en raison de leur capacité de trouver des frontières non-linéaires séparant les états émotifs. [Fernandez et al, 2003] ont également obtenu un taux de reconnaissance de 50.6 % pour ces 4 états émotionnels en utilisant les mêmes paramètres de caractéristiques que ceux qui sont testés avec le SVM.

Avec les ANN, [Schüller et al, 2004] ont aussi obtenu un taux de reconnaissance de 90 % pour les 7 états émotionnels (la colère, le dégoût, la peur, la joie, le neutre, la tristesse et la surprise) dans le cas de reconnaissance dépendante du locuteur sur le total de 2829 énoncés émotionnels actés par 12 locuteurs et une locutrice, et 73,15 % pour la reconnaissance indépendante du locuteur. Pour ce résultat, ils ont utilisé l'ensemble des paramètres globaux statistiques de la prosodie et du spectre de signal de la parole.

Le Tableau 5 donne la comparaison de l'efficacité de plusieurs modèles (réalisés par plusieurs équipes de recherche sur la même base de données quasi-réelle AIBO avec 4 états émotionnels d'enfants : la colère, « motheres », l'emphase et le neutre) reportée par les études de [Batliner et al, 2006]. Selon les auteurs de ces études, les taux de reconnaissance sont faibles et peuvent être encore largement améliorés. L'objectif de ces études est d'établir que la fusion des meilleurs paramètres ou des sorties des différents modèles peut conduire à des améliorations de la performance de classification (voir la citation pour le détail).

Equipe de recherche	***N. de paramètres selectionnés / N. de paramètres original***	***Type de parmètres***	***Modèle***	***Taux de classification***
FAU	87 / 303	posodique, POS, lexical	Réseau neurone	55,3
TUM	103 / 980	prosodique, spectral, MFCC, POS, lexical, recherche génétique	SVM	56,4
ITC	32 / 32	prosodique, POS	Random Forest	55,8
UKA	25 / 1320	prosodique, MFCC, lexical	Linear Regressor	54,8
UA	84 / 1289	prosodique, spectral, MFCC	Naive Bayes	52,3
LIMSI	26 / 76	prosodique, spectral, POS, lexical	SVM	56,6
TAU	24 / 24	prosodique	Rule-based	46,6

Tableau 5 : Paramètres et modèles de classification [Batliner et al, 2006]

Un grand problème rencontré lors du travail avec les données réelles est l'existence de plusieurs émotions dans le même énoncé. [Vidrascu & Devillers, 2005-1] et [Vidrascu & Devillers, 2005-2] font face à ce problème en permettant aux annotateurs de choisir en même temps deux états émotionnel : « major » et « minor » pour chaque énoncé. Le traitement de la cohérence des annotations intra-annotateur et inter-annotateur est ensuite proposé lors de fusion des annotations. Les auteurs de ces expérimentations donnent la conclusion de la haute performance

du système si on prend en compte de la cohérence des annotations (en choisisant seulement des énoncés qui n'a pas l'ambigüité entre des annotations) : 80% est le taux de détection entre les deux états négatif contre neutre ou peur contre neutre.

Le réseau des neurones permet non seulement d'exploiter l'aspect statistique du signal, il permet aussi de capturer l'évolution du signal en fonction de temps par l'analyse des vecteurs des caractéristiques à court terme. Cette technique a obtenu un taux élevé de classification du type émotif/neutre (plus de 90 %) des trois états (colère, *lombard* et bruyant) parmi 10 états affectifs en utilisant des régions de « *cross-sectional* » du conduit vocal [Womack et al, 1996]. *Cross-sectional* est une mesure de la distance entre le palais souple et le palais dur des parties du conduit vocal (voir [Womack et al, 1996] pour le détail). Grâce au traitement de l'émotion, une amélioration plus de 10 % du taux de reconnaissance d'un système de reconnaissance de la parole a aussi été retenue dans ce travail. Le corpus SUSAS sur lequel [Womack et al, 1996] ont les résultats contient 16 000 énoncés produits par 44 locuteurs (14 femmes et 30 hommes).

Dans le même objectif, [Womack et al, 1999] ont utilisé un HMM multi-canal, premièrement pour la classification de la parole affective et, deuxièmement, pour améliorer la reconnaissance de la parole avec une collection ne contenant que 35 mots parlés dans 4 états affectifs. L'idée de cette approche est d'utiliser des modèles HMM traditionnel (canal unique) correspondant avec des états affectifs différents pour construire le nouveau modèle HMM (multi-canal) où la transition entre des canaux (entre des états) peut être obtenue par l'entrainement proposé par [Womack et al, 1999]. Effectivement, l'étude s'appuie sur le fait que sous des conditions spécifiques, les phonèmes peuvent se trouver sous différentes formes et le modèle HMM traditionnel n'est pas conçu pour pouvoir s'adapter à ce changement. Par exemple, si l'on utilise ce modèle pour la reconnaissance du mot HELP : sous l'effet *lombard*, les phonèmes /H/ et /P/ peuvent se trouver dans les autres canaux (états) que les phonèmes /E/ et /L/ et la possibilité de transition entre les canaux (états différents) permet au système de capturer toujours ce mot. Le taux correct de classification de l'émotion avec ce modèle est de 57.6 % en utilisant des coefficients MFCCs, ce qui est presque égal à 58.6 % de taux de classification de la parole affective obtenue en utilisant un HMM à canal unique et avec les mêmes coefficients MFCCs. La raison de la détérioration de la performance peut s'expliquer par la petite taille de la collection. Effectivement, le HMM multi-canal a pu obtenir le taux de reconnaissance de la parole jusqu'à 94,4 % tandis que le HMM à canal unique n'a obtenu que 78,7 % pour la même tâche. Cette grande différence de performance dans la reconnaissance de la parole peut être une indication de l'efficacité du modèle multi-canal en travaillant avec des grandes collections.

Une autre technique utilisée par [Fernandez et al, 2003] pour classifier les 4 états des conducteurs (voir ci-dessus) est le mélange des HMMs. L'idée de ce modèle est d'utiliser un ensemble de N modèles HMM pour capturer l'information de l'évolution temporelle des N partitions de données (voir la Figure 8) où $\sum_{1}^{N} \alpha_i = 1$.

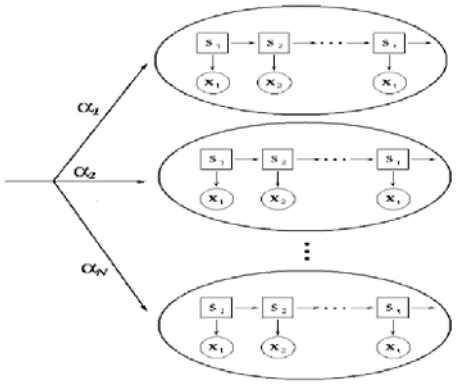

Figure 8 : Mélange des HMMs [Fernandez et al, 2003]

Le taux correct de la classification obtenu avec cette technique est de 62 %. Avec les mêmes caractéristiques, un modèle HMM à canal unique a donné un taux correct de 10 % inférieur.

Dans cette partie, les principales approches pour la classification des états émotionnels dans la parole et la description des résultats obtenus ont été passées en revue. En général, les résultats de la littérature ne sont pas directement comparables les uns avec les autres parce qu'ils ont été expérimentés sur plusieurs collections de données en utilisant des protocoles différents. Par conséquent, sans compter sur l'indisponibilité d'une telle collection commune de volume adéquat, des protocoles expérimentaux communs devraient être considérés et définis dans le futur. Des campagnes de compétitions de ce type mériteraient une attention, comme par exemple le lancement des campagnes de NIST comme TREC, TRECVID, FERET, etc.

Essayer de combiner plusieurs modalités pour augmenter la performance de la reconnaissance émotionnelle est une direction naturelle. Dans la partie suivante, nous parlerons de quelques recherches récentes portant sur cette tentative de combinaison.

3.3.4. Multimodalité

Comme précédemment présenté, on traite la parole séparément par deux modalités : la modalité du signal acoustique et celle de la transcription ou du texte. La combinaison entre ces deux aspects est une approche naturelle, mais en réalité, cette combinaison n'a attiré l'attention des chercheurs que depuis quelques années seulement.

L'information extraite à partir du texte peut être de plusieurs niveaux : le niveau lexical, le niveau grammatical, le niveau sémantique, etc. Des études récentes montrent que l'information émotionnelle existe pour tous les niveaux (voir la section précédente).

[Devillers et al, 2003], [Devillers et al, 2004] ont effectué des premiers essais portant sur la détection basée sur plusieurs niveaux de l'information mais séparément : le signal acoustique et l'information lexicale. L'information de l'émotion dans le texte de la collection est annotée manuellement par des annotateurs sans écouter le signal acoustique qui entraîne le modèle markovien unigramme. L'émotion portée par une phase inconnue *u* est déterminée par le modèle *E* qui obtient la meilleure probabilité à posteriori $P(u/E)$:

$$\log P(u/E) = \frac{1}{L_u} \sum_{w \in u} tf(w,u) \log \frac{\lambda P(w/E) + (1-\lambda) P(w)}{P(w)} \qquad (3.4)$$

où $P(w/E)$ est la probabilité d'un mot w sachant le modèle d'émotion E, $P(w)$ est la fréquence d'un mot dans le modèle général obtenu sur l'ensemble du corpus d'entraînement, $t_f(w,u)$ représente la fréquence d'un mot dans la phrase, et L_u est la longueur de la phrase en nombre de mots.

Au niveau acoustique, ces auteurs ont employé les paramètres relatifs aux variations du contour mélodique de la phrase (variation de F_0). Les derniers résultats obtenus par [Devillers et al, 2006] sont de 78 % et 60 % respectivement en utilisant l'information lexicale et l'information acoustique pour les 4 états émotionnels (la colère, la peur, la tristesse et le « soulagement » (*relief* en anglais) d'un corpus appelé CEMO. C'est un corpus de données réelles enregistrées dans un centre d'appel pour l'assitance medicale. Les auteurs supposent qu'il y a un moyen pour fusionner ces deux modalités para-lingusitque et textuelle afin d'améliorer la performance de classification car ils constatent que la colère n'est pas très bien reconnue par le modèle lexical et la tristesse n'est pas bien reconnue par le modèle para-linguistique. Effectivement, [Devillers et al, 2005] ont experimentalement contaté une amélioration de 5 % sur le taux de classification les deux états neutre / négative en combinant linéairement les deux modalités : textuelle et para-linguistique ; les auteurs ont travaillé avec un corpus réel de 100 dialogues obtenus à partir d'un centre d'appel.

[Schüller et al, 2004] ont combiné l'information acoustique et l'information linguistique pour extraire l'état émotionnel d'une expression d'entrée parmi les 7 émotions standards : la colère, la joie, le dégoût, la tristesse, la surprise et le neutre sur le total de 2829 énoncés émotionnels actés de 12 locuteurs et une locutrice. Un réseau neuronal multicouche MLP a été employé, possédant 14 nœuds d'entrée dont 7 nœuds sont la sortie du modèle acoustique et les 7 autres nœuds sont la sortie du modèle linguistique. Ce modèle MLP a également 7 nœuds de sortie correspondants aux 7 états émotionnels pour la reconnaissance. Un taux d'erreur de 8 % est la performance que la combinaison des deux aspects permet d'obtenir, à comparer avec les taux d'erreur de 25,8 % et 40,4 % qui sont respectivement les résultats obtenus en utilisant séparément l'aspect acoustique et l'aspect linguistique.

De même, [Lee et al, 2005] a combiné trois aspects de la parole : l'information acoustique, l'information lexicale et l'information du discours comme montré dans la Figure 9. En utilisant des conversations émises dans le cadre d'un centre d'appel, les étiquettes de discours sont basées sur la classification des réponses de l'utilisateur en 5 catégories : le rejet, la répétition, la reconstruction de la phrase, la demande d'aide ou la demande de recommencement, et aucune de celles qui précèdent. Le corpus sur lequel ces auteurs ont conduit leus expérimentations contient des données réelles obtenues à partir de 1187 appels à un centre d'appel automatique, au total 7200 énonés.

L'équation (3.5) illustre le calcul correspondant à la Figure 9 où x est un énoncé d'entrée, y_n est la sortie correspondante du $n^{\text{ème}}$ canal, $P(E_k|x)$ est la probabilité pour que x appartienne à l'émotion E_k. $P(E_k|x)$ a été simplement calculée par la moyenne des sorties y_n de tous les canaux combinés.

$$P(E_k \mid x) = \frac{1}{N}\sum_{n=1}^{N} y_n(E_k \mid x) \qquad (3.5)$$

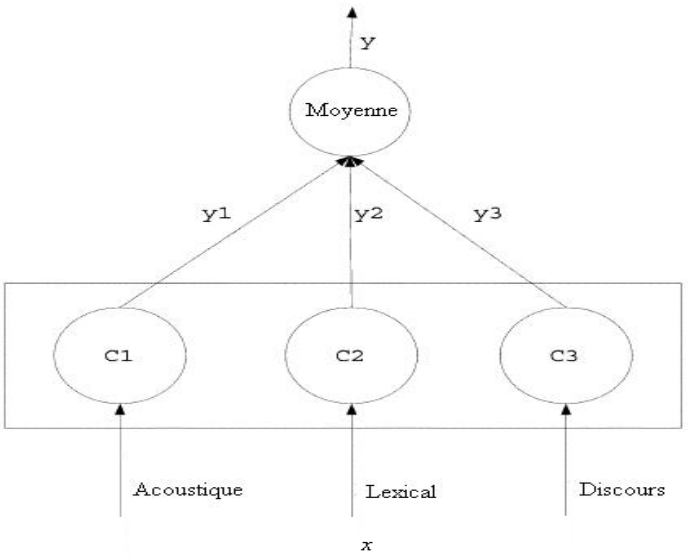

Figure 9 : Combinaison des trois canaux d'information en prenant la moyenne de chaque canal. [Lee et al, 2005]

[Lee et al, 2005] ont constaté que de façon générale, la combinaison d'autres sources d'informations avec l'information acoustique apporte des améliorations de la performance de la reconnaissance de l'émotion et que le cas de la combinaison d'informations acoustiques et lexicales a montré la meilleure performance dans presque tous les cas. L'inclusion d'information de discours ne semble pas donner une amélioration significative une fois utilisée en même temps que les informations acoustiques et lexicales. Ceci peut être dû au fait que l'information lexicale est fortement corrélée avec l'information mesurée de discours. 40.7 % pour des hommes et 36.4 % pour des femmes en travaillant avec les deux états négatif/non-négatif de l'émotion sont les performances de la reconnaissance que ces auteurs ont pu améliorer.

De même, il y a des essais sur la combinaison des deux modalités visuelle et audio depuis 1996 pour l'aspect émotionnel. [Pelachaud et al, 1996] ont essayé de construire un système qui produit des expressions faciales animées pour la parole synthétique. [Tao et al, 2004] ont aussi essayé de produire l'émotion par la combinaison des deux aspects : audio et visuel. Mais, comme nous le voyons, ces travaux ont seulement porté sur l'aspect de reproduction synthétique et pas sur la reconnaissance des émotions.

L'étude de [Chen et al, 1998] est l'une des premières études sur la combinaison les deux modalités en utilisant 16 paramètres de prosodie du canal audio et 15 paramètres des mouvements des yeux, des joues, et de la bouche du canal vidéo. Les auteurs ont combiné les deux modalités en concaténation simple des deux vecteurs des caractéristiques des deux canaux pour obtenir un nouveau vecteur bimodal des caractéristiques. Les techniques des k plus proches voisins et le modèle gaussien ont été utilisés, et ont respectivement donné 97,2 % et 94,4 % de taux de reconnaissance correcte. D'après les expérimentations de ces auteurs, ces résultats ont été énormément améliorés grâce à l'utilisation de l'information bimodale (en monomodal : 75 % pour l'audio et 69.4 % pour la vidéo). Ces résultats ont été obtenus avec les 6 émotions primaires : le bonheur, la tristesse, le dégoût, la colère, la surprise et la peur.

Travaillant sur les mêmes émotions mais sur un nouveau corpus construit pour eux-mêmes, [De Silva et al, 2000] ont proposé une méthode basée sur les règles pour une classification des données audiovisuelles d'entrée dans l'une de ces six catégories d'émotions. Les mouvements

et la vitesse de certains signes faciaux (des lèvres, de la bouche, des sourcils) sont détectés par la technique des flots optiques. La fréquence fondamentale est la caractéristique utilisée dans le canal audio. D'après les expérimentations de ces auteurs, la classification sur le canal vidéo est meilleure que celle sur le canal acoustique, et l'approche bimodale donne toujours les meilleurs résultats.

A la même époque, [Chen et al, 2000] ont également proposé un ensemble de méthodes pour la classification des données audiovisuelles dans l'une des six émotions primaires. Ils ont rassemblé les données venant de l'interprétation de cinq sujets. Considérant le fait que dans les données enregistrées, une expression faciale pure peut se produire bien avant ou après la phrase parlée, les auteurs ont utilisé principalement des paramètres acoustiques pour déterminer l'état émotionnel du sujet (classification uni-modale), mais le traitement dans le canal vidéo est aussi fusionné pour donner la reconnaissance finale. Avec les six émotions primaires, la fusion a donné un taux de reconnaissance d'environ 50 % en moyenne. La prosodie (l'énergie, la fréquence fondamentale et la vitesse de la parole) est la caractéristique utilisée dans le canal audio. Les six unités d'action (AUs) sont des paramètres extraits à partir des mouvements de la bouche, des lèvres, des sourcils, des joues, des paupières dans le canal vidéo.

Jusqu'ici, la technique utilisée est le traitement séparé uni-modal des canaux d'information, la combinaison n'étant effectuée qu'à la fin du traitement. Ceci est presque certainement incorrect car, dans la conversation quotidienne, les aspects audio et visuel sont des signaux communicatifs d'une façon réciproquement complémentaire mais aussi redondante. [Chen et al, 1998] ont expérimentalement montré cette conclusion. Pour les expérimentations, les auteurs ont travaillé sur un corpus de données actées par les 5 sujets qui fait au total 36 clips vidéo/audio synchronisés (6 clips pour un état émotionnel) de 6 émotions de base : la joie, la tristesse, la colère, le dégoût, la surprise et la peur.

En effet, afin d'accomplir une analyse multimodale comme l'être-humain, les signaux multimodaux à l'entrée ne peuvent pas être considérés comme mutuellement indépendants et ne peuvent pas être combinés de manière non-contextuelle. Au contraire, les données d'entrée devraient être traitées dans un espace commun et avec un modèle contexte dépendant. Sans compter les problèmes du contexte et les problèmes du développement de modèles de contexte-dépendants, on doit faire face au problème de la taille de l'espace des caractéristiques : une grande dimensionnalité, une grande différence de format et une grande différence dans la résolution du temps. Comme solution à ce problème, [Sebe et al, 2005] ont théoriquement proposé une topologie d'un réseau bayésien qui combine les deux modalités de la manière illustrée dans la Figure 10.

Le nœud supérieur est une variable de classe émotionnelle. Il est affecté par les expressions faciales identifiées, par les expressions vocales identifiées, par des mots-clés identifiés comme possédant une signification affective, et par le contexte dans lequel le système fonctionne (si c'est disponible). La reconnaissance émotionnelle de l'expression faciale est affectée par une variable qui indique si la personne est en train de parler ou non en se basant sur les mouvements de la bouche et sur des caractéristiques extraites à partir du canal acoustique. Les paramètres du réseau peuvent être appris par des données ou peuvent être manuellement configurés pour quelques variables. Avec cette approche, l'état émotionnel du sujet peut être détecté même lorsque certaines informations sont absentes, par exemple, le signal acoustique est trop bruyant ou la trace du dépistage de visage est perdue.

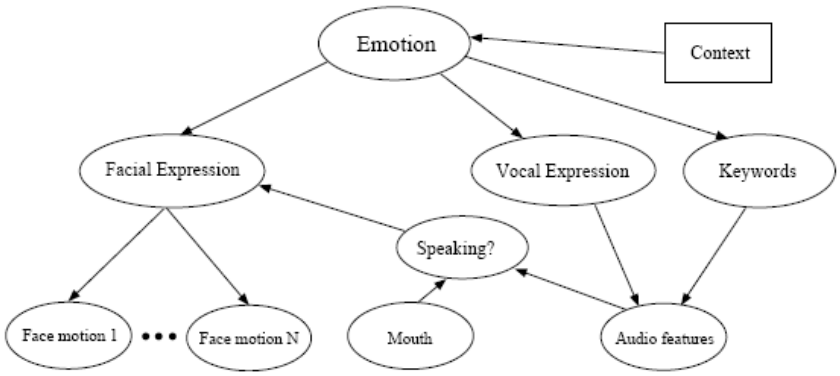

Figure 10 : Typologie du réseau bayesien pour la reconnaissance émotionnelle bimodale proposée par [SEBE et al, 2005]

3.3.5. Conclusion

Dans ce chapitre, nous avons passé en revue plusieurs approches et techniques appliquées dans la reconnaissance émotionnelle. Elles peuvent être conçues pour une seule modalité : pour le signal audio, pour l'information textuelle, linguistique, ou pour l'information imagière, etc. Mais elles peuvent aussi être des techniques qui cherchent à améliorer la performance de reconnaissance en utilisant l'information multimodale comme la combinaison entre le signal audio avec l'information linguistique, la combinaison entre l'audio et la vidéo, ou la combinaison des trois aspects.

En conclusion, l'émotion est actuellement toujours une nouvelle branche dans tous les domaines de recherches avec des progrès ces dernières années dans le traitement de l'audio ainsi que le traitement d'image, de la vidéo, mais il faut encore beaucoup de travaux dans chaque domaine, chaque modalité pour pouvoir améliorer systématiquement la performance des systèmes travaillant avec l'émotion. Notre travail s'intéresse au signal acoustique. Dans le Chapitre 4 suivant, nous présenterons une étape de base mais indispensable pour tous les systèmes de reconnaissance : la collection et la préparation des données.

Chapitre 4. **Corpus**

4.1. Introduction

Un corpus est un ensemble de documents, artistiques ou non (textes, images, audio, vidéos, etc.), regroupés dans une optique précise. On peut utiliser des corpus dans différents domaines : études littéraires, linguistiques, scientifiques, etc.

Les corpus sont essentiels pour l'entraînement et l'évaluation des systèmes de reconnaissance en général et des systèmes de reconnaissance de la parole en particulier. Dans le domaine de reconnaissance de la parole, des corpus et des systèmes sont déjà disponibles pour la plupart des langues occidentales et orientales comme l'anglais, le français, l'espagnol, le chinois, le japonais, le coréen, l'arabe, etc. Cependant, les aspects extralinguistiques, comme l'émotion, ont été assez peu considérés jusqu'ici. Nous avons besoin de corpus spécifiques pour notre travail sur la reconnaissance des états émotionnels du locuteur. Construire des corpus contenant les émotions est une tâche difficile et fastidieuse. Les corpus doivent être à la fois importants et naturels. Ceux qui sont naturels sont aussi difficiles à produire car il faut arriver à mettre le locuteur dans une situation dans laquelle il éprouve réellement les émotions visées, ce qui peut être complexe et aussi parfois désagréable pour celui-ci.

Plusieurs corpus ont été produits pour la recherche sur la reconnaissance des émotions mais beaucoup ne sont pas disponibles pour des raisons de propriété ou de manque de finition. Les quelques corpus disponibles peuvent être classés en trois groupes en fonction de la méthodologie utilisée lors de leur collection : les corpus naturels, les corpus « extraits » et les corpus simulés.

Les corpus naturels contiennent des émotions associées à de la parole spontanée ou quasi-spontanée obtenue lors d'interactions dans la vie quotidienne. Pour enregistrer ce type de

données, des volontaires sont équipés avec un matériel d'enregistrement adapté, suffisamment léger pour qu'ils ne se sentent pas mal à l'aise [Campbell 2002]. La qualité de la parole obtenue avec cette approche n'est pas toujours très bonne à cause des bruits qui peuvent venir des équipements eux-mêmes, de l'environnement, des autres locuteurs et de la distance entre la bouche et le microphone. Le microphone lui-même est aussi une cause de distorsion.

L'annotation et l'étiquetage de ce type de corpus sont également des étapes vraiment difficiles et fastidieuses. La détermination de l'émotion ressentie ou non est subjective et correspond à un problème mal posé. Il n'y a en général pas de situation contrôlée permettant de savoir ce que l'on doit chercher.

Pour assurer la correction de l'annotation, Nick Campbell a encouragé les locuteurs originaux à attribuer des étiquettes à leurs énoncés [Campbell 2002]. La transcription est ensuite effectuée manuellement avant un alignement automatique forcé en utilisant des modèles de Markov cachés [Campbell 2002]. Bien que le niveau naturel de la parole obtenue avec cette approche soit le meilleur, ce type de corpus présente l'inconvénient de ne pas fournir une même phrase (mot, passage, …) interprétée selon plusieurs états émotionnels différents (d'un seul locuteur ou même de plusieurs locuteurs) afin de pouvoir les comparer entre eux et de trouver les caractéristiques les plus représentatives ou les plus discriminantes. Ce type de corpus présente aussi l'inconvénient de ne pas isoler clairement les types d'émotions impliqués. Plusieurs émotions peuvent être présentes sur un même segment et/ou une émotion présente est difficile à associer à une catégorie prédéfinie.

En raison de ces difficultés, Nick Campbell est la seule personne que nous ayons trouvée dans littérature à avoir expressément construit un tel corpus. Quelques rares corpus utilisés par [Ang et al, 2002], [Devillers et al, 2003], [Chateau et al, 2004], [Blouin et al, 2005], [Lee et al, 2005], obtenus à partir des centres automatiques où des clients communiquent avec la machine peuvent aussi être considérés comme des corpus naturels. Mais la plupart des autres études actuelles sur l'émotion, y compris le nôtre, considère des corpus faisant partie de l'un des groupes suivants.

Les corpus « extraits » sont construits par extraction de segments dans des films, des interviews ou des journaux télévisés. Dans ce groupe, nous pouvons citer [Cowie et al, 2000] qui ont construit un corpus en enregistrant des discussions portant sur des sujets sensibles ou en extrayant des segments de programmes télévisées. [Schüller et al, 2005] ont également construit une partie de leur corpus en utilisant des segments émotionnellement chargés dans les 7 films américains : Alien, Annie Hall, Five Easy Pieces, Notting Hill, Scream, 10 things I hate about you, and Toy Story.

Dans ce groupe, les émotions ne sont pas aussi spontanées que dans le premier, mais elles le sont encore suffisamment parce qu'elles sont exprimées en contexte. De plus, dans les films, elles sont interprétées par des acteurs professionnels.

A côté de cet avantage de spontanéité, cette approche possède celui d'une construction (relativement) facile. En effet, les films contiennent souvent des passages émotionnellement chargés. L'obtention d'un grand corpus n'est plus très difficile mais le problème de la répétition reste présent dans ce deuxième groupe. L'annotation des données est aussi plus difficile pour ce groupe que pour le premier en raison de l'indisponibilité des locuteurs originaux qui empêche le préapprentissage ou la pré-annotation des locuteurs originaux pour améliorer la qualité de l'annotation.

Comme le premier groupe, ce type de corpus peut difficilement être utilisé pour l'étude des caractéristiques discriminatives en raison de son inhomogénéité. Il est par contre bien adapté pour les tests et la validation.

Comme pour l'anglais, les corpus existants cités ci-dessus ne sont pas librement disponibles et, pour le vietnamien aucun corpus de ce type existe, nous avons nous-mêmes construit pour notre travail deux corpus « extraits » : EnEmo (corpus en anglais) et VnEmo (corpus en vietnamien), voir le chapitre 8). Cependant, en raison de manque de temps, nous n'avons pas pu utiliser ces deux corpus pour nos expérimentations.

Le troisième groupe se compose des corpus construits par simulation. L'approche la plus répandue est d'employer des volontaires pour interpréter des expressions émotionnelles avec des scénarios préparés. Cette approche présente l'avantage de la facilité de la réalisation mais elle est aussi celle la moins naturelle des trois. En effet, l'évocation des émotions chez des amateurs n'est pas une tâche facile. L'utilisation d'acteurs ou d'actrices à leur place est une bonne solution comme [Engberg et al, 1996] [Engberg et al, 1997] [Fek et al, 2004] bien que les professionnels aient tendance à exagérer leurs expressions. D'autres techniques permettant d'améliorer le naturel des émotions produites utilisent des provocations contextuelles : on montre aux locuteurs des images ou des vidéos génératrices d'émotions ou on les fait jouer à des jeux provoquant des émotions pendant qu'on les enregistre. La lecture de certains textes sémantiquement émotionnels peut également produire des états émotionnels chez les locuteurs comme [Iida 2002] l'a fait.

L'avantage le plus important des corpus de ce groupe, et que l'on ne trouve pas dans les précédents, est la possibilité de produire des données avec distribution uniforme vis à vis de chaque locuteur et vis-à-vis de chaque émotion. En effet, nous pouvons y trouver les mêmes phrases (mots, passages …) interprétées avec des émotions différentes par des locuteurs différents. Ceci favorise les études portant sur la différence des caractéristiques de la parole selon les émotions et selon les locuteurs (femmes et hommes par exemple). Pour notre travail, ces études sont indispensables afin d'avoir une vue globale sur la variabilité des caractéristiques lors de l'expression émotionnelle en parole afin d'avoir de bonnes bases pour les étapes suivantes de la reconnaissance des émotions.

Nous utilisons dans nos travaux trois corpus simulés disponibles sur Internet. Le premier est le Danish Emotional Speech Database (DES) [Engberg et al, 1996], le second est le Berlin Database of Emotional Speech (BES) [Burkhardt et al, 2005] et le dernier est Orator [Quast 2002]. Le Tableau7 montre une vue globale de ces trois corpus et le Tableau 6 présente une synthèse effectuée par [Zeng et al, 2009] sur les autres corpus existants dans le domaine de reconnaissance de l'émotion (voir [Zeng et al, 2009] pour les références et le détail).

Les sections 4.2 ; 4.3 et 4.4 détaillent respectivement les corpus DES en danois, BES et Orator ainsi que les méthodologies pour la construction de ces corpus, et pour la validation de leurs contenus par le test de perception.

La dernière section du chapitre décrira brièvement notre approche pour construire nos deux corpus qui appartiennent au deuxième groupe – des corpus « extraits », un est en anglais et l'autre est en vietnamien.

Auteurs	*Type d'émotions*	*Taille*	*Type de données*	*Description des émotions*	*Annotation*	*Accès*
Cohn-Kanade 2000	Emotions actées	210 adultes ; 3 races ; 480 vidéos	Vidéo	6 émotions primaires	FACS	Oui
Sebe et al. 2004	Emotions naturelles : Sujets regardent les scenarios vidéo pour évoquer l'émotion	28 adultes	Vidéo	4 émotions : Neutre, Joie, Surprise, Dégoût	Reporté par les mêmes locuteurs	Non
MMI 2005	Emotions actées : images statiques, vidéo enregistré en vue frontale et de profil. Emotions naturelles : Enfants en interaction avec un comédien.	Actées par : 61 adultes. Naturelles : 11 enfants et 18 adultes. Au total : 3 races ave 1250 videos et 600 images statiques	Vidéo et Images statiques	6 émotions primaires	FACS et Jugement des évaluateurs.	Oui
UT Dallas 2006	Emotions naturelles : Sujets regardent les scenarios vidéo pour évoquer l'émotion	229 adultes.	Vidéo	6 émotions primaires avec perplexité, dérision, ennuie et incrédulité	Jugement des évaluateurs.	Oui
BU-3DFE 2006	Emotions actées	100 adultes.	Vidéo	6 émotions primaires avec quatre niveaux de l'intensité	Non	Oui
FABO	Emotions actées avec visage et posture	23 adultes, 210 vidéos	Vidéo	6 émotions primaires avec neutre, incertain, anxiété et ennuie	Non	Oui
Banse-Scherer	Emotions actées	6 acteurs / 6 actrices 1344 énoncés audios.	Audio	« Hot/Cold anger», panique, peur, anxiété, désespoir, tristesse, exultation, joie, intérêt, ennuie, honte, fierté, dégoût, mépris.	Jugement des évaluateurs.	Oui
ISL meeting corpus 2002	Emotions naturelles : corpus des réunions	18 réunions, 5 participants par réunion en moyenne.	Audio	Trois états : positif, neutre, ou négatif.	Jugement des évaluateurs.	Oui
CSC corpus	Emotions naturelles	32 aldultes, 15.2 heures, 3882 tours de locuteurs, 9687 SUs	Audio	Mensonge / Vérité	Reporté par les mêmes locuteurs	Non
Automatic call center (ACC) 2005	Emotions naturelles : Dialogue homme-machine dans un centre d'appel de commerce	1187 appels, 7200 énoncés	Audio	Etat négatif et non-négatif	Jugement des évaluateurs.	Non
Bank and Stock Service 2004	Emotions naturelles : Dialogue homme-homme dans un centre d'appel	350 dialogues, 10000 tours de parole.	Audio	Peur, colère, stresse	Jugement des évaluateur	Non
AIBO 2004	Emotions naturelles : Interaction entre des enfants et des robots.	110 dialogues, 29200 mots	Audio	Joie, emphases, surprise, ironie, incapacité, susceptibilité, colère, ennuie, « motherese », réprimande et tranquillité	Jugement des évaluateurs	Non
Chen-Huang 2000	Emotions actées	100 adultes, 9900 expressions audiovisuelles.	Audio / Vidéo	6 émotions primaires et 4 états cognitifs (intérêt, perplexité, ennuie, déception)	Non	Non
Adult Attachment Inteview 2004	Emotions naturelles : Sujets sont interviewés pour décrire son enfance.	60 adultes, 30 à 60 minutes pour chaque adulte.	Audio / Vidéo	6 émotions primaires et embarras, mépris, honte, généralement négative et positive	FACS	Non
RU-FACS 2005	Emotions naturelles : Sujets essaient à persuader l'intervieweur de la vérité	100 adultes	Audio / Vidéo	33 AUs	FACS	Non
Belfast database 2003	Emotions naturelles : Clips coupés à partir de la télévision et des interviews	125 sujets, 209 séquences de la télévision, 30 des interviews.	Audio / Vidéo	Annotation dans les dimensions et aussi par la catégorie	FEEL-TRACE	Oui

Tableau 6 : Synthèse des corpus existants par [Zeng et al, 2009]

Corpus	*Type d'émotions*	*Taille*	*Type de données*	*Description des émotions*	*Annotation*	*Accès*
DES [Engberg et al, 1996]	Emotions actées	2 acteurs et 2 actrices ; 2 mots, 2 phrases, 2 passages ; 10 minutes.	Audio en danois.	Neutre, Surprise, Joie, Tristesse, Colère	Tests d'écoute avec 20 auditeurs.	Publique
BES [Burkhardt 2005]	Emotions actées	10 acteurs / actrices ; 10 phases ; ~50 minutes.	Audio en allemand.	Neutre, Peur, Joie, Tristesse, Colère, Dégout, Ennui.	Jugements en échelle binaire discret par 20 auditeurs	Publique
Orator [Quast 2002]	Emotions actées	27 acteurs / actrices ; 150 monologues ; 70 minutes.	Audio en allemand.	Non-confiance / Confiance, Calme / Agité, « Leadership », Non-joie / Joie, Faiblesse / Force, Non-Colère / Colère, Non-Plaisance / Plaisance.	Jugements en échelle « continue » par 20 anglophones natifs.	Publique

Tableau7 : Vue globale des corpus utilisés

4.2. Le corpus Danish Emotional Speech Database (DES)

4.2.1. Introduction

Le corpus Danish Emotional Speech Database (DES) [Engberg et al, 1996] a été enregistré par le centre de communication humaine PersonKomunikation (CPK) de l'université d'Aalborg du Danemark dans le cadre du projet de VAESS (voix, attitudes et émotions dans la synthèse de la parole). Le travail a été effectué par Gudrun Klasmeyer, de l'université technique de Berlin (TUB) à CPK, entre juillet et octobre 1995.

Le but du projet VAESS est d'améliorer la qualité de la synthèse de parole en lui ajoutant une capacité émotionnelle. Cette synthèse de parole améliorée sera incluse dans un communicateur personnel pour aider aux personnes handicapées de la parole. Ces personnes pourront tenir le communicateur dans une main et commander le contenu ainsi que les émotions de la voix synthétisée.

Un des objectifs du projet VAESS est donc de fournir des données de parole avec des étiquettes complètes et précises pour permettre une étude systématique des variations interlocuteur et inter-attitude dans la parole. La base de données en danois EUROM.1 a été construite pour l'étude des variations interlocuteur. Afin d'étudier des variations inter-attitude, le corpus DES été crée. Il contient la voix de 4 locuteurs (2 hommes et 2 femmes) qui expriment 5 émotions (le neutre, la surprise, la joie, la tristesse et la colère), chacun pendant 30 secondes. Cela représente au total 10 minutes de parole émotive en danois.

4.2.2. Choix des locuteurs et affichage du texte

4.2.2.1 Locuteurs

Pour l'analyse des paramètres, de bons signaux sans bruit du fond sont exigés. Afin d'étudier la parole émotive et sa différence par rapport à la parole neutre, il est nécessaire d'enregistrer les mêmes énoncés dans différentes situations émotives. En conséquence, l'enregistrement doit être fait systématiquement dans des conditions de laboratoire.

Dans quelques expériences psychologiques, on a essayé d'induire des émotions chez des personnes cobayes mais, pour des raisons éthiques, il est indésirable d'induire des émotions négatives chez ces personnes. Par conséquent, la parole émotive a dû être exprimée par des acteurs. La bonne approximation de la parole émotive naturelle par de la parole simulée par des acteurs est montrée dans [Williams et al, 1972]. Des enregistrements d'un présentateur rapportant un événement dramatique ont été comparés à des enregistrements d'un acteur simulant l'état émotif du journaliste durant la présentation de l'événement. Des différences entres les enregistrements ont été trouvées, mais, en général, la façon de parler et la prosodie étaient semblables.

Il est cependant non recommandé d'employer des acteurs stagiaires, parce que comme nous l'avons mentionné ci-dessus, ils ont tendance à exagérer quelques caractéristiques de la parole pour rendre le contenu émotif très clair, ce qui rend les expressions artificielles. Quatre acteurs familiers avec le théâtre de radio ont été employés pour l'enregistrement du corpus DES (voir Tableau8). Le profil de chaque acteur est donné dans l'annexe C.

Initiales	***Sexe***	***Age***
DHC	Femme	34
KLA	Femme	52
JZB	Homme	38
HO	Homme	52

Tableau8 : Le genre et l'âge des 4 acteurs employés dans la collection du DES

4.2.2.2 Affichage du texte

Les émotions peuvent être identifiées de manière fiable à partir d'énoncés très courts comme « oui » ou « non » [Klasmeyer 1995]. Ceci signifie que les phrases courtes ou même les mots simples sont appropriés pour analyser les caractéristiques émotives de la parole mais il peut être intéressant d'analyser des passages plus longs pour étudier les pauses et les bruits émotifs spécifiques comme le rire ou les soupirs. L'idéal semble être de choisir des expressions qui apparaissent souvent dans la communication quotidienne.

Jusqu'à présent, la façon dont les émotions sont perçues par les auditeurs est très peu claire. Dans des situations normales de communication, l'auditeur ne prend pas des décisions vraiment conscientes au sujet de l'état émotif du partenaire. Ceci doit être pris en compte lors de la conception des tests d'écoute. S'ils sont invités à juger la teneur émotive d'une expression, une décision consciente est nécessaire. Les auditeurs essaieront d'abord de prendre la décision à partir de la signification sémantique. Si ce n'est pas possible, ils essaient d'imaginer les situations dans lesquelles l'expression pourrait apparaître. Si ce n'est toujours pas possible, il devient plus difficile pour les auditeurs de prendre des décisions avec confiance. C'est la raison pour laquelle le texte d'incitation doit être sémantiquement neutre. Ceci signifie qu'il ne doit pas provoquer un état émotif spécifique.

4.2.2.3 Enregistrement du DES

Selon les spécifications de la base de données indiquée dans [Tide Project1995] et décrite dans les sections précédentes, l'enregistrement et la validation du corpus DES ont été effectués dans les conditions suivantes :

Il est enregistré :

- 2 mots simples ;
- 9 phrases et ;
- 2 passages.

Ces énoncés sont prononcés par les quatre acteurs pour les cinq émotions.

Les mots, les phrases et les passages sont donnés dans l'annexe A. Une transcription phonétique du texte danois est également incluse dans cette annexe ainsi qu'une traduction.

4.2.2.4 Conditions d'enregistrement

Le corpus DES a été enregistré dans une chambre sourde au théâtre d'Aarhus avec un microphone de haute qualité qui préserve les caractéristiques spectrales d'amplitude et de phase du son articulé. Des détails sur les enregistrements et l'équipement peuvent être trouvés dans l'annexe de [Engberg et al, 1996].

4.2.3. Tests d'écoute

Comme mentionné dans la section 3.2.2, il est très important de rassembler des enregistrements de parole avec un contenu émotif non-ambigu. Ceci est normalement garanti par des tests d'écoute dans lesquels des auditeurs évaluent la teneur émotive des expressions enregistrées. On trouve des différences considérables lors de la reconnaissance des émotions dans la parole. Les résultats d'une étude de Scherer en 1995 avec 14 émotions différentes montrent que le taux d'identification des émotions dans la parole peut varier de 81 % pour la colère forte à 15 % pour le dégoût [Scherer 1995].

4.2.3.1 Réalisation des tests d'écoute.

Un test a été réalisé pour examiner si les auditeurs pouvaient identifier le contenu émotif des expressions enregistrées. 20 auditeurs (10 hommes et 10 femmes principalement des personnels à CPK) ont été employés. Leur âge moyen est de 38 ans, s'étendant de 18 à 59 ans. Le Tableau 9 résume les résultats des tests effectués. Les taux dans les colonnes au dessous des initiales des acteurs sont le taux d'émotions identifiées correctes sur 65 échantillons de chaque locuteur. La dernière colonne contient le pourcentage moyen des émotions identifiées correctes pour chaque auditeur.

Auditeur	*Sexe*	*Age*	*Date*	*Ordre d'écoute*	*HO %*	*DHC %*	*JZB %*	*KLA %*	*Moyenne %*
AVH	Femme	28	1-8	HO-DHC-JZB-KLA	74	74	80	77	76
SVA	Homme	27	8-8	HO-DHC-JZB-KLA	69	68	74	68	70
OA	Homme	33	9-8	JZB-KLA-HO-DHC	55	75	71	75	69
PLE	Homme	26	9-8	HO-DHC-JZB-KLA	54	58	69	62	61
HE	Homme	59	14-8	JZB-KLA-HO-DHC	51	58	68	45	55
JPM	Homme	34	15-8	JZB-KLA-HO-DHC	75	85	85	77	80
ILW	Femme	57	21-8	JZB-KLA-HO-DHC	63	77	71	71	70
HC	Femme	26	2-9	HO-DHC-JZB-KLA	62	68	68	66	66
HEB	Homme	45	4-9	DHC-JZB-KLA-HO	62	62	69	69	65
POR	Homme	41	5-9	KLA-HO-DHC-JZB	58	65	68	62	63
GE	Femme	57	5-9	KLA-HO-DHC-JZB	57	82	72	69	70
BLR	Femme	36	6-9	HO-DHC-JZB-KLA	65	71	83	75	73
PE	Homme	39	9-9	DHC-JZB-KLA-HO	66	69	71	63	67
VJP	Femme	39	10-9	JZB-KLA-HO-DHC	72	69	74	72	72
SPJ	Homme	47	10-9	KLA-HO-DHC-JZB	54	65	72	62	63
JEB	Femme	53	11-9	DHC-JZB-KLA-HO	51	65	66	55	59
PED	Homme	18	11-9	KLA-HO-DHC-JZB	66	62	71	60	65
JFV	Femme	24	12-9	KLA-HO-DHC-JZB	75	68	69	62	68
IHH	Femme	19	18-9	DHC-JZB-KLA-HO	49	62	62	68	60
JT	Femme	47	20-9	DHC-JZB-KLA-HO	69	66	82	74	73
Moyenne		38			63	68	72	66	67%

Tableau 9 : Ages, genre et taux de reconnaissance pour 20 auditeurs indigènes [Engberg et al, 1996].

Quatre tests, un pour chacun des acteurs, ont été conçus. Chaque test se compose de 13 expressions (2+9+2) parlées avec 5 émotions différentes. Cela fait 4 tests avec 65 expressions pour chacun. Les auditeurs réalisent le test en écoutant séquentiellement phrase par phrase.

Aucune session de formation n'a été offerte aux auditeurs avant les tests. Le test n'a pas toujours commencé par le même acteur, mais selon les résultats, les successions des acteurs étaient identiques.

On a demandé aux auditeurs de juger le contenu émotif de l'expression par un choix obligatoire. On leur a permis d'entendre une expression plusieurs fois avant de décider de la catégorie émotive. On ne leur a cependant pas permis de revenir en arrière pour pouvoir comparer avec des expressions précédentes ni non plus de modifier des choix précédents. Après chaque test d'écoute, les auditeurs ont été invités à dire si ils ont trouvé la tâche de reconnaissance des états émotionnels très facile, facile, ni facile ni difficile, difficile ou très difficile. Ils ont aussi été invités à dire quels facteurs les ont fait choisir les différentes émotions. Enfin, on leur a demandé s'ils avaient d'autres choses à dire à propos du test avec ce locuteur. Le questionnaire pour le test peut être trouvé dans l'annexe D.2.

4.2.3.2 Résultats des tests

Les émotions ont été correctement identifiées dans 67 % des cas, s'étendant de 55 % à 80 %, voir le Tableau 9. La surprise et la joie étaient souvent confondues aussi bien que le neutre et la tristesse. Les confusions entre les émotions sont montrées dans le Tableau 10, dans lequel les

émotions en verticale sont les émotions que les acteurs essaient d'induire et les émotions en horizontale sont celles reconnues par les auditeurs.

Les auditeurs n'ayant pas reçu de formation avant le test, une vérification pour savoir si les auditeurs obtiennent de meilleurs résultats sur les 20 dernières expressions que sur les 20 premières a été effectuée. Effectivement, 63 % des 20 premières expressions ont correctement été reconnues, et 73 % des 20 dernières expressions ont été correctement perçus. Les auteurs trouvent donc que cette différence de 10 % entre les 20 premières et les 20 dernières expressions prouve que les auditeurs se sont adaptés à la voix ou à la méthode.

On a également examiné si les auditeurs féminins étaient de meilleurs auditeurs que les auditeurs masculins. D'après les statistiques, les auditeurs féminins ont correctement perçu 69 % des cas alors que les auditeurs masculins ont correctement perçu 66 % des cas. Et cette différence de 3 % n'est pas significative selon les auteurs dans ce cas de test.

Le DES contient des mots simples, des phrases et des passages en même temps. La variable «type d'expressions » a été trouvée significative par rapport au score donné par des auditeurs. Les phrases ont été les items dont le contenu émotif a été le plus facilement identifiable (correctement reconnu dans 76 % des cas). Les émotions pour les mots simples ont correctement été perçues dans 65 % des cas, 68 % pour les passages. Les auteurs expliquent que ces différences peuvent être dues à la longueur des expressions. Effectivement, les mots simples sont courts et leur durée faible peut être la cause de doutes chez l'auditeur, surtout entre des émotions confuses comme la surprise et la joie, alors que les passages sont, quant à eux, à l'inverse très longs et peuvent générer des hésitations.

Auditeurs ⇒ Acteurs ⇓		REPONSE en %				
		Neu.	Sur.	Joi.	Tri.	Col.
	Neu.	**60,8** 57,8-63,7	**2,6** 1,8-3,8	**0,1** 0,0-0,6	**31,7** 29,0-34,6	**4,8** 03,7-06,3
	Sur.	**10,0** 8,3-12,0	**59,1** 56,1-62,1	**28,7** 26,0-31,5	**1,0** 0,5-1,8	**1,3** 0,7-2,1
	Joi.	**8,3** 6,8-10,1	**29,8** 27,1-32,7	**56,4** 53,4-59,4	**1,7** 1,1-2,7	**3,8** 2,8-5,1
	Tri.	**12,6** 10,7-14,8	**1,8** 1,2-2,8	**0,1** 0,02-0,6	**85,2** 82,9-87,2	**0,3** 0,1-0,9
	Col.	**10,2** 8,5-12,2	**8,5** 6,9-10,3	**4,5** 3,4-6,0	**1,7** 1,1-2,7	**75,1** 72,4-77,6
?		**20,4** 19,3-21,5	**20,4** 19,3-21,5	**18,0** 16,9-19,0	**24,3** 23,1-25,5	**17,0** 16,0-18,1

Tableau 10 : Tableau de confusion entre les émotions pour tous les locuteurs et auditeurs. Neu est abréviation de Neutre, Sur. pour Surprise, Joie. pour Joie et Tri. Pour Tristesse, Col pour Colère. Le total montre combien de fois les différentes émotions ont été choisies par les auditeurs [Engberg et al, 1996].

4 des 9 phrases dans le DES sont des phrases interrogatives. Les auteurs ont aussi examiné si les phrases interrogatives étaient reconnues comme de la surprise plus souvent que les phrases non-interrogatives. Etant donné λ la représentation relative de la surprise dans les phrases interrogatives, ξ la représentation relative de la surprise dans les phrases non-interrogatives, elles sont calculées comme suivant :

$$\lambda = \frac{Phases\ \ interrogatives\ reconnues\ comme\ surprise}{Toutes\ les\ \ phrases\ interrogatives\ qui\ sont\ en\ surprise} \quad (4.1)$$

$$\xi = \frac{Phrases\ NonInterrogatives\ \ reconnues\ \ comme\ \ surprise}{All\ non\ inquring\ sentences\ that\ were\ surprised} \quad (4.2)$$

La représentation relative de la surprise λ pour des phrases interrogatives est 168 %. Au contraire, ξ la représentation relation de la surprise pour des phrases non-interrogatives n'est que 67 %. A partir de ce résultat, on peut constater que les phrases interrogatives sont plus souvent perçues comme la surprise même si elles contiennent une autre émotion. Pour les constructions des corpus dans le futur, les questions ne doivent donc pas être employées en tant que phrases neutres.

Initiaux	***HO***	***DHC***	***JZB***	***KLA***	***Total***
Taux de reconnaissance	62,4 %	68,3 %	72,1 %	66,5 %	67,3 %
Très difficile	3	1	-	3	7
Difficile	11	4	4	10	29
Ni / Ni	5	9	8	6	28
Facile	-	5	5	1	11
Très facile	-	-	1	-	1
Total	19	19	18	20	76

Tableau 11 : Les jugements de la difficulté avec les scores réels pour les 4 locuteurs. « Ni/Ni » représente « ni facile ni difficile » et le tiret « - » veut dire que personne ne choisit cette option [Engberg et al, 1996].

Le jugement de la difficulté ainsi que les scores réels pour les différents acteurs sont montrés dans le Tableau 11. Précisément, ce tableau montre le nombre d'auditeurs qui ont donné leurs avis auprès chaque locuteur. La dernière colonne est le nombre total d'avis pour tous les quatre locuteurs.

Les auditeurs diffèrent beaucoup sur l'opinion de l'exagération des acteurs. Certains ont déclaré que les interprétations ont l'air amateur particulièrement celle de JZB, mais d'autres ont déclaré que les émotions n'étaient pas assez claires. Les auditeurs ont trouvé plus facile de juger JZB qui a effectivement le plus haut score (voir le Tableau 9). En général, les émotions chez les deux plus jeunes acteurs sont jugées plus faciles à identifier en comparaison avec les deux autres acteurs plus âgés. Les auteurs soulignent que ceci peut indiquer qu'il est plus facile identifier des émotions exprimées par des personnes plus jeunes.

4.2.4. Traitements sur le corpus DES

4.2.4.1 Etiquetage phonétique

Un modèle de Markov caché [Young et al, 1996] a été employé pour étiqueter des données du corpus : modèle à trois états entrainé par l'EUROM.1 [Lindberg 1995].

Le format d'étiquette est XWAVES (le format d'étiquette utilisé dans les xwaves d'Entropic). Le format des textes dans les fichiers est basé sur le format ASCII, donc il est assez facile de convertir les fichiers d'étiquettes dans d'autres formats. Pour un exemple d'un fichier XWAVES, voir le Tableau 13, et le format des fichiers d'étiquette de XWAVES est montré dans le Tableau 12.

#
end ccode label_name
.
.
.
end ccode label_name

Tableau 12 : Format des fichiers XWAVES où « end » est le repère de l'étiquette « label_name » et « ccode » est un code de couleur utilisé par xwaves. « end » est spécifié en seconds. N'importe quelle valeur peut être employée pour le « ccode » mais souvent une valeur de 121 est employée.

#
0.052436 121 SIL
0.165000 121 f
0.264999 121 A
0.310750 121 v
0.346749 121 SIL

Tableau 13 : Un exemple d'un fichier XWAVES

Pour une étude de la distribution phonétique du DES, les durées moyennes des phonèmes dans les phrases et dans les passages du DES sont listées dans le Tableau 14 en comparaison avec les durée moyenne des phonèmes du corpus EUROM.1 (la partie de plusieurs locuteurs). A partir du Tableau 14 les auteurs voient que la durée moyenne de phonème dans le DES est proche de celle de EUROM.1.

Le modèle de SIL a été employé pour modéliser des périodes initiales et finales de silence, le modèle appelé SP est pour optimiser la pause entre les mots.

	DES		EUROM.I plusieurs locuteurs			EUROM.1 plusieurs locuteurs		EUROM.I plusieurs locuteurs	
Etiquette	Nombre	Durée et moyenne [ms]	Nombre	Durée et moyenne [ms]	**Etiquette**	Nombre	Durée et moyenne [ms]	Nombre	Durée et moyenne [ms]
SIL	1632	322,4	1106	563,4	***l***	794	43,7	3440	53,5
SP	1053	81,4	3804	321,9	***R***	350	58,4	1671	68,5
p	329	82,9	776	83,1	***i***	908	71,8	3420	74,8
b	394	63,1	1548	64,8	***e***	1064	65,8	3379	65,8
t	397	95,3	1505	94,8	***E***	817	71,0	2424	69,7
d	1548	54,2	4396	52,6	***{:***	154	124,4	551	124,4
k	309	104,2	989	92,3	***a***	1110	84,2	3681	75,2
g	765	60,8	2430	58,4	***A***	916	90,6	2727	87,8
f	411	80,3	1625	92,7	***@***	978	44,4	3182	57,2
s	1394	96,3	4945	95,0	***y***	114	84,6	490	89,7
v	451	58,4	1558	57,7	***2***	183	81,3	695	93,3
D	550	73,7	1867	75,9	***9***	121	101,0	579	91,6
j	464	61,2	1274	58,4	***u***	462	84,6	1689	94,6
h	334	50,3	1203	62,0	***o***	289	91,1	1032	109,8
m	789	66,8	2457	75,8	***O***	292	88,3	1133	100,2
n	1421	66,0	5416	67,0	***Q***	1494	68,7	5381	82,4
N	194	87,0	648	88,1					

Tableau 14 : Statistique des étiquettes pour le corpus DES en comparaison avec EUROM.1 (la partie de plusieurs locuteurs : 60 [Eurom I 1995]).

4.2.4.2 Stockage des données

Les données sont codées avec 16 bits/échantillon et une fréquence d'échantillonnage de 20 kHz.

Au total, le corpus se compose de 350 segments de signal de parole. Chaque segment contient soit un mot simple, soit une phrase, soit un passage de discours. La taille du segment le plus court (mot simple) est : 22,5 kB équivalente à 0,576 s. La taille du segment le plus long (passage) est de 1225 kB pour une durée de 31,36 s.

La taille moyenne des phrases est d'environ 45 kB équivalente à 1,152 s

La longueur totale du corpus est de 77,5 MB équivalent à environ 2034 s ≈ 34 minutes

Le rapport moyen parole/silence est d'environ 933 s / 1101 s ≈ 0,85. Le tableau suivant montre cette proportion :

Emotions	***Parole***	***Silence***	***Total***	***Proportion***
Colère	94 s	99 s	193 s	0,949
Joie	106 s	101 s	207 s	1,050
Neutre	692 s	513 s	1205 s	1,349
Tristesse	105 s	118 s	223 s	0,890
Surprise	104 s	101 s	205 s	1,030

Tableau 15 : Proportion moyenne entre parole/silence

Une première remarque peut être retenue : en états émotionnels, on a tendance à prendre moins de temps pour produire la parole ou à augmenter la durée des silences par rapport à l'état de neutre.

A partir de ce tableau, on peut grouper la tristesse et la colère comme montrant des taux de parole/silence les plus petits, mais si on considère le temps total de ces deux états nous trouvons que la nature de ces deux effets n'est pas la même. En effet, nous constatons également, les locuteurs ont tendance à augmenter la durée des silences pour la tristesse, cette remarque veut dire qu'ils parlent plus lentement ou ils hésitent simplement plus longuement ou bien ils augmentent juste les pauses ? Au contraire, en situation de colère, ils ont tendance à diminuer le temps de la parole (ils parlent donc plus rapidement ?). Pour avoir des conclusions plus fiables, dans le chapitre suivant nous allons mesurer et utiliser le débit phonétique.

Le deuxième groupe se compose de la surprise et la joie qui a le même taux de la parole que celui du silence. Et le neutre constitue le dernier groupe où la parole est la majorité. Dans le Chapitre 5, nous analyserons et discuterons des opérateurs nous permettant de capturer cette caractéristique et d'autres caractéristiques pour discriminer les états émotionnels.

4.2.4.3 Utilisation de ce corpus

Avec 4 locuteurs et 13 expressions par émotion pour chaque locuteur, clairement ce corpus n'est pas assez grand pour servir à entraîner des modèles globaux indépendants du locuteur. Pourtant, il est très utile pour des études du comportement des caractéristiques de la parole en états affectifs. Effectivement, comme on peut trouver dans ce corpus une même expression interprétée en plusieurs émotions pour tous les locuteurs, la comparaison est plus facilement réalisée, entre des émotions, entre des hommes et femmes. Des résultats du comportement des caractéristiques de la parole non seulement peuvent être utilisés dans les systèmes de synthèse de la parole avec la capacité émotionnelle intégrée, mais ils sont également utiles pour des systèmes de reconnaissance des émotions en se basant sur ces comportements.

Pour notre travail, nous avons utilisé essentiellement ce corpus pour notre étude de la reconnaissance de l'émotion.

Un autre plus grand corpus, qui peut être aussi utilisé pour des études du comportement des caractéristiques de la parole en émotion, est le corpus de Berlin (BES) que nous allons présenter ci-dessous.

4.3. Berlin Database of Emotional Speech (BES)

4.3.1. Introduction

Le corpus « Berlin Database of Emotional Speech » (BES) [Burkhardt 2005] contient dix phrases exprimées en 7 émotions par 10 acteurs, et cette base a été enregistrée dans le but d'étudier les caractéristiques prosodiques, les caractéristiques articulatoires et la vérification des résultats obtenus au moyen de ré-synthèse.

Pour assurer la qualité du corpus et des expressions émotionnelles, en construisant le corpus, les auteurs ont tenu en compte les points suivants :

- ils ont utilisé un assez grand nombre de locuteurs ;

- les locuteurs ont exprimé toutes les émotions pour assurer une distribution uniforme du corpus;
- tous les locuteurs ont exprimé le même contenu verbal afin de favoriser la comparaison entre les émotions et entre les locuteurs ;
- les enregistrements ont été de haute qualité audio avec un bruit de fond minimum.

4.3.2. Choix des émotions

Dans la littérature, les émotions sont habituellement décrites soit en utilisant des « dimensions émotionnelles » comme « l'activation », le « plaisir » soit par les concepts discrets tels que la « colère », la « peur » (voir le Chapitre 2). Les concepts distincts sont logiquement le choix si on étudie des émotions simulées car ces concepts sont facilement compris par des locuteurs ainsi que par des auditeurs.

En tenant compte de la possibilité de comparer les résultats avec d'autres études, les 7 émotions souvent rencontrées ont été choisies par les auteurs pour construire ce corpus : le neutre, la colère, la peur, la joie, la tristesse, le dégoût et l'ennui.

4.3.3. Choix des locuteurs

Considérant que les acteurs apprennent à exprimer des émotions de manière assez exagérée, les auteurs n'ont pas été convaincus que les acteurs entraînés seraient le meilleur choix pour interpréter des expressions émotionnelles naturelles. Par conséquence, ils ont décidé de laisser cette question ouverte et de rechercher des interprètes par l'intermédiaire d'un journal. Les trois experts auditeurs ont alors choisit 10 personnes parmi 40 personnes qui ont répondu à l'annonce avec un nombre égal d'hommes et de femmes, en jugeant le niveau de naturel et le niveau reconnaissable des expressions enregistrées par ces 40 personnes. Par hasard, l'une de ces personnes choisies avait suivi une école de théâtre.

4.3.4. Choix des textes

La construction du corpus de la parole émotionnelle par la simulation des acteurs ou actrices présente l'avantage qu'il est possible de contrôler les phrases à interpréter. Pourtant, il est important que toutes ces phrases soient être interprétables dans toutes émotions de cibles et elles ne doivent impliquer aucune tendance émotionnelle. Deux genres différents de textes répondent habituellement à ces exigences :

- des textes vides de sens, par exemple comme une série aléatoire de figures ou de lettres, ou des mots imaginaires (exemples décrits dans [Banse et al, 1996]) ;
- ou des phrases normales de la vie quotidienne.

Des phrases vides de sens sont garanties neutres. Cependant, l'inconvénient de ce type de phrases réside dans le fait que les acteurs les trouveront difficiles à prononcer et auront des difficultés à imaginer une situation émotive afin de pouvoir produire naturellement les expressions émotionnelles souhaitées. Selon [Scherer et al, 1981], c'est la raison pour laquelle des textes vides de sens donnent plutôt des résultats rigidement excessifs *(stereotyped overacting*).

En comparaison avec des textes très riches de sens comme la poésie ou à contrario avec les phrases vides de sens, l'utilisation de phrases quotidiennes s'avère en fait la meilleure [Scherer et al, 1981], parce que c'est une forme normale de la parole qui contient des états émotionnels.

D'ailleurs, les acteurs peuvent immédiatement exprimer l'émotion de mémoire avec de type de phrases. En effet, l'utilisation de textes non familiers peut exiger un processus de mémorisation du contenu du texte, ce qui peut causer la dégradation du niveau naturel de la parole obtenue. C'est la raison par laquelle, les phrases quotidiennes ont été employées par les auteurs. Au total, le corpus possède 10 expressions dont 5 sont des phrases simples et les 5 autres phrases sont des passages de deux phrases consécutives.

Un des objectifs du développement de ce corpus est de faciliter l'analyse de la « réduction articulatoire », particulièrement dans les états émotionnels, [Kienast et al, 1999], ainsi, la conception phonologique des phrases a dû tenir compte de la possibilité de différentes formes de réductions. Les phrases du corpus ont été construites de sorte qu'elles permettent toutes les suppressions et assimilations possibles des segments, selon [Kohler 1995]. Ainsi, dans le but d'effectuer l'analyse des formants, les phrases doivent contenir autant de voyelles que possible. Les phrases utilisées sont présentées dans le Tableau 16 avec la traduction en français et en anglais.

Echantillons	Allemand	Anglais
A01	Der Lappen liegt auf dem Eisschrank.	Le tissu se trouve sur le réfrigérateur. *The cloth is lying on the fridge.*
A02	Das will sie am Mittwoch abgeben.	Elle va le remettre le mercredi. *She will hand it in on Wednesday.*
A04	Heute Abend könnte ich es ihm sagen.	Ce soir, je peux lui dire. *Tonight I could tell him.*
A05	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.	La feuille de papier noir est là-haut à côté d'un morceau de bois. *The black sheet of paper is up there beside the piece of timber.*
A07	In sieben Stunden wird es soweit sein.	Dans sept heures, il sera prêt. *In seven hours the time will have come.*
B01	Was sind denn das für Tüten, die da unter dem Tisch stehen?	Qu'est-ce que c'est, les sacs sous la table? *What are the bags standing there under the table?*
B02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	Ils viennent de le monter et maintenant ils vont le redescendre. *They have just carried it upstairs and now they are going down again.*
B03	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	Le weekend, je suis rentré souvent chez moi et j'ai vu Agnes. *At the weekends I have always gone home now and seen Agnes.*
B09	Ich will das eben wegbringen und dann mit Karl was trinken gehen.	Je veux juste l'enlever et puis aller boire un verre avec Karl. *I just want to take this away and then go for a drink with Karl.*
B10	Die wird auf dem Platz sein, wo wir sie immer hinlegen.	Il sera à la place où nous le mettons régulièrement. *It will be in the place where we always put it.*

Tableau 16 : Contenu en texte des échantillons du corpus BES

4.3.5. Enregistrement des données

Pour la haute qualité audio, les enregistrements ont eu lieu dans la chambre sourde de l'Université Technique de Berlin (Techinical University Berlin). Les enregistrements ont été réalisés avec une fréquence d'échantillonnage de 48 kHz, suivi d'un sous-échantillonnage à 16 kHz.

Il y a eu une session d'enregistrement avec chaque acteur sous la surveillance des trois phonéticiens. Chaque session a duré environ deux heures. Le texte de chaque expression a été donné oralement à l'acteur pour éviter une intonation de type lecture. Ensuite, l'acteur a pu choisir les émotions l'une après l'autre. Chaque acteur a entendu un court exemple de cette émotion, par exemple le bonheur après avoir gagné une grande somme d'argent à la loterie ou la tristesse provoquée par la perte d'un bon ami ou d'un proche. Les acteurs ont également du temps pour se préparer à l'émotion spécifique. Il a été demandé aux acteurs de se rappeler une vraie situation de leur passé où ils ont vécu cette émotion. De cette façon, les enregistrements sont obtenus avec des acteurs qui ont ré-expérimenté les émotions dans des conditions favorables pour développer les même effets physiologiques qu'en situation vraie.

Les acteurs ont produit chacune des phrases autant de fois qu'ils l'ont souhaité, et pour quelques combinaisons, un certain nombre de variantes a été enregistré. Il a été demandé aux acteurs de ne pas crier en exprimant la colère et de ne pas chuchoter en exprimant l'anxiété. Cela est nécessaire afin d'obtenir des données encore analysables au niveau de la qualité de la voix.

Selon les auteurs, il reste toujours au moins les problèmes suivants pour ce corpus : premièrement, puisque les acteurs ne se tenaient pas immobiles mais se déplaçaient devant le microphone, la distance entre la bouche et le microphone n'était pas constante et c'est pourquoi l'analyse de l'énergie du signal est incertaine. En outre, le niveau d'enregistrement a dû être ajusté entre la parole très forte (la plupart du temps en colère) et la parole très faible (la plupart du temps pour la tristesse). Enfin, au niveau du contour de l'intonation, les acteurs ont choisi des mots différents pour réaliser l'accent de la phrase, ce qui fait que la comparaison entre les contours de la fréquence fondamentale s'avère plus compliquée.

4.3.6. Evaluation des données

Pour s'assurer de la qualité et du niveau de naturel des expressions, un test de perception a été effectué. 20 personnes ont participé à ce test. Des expressions ont été affichées en ordre aléatoire sur un écran d'ordinateur. Il n'a été permis aux auditeurs de n'écouter chaque expression qu'une seule fois avant de décider de l'état émotionnel et du niveau de naturel de l'expression.

Les taux moyens de reconnaissance par les auditeurs sont montrés sur la Figure 11. Des lignes connectant deux émotions montrent qu'il y a une différence significative entre ces deux émotions. Les expressions avec un taux de reconnaissance supérieur à 80 % et un niveau naturel supérieur à 60% ont été choisies pour davantage d'analyse. Au total, environ 500 expressions sur 800 ont été ainsi sélectionnées.

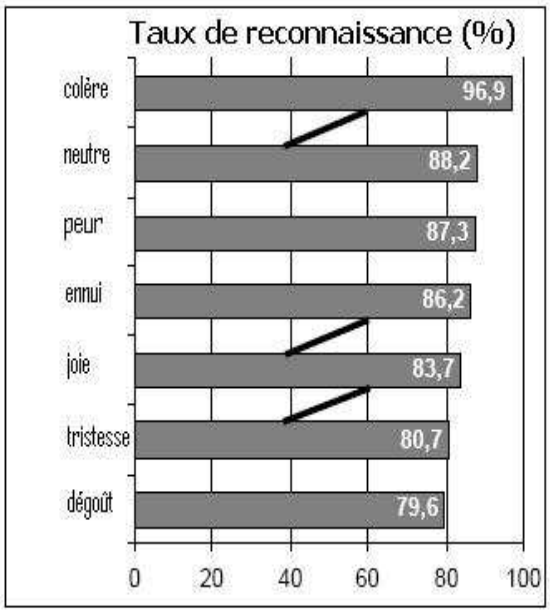

Figure 11 : Taux de reconnaissance [Burkhardt et al, 2005]

Deux autres tests supplémentaires ont été réalisés. Dans le premier, il a été demandé au sujet d'évaluer la force de l'émotion montrée dans chaque expression. Pour le deuxième test, le sujet a dû juger l'accent de syllabe (*syllable stress*) de chaque expression, ce test était le seul dans lequel seulement des personnes phonétiquement formées pouvaient participer parce que les autres personnes se sont senties incapable d'évaluer les accents. Dans les deux tests, l'évaluateur a eu la possibilité d'écouter les expressions aussi souvent que souhaité avant donner son estimation. La plupart des émotions a été jugé forte. Cette évaluation de force a été alors employée comme variable de contrôle dans les statistiques.

4.3.7. Etiquetage des données

Les expressions ont été annotées en utilisant ESPS/waves+. Deux fichiers d'étiquetage dans le format ASCII ont été créés pour chaque expression. Le premier contient une transcription phonétique étroite qui est basée sur un jugement auditif soutenu par une analyse visuelle d'oscillogramme et de spectrogramme (voir la figure 11 pour un exemple). Pour la transcription, l'alphabet phonétique SAMPA a été employé. Des caractéristiques de l'émotion et la manière de parler ont été marquées par des notations additionnelles, à savoir, les caractéristiques articulatoires de la voix comme : forte où chuchotée. Tandis que les segments phonémiques étaient annotés avec les symboles SAMPA, les caractéristiques et les diacritiques ont été marqués avec des abréviations en allemand (par exemple « NAS » pour nasal).

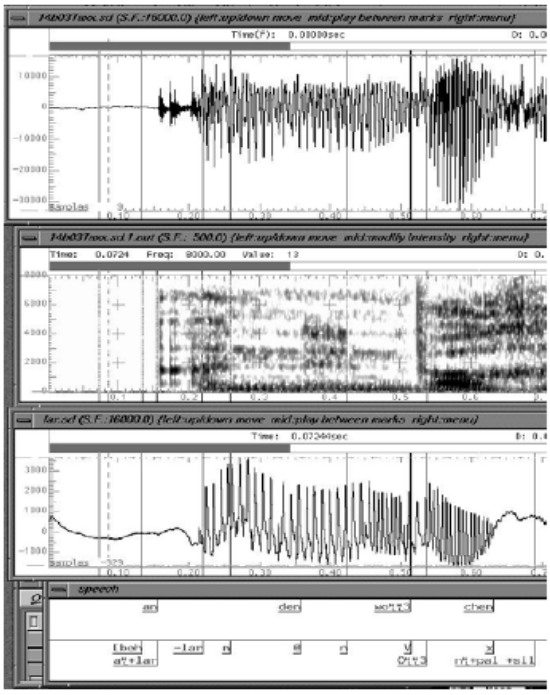

Figure 12 : De haut en bas : la capture d'écran des signaux : oscillogramme, spectrogramme, électro-glottogramme [Burkhardt et al, 2005].

Les frontières des segments et des pauses ont été également étiquetées. Chaque son a été transcrit en un symbole, excepté les diphtongues et les plosives. Les diphtongues ont été traitées comme un segment. Des symboles additionnels ont été assignés aux phases d'éclat et d'aspiration des plosives. Les temps exacts d'apparition et de disparition ont été marqués avec des signes « +/- » (par exemple « +nas »jusqu'à « -nas » pour la partie nasalisée).

Le deuxième fichier contient une segmentation des syllabes et des annotations de quatre niveaux de l'accent (accent de phrase, primaire, secondaire, et sans-accent). Ces niveaux ont été vérifiés dans un test de perception par huit phonéticiens qualifiés pour rendre les données plus fiables. La segmentation des syllabes est basée sur les frontières de la transcription phonétique étroite du premier fichier d'étiquetage. Seulement dans le cas des segments « ambisyllabiques », la frontière de syllabe est placée au milieu du son.

Tous les fichiers de données (la parole et l'annotation) peuvent être récupérés sur le site web à l'adresse http://pascal.kgw.tu-berlin.de/emodb/.

En comparaison avec le DES (2 hommes et 2 femmes), ce corpus contient plus de locuteurs (5 femmes et 5 hommes). Cela permet des résultats plus généraux pour des études statistiques sur le comportement des caractéristiques de la parole ainsi que pour la reconnaissance des émotions indépendante du locuteur. BES est choisi et utilisé pour la validation de nos résultats qui sont obtenus avec le corpus DES.

4.4. Orator

Les deux corpus DES et BES sont des corpus construits suivant l'approche discrète où chaque échantillon est supposé contenir un seul état émotionnel et où, pour le jugement, les évaluateurs doivent choisir un seul état émotionnel parmi un certain nombre d'états donnés. Bien que cette

approche permette de créer des énoncés assez « purs » pour chaque émotion, elle n'est pas souvent correcte pour les expressions émotionnelles quotidiennes. En effet, l'existence de plusieurs états émotionnels dans un seul énoncé est toujours possible et ceci exige une autre approche pour l'évaluation de ceux-ci.

Ces dernières années, [Devillers et al, 2005] et [Devillers et al, 2006] ont fait face à ce problème en permettant aux annotateurs d'annoter un autre état émotionnel « minor » à côté de l'état émotionnel dominant « major », cette nouvelle approche de l'annotation est appelée l'annotation complexe. Ces auteurs ont utilisé l'annotation complexe comme un moyen pour analyser des énoncés avec les trois classes : classe des énoncés contenant des émotions ambiguës, classe des énoncés contenant des émotions contradictoires et classe des énoncés contenant des émotions non-contradictoires (voir [Devillers et al, 2005] pour le detail). Cependant, les autres exploitations plus loins de cette nouvelle méthode d'annotation dans la reconnaissance de l'émotion sont encore des perspectifs de l'étude.

Orator est un corpus de parole affective enregistré par [Quast 2002] en utilisant l'approche continue pour évaluer le degré des émotions contenues dans les énoncés en 7 dimensions psycholinguistiques : le plaisir, la joie, la confiance, la force, l'inquiétude, l'autorité, la colère. Cette approche est donc aussi une solution pour le problème d'émotions complexes.

4.4.1. Introduction

L'intention globale du projet dans le cadre duquel le corpus Orator a été construit est de concevoir et de tester une plateforme pour doter la machine d'une fonction de perception des aspects non-verbaux de la parole, ce qui pourrait être employé dans des applications comme l'aide aux personnes sourdes innées à parler avec une prosodie plus naturelle ou l'aide aux étudiants lors des cours de langue étrangère au niveau de la prononciation ainsi que de la prosodie. Plus spécifiquement, les objectifs ont été les suivants :

- créer une base de données flexible qui contient des enregistrements de parole naturelle qui peuvent être employés pour produire sur l'auditeur une variété d'impressions différentes ;
- extraire des paramètres acoustiques et/ou spectraux qui contiennent probablement des informations non-verbales dans les échantillons de la parole ;
- évaluer le contenu non-verbal des enregistrements ;
- et exécuter des tâches de reconnaissances des formes et de traitement des signaux pour voir si le contenu non-verbal du signal de la parole peut être reconnu par des techniques de reconnaissance des formes.

Les enregistrements du corpus sont interprétés par des acteurs et des non-professionnels à la fois. Ils sont également évalués par des auditeurs américains et allemands qui donnent leurs impressions pour chaque enregistrement avec 7 catégories d'émotions. Les mesures obtenues sont finalement normalisées.

4.4.2. Locuteurs et textes

L'ensemble des données utilisées dans ce corpus contient 145 enregistrements d'un monologue allemand, ce monologue contient les 8 phrases suivantes :

(1) In der Vergangenheit ist schon einiges an guter Vorarbeit geleistet worden. (2) Die Ziele, die wir jetzt verfolgen, sind die gleichen und müssen auch auf die gleiche Weise behandelt werden. (3) Unsere Aufgabe ist nun, noch einmal die Zeiteinteilung durchzusehen. (4) Sie überprüfen dann das Weitere. (5) Bitte notieren Sie die Punkte, die Sie heraussuchen, und tragen Sie uns diese vor! (6) Wir erledigen alles Andere. (7) Glauben Sie, daß Sie das schaffen? (8) Gut!

L'auteur a essayé de combiner une variété de différentes structures de la phrase sans compromettre la cohérence du monologue. Le monologue contient deux exclamations, dont l'une est une interjection (8) et l'autre une demande (5). Une question (7) et le reste se compose de phrases régulières : deux parmi elles (4,6) ont le même nombre de syllabes et la même structure d'intonation pour tenir compte de l'apprentissage / le test des phrases croisées.

Au total, 27 acteurs professionnels/non-professionnels ont produit leurs expressions de monologue. L'âge, le sexe et la profession des locuteurs sont aussi sauvegardés.

4.4.3. Enregistrement

Il n'a pas été demandé aux acteurs de produire une expression spécifique, mais de s'imaginer eux-mêmes dans différentes situations et exprimer le monologue correspondant avec ces situations. De cette façon, les émotions désirées n'ont pas été présentées d'une manière artificielle. Ceci aide également l'acteur à se trouver à l'aise pour les interprétations. [Bänziger et al, 2006] ont effectué la même stratégie pour assurer la spontanéité des énoncés produits par les locuteurs lors de la construction de leur corpus : GEMEP (*GEneva Multimodal Emotion Portrayals*). La différence entre ces deux études est que les énoncés du corpus Orator sont annotés sans tenir en compte du contexte (dans lequel le locuteur s'est mis) comme GEMEP (voir la section suivante).

La longueur moyenne des enregistrements est d'environ 30 secondes. L'enregistrement a été effectué avec une fréquence d'échantillonnage de 48 kHz et stocké dans des fichiers .raw, mono, 16 bits, PCM linéaire.

4.4.4. Evaluation des enregistrements

L'évaluation est effectuée sur les dimensions psycholinguistiques qui sont le plaisir, la joie, la confiance, la force, l'inquiétude, l'autorité, la colère. Chaque enregistrement est évalué par chaque auditeur en complétant un vecteur d'évaluation tel que 2, 1, -2, 1, 2, -1, 1, 0. Comme nous voyons, chaque dimension peut prendre des valeurs différentes de -2 à 2, ces valeurs dépendent de la sensation de chaque auditeur. L'utilisation de cette échelle de valeurs de -2 à 2 présente un avantage en comparaison avec le procédé binaire précédent où le choix de l'auditeur est forcé entre oui ou non. En effet, une expression en parole n'est pas nécessairement heureuse ou malheureuse, elle peut tout-à-fait assigner différents degrés de bonheur. D'ailleurs, avec cette approche, les enregistrements sont évalués dans plusieurs dimensions, et clairement, le point désigné par le vecteur d'évaluation est présenté dans toutes les dimensions par leur degré. Cela veut dire que chaque expression peut contenir plusieurs aspects émotionnels, seulement dans de rares cas les émotions pures se produisent dans la parole naturelle et la plupart du temps, des affections de multiples catégories contribuent au contenu vocal d'une expression [Quast 2002].

150 enregistrements ont été choisis en tenant compte de la taille, de la diversité du corpus mais aussi de la lassitude des évaluateurs lors de leurs tests de perception. Les dimensions sont bipolaires, c'est-à-dire des axes opposés représentent des impressions opposées comme heureux et non-heureux.

Le processus d'évaluation a été réalisé avec 20 anglophones natifs (10 hommes et 10 femmes). Chacun a évalué 150 enregistrements du corpus avec les 7 catégories : non-confiant / confiant ; calme / agité ; « leadership » fort / fable ; non-joie / joie ; faiblesse / force ; non-colère / colère ; non-plaisant / plaisant. Les évaluateurs ont de 18 à 54 ans avec un âge moyen de 27 ans.

L'échelle d'évaluation contient les 5 choix possibles de –2 à +2, y compris la valeur neutre 0.

Lors de l'évaluation, l'auditeur peut rejouer l'énoncé autant de fois qu'il le veut. Une fois que les échantillons ont été tous traités ou quand l'utilisateur a décidé de finir le processus, il est invité à compléter un formulaire court (volontaire) par son l'âge, son langage, et ses commentaires. Le score et les informations de l'évaluateur seront ensuite envoyés au centre de traitement. L'interface d'évaluation peut être affichée par des navigateurs Web mais l'évaluation peut être aussi téléchargée pour la lancer hors ligne. De ce fait, un même corpus peut être évalué par plusieurs personnes, et cela permet aussi la comparaison entre des auditeurs dans le monde.

La méthode de téléchargement s'est avérée bien fonctionner en rassemblant des scores de 20 auditeurs allemands (10 hommes, 10 femmes de 22 ans à 58 ans, âge moyen : 29) qui ont effectué le même processus d'évaluation sur leur propre ordinateur. Les résultats des évaluations peuvent être consultés dans [Quast 2002].

4.4.5. Post-Traitement du score d'évaluation

Pour tenir compte de la différence du comportement d'évaluation, les scores de chaque personne sont normalisés par rapport à la moyenne et l'écart type de chaque catégorie.

On a montré que l'humeur affecte le jugement, [Niedenthal et al, 1994] (l'humeur – par opposition à l'émotion – décrit des états affectifs à plus long terme qui persistent des heures, voire des jours, et même probablement plus longtemps ; les phénomènes décrits ici comme étant des émotions ne durent que quelques minutes). L'humeur donne des préjugés, par exemple, une personne avec une prédisposition positive pensera probablement plus aux résultats positifs que négatifs et vice-versa [Mayer et al, 1993]. Cela influence également la qualité des jugements dans le sens que l'humeur d'une personne est reflétée dans ses évaluations [Clore 1992]. Pour éliminer l'influence de l'humeur, les points de chaque évaluateur sont décalés de sorte que les réponses moyennes pour toutes les catégories soient zéro.

Une autre question de la normalisation qui doit être abordée est la différence de l'amplitude des points donnés par des évaluateurs. En fonction du caractère, de l'humeur et de l'expérience, certains ont tendance de dévier plus loin de la ligne zéro que des autres. On a choisi alors la valeur 0.5 pour ajuster les écarts types. Les deux histogrammes de la Figure 13 montrent un exemple de la distribution des scores donnés par les évaluateurs avant et après la normalisation.

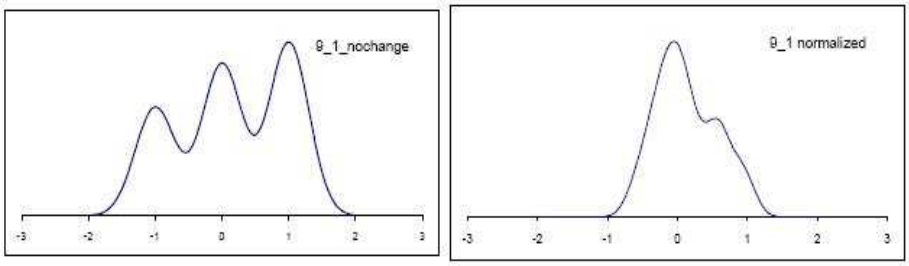

Figure 13 : a) Une de 7 x 150 = 1050 histogrammes (chacune contient 20 points de scores donnés par 20 évaluateurs) pour une catégorie d'émotion avant la normalisation b) Les mêmes scores après la normalisation

On constate aussi que la valeur moyenne de l'écart type avant la normalisation de tous les évaluateurs, de tous les enregistrements et pour toutes les catégories, est un bon indicateur pour savoir combien d'information affective est contenue dans chaque catégorie. Il est alors possible de les classer comme suit :

Catégories Psycholinguistiques	***Ecart type***
Agité	1.04
Confiance	0.99
Conduite	0.93
Force	0.86
Colère	0.85
Plaisant	0.74
Joie	0.66

Tableau 17 : L'écart type des catégories psycholinguistiques.

La différence de l'évaluation entre des évaluateurs natifs et des évaluateurs non-natifs a été également prise en compte et quelques petites tendances sont perceptibles :

- S'il y avait de la différence spécifique de la culture et du langage dans la communication affective, elle pourrait se manifester dans la perception. Par exemple, les sons fricatifs comme [x] du mot Dach en allemand sont habituellement perçus comme criards par les auditeurs américains, et pourraient donc toujours leur donner une impression de parole plus fâchée que dans leur propre langue.
- A la différence de la moyenne, l'écart-type chez les auditeurs américains est plus important que celui des allemands, excepté le plaisir et la force qui sont marqués presque de la même manière. Les raisons possibles de ceci pourraient être que l'information linguistique a été cachée aux américains, et ils étaient donc plus sensibles au contenu affectif. Une autre explication très simple est que les américains ont tendance à donner des grandes valeurs (voir [Quast 2002]).
- En regardant et en comparant des histogrammes d'évaluation des allemands et celles des américains pour toutes catégories confondues, on a constaté que les allemands se mettent d'accord plus fortement sur leurs jugements en comparaison avec les américains. Ces

meilleurs résultats chez les allemands peuvent s'expliquer par leurs tests dans la langue maternelle.

D'autres statistiques plus précises pour toutes 7 catégories peuvent se trouver dans [Quast 2002].

Chapitre 5. **Etudes des paramètres**

L'étude des paramètres est une partie indispensable de tous les systèmes de classification et de reconnaissance. Dans un domaine nouveau comme la reconnaissance de l'émotion. Il nous faut donc une analyse précise des paramètres pour l'estimation et pour une sélection efficace. Afin d'obtenir des conclusions sur l'importance des paramètres, nous proposons d'utiliser la comparaison relative entre des états émotifs en utilisant la normalisation par rapport au neutre. Cette méthode ainsi que ses avantages seront présentés dans la première partie. Dans la deuxième partie, nous parlerons des techniques pour l'évaluation et pour la sélection des ensembles de paramètres. Nous expérimenterons également les difficultés posées par les variétés interlocuteurs dans les deux modes : la reconnaissance multi-locuteur et la reconnaissance indépendante du locuteur et dans chaque cas, nous proposons les ensembles les plus efficaces selon notre expérimentation. Il faut noter que nos résultats et nos conclusions sont obtenus sur des corpus particuliers qui ont certaines limitations comme le niveau de naturel ou la taille (le nombre de locuteurs et le nombre d'énoncés). On ne peut donc pas en déduire que ces résultats se généraliseraient dans un contexte plus général, notamment à d'autres applications ou à d'autres types de données.

5.1. Paramètres étudiés

Parmi plusieurs aspects de la parole, la prosodie est la première candidate en raison de son rôle important dans l'expression émotionnelle, lequel a été prouvé par plusieurs systèmes de synthèse de la parole [Schröder 1999], [Oudeyer 2002], [Mozziconacci 2004]. L'étude précise des paramètres de prosodie sera présentée dans la section 5.3.1.4

Un deuxième aspect aussi important que la prosodie est l'aspect spectral du signal qui peut être représenté par les paramètres spectraux (MFCCs, LFCCs et LPCs). Ces trois ensembles de paramètres seront discutés dans la section 5.3.1.6.

Enfin, d'autres paramètres comme le nombre de passage par zéro, le nombre d'extrémités (points maxima et minima locaux) seront présentés ainsi que les raisons pour lesquelles nous les avons introduits dans cette étude.

5.1.1. Paramètres de prosodie

Dans la conversation quotidienne, mettons de côté la facette visuelle du visage, du geste et de la posture et ne travaillons qu'avec le son. Dans ce cas, la colère nous fait immédiatement penser à une voix forte, haute, et « lourde », au contraire la tristesse nous évoque une voix basse, faible et lente. Il est aisé de remarquer que ces éléments : forte/faible, rapide/lente, haute/basse, … appartiennent à un aspect qui est appelé la prosodie dans le traitement de la parole. C'est aussi la raison pour laquelle, la prosodie est la caractéristique la plus étudiée dans la littérature comme dans [Vroomen et al, 1993], [Rank et al, 1998], [Montero et al, 1999], [Schröder 1999], [Oudeyer 2002] ou [Mozziconacci 2004]. Nous allons étudier les comportements de la prosodie montrés dans différentes émotions pour avoir une vue globale du rôle et de l'influence distinctive de cette caractéristique afin d'améliorer la reconnaissance de l'émotion dans le cadre des corpus utilisés.

Il existe différentes manières de définir les paramètres prosodiques, selon qu'on les considère sur le plan de la production, sur le plan acoustique, ou sur le plan perceptif. Pour notre système de reconnaissance des émotions dans le signal acoustique, la substance acoustique des paramètres prosodiques se définit par : la fréquence fondamentale, l'intensité qui correspond à l'énergie contenue dans le signal au cours d'un intervalle de temps donné, et le débit qui est le débit phonétique dans notre cas.

Nous commençons par faire des analyses statistiques de la distribution des paramètres de prosodie pour examiner leurs capacités discriminatives pour la reconnaissance des émotions. Ces résultats seront ensuite appliqués pour la reconnaissance. Le corpus DES avec cinq émotions primaires est choisi pour les raisons suivantes :

- premièrement, la prosodie est une caractéristique qui dépend fortement de l'aspect physique ainsi que de l'habitude de chaque locuteur ; par exemple, la prosodie d'une femme est différente de celle d'un homme, celle d'un enfant est différente de celle d'un adulte. Donc, si nous voulons étudier ces caractéristiques, il nous faut tout d'abord le faire avec des locuteurs précis ;
- deuxièmement, les trois aspects de la prosodie : la fréquence fondamentale, l'énergie et le débit phonétique sont aussi affectés par l'aspect articulatoire phonétique du mot ou de la phase. Particulièrement, si c'est dans le cas d'un langage tonal, la fréquence fondamentale (ou la F_0) change plus fortement en raison de l'existence du ton. Il nous faut donc étudier ces caractéristiques avec des locuteurs précis et avec des transcriptions précises ;
- et enfin, le corpus DES répond bien à ces besoins avec cinq émotions primaires, interprétées sur 13 segments de texte par les mêmes quatre locuteurs (voir le chapitre 4).

5.1.1.1 Fréquence fondamentale

Au niveau acoustique, la fréquence fondamentale F_o est un paramètre primordial pour la construction des accents (particulièrement dans les langues tonales), ainsi que pour la prosodie. C'est la raison pour laquelle il y a beaucoup d'études [Rank et al, 1998], [Montero et al, 1999], [Schröder 1999], [Oudeyer 2002], [Mozziconacci 2004], etc. qui cherchent à se servir de cette caractéristique pour caractériser les émotions. D'autres études comme [Vu et al, SFC2005] [Vu

et al, MajeSTIC2005] trouvent que F_0 est aussi un bon indice pour détecter les questions / non-questions. F_0 est encore un élément fondamental de la technique de synthèse appelée PSOLA qui est présentée par [Dutoit et al, 1993] et le projet très connu [MBROLA Project, 1996] est un des exemples qui utilise cette technique.

Cependant, malgré cette importance des composants de la prosodie on trouve peu de systèmes considérant explicitement ces composants comme un paramètre d'entrée dans les vecteurs caractéristiques. En effet, dans la plupart des systèmes de reconnaissance de la parole [Le V. B. 2006], [Shannon et al, 2004], [Zhu Xuan et al, 2002] etc., les coefficients de MFCCs sont toujours des paramètres uniques. Quelques études comme [Xu Shaoal 2004], [Zhu Xuan et al, 2002] ont réussi à expliquer cette contradiction en montrant que l'information de la prosodie comme F_0 ou l'énergie existe implicitement dans l'ensemble des coefficients MFCCs, donc l'utilisation des coefficients MFCCs entraine l'utilisation implicite de la prosodie. Nous verrons que c'est aussi le cas dans nos résultats dans la partie d'expérimentation avec la reconnaissance de l'émotion.

L'estimation de la fréquence fondamentale est un problème qui n'est toujours pas complètement résolu de nos jours. En effet, malgré une cinquantaine d'années d'étude environ, les techniques courantes ne sont toujours pas à un niveau idéal d'exactitude et de robustesse. Beaucoup de tentatives ont été faites dans des contextes spécifiques et un bon nombre d'entre elles fonctionnent bien dans leur contexte, mais développer un estimateur de F_o indépendant du contexte reste encore un vrai défi : jusqu'à aujourd'hui, on peut affirmer qu'il n'y a pas encore de détecteurs fonctionnent d'une manière satisfaisante quel que soit le domaine ou l'application.

Dans l'annexe A, nous faisons une synthèse des techniques appliquées pour l'extraction de la fréquence fondamentale : les approches dans le domaine temporel, les approches dans le domaine fréquentiel et les approches statistiques.

Dans notre travail, en utilisant PRAAT, nous utilisons également une version robuste par l'algorithme d'autocorrélation dans le domaine temporel pour calculer le contour de la fréquence fondamentale de tous les énoncés.

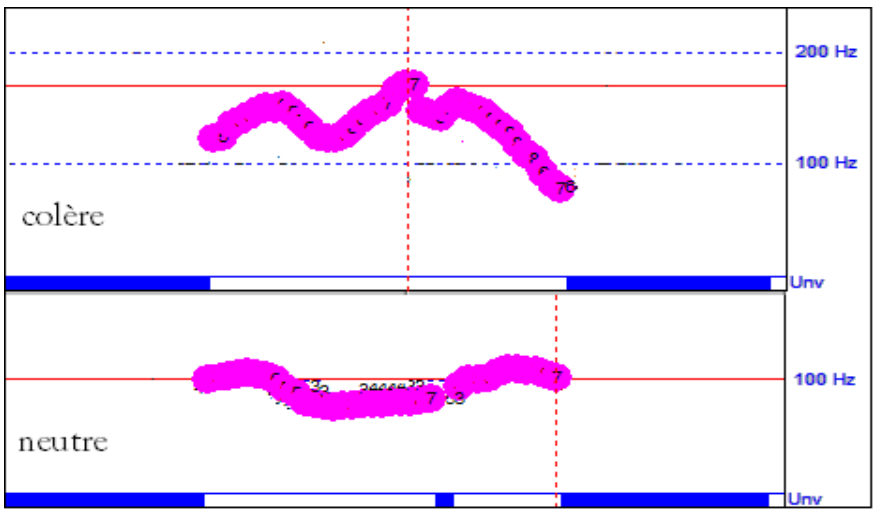

Figure 14 : Un exemple des contours de la fréquence fondamentale d'une phrase en colère et en neutre du corpus BES

La Figure 14 nous montre un exemple des contours de F_0 (mis en comparaison) d'un couple de deux états (colère et neutre) de la cinquième phrase du corpus BES ; bien que cela n'ait pas de sens en termes de statistique, une remarque peut être retenue : il existe une différence importante de la moyenne et de la variation de la F_o. C'est la raison pour laquelle, nous avons utilisé un ensemble d'opérateurs qui seront présentés dans la section 5.3.1.1 pour capturer ces différences.

5.1.1.2 Intensité

L'intensité a des liens évidents avec les états émotifs de la parole. Nous pouvons l'observer nous-même par des contours de l'intensité extraits à partir des phases du corpus BES dans des états émotionnels différents. La Figure 15 met en comparaison les deux contours de l'intensité de la cinquième phrase de BES.

En observant ces deux contours, on remarque que la vitesse de variation, et la moyenne de l'intensité pourraient être des bons paramètres pour discriminer ces deux états émotionnels. Avec PRAAT, nous utilisons également son algorithme par default qui convolute le carré du signal avec une fenêtre d'analyse gaussienne (Kaiser-20) pour calculer l'intensité. Les dérivées en premier et en deuxième ordre de l'intensité sont également introduites dans notre ensemble de paramètres.

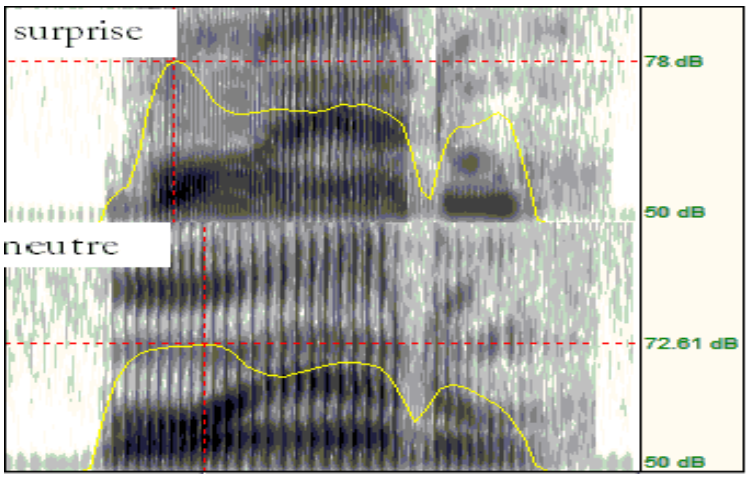

Figure 15 : Un exemple des contours de l'intensité d'une phrase du corpus BES en neutre et en surprise

5.1.1.3 Débit phonétique

Actuellement, à cause de la difficulté de l'extraction du débit phonétique, les études soit n'utilisent pas ce paramètre, soit utilisent une approximation dans sa mesure. Par exemple [Noble 2003] a remplacé l'extraction du débit par le calcul du taux des régions voisées / non-voisées. C'est pourquoi nous ne trouvons pas beaucoup de résultats portant simultanément sur les trois aspects de la prosodie.

En profitant d'annotations fournies par les deux corpus DES et BES, nous définissons alors le débit phonétique par la formule suivante :

$$DébitP = \frac{F_e}{E_p} \qquad (5.1)$$

où E_p est la durée en nombre d'échantillons d'un phonème et F_e est la fréquence d'échantillonnage. E_p est calculé en déterminant le nombre d'échantillons existants entre les deux frontières de chaque phonème.

5.1.1.4 Rapports relatifs

En tenant compte de la tolérance aux variations interlocuteur, en travaillant avec les trois caractéristiques de la prosodie, nous avons utilisé des nouveaux paramètres relatifs à la place des paramètres originaux. Contrairement aux valeurs absolues des paramètres originaux, les paramètres relatifs portent les informations en rapport instantané avec les moyennes de chaque paramètre, les valeurs des paramètres relatifs sont ainsi définies comme suit :

$$r_i = \frac{p_i}{\overline{P}} \tag{5.2}$$

Ces paramètres relatifs sont alors traités de la même manière que des paramètres originaux.

5.1.2. Paramètres spectraux

Parmi les paramètres spectraux, nous nous intéressons particulièrement aux trois ensembles de paramètres suivants : les coefficients MFCCs (*Mel Frequency Cepstral Coefficients*), les coefficients LFCCs (*Linear Frequency Cepstral Coefficients*) et les coefficients LPC (*Linear Predictive Coding*).

5.1.2.1 MFCC - Mel Frequency Cepstral Coefficients

Les coefficients MFCC sont des coefficients cepstraux très souvent utilisés en reconnaissance automatique de la parole. En effet, dans ce domaine, [Davis et Mermelstein 1980] ont prouvé que les coefficients MFCCs présentent les caractéristiques les plus robustes.

La Figure 16 montre les étapes pour extraire les coefficients MFCC (voir [Davis et Mermelstein 1980] et [Rabiner et al, 1993] pour le détail) :

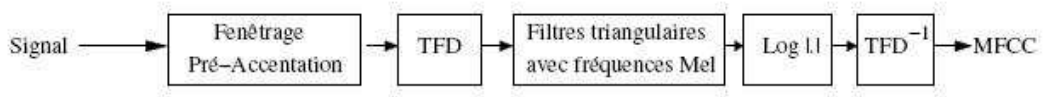

Figure 16 : processus de calcul des coefficients MFCC

Le calcul des paramètres MFCC utilise une échelle fréquentielle non-linéaire qui tient compte des particularités de l'oreille humaine.

Cette échelle de fréquence Mel est définie par :

$$B(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{5.1}$$

où f représente la fréquence en Hz et $B(f)$ est la fréquence correspondante en échelle de fréquence Mel.

Le spectre du signal est filtré par des filtres triangulaires (voir Figure 17) dont les bandes passantes sont équivalentes en domaine de fréquences Mel. Les points de frontières $B(m)$ des filtres en échelle de fréquence Mel sont calculés à partir de la formule (5.2)

. $$B(m) = B(f_b) + m\frac{B(f_h) - B(f_b)}{M+1} \qquad 0 \le m \le M+1 \tag{5.2}$$

où M désigne le nombre de filtres, f_h la fréquence la plus haute et f_b la fréquence la plus basse du signal.

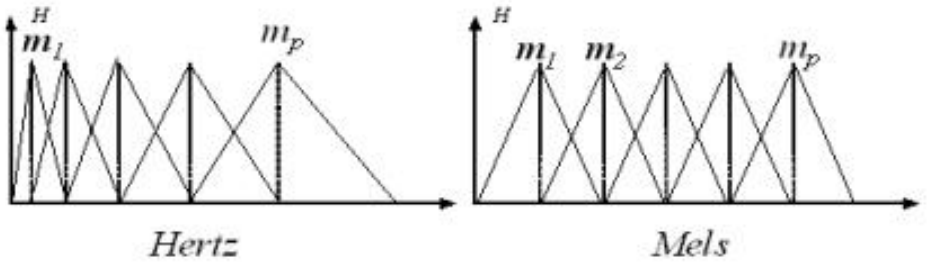

Figure 17 : Filtres triangulaires [Rabiner et al, 1993]

Dans le domaine fréquentiel, les points $f(m)$ discrets correspondants sont calculés d'après :

$$f_{b_m} = \left(\frac{N}{F_s}\right) B^{-1}\left(B\left(f_b\right) + m\frac{B(f_h) - B(f_b)}{M+1}\right) \tag{5.3}$$

où N est le nombre de points de FFT, M est le nombre de filtres, $B(.)$ signifie la transformation(5.1), et $B^{-1}(.)$ signifie la transformation inverse de (5.1) qui est formulée par :

$$B^{-1}\left(f_{mel}\right) = 700.\left[\exp\left(\frac{f_{mel}}{2595}\right) - 1\right] \tag{5.4}$$

Les MFCCs devraient être le premier candidat pour notre étude sur la reconnaissance de l'émotion parce que l'utilisation de ces coefficients pourrait donc donner de bonnes performances dans un système comme le nôtre qui cherche aussi à faire de la reconnaissance automatique des émotions à la place des êtres humains. En plus, les MFCCs sont aussi une technique standard, et leur performance a été bien vérifiée par plusieurs systèmes réels dans le domaine de reconnaissance de la parole comme [Vaufreydaz 2002] [Lima et al, 2004] [Le V. B. 2006] [Linarès et al, 2007].

Dans les systèmes de reconnaissance de la parole , les 12 MFCCs sont souvent utilisés, nous imposons aussi ces 12 MFCCs dans l'ensemble de parametres mais aussi avec $MFCC_0$ et une extension de 4 autres MFCCs qui font au total 17 coefficients parce que selon nos observations, $MFCC_0$ est un coefficient efficace et que l'ensemble de 16 coefficients MFCCs est aussi un ensemble de paramètres étudié dans les plusieurs systemès de reconnaissance de la parole, reconnaissance des traits extralinguistiques et aussi de reconnaissance de l'émotion comme [Akbar et al, 1998], [Vacher et al, 2003], [Zhou et al, 2007], [Beritelli et al, 2005].

Effectivement, dans le cas des MFCCs, en étudiant les coefficients nous constatons que l'information de l'énergie contenue dans le coefficient $MFCC_0$ est plus riche que l'information de l'énergie des trames (ou l'intensité) que nous avons présentée ci-dessus dans la section sur la

prosodie. La Figure 18 montre un exemple de la similarité entre le contour du premier coefficient $MFCC_0$ et le contour de l'intensité d'un segment de la parole.

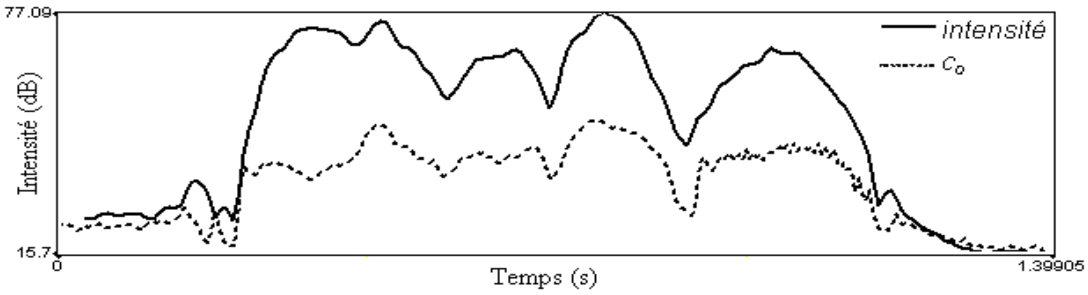

Figure 18 : Un exemple de $MFCC_0$ et l'intensité

Effectivement, plus exactement, $MFCC_0$ ne corresponds pas à l'énergie des trames, mais il est considéré comme la moyenne des énergies des sous-bandes fréquentielles (FBE ou *frequency band energy* en anglais) du signal, il contient alors une information différente et plus proche de la perception humaine de l'intensité [Zheng et al, 2001].

Bien que des systèmes de reconnaissance automatiques de la parole n'utilisent pas le FBE (ou $MFCC_0$) en raison de son instabilité, [Zheng et al, 2001] ont montré que si on voulait intégrer l'information de l'énergie au système, parmi les différents types d'informations sur l'énergie, le FBE donnait le meilleur résultat en combinaison avec les autres coefficients MFCC. C'est aussi la raison pour laquelle nous proposons d'étudier et d'utiliser ce paramètre dans la reconnaissance de l'émotion en prenant en compte ce FBE dans l'ensemble de 16 coefficients de MFCCs ainsi que leurs dérivées premières et deuxièmes pour produire un ensemble de 51 coefficients basés sur des MFCCs.

5.1.2.2 LFCC

En simulant la perception humaine par l'échelle Mel, les MFCCs amplifient le rôle des composants à basses fréquences, tout en atténuant le rôle des composants à hautes fréquences qui sont peut-être significatives pour la discrimination des émotions. C'est la raison pour laquelle, nous introduisons également des coefficients LFCCs dans notre ensemble de paramètres à étudier parce que ces coefficients sont calculés de la même manière que les MFCC, mais avec l'échelle linéaire et pas avec l'échelle Mel. La performance de ces paramètres LFCCs peut se trouver dans la section de l'expérimentation.

5.1.2.3 LPC

A côté des MFCCs et des LFCCs, les coefficients LPCs, qui sont une autre représentation de l'enveloppe spectrale du signal, sont aussi des candidats pour la recherche de l'ensemble des paramètres les plus efficaces.

Les coefficients LPC sont basés sur le modèle de production de la parole qui considère en première approximation que l'appareil de production de la parole (cordes vocales et conduit vocal complet) est constitué d'une source (source pseudopériodique ou source de bruit) et d'un filtre se comportant comme un résonateur (conduit vocal).

Le signal de parole peut être ainsi modélisé comme étant le signal en sortie d'un filtre $H(z)$ dont la source d'excitation à l'entrée du filtre $u(t)$ est, soit une série d'impulsions quasi-périodiques, soit une source de bruit aléatoire.

L'analyse LPC repose sur l'hypothèse que le filtre est un filtre tous-pôles, ce qui est une bonne approximation pendant la production des voyelles.

$$H(z) = \frac{S(z)}{G.U(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)} \qquad (5.5)$$

où G est le coefficient de gain, a_k sont les coefficients LPC et p est l'ordre du filtre.

Avec cette hypothèse, le signal de la parole peut être considéré comme un signal auto régressif :

$$s(n) = \sum_{k=1}^{p} a_k .s(n-k) + G.u(n) \qquad (2.6)$$

Les coefficients a_k et le gain G sont calculés grâce à des méthodes fondées sur le calcul de la matrice de covariance ou grâce à des méthodes fondées sur le calcul de la matrice d'autocorrélation (la méthode utilisée dans notre travail est basée sur la matrice d'auto corrélation).

Pour l'étude, nous utilisons PRAAT pour extraire 17 coefficients de MFCCs (y compris $MFCC_0$) et 16 coefficients de chacun des deux ensembles de LFCCs et de LPC avec leurs dérivées premières et deuxièmes. Donc, au total, nous allons utiliser soit 51 coefficients MFCCs, soit 48 coefficients LFCCs ou soit 48 coefficients LPCs.

5.1.3. Autres paramètres

5.1.3.1 Nombre de passages par zéro

Le nombre de passages par zéro et ses dérivés (ou *Zero Crossing* en anglais) est un paramètre intéressant qui est aussi introduit dans plusieurs systèmes de reconnaissance de la parole [Vaufreydaz 2002] [Le V. B. 2006].

Depuis qu'il a été popularisé par [Kedem 1986], l'utilisation du taux de passage par zéro (ZCR ou *Zero Crossing Rate* en anglais) a souvent été discutée, y compris dans le domaine de détection de la fréquence fondamentale (F_0, voir l'Annexe A). Comme son nom l'indique, il est défini par le nombre de passages par zéro dans une région définie de signal, divisé par le nombre d'échantillons de cette région [Gouyon et al, 2000] :

$$ZCR = \frac{1}{N-1} \sum_{n=1}^{N-1} \text{sign}(s(n)s(n-1)) \qquad (5.7)$$

où :

$$\text{sign}(x) = \begin{cases} 1 & \text{si } x < 0 \\ 0 & \text{si } x \geq 0 \end{cases} \qquad (5.8)$$

Beaucoup de chercheurs ont examiné les caractéristiques statistiques du ZCR pour l'utiliser ou le combiner avec des autres paramètres pour améliorer la qualité de leur système. [Scheirer et al, 1997] [Panagiotakis et al, 2005] ont utilisé le ZCR pour la discrimination entre la parole et la musique. [Le V. B. 2006] a également combiné le ZCR avec d'autres paramètres pour améliorer la qualité d'un système de reconnaissance de la parole. Le ZCR a aussi récemment été employé dans la détection de F_o comme [Rossignol et al, 1998] [Rossignol 2000], (voir Annexe A).

5.1.3.2 Nombre d'extrémités

Par l'observation sur le corpus DES et BES, nous constatons que le nombre d'extrémités (des maxima et des minima locaux) est aussi une caractéristique portant des informations émotionnelles. Il est donc un candidat de notre étude.

5.2. Outils d'extraction des paramètres

Parmi plusieurs logiciels disponibles, nous avons choisi PRAAT [Boersma et Weenink 2005] comme outil principal pour extraire la plupart de nos paramètres primaires parce que, d'une part, ce logiciel est très connu dans le domaine du traitement de la parole par sa simplicité dans l'utilisation, sa rapidité dans les calculs et sa flexibilité dans les modes de fonction, (en effet, le contrôle et l'interaction directe avec le logiciel par un langage script nous permet d'utiliser ses fonctionnalités dans notre système en temps réel), d'autre part, PRAAT est aussi un logiciel « Open Source », ainsi la méthodologie et l'implémentation de chaque l'algorithme sont bien vérifiées et validées par des utilisateurs dans le domaine.

Les étapes de traitement sont numérotées comme sur la Figure 19 :

(1) le système reçoit un morceau de signal en entrée ;

(2)(3) le système crée les données de configuration et les scripts et envoie la commande vers PRAAT pour que PRAAT lance les scripts créés par l'étape (2) ;

(4)(5) PRAAT charge les scripts et les modules préparés ;

(6)(7) PRAAT calcule les paramètres selon la configuration choisie et avertit le système quand sa tâche est terminée ;

(8)(9) le système reçoit les paramètres et les utilise pour la reconnaissance, puis renvoie les résultats vers l'utilisateur.

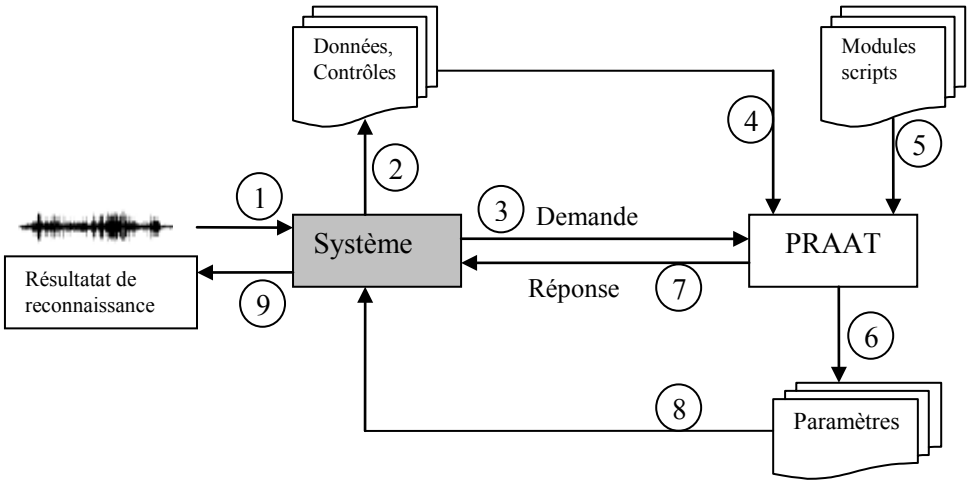

Figure 19 : Exploitation de PRAAT

A côté de PRAAT, nous utilisons les deux autres bibliothèques « Open source » appelée LtiLib [LtiLib 2006] et LibSVM [Hsu et al, 2003] pour réaliser des algorithmes de traitement plus complexe comme l'Analyse en Composantes Principales, l'Analyse Discriminante Linéaire, le KNN, le modèle de machine à vecteur de support etc. La LtiLib est développée au « Chair of Technical Computer Science LTI (Lehrstuhl für Technische Informatik) » à l'université de technologie Aachen en tant qu'une partie de plusieurs projets de recherche portant sur la vision par ordinateur pour la robotique, l'identification d'objets, la reconnaissance de la langue chantante et l'identification des gestes. Les détails peuvent se trouver dans la référence citée : [LtiLib 2006]. La LibSVM est développée et destinée particulièrement aux problèmes de classification utilisant la technique de machine à vecteurs de support. Les détails peuvent se trouver dans la référence citée : [Hsu et al, 2003].

Contrairement à l'utilisation de PRAAT, les deux bibliothèques LtiLib et LibSVM sont intégrées dans le noyau du système qui est développé sous l'environnement Windows avec Visual C++ 6.0.

5.3. Performance des paramètres

En travaillant avec les paramètres de parole, nous distinguons deux types d'informations qui peuvent être obtenus et que nous appelons les paramètres locaux et les paramètres globaux. Les paramètres locaux sont des paramètres instantanés qui sont extraits tout au long du signal (trame par trame), et les paramètres globaux sont des paramètres qui ne peuvent être obtenus que sur le signal entier ou sur un ensemble de paramètres locaux.

Notre méthodologie pour la reconnaissance des émotions commence par l'étude du comportement des paramètres au niveau global.

5.3.1. Paramètres globaux

5.3.1.1 Opérateurs

Pour pouvoir extraire les comportements globaux des paramètres, nous proposons un jeu de paramètres construits à partir des opérateurs suivants :

No	*Opérateurs*	*Description*	*Remarque*
1	Max	Valeur maximale du paramètre dans un énoncé	Pour éviter des occurrences anormales, 5% des valeurs les plus élevées de ce paramètre sont éliminées et cette valeur maximale est le maximum du reste (95%).
2	Min	Valeur minimale du paramètre dans un énoncé	5% des valeurs les plus basses de ce paramètre sont éliminées et cette valeur minimale est le minimum du reste (95%)
3	Mean	Valeur moyenne du paramètre dans un énoncé	5% des valeurs les plus basses et 5% des valeurs les plus élevés de ce paramètre ont été éliminées avant d'extraire ce paramètre. La valeur obtenue par cet opérateur montre l'orientation générale du paramètre en état émotionnel spécifique
4	Median	Valeur médiane du paramètre dans un énoncé	
5	Variance	Ecart-type du paramètre dans un énoncé	Cet opérateur nous permet de savoir si ce paramètre est plus convergent ou plus reparti.
6	Range	Gamme du paramètre pour un énoncé (Max-Min)	Grâce à cet opérateur, on peut deviner le dégrée de variété de ce paramètre
7	RisingFallingCountRatio [Vu 2007]	Rapport entre le nombre de pas croissants / le nombre de pas décroissants	Est-ce que, en général, la forme de ce paramètre est montante ?
8	RisingFallingSumRatio [Vu 2007]	Rapport entre la somme des croissances / la somme des décroissances	Est-ce que, au total, la forme de ce paramètre est montante ?

Tableau 18 : Les 8 opérateurs imposés sur l'ensemble de paramètres originaux

Pour l'étude, nous nous servons des valeurs obtenues avec ces 8 opérateurs comme paramètres dans les deux processus : l'apprentissage et la reconnaissance.

Ces 8 opérateurs peuvent être groupés en 3 types :

- le premier type comprend les 4 opérateurs Max, Min, Mean et Médian, qui donnent essentiellement des informations sur la valeur du paramètre ;
- le deuxième type se compose des 2 opérateurs Variance et Range qui portent essentiellement sur la caractéristique de la distribution des valeurs du paramètre ;
- et le troisième type contient les 2 operateurs restants : RisingFallingCountRatio et RisingFallingSumRatio qui ont récemment été utilisés pour la détection Question/Non-Question dans le travail de [Vu 2007]. Comme leur nom l'indique, ces deux opérateurs

caractérisent la forme des paramètres ou autrement dit, ils capturent les caractéristiques en évolution du contour. La Figure 20 illustre la différence de la mesure de RisingFallingCountRatio et de RisingFallingSumRatio. Comme dans l'illustration, RisingFallingCountRatio prend en compte la durée où il y a de la croissance/décroissance, alors que RisingFallingSumRatio porte sur la quantité de croissance/décroissance (ou la grandeur des flèches). Donc, ces deux derniers opérateurs forment un couple pour caractériser la forme des contours des paramètres. Ainsi, les conclusions obtenues sur les valeurs de ces deux opérateurs portent plutôt sur les comportements globaux des contours au cours de leur évolution temporelle.

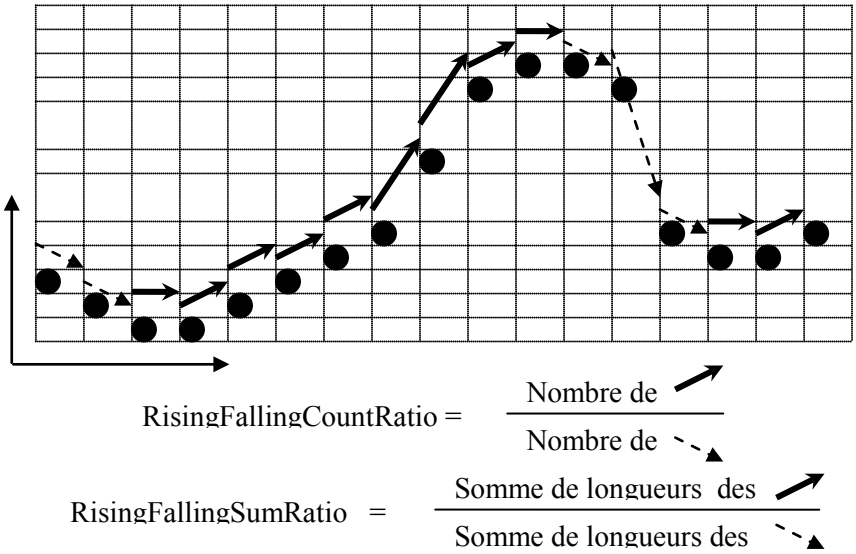

Figure 20 : Explication des deux opérateurs RisingFallingCountRatio et RisingFallingSumRatio

Pour enrichir l'information sur l'évolution temporelle, nous avons également appliqué ces 8 opérateurs sur les dérivées du premier et du deuxième ordre de tous les paramètres étudiés.

Donc, théoriquement nous obtiendrons au maximum les 8 paramètres globaux sur chaque paramètre local. Cependant en raison de la répétition de l'information ou des cas non nécessaires (par exemple le maximum d'une dérivée de nombre de passages par zéro), notre ensemble pratique de paramètres n'est pas toujours complet. Nous préciserons ce nombre de paramètres lors de chaque cas d'études.

5.3.1.2 Normalisation par rapport au neutre

Avant d'entrer dans les détails, nous voulons discuter un peu de la dépendance des paramètres globaux par rapport au locuteur et au contexte. En effet, par exemple, la fréquence fondamentale (F_0) dépend fortement de l'état physique et de l'articulation des phonèmes. La Figure 21 montre la variété des valeurs maximales de F_0 en fonction des 13 énoncés pour les 5 émotions avec deux locuteurs différents.

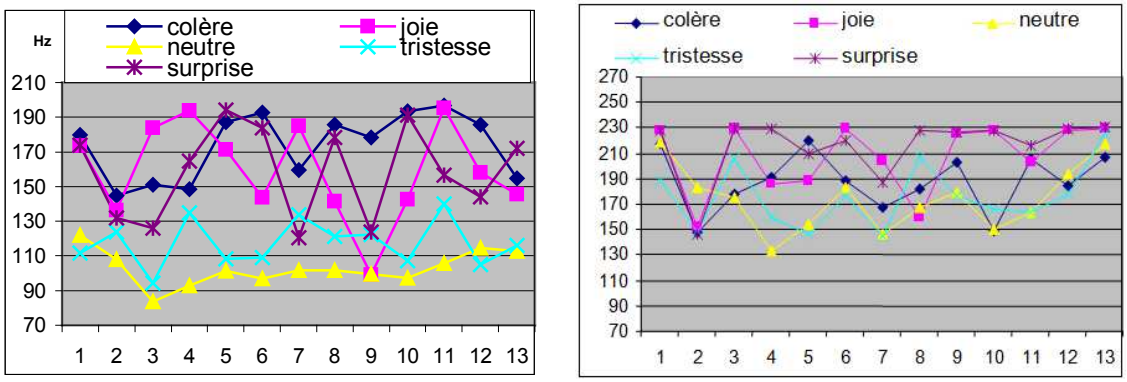

Figure 21 : Les valeurs maxima de F_0 en fonctions des 13 énoncés des deux locuteurs HO (homme) et DHC (femme) du corpus DES

Donc, à cause de cette variété, les résultats des études basées sur des valeurs absolues des paramètres globaux risquent de ne pas être généraux et de ne pas être applicables dans la plupart des autres cas. Nous proposons donc, dans cette étape d'étude, d'utiliser une normalisation par rapport au neutre.

Pour chaque locuteur, pour chaque énoncé, et pour chaque paramètre mesuré, nous utilisons la mesure obtenue en état neutre comme référence pour les autres mesures de ce paramètre pour la colère, la joie, la tristesse et la surprise.

En appliquant la normalisation par rapport au neutre, nous trouvons deux avantages qui peuvent améliorer la généralité et la globalité des résultats obtenus :

- premièrement, la dépendance des paramètres par rapport au locuteur est fortement réduite grâce au rapport relatif sur le même locuteur.
- deuxièmement, la dépendance des paramètres par rapport aux phénomènes de co-articulation est également fortement réduite grâce à l'utilisation d'énoncés de même structure phonémique pour calculer les rapports durant la normalisation.

5.3.1.3 Analyse de la performance d'un paramètre

La Figure 22 montre les maxima de l'intensité pour différentes émotions en pourcentage par rapport à l'état neutre. La variance de ce paramètre, également normalisée par rapport au neutre, est aussi montrée. Ceci nous donne des indications intéressantes sur le comportement, ainsi que sur la capacité de la discrimination des émotions, de chaque paramètre. Par exemple dans cette illustration, nous constatons qu'il y a une différence significative de l'intensité entre la tristesse et les autres émotions avec des petits écart-type (montrées par VarianceVsNeutre). Cela nous permet de conclure que, pour les corpus utilisés, l'intensité peut être utilisée pour discriminer les deux groupes : tristesse / colère + joie + surprise.

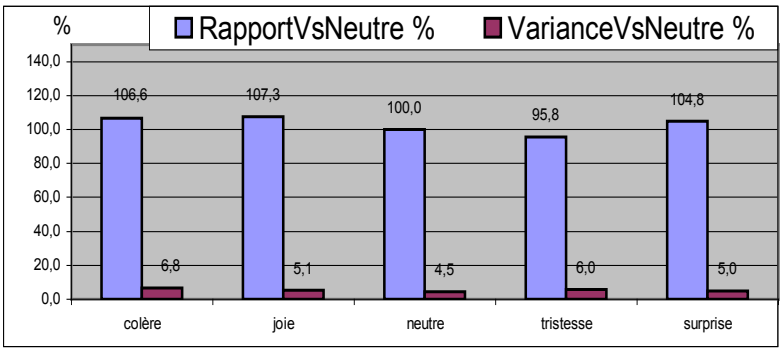

Figure 22 : Rapports des maxima de l'intensité en différentes émotions avec l'état neutre

Cette capacité de discrimination est appelée la performance du paramètre. Donc, un paramètre peut posséder une capacité de discrimination des cinq émotions, ou simplement une capacité de discrimination de deux émotions quelconques.

Nous avons vérifié la capacité prédite en effectuant des tests de performance avec l'algorithme de construction d'arbre C4.5 [Quinlan 1993]. L'implémentation de cet algorithme provient du logiciel *open-source* Weka[5] qui comprend les algorithmes de *classification*, *régression*, *clustering*, *règles d'association* écrits en Java. Plus d'informations de Weka peuvent se trouver dans la citation. Le Tableau 19 montre un exemple d'un arbre de décision entrainé seulement par des rapports de variance de F_0.

```
varianceF0Ratio <= 1.049977
|   varianceF0Ratio <= 0.999249
|   |   varianceF0Ratio <= 0.41633 : colère (5.0/1.0)
|   |   varianceF0Ratio > 0.41633 : tristesse (35.0/11.0)
|   varianceF0Ratio > 0.999249
|   |   varianceF0Ratio <= 1 : neutre (52.0)
|   |   varianceF0Ratio > 1
|   |   |   varianceF0Ratio <= 1.023613 : colère (2.0)
|   |   |   varianceF0Ratio > 1.023613 : tristesse (3.0)
varianceF0Ratio > 1.049977
|   varianceF0Ratio <= 1.229065
|   |   varianceF0Ratio <= 1.083727 : joie (4.0/1.0)
|   |   varianceF0Ratio > 1.083727
|   |   |   varianceF0Ratio <= 1.16065 : tristesse (5.0)
|   |   |   varianceF0Ratio > 1.16065
|   |   |   |   varianceF0Ratio <= 1.205336 : colère (6.0/1.0)
|   |   |   |   varianceF0Ratio > 1.205336 : joie (3.0/1.0)
|   varianceF0Ratio > 1.229065 : surprise (145.0/95.0)
```

Tableau 19 : L'arbre de décision construit par des rapports de variance de F_0

Parmi les corpus étudiés dans le chapitre 4, nous nous servons du corpus DES pour nos études des paramètres globaux car ce corpus répond bien à nos requis de normalisation neutre avec les énoncés de même structure phonémique interprétés avec différentes émotions par tous les locuteurs (voir la section 4.2).

[5] http://www.cs.waikato.ac.nz/~ml/weka/

Bien qu'il y ait toujours des différences de valeur, de tendance des paramètres chez des locuteurs différents, nous ne nous intéressons qu'aux tendances générales des paramètres. C'est la raison pour laquelle les rapports moyens de tous les locuteurs sont toujours utilisés pour les analyses dans la section suivante. Les variances de ces rapports sont également considérées pour apprécier la sélectivité de ces paramètres (voir les colonnes VarianceVsNeutre dans la Figure 22).

Il faut noter qu'en appliquant la normalisation par rapport au neutre sur un énonce d'entrée *E*, l'énoncé en neutre *EN* (celui qui est du même locuteur, et qui a la même structure phonémique que *E*) correspondant avec cet énoncé *E* sera pris comme la référence pour le calcul du rapport par rapport au neutre de *E* :

$$RP_E = P_E / P_{EN}$$

où *RE* est le rapport par rapport au neutre de *E*, P_E, P_{EN} sont respectivement les paramètres extraits à partir de *E* et à partir de *EN*.

Donc, si *E* est un énoncé en neutre, RP_E sera toujours égal à 1 car *E* et *EN* sont le même énoncé.

Autrement dit, les énoncés en neutre sont toujours correctement reconnus avec une simple règle dans le modèle de l'arbre de décision : RP_E =1. C'est la raison pour laquelle, nous ne prenons pas en compte le neutre dans la mesure du taux de reconnaissance globale du système. Cependant, l'état neutre est toujours une branche des arbres de décisions pour capturer les confusions des autres états émotionnels avec le neutre. Par exemple l'arbre de décision du Tableau 19 va classifier de « neutre » tous les énoncés qui présenteront un rapport de variance de F_0 (par rapport au neutre) situé dans l'intervalle (0,999249 ; 1].

De ce fait, la performance d'un paramètre n'est estimée que par le taux de reconnaissance des énoncés en quatre émotions non-neutres du corpus DES : la colère, la joie, la tristesse, la surprise. Parce qu'un énoncé d'entrée peut être reconnu comme un des 5 états émotionnels (le neutre reste toujours une branche pour la sortie du modèle), donc 20 % représentera le taux aléatoire de classification correcte.

5.3.1.4 Expérimentations avec des paramètres de prosodie

Parce que la fréquence fondamentale est un composant assez important de la prosodie pour des études de l'émotion, nous étudions en détail la fréquence fondamentale dans la première section. Dans les deuxième et troisième sections, nous présenterons des analyses et des résultats pour les deux autres composantes : l'intensité et le débit phonétique. Enfin, un essai de fusion des paramètres pour tester la performance sera discuté.

5.3.1.4.1 Analyse de la fréquence fondamentale

En analysant les paramètres obtenus par les opérateurs sur F_0 et en utilisant le modèle de l'arbre de décision pour la vérification des remarques obtenues (avec l'utilisation des paramètres de F_0 pour la reconnaissance des émotions sur le corpus DES), les conclusions suivantes ont été tirées dans le contexte du corpus DES :

- en terme de l'amplitude de F_0, la surprise montre toujours un état le plus « excité » (F_0 est la plus haute) parmi les cinq états émotionnels ; au contraire, la tristesse est un état présentant le moins de différences par rapport au neutre ;

- en se basant seulement sur l'amplitude de F_0 (évaluée par les valeurs obtenues avec des opérateurs du premier type), nous pouvons grouper les 5 émotions en 2 groupes : les émotions « hautes » et les émotions « basses » La colère, la joie, la surprise appartiennent aux émotions hautes car elles donnent toujours les plus grandes valeurs. Au contraire, la tristesse et le neutre se trouvent toujours dans la zone des émotions basses dont les valeurs sont assez petites. Pour illustration, nous donnons ici un exemple des valeurs RP_E des maxima de F_0 entre la colère, la joie, la surprise, la tristesse par rapport au neutre, qui sont respectivement de 1,18 ; 1,23 ; 1,26 et 1,04. A partir de ces rapports, il est aisé de constater que ces informations sur l'amplitude de F_0 sont assez discriminatives pour les deux groupes d'émotions : la tristesse + le neutre / la colère + la surprise + la joie. Par l'expérimentation, la capacité de discrimination de la tristesse (tristesse/colère + joie + surprise) par la grandeur de F_0 a été vérifiée avec un taux de classification correcte 78 % en utilisant seulement les 4 paramètres : Max, Mean, Moyenne et Median. Nous ne tenons pas compte du neutre car comme mentionné, le neutre est toujours reconnu à 100 % ;
- nous constatons également que ces deux groupes d'émotions deviennent plus séparables avec les valeurs obtenues par les opérateurs du deuxième type *Variance* et *Range*. En effet, le taux de classification correcte obtenu pour la classification tristesse / colère + joie + surprise en utilisant la combinaison des valeurs de *Variance* et de *Range* est de 81 % ;
- les opérateurs du troisième type, *RisingFallingCountRatio* et *RisingFallingSumRatio*, appliqués sur le paramètre F_0 nous permettent d'obtenir des valeurs avec une différence significative pour classifier les 5 émotions en trois groupes : la colère / la joie + la tristesse / la surprise dont les rapports par rapport au neutre RP_E sont respectivement ~0,88 / ~1,14 ; ~1,09 / ~1,28. En effet, le fait que la colère possède les valeurs de RisingFallingCountRatio et de RisingFallingSumRatio beaucoup plus faibles que ceux du neutre (RP_E = 0,88) veut dire que malgré la valeur moyenne de F_0 assez élevée chez la colère (RP_E =1,18), généralement en cet état, F_0 a tendance à descendre plus que monter.

Ce résultat correspond aussi à nos observations dans les conversations quotidiennes où la colère possède une forte et rapide croissance de F_0 dans une petite partie au début des énoncés, et suivi d'une décroissance. La Figure 23 est un exemple très net de cette tendance du contour de F_0 d'une phrase du corpus DES.

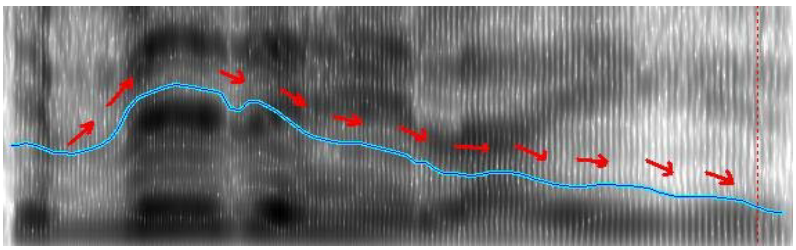

Figure 23 : Exemple de la tendance de décroissance de F_0 en colère

L'essai de classification de ces trois groupes a été effectué en fusionnant les deux paramètres *RisingFallingCountRatio* et *RisingFallingSumRati*o, mais le résultat (24 %) n'est pas comme attendu. Cela peut s'expliquer par le fait que ces paramètres ne permettent pas encore de capturer l'évolution en fonction de temps de la F_0 (la forme) ou par le chevauchement important

des paramètres en différentes émotions qui est causé par la variance importante montrée dans la Figure 24 suivante (0,589 ; 0,670 ; 0,540 ; 0,582 et 1,131).

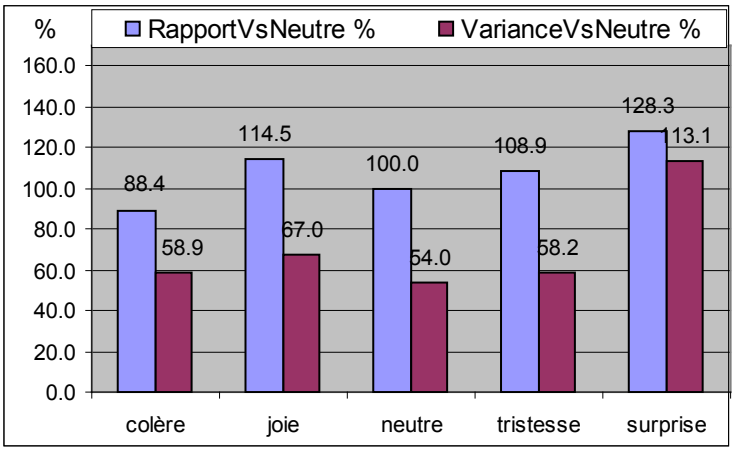

Figure 24 : RisingFallingSumRatio des différentes émotions par rapport au neutre

- pour les 8 valeurs obtenues par 8 opérateurs sur F_0, malgré des petites différences, nous ne trouvons pas encore un moyen fiable pour discriminer la tristesse et le neutre ;

Le Tableau 20 montre les résultats obtenus en faisant la reconnaissance avec les 8 paramètres isolés. La dernière colonne donne le résultat en fusionnant les 8 paramètres. Ces résultats sont obtenus sans compter les 100 % correctement reconnu du neutre.

	Max (%)	***Min (%)***	***Mean (%)***	***Median (%)***	*Variance* ***(%)***	***Range (%)***	***Rising-Falling-CountRat (%)***	***Rising-Falling-SumRat (%)***	***Tous les 8 paramètres (%)***
F_0	33,8	26,0	37,5	31,8	40,4	35,6	36,1	30,1	39,7
ΔF_0	30,8	29,8	26,2	28,5	35,1	38,5	32,0	28,8	32,5
$\Delta\Delta F_0$	24,6	27,9	25,4	24,5	27,4	26,4	26,9	24,1	30,8

Tableau 20 : Résultats de classification des 4 émotions

Selon le Tableau 20 ; 40,4 % est le taux de classification le plus élevé en utilisant uniquement que la variance de F_0. On peut dire que ce résultat est assez élevé avec un seul paramètre (en comparaison avec le taux aléatoire 20 %), cependant d'après la matrice de confusion du Tableau 21, en utilisant seulement la variance, presque toutes les émotions *hautes* sont mal reconnues comme, par exemple, la surprise. Donc, ce paramètre de variance ne sera utile qu'à discriminer les deux groupes émotions *hautes* / *basses* comme nous l'avons remarqué ci-dessus.

Moins efficace, selon notre observation avec le corpus DES, la fusion des tous les paramètres de F_0 ne donne pas d'amélioration avec 39,7 % de classification correcte. Cependant, en observant les matrices de confusion du Tableau 21, nous constatons que la confusion devient plus uniforme, cela veut dire que les autres aspects de F_0 participent aussi au processus de classification.

Variance de F_0	***colère (%)***	***joie (%)***	***neutre (%)***	***tristesse (%)***	***surprise (%)***
Colère	7,7	3,8	1,9	23,1	63,5
Joie	0,0	5,8	0,0	13,5	80,8
Tristesse	5,8	3,8	0,0	55,8	34,6
Surprise	1,9	3,8	0,0	1,9	92,3
Moyenne	40,4				
8 paramètres	***colère (%)***	***joie (%)***	***neutre (%)***	***tristesse (%)***	***surprise (%)***
Colère	35,8	15,1	0,0	32,1	17,0
Joie	19,2	36,5	0,0	17,3	26,9
Tristesse	34,6	5,8	1,9	42,3	15,4
Surprise	11,5	32,7	0,0	11,5	44,2
Moyenne	39,7				

Tableau 21 : Résultats de classification des 4 émotions par la variance de F0 et par la fusion de tous les 8 paramètres.

A partir du Tableau 20, nous remarquons encore que les paramètres de F_0 sont les plus efficaces, les paramètres de ΔF_0 sont au deuxième rang, et ceux de $\Delta\Delta F_0$ ne montre pas beaucoup leur efficacité en mode isolé ; en mode de fusion, ces 8 paramètres de $\Delta\Delta F_0$ nous donne un signe faible de leur capacité de la discrimination par un taux de classification pas très élevé : 30,8 %.

Le test de la combinaison de ces 24 paramètres de F_0, de ΔF_0 et de $\Delta\Delta F_0$ n'améliore pas beaucoup le taux de classification : 39,9 % qui sont montré dans le Tableau 22. Cela peut s'expliquer par l'existence de corrélations fortes entre les paramètres et que la concaténation directe des paramètres n'est pas une bonne méthode pour sélectionner les paramètres car elle ajoute non seulement des bonnes informations mais aussi beaucoup de bruits, ce qui baisse la performance du système. Une autre approche plus adéquate pour sélectionner les paramètres sera expérimentée dans la section portant sur des paramètres locaux.

	colère (%)	***joie (%)***	***neutre (%)***	***tristesse (%)***	***surprise (%)***
Colère	42,3	17,3	0,0	25,0	15,4
Joie	21,2	44,2	0,0	11,5	23,1
Tristesse	26,9	23,1	1,9	40,4	7,7
Surprise	15,4	34,6	0,0	17,3	32,7
Moyenne	39,9				

Tableau 22 : Taux de reconnaissance en utilisant 24 paramètres de F0, de ΔF0 et de ΔΔF0

Dans cette section, nous avons analysé en détail des aspects qui peuvent être extraits sur F_0 et qui peuvent donner des informations discriminatives pour identifier les quatre émotions. Il y a des paramètres assez efficaces comme le rapport des moyennes, le rapport des variances de F_0 mais il y a aussi des paramètres dont nous ne voyons pas beaucoup leurs efficacités en les utilisant isolement comme les minima de F_0, particulièrement, ou comme les dérivées secondaires $\Delta\Delta F_0$.

Cependant, nos résultats sur les comportements de F_0 nous permettent de réaffirmer le rôle important de F_0 dans l'expression des émotions sur le corpus étudié, de nous donner une vue globale de F_0 dans les différents états émotionnels.

Et enfin, la normalisation par rapport au neutre se montre une méthode efficace qui nous permet de sortir des informations utiles des paramètres en éliminant des aspects qui influencent la fréquence fondamentale comme la variabilité entre locuteurs, la variété de la structure articulatoire des énoncés, etc. Cela est aussi bien démontré par la différence significative en comparaison des résultats obtenus avec les paramètres bruts de F_0 (sans normalisation par rapport au neutre) : 26,4 % (dans le Tableau 23 suivant) et 39,9 % par notre méthode de normalisation par rapport au neutre (Tableau 22).

	Colère (%)	***joie (%)***	***neutre (%)***	***Tristesse (%)***	***surprise (%)***
Colère	32,7	19,2	17,3	13,5	17,3
Joie	17,3	17,3	19,2	13,5	32,7
Tristesse	26,9	17,3	19,2	28,8	7,7
surprise	23,1	36,5	3,8	9,6	26,9
moyenne	26,4				

Tableau 23 : Taux de reconnaissance en utilisant 24 paramètres bruts de F0, de ΔF0 et de ΔΔF0

Dans la section suivante, nous continuons avec d'autres aspects de la prosodie, l'intensité et le débit phonétique, de même manière.

5.3.1.4.2 Analyse de l'intensité de prosodie

L'intensité est un paramètre reconnu comme peu fiable pour une utilisation dans la reconnaissance car elle est affectée par plusieurs aspects comme l'aspect physique du locuteur, les dispositifs d'enregistrement et de transmission, etc. Dans cette partie de l'analyse, nous espérons par notre approche pouvoir faire ressortir des comportements globaux de l'intensité.

Le Tableau 24 contient les résultats statistiques et les résultats de classification en utilisant isolement les paramètres de l'intensité.

	Max (%)	***Min (%)***	***Mean (%)***	***Med. (%)***	***Var. (%)***	***Range (%)***	***Rising-Falling-CountRatio (%)***	***Rising-Falling-SumRatio (%)***	***Tous les 8 param. (%)***
Intensité	***40***	***37***	***40***	***43***	***16***	***22***	***30***	***25***	***40***
Δ Intensité	***37***	29	24	26	26	31	21	20	34
ΔΔ Intensité	***33***	29	31	31	28	31	26	31	32

Tableau 24 : Taux de reconnaissance en utilisant les rapports de l'intensité

En comparaison avec la fréquence fondamentale du Tableau 20, l'intensité semble plus discriminante pour classifier les quatre émotions du corpus DES. Effectivement, nous obtenons

le taux de classification le plus élevé 43 % avec la médiane de l'intensité. La Figure 25 nous donne une vue globale des valeurs médianes de l'intensité pour différentes émotions par rapport au neutre.

Par cette figure, l'efficacité du paramètre médian (43,3 %) peut s'expliquer par la dissociation des états émotionnels qui est caractérisée par des valeurs moyennes assez différentes pour les 4 groupes : colère (RP_E=1,066) + joie (RP_E=1,073) / neutre (RP_E=1) / tristesse (RP_E=0,958) / surprise (RP_E=1,048) et des petites valeurs de variance (0,068 ; 0,051 ; 0,045 ; 0,060 et 0,050).

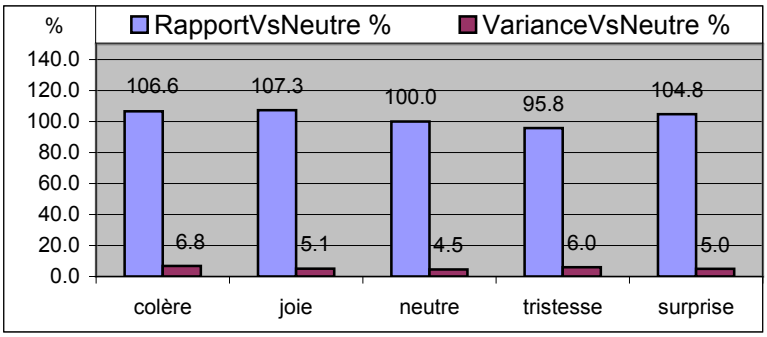

Figure 25 : Rapport des médians de l'intensité au neutre

La conclusion de l'importance de l'intensité dans l'expression émotionnelle peut être aussi reconfirmée sur les corpus étudiés par ces résultats.

Quelques autres remarques :

- semblablement à la fréquence fondamentale, les trois émotions, colère, joie et surprise possèdent des valeurs de l'intensité normalement et relativement plus élevées que celles du neutre et de la tristesse, mais en termes de l'intensité, la joie devient l'émotion la plus forte, la surprise n'est qu'au troisième rang après la colère ;
- si en analysant la fréquence fondamentale, nous ne trouvons pas encore des caractéristiques discriminatives entre la tristesse et le neutre, par contre, en considérant la grandeur de l'intensité, la tristesse se démarque par ses petites valeurs, qui sont toujours les plus basses par rapport aux quatre autres émotions. Nous avons profité de ce constat pour sortir la tristesse des 4 émotions et 81,3 % est le taux de classification correcte que nous avons obtenu :

	autre émotions (%)	***tristesse (%)***
autre émotions (%)	80,1	19,9
tristesse (%)	15,4	84,6
moyenne	81,3	

Tableau 25 : Classification tristesse / colère + joie + surprise en utilisant la médiane de l'intensité

- les dérivées première ΔIntensité et seconde ΔΔIntensité ne sont pas non plus très efficaces en mode isolé ainsi qu'en mode de fusion par rapport à l'intensité ;
- en général, les contours de l'intensité, de ΔIntensité et de ΔΔIntensité ne montrent pas des comportements spécialement efficaces à travers les valeurs des deux paramètres

RisingFallingCountRatio et RisingFallingSumRatio comme dans le cas de la fréquence fondamentale ;

- 40,5 % est le taux de classification correcte obtenu en fusionnant les 24 paramètres de l'intensité, de ΔIntensité et de ΔΔIntensité. Comme nous avons expliqué ci-dessus, la dégradation du taux de reconnaissance en mode de fusion peut être causée par la corrélation entre les paramètres ainsi que par l'introduction de plus de bruits que d'informations utiles.

Nous avons donc analysé et étudié la performance de deux éléments importants de la prosodie. Dans la partie suivante, nous travaillerons avec le débit phonétique.

5.3.1.4.3 Débit phonétique

Parmi les trois éléments de la prosodie, le débit phonétique est le paramètre le plus difficile à obtenir automatiquement et précisément, en effet, pour extraire ce type d'informations, il nous faut un système de reconnaissance automatique de la parole en arrière-plan.

Dans le cadre du corpus DES, nous profitons de repères des phonèmes dans les fichiers d'annotation disponibles pour la mesure du débit phonétique. Voir la section 4.2.4.1 pour plus de détails.

Le Tableau 26 contient les résultats de classification des 4 émotions en utilisant isolément les paramètres extraits à partir du débit phonétique.

	Max (%)	***Min (%)***	***Mean (%)***	***Med. (%)***	***Var. (%)***	***Range (%)***	***Rising-Falling-CountRatio (%)***	***Rising-Falling-SumRatio (%)***	***Tous les 8 param. (%)***
DébitP	26,9	17,8	26,8	19,2	27,4	20,2	16,8	24,5	26,9
ΔDébitP	24,5	16,6	26,0	20,2	24,5	23,3	21,6	17,3	28,5
ΔΔDébitP	25,5	27,9	25,5	24,5	27,4	26,4	26,9	22,1	30,8

Tableau 26 : Taux de reconnaissance en utilisant les rapports du débit phonétique

Les mauvais résultats du Tableau 26 peuvent s'expliquer essentiellement par la grande variance des paramètres du débit phonétique qui est causée par l'inexactitude dans la détermination des phonèmes ainsi que les frontières phonémiques du corpus DES. Malgré ce fait, les petites améliorations obtenues par la fusion des 8 paramètres nous permettent de réaffirmer la contribution du débit phonétique aux expressions émotionnelles.

De plus, d'autres comportements du débit phonétique sont constatés :

- parmi les cinq émotions, la tristesse présente toujours un état le plus « *lent* » en terme de vitesse (RP_E=~0,95). La surprise et la joie sont les deux états les plus « *rapides* » dont les débits phonétiques sont les plus grands : RP_E=~1,99 et RP_E=~112 %, respectivement par rapport au neutre ;
- en comparaison avec le neutre, la colère semble de ne pas présenter trop de différences pour les 5 paramètres qui sont la valeur maximale, la valeur minimale, la valeur moyenne, la variance et la range, à l'exception de la durée de croissances du débit en

colère qui est plus grande que celle du neutre (capturée par l'opérateur : RisingFallingCountRatio) ;

- la fusion des 24 paramètres du débit phonétique nous donne un taux de classification correcte d' environ 28,1 % pour les 4 émotions.

Jusqu'ici, dans le cadre du corpus étudié, nous avons analysé isolément les comportements des trois composants principaux de la prosodie pour la classification des émotions en comparaison relative des paramètres de ces trois composants avec l'état neutre. Parmi ces trois composants, la fréquence fondamentale et l'intensité se montrent les plus nettes pour les expressions des émotions et donc les plus efficaces pour la discrimination entre ces émotions. Le débit phonétique est un paramètre difficile à obtenir et aussi difficile à utiliser, mais il porte aussi des informations émotives. La fusion de ces trois aspects permettrait-elle des améliorations de la discrimination ?

5.3.1.4.4 Fusion des aspects de la prosodie

La sélection des paramètres efficaces pour la fusion peut être effectuée en appliquant des techniques comme l'Analyse en Composantes Principales, ou en se basant sur des critères comme celui de Fisher. Ces techniques seront testées avec notre approche utilisant des paramètres locaux. Dans cette partie, comme notre objectif principal est de chercher à comprendre pour obtenir une vue globale des comportements des paramètres pour une meilleure utilisation de ces paramètres dans notre système, nous appliquerons simplement la concaténation directe des paramètres pour tester leur performance.

La première fusion que nous avons étudiée est la fusion entre la fréquence fondamentale et l'intensité. Au total, 24 paramètres de chaque composant sont mis en ensemble dans un vecteur de 48 paramètres. L'ensemble des vecteurs des caractéristiques sont ensuite utilisés pour entraîner un modèle d'arbre de décision C4.5 avec une approche de validation croisée par la division en 10 plis (« folds »). Le résultat est montré dans le tableau suivant :

	colère (%)	***joie (%)***	***neutre (%)***	***tristesse (%)***	***surprise (%)***
Colère	30,0	32,0	2,0	10,0	26,0
joie	20,0	36,0	0,0	8,0	36,0
tristesse	10,0	2,0	4,0	76,0	8,0
surprise	18,0	34,0	2,0	12,0	34,0
moyenne sans neutre	44,0				

Tableau 27 : Fusion de la fréquence fondamentale (24 paramètres) et l'intensité (24 paramètres)

La combinaison entre la fréquence fondamentale et l'intensité donne un taux de classification de 44,0 % pour les quatre émotions. Ce taux est plus élevé si on le compare avec les résultats obtenus en utilisant isolement soit F_0 ou soit l'intensité. Cependant, en observant la matrice de confusion du Tableau 27, nous constatons que ces deux paramètres ne sont pas suffisants pour séparer les deux états joie et surprise. En effet, la confusion entre la joie et la surprise est toujours très élevée (34 % et 36 %). Une combinaison ajoutant le débit phonétique donnerait-elle de meilleurs résultats diminuant cette confusion ?

	colère (%)	***joie (%)***	***neutre (%)***	***tristesse (%)***	***surprise (%)***
colère	37,5	29,2	0,0	6,3	27,1
joie	20,8	41,7	0,0	8,3	29,2
tristesse	12,5	4,2	4,2	66,7	12,5
surprise	29,5	30,8	0,0	14,6	25,0
moyenne sans neutre	42,7				

Tableau 28 : Fusion de la fréquence fondamentale (24 paramètres), l'intensité (24 paramètres) et le débit phonétique (24 paramètres)

En réponse, nous constatons que la fusion de ces trois composants, fréquence fondamentale, intensité et débit phonétique, semble diminuer la confusion entre surprise et joie (30,8 % et 29,2 %), comme présenté dans le Tableau 28, mais elle augmente en même-temps malheureusement la confusion entre d'autres états (Tableau 29). En définitive, cet ensemble de 72 paramètres devient moins efficace que l'ensemble précédent des 48 paramètres de F_0 et d'intensité.

Cet effet se produit également dans le cas de la combinaison des 24 paramètres de F_0 et des 24 pramètres du débit phonétique (Tableau 29). Le taux de classification correcte obtenu de 36,7 % est beaucoup plus bas que celui obtenu avec des seuls paramètres de F_0 utilisés isolément.

	colère (%)	***joie (%)***	***neutre (%)***	***tristesse (%)***	***surprise (%)***
colère	34,7	22,4	0,0	30,6	12,2
joie	26,5	28,6	0,0	16,3	28,6
tristesse	16,3	18,4	4,1	42,9	18,4
surprise	26,5	20,4	0,0	12,2	40,8
moyenne sans neutre	36,7				

Tableau 29 : Fusion de la fréquence fondamentale (24 paramètres) et le débit phonétique (24 paramètres)

En conclusion, au travers des résultats obtenus avec les fusions présentées nous pouvons conclure, dans ce contexte de travail, que :

- la prosodie, plus particulièrement la fréquence fondamentale et l'intensité, joue un rôle important dans l'expression émotionnelle avec des comportements distinctifs qui sont montrés tout au long de cette partie.
- bien que des variations de comportement des paramètres puissent exister entre des locuteurs différents et entre des locuteurs et des locutrices - mais leur étude n'est pas l'objectif de cette partie, ni prévue dans le cadre de la thèse - nos résultats ont été obtenus par une analyse statistique globale, sur les valeurs moyennes, vérifiés en appliquant la reconnaissance sur les quatre locuteurs du corpus DES et nous estimons alors que nos conclusions pourraient être utiles pour d'autres études dans le domaine sous réserve de vérifications dans des conditions plus générale.

- l'influence de la vitesse - le débit phonétique - n'est pas aussi claire que celle de la fréquence fondamentale ou l'intensité pour notre cas, mais à certains degrés, cette caractéristique contribue aussi à l'expression des émotions.

Dans la section suivante, nous proposons l'analyse d'autres paramètres globaux dans le domaine temporel ainsi que dans le domaine fréquentiel.

5.3.1.5 Expérimentations avec d'autres paramètres

En dehors de la prosodie; nous intéressons aussi à deux paramètres qui sont le nombre de passages par zéro et le nombre de sommets avec l'étude de la même manière : normalisation par rapport au neutre.

5.3.1.5.1 Nombre de passages par zéro (ZRC – zéro crossing

Le Tableau 30 contient les résultats de classification obtenus en appliquant isolement chaque paramètre de ZRC sur les quatre émotions du corpus DES. Nous n'étudions pas les dérivées car elles n'ont pas de signification dans ce cas.

	Max (%)	***Min (%)***	***Mean (%)***	***Med. (%)***	***Var. (%)***	***Range (%)***	***Rising-Falling-CountRatio (%)***	***Rising-Falling-SumRatio (%)***	***Tous les 8 param. (%)***
ZCR	29,8	23,1	38,0	26,5	21,2	29,3	23,6	32,2	31,8

Tableau 30 : Taux de reconnaissance de quatre émotions du corpus DES en utilisant le nombre de passage par zéro

Selon les résultats présentés, bien que le taux de classification (31,8 %) ne soit pas aussi élevé que celui des autres paramètres étudies, nous pouvons constater que le nombre de passages par zéro contient aussi des informations reflétant les émotions. Selon nos analyses, la joie, la colère et la surprise sont aussi notées comme des émotions ayant le plus grand nombre de passages par zéro ainsi que leur intervalle de la variété du ZCR en joie, en surprise et en colère est aussi beaucoup plus large par rapport celui en neutre (RP_E ~135%).

En espérant que la combinaison entre des aspects de ZCR avec la prosodie peut améliorer la performance, nous avons choisi les quatre paramètres les plus efficaces en mode isolé, Mean, Max, RsisingFallingSumRatio et Range, pour les fusionner avec l'ensemble de 72 paramètres de la prosodie (24 paramètres de F_0, 24 paramètres de l'intensité et 24 paramètres du débit phonétique) qui fait au total un ensemble de 76 paramètres. Cependant, la combinaison des meilleurs paramètres ne donne pas toujours le meilleur résultat ; le résultat obtenu est moins efficace que celui de 72 paramètres de la prosodie. La raison pourrait être que l'ajout des paramètres de ZCR dans l'ensemble des paramètres prosodiques fait baisser le taux de reconnaissance de la joie et de la tristesse, et donc, la performance globale a été atténuée. Une stratégie de recherche d'un ensemble de paramètres les plus efficaces sera étudiée dans les parties suivantes.

5.3.1.5.2 Nombre de sommets

Le nombre de sommets est le nombre de valeurs localement extremum dans une trame choisie du signal.

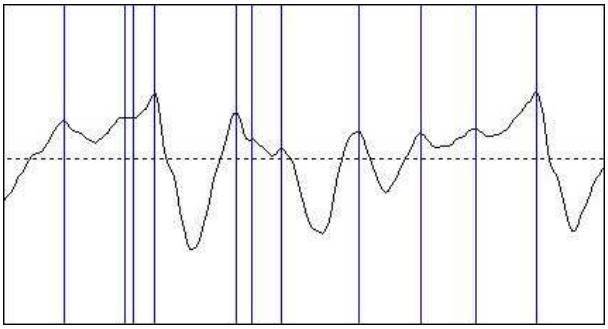

Figure 26 : Sommets du signal

Nous constatons également que, cette information est aussi un signe pour la classification des émotions avec la performance montrée dans le Tableau 31.

	Max (%)	***Min (%)***	***Mean (%)***	***Median (%)***	***Variance (%)***	***Range (%)***	***Rising-Falling-CountRatio (%)***	***Rising-Falling-SumRatio (%)***	***Tous les 8 param. (%)***
Nombre de sommets	33,2	30,3	31,7	32,3	28,8	25,5	32,7	29,8	31,3

Tableau 31 : Taux de reconnaissance en utilisant le nombre de sommets

Par contre, ce paramètre marche bien pour la reconnaissance de la tristesse. Selon les analyses, nous constatons que, dans ce corpus, en tristesse le locuteur a tendance à produire beaucoup plus de vibrations qui reflexe par l'augmentation forte du nombre de sommets. Cela est montré par les valeurs des quatre type de paramètres : Max de nombre de sommets (RP_E ~ 112,3%), Min (RP_E ~ 129,7%), Mean (RP_E ~ 123,2%) et Médiane (RP_E ~ 120,5%). L'intérêt le plus important ici est que le nombre de sommets contient des informations permettant distinguer la tristesse et le neutre. En effet, jusqu'ici ces deux états émotions (la tristesse et le neutre) ont presque toujours les mêmes comportements des paramètres. Pour l'expérimentation, la fusion de ces quatre paramètres nous donne un bon résultat de classification pour les deux groupes : tristesse / autres émotions : 81,9%.

Cependant, comme le ZCR, la combinaison de ce paramètre avec la prosodie n'améliore pas la performance de classification.

En conclusion, dans le contexte du corpus DES étudié, à côte de la prosodie, l'état émotionnel du locuteur influe également aux autres aspects de la parole comme le nombre de passages par zéro et le nombre de sommets. L'influence peut être plus nette comme dans le cas du nombre de sommets mais elle peut aussi floue comme dans le cas de nombre de passages par zéro.

5.3.1.6 Expérimentations avec les paramètres du domaine fréquentiel

5.3.1.6.1MFCC – Mel Frequency Cepstral Coefficients

Parmi les 17 coefficients *MFCC* que nous utilisons, nous nous intéressons particulièrement au premier coefficient $MFCC_0$ qui est considéré comme une autre manière de caractériser l'énergie

du signal (voir section 5.1.2.1). De plus, l'utilisation de ce coefficient est plus de sûr que l'utilisation de l'intensité, car $MFCC_0$ est plus stable.

En effet, la performance de classification du paramètre $MFCC_0$ est significative comme montrée dans le tableau suivant (pour classifier les 4 émotions):

	Max (%)	*Min (%)*	*Mean (%)*	*Median (%)*	*Variance (%)*	*Range (%)*	*Rising-Falling-CountRatio (%)*	*Rising-Falling-SumRatio (%)*	*Tous les 8 paramètres (%)*
$MFCC_0$	39,0	35,3	43,8	39,0	37,5	31,3	28,9	27,9	38,5

Tableau 32 : Taux de reconnaissance en utilisant $MFCC_0$

Si nous utilisons seulement la moyenne du paramètre$MFCC_0$, le taux obtenu est assez élevé et atteint les 43,8 %.

La différence entre la tristesse et la colère, la joie ou la surprise, devient aussi très nette avec ce coefficient $MFCC_0$; en effet, les taux très élevé que nous pouvons obtenir en classifiant les deux groupe tristesse/colère + joie + surprise, sont de 88,0 % et 84,1 % en utilisant respectivement les maxima de $MFCC_0$ et les minima de $MFCC_0$.

Le Tableau 33 est une synthèse de l'efficacité des 17 paramètres MFCCs.

	Max (%)	*Min (%)*	*Mean (%)*	*Median (%)*	*Variance (%)*	*Range (%)*	*Rising-Falling-CountRatio (%)*	*Rising-Falling-SumRatio (%)*	*Tous les 8 paramètres (%)*
MFCC0	39,0	35,3	43,8	23,8	37,5	31,3	28,9	27,9	38,5
MFCC1	31,3	26,0	27,4	28,4	28,9	16,9	25,5	26,9	29,9
MFCC2	33,1	29,4	32,3	32,3	43,3	44,8	30,8	26,0	40,9
MFCC3	31,8	30,3	27,9	29,4	30,3	36,0	18,3	29,8	28,9
MFCC4	34,1	36,0	23,1	24,5	25,0	36,8	30,8	23,1	30,8
MFCC5	27,9	34,1	36,0	37,5	35,1	31,3	28,9	20,4	27,9
MFCC6	25,0	31,8	26,9	27,4	27,9	33,1	18,3	29,8	26,5
MFCC7	19,3	27,4	25,0	25,0	22,1	28,9	21,9	26,9	20,6
MFCC8	21,1	28,9	29,8	26,5	30,3	31,3	27,9	30,8	24,5
MFCC9	26,0	29,8	23,1	28,9	30,3	30,3	28,4	28,9	21,6
MFCC10	22,6	26,9	22,1	21,1	24,5	23,1	20,6	24,0	21,1
MFCC11	31,8	28,9	25,0	24,0	32,3	30,8	23,1	22,1	26,0
MFCC12	25,5	30,3	22,1	27,4	29,4	29,4	24,5	15,9	26,5
MFCC13	25,0	29,8	27,9	24,5	31,3	37,0	22,6	17,3	29,8
MFCC14	27,9	33,6	25,4	28,4	22,6	34,6	20,6	28,4	28,6
MFCC15	21,6	25,5	23,5	25,0	33,1	27,4	19,3	26,0	26,0
MFCC16	22,6	29,6	25,5	23,8	26,5	30,3	24,0	21,1	22,6

Tableau 33 : Taux de reconnaissance en utilisant 17 paramètres globaux de MFCC. Les cellules grises marquent les meilleurs résultats

Selon les résultats du Tableau 33, nous constatons que, dans notre cas :

- MFCC0 est un des coefficients les plus efficaces que nous pouvons observer en mode statistique globale ;
- les meilleurs résultats de classification se concentrent sur les premiers coefficients. Cela peut s'expliquer par le renforcement de la partie des basses fréquences de l'échelle Mel où se trouvent beaucoup d'informations concernant la prosodie qui sont très utiles pour la discrimination des émotions ;
- selon nos résultats, en mode global, les dérivées premières ΔMFCCs et secondaires ΔΔMFCCs contiennent aussi des informations utiles mais ils sont beaucoup moins efficaces que des MFCCs ; nous l'expliquons par le fait que ΔMFCCs et ΔΔMFCCs caractérisent essentiellement l'aspect dynamique (vitesse et accélération) et cet aspect dynamique pour l'émotion montre des variations lentes qui portent plutôt sur toute la phrase ;
- et enfin, nous pouvons conclure que les aspects fréquentiels jouent aussi un rôle important pour l'expression et pour la reconnaissance des émotions parce qu'ils nous fournissent non seulement des informations des bandes fréquentielles mais ils impliquent également les autres informations de la prosodie, par exemple le timbre de la voix, les harmoniques du signal qui sont des multiples de la fréquence fondamentale.

5.3.1.6.2LFCC – Linear Frequency Cepstral Coefficients

Les coefficients LFCC sont calculés de la même manière que les MFCC, mais avec la différence que les fréquences des filtres sont uniformément réparties sur l'échelle linéaire des fréquences, et non plus sur une échelle Mel. Cet effet a pour conséquence une distribution uniforme en termes de la performance des coefficients de LFCCs. Cependant, en comparaison des meilleurs resultats avec ceux des MFCCs, des LFCCs sont moins efficaces (Tableau 34).

	Max (%)	***Min (%)***	***Mean (%)***	***Median (%)***	***Variance (%)***	***Range (%)***	***Rising-Falling-CountRatio (%)***	***Rising-Falling-SumRatio (%)***	***Tous les 8 paramètres (%)***
LFCC0	36,1	24,5	34,1	33,6	32,3	37,5	24,0	28,4	37,0
LFCC1	29,9	27,9	32,8	23,1	25,0	28,9	27,4	22,6	29,4
LFCC2	24,5	19,8	30,8	29,8	37,0	31,4	24,0	33,3	29,9
LFCC3	27,5	26,5	31,3	24,5	31,8	32,8	26,5	21,8	26,5
LFCC4	27,5	28,4	31,8	26,0	31,8	30,8	26,5	27,0	26,0
LFCC5	28,4	26,0	27,4	17,8	30,4	38,1	28,4	31,3	29,9
LFCC6	22,1	25,0	22,6	25,5	33,3	29,9	27,4	26,5	28,9
LFCC7	24,0	30,4	25,5	25,0	30,3	35,1	23,1	27,0	19,8
LFCC8	20,8	23,6	27,5	28,9	32,3	31,3	28,9	23,6	28,9
LFCC9	25,5	27,9	26,4	20,6	28,4	27,9	22,6	22,6	21,6
LFCC10	31,8	30,8	26,0	26,5	26,5	26,1	23,6	27,5	24,0
LFCC11	27,0	20,6	25,1	26,5	30,6	25,5	25,0	22,6	19,8
LFCC12	30,4	32,8	34,1	30,3	29,4	34,6	22,6	28,9	29,4
LFCC13	28,4	39,5	27,0	27,9	34,6	34,6	21,6	29,4	25,5
LFCC14	29,9	33,8	26,5	28,4	35,1	28,9	13,5	25,0	27,5
LFCC15	22,6	27,5	31,4	24,0	32,8	28,9	26,0	13,6	26,5

Tableau 34 : Taux de reconnaissance en utilisant des paramètres globaux de LFCCs

5.3.1.6.3 LPC – Linear Predictive Coding

Parmi les trois ensembles de paramètres : MFCC, LFCC et LPC, bien que les coefficients LPC ne donnent pas des meilleurs taux de classification (maximum 34,6 %), tous les coefficients LPC montrent plus clairement leurs contributions aux discriminations des émotions (Tableau 35). Particulièrement, leurs minima sont assez stables pour tous les 16 coefficients LPC. La combinaison des minima de ces paramètres nous donne un taux de classification 35,1%.

	Max (%)	***Min (%)***	***Mean (%)***	***Median (%)***	***Variance (%)***	***Range (%)***	***Rising-Falling-CountRatio (%)***	***Rising-Falling-SumRatio (%)***	***Tous les 8 paramètres (%)***
LPC0	25,5	35,6	33,6	27,4	37,0	24,5	22,1	31,8	34,6
LPC1	35,6	32,8	24,0	28,4	28,9	24,5	24,5	35,1	25,0
LPC2	21,6	32,6	23,5	26,5	27,9	26,5	31,3	31,8	29,3
LPC3	31,8	35,6	25,0	21,6	26,9	21,1	18,3	34,1	30,3
LPC4	27,9	32,3	30,8	33,6	37,5	20,8	26,5	35,1	30,3
LPC5	31,3	33,1	26,0	27,4	32,3	25,5	29,8	33,6	26,9
LPC6	24,5	34,6	31,8	32,3	31,3	26,9	26,9	29,3	32,8
LPC7	26,9	38,0	26,5	28,4	19,8	18,8	26,5	37,0	31,8
LPC8	24,5	32,8	24,0	26,0	32,8	26,9	33,1	29,8	26,5
LPC9	39,4	31,3	29,8	35,6	26,0	21,6	15,4	36,0	27,4
LPC10	20,6	31,8	36,0	26,0	31,8	30,8	25,0	29,8	26,0
LPC11	31,3	31,5	29,4	27,9	21,6	28,9	23,1	31,9	32,3
LPC12	21,6	32,3	26,9	28,9	29,8	26,0	23,5	30,3	31,3
LPC13	36,0	34,1	30,8	31,3	24,5	23,0	24,0	28,9	29,8
LPC14	27,4	27,9	30,8	30,8	32,8	12,5	24,0	34,1	29,8
LPC15	21,6	35,6	30,3	12,6	26,5	25,0	26,9	27,4	25,5

Tableau 35 : Taux de reconnaissance en utilisant des paramètres de LPCs

En conclusion, dans cette partie, nous avons analysé en détail les caractéristiques qui pourraient porter les informations des émotions comme les trois composants principaux de la prosodie (F_0, l'intensité et le débit), le nombre de passage par zéro, le nombre de sommets et les coefficients connus dans le domaine fréquentiel tels que les MFCCs, LFCCs, LPCs.

Pour limiter maximalement les variétés causées par la dépendance des paramètres du locuteur et du contexte, nous avons proposé d'utiliser l'état neutre comme un état standard, et en utilisant la comparaison avec l'état neutre, nous avons essayé d'extraire des caractéristiques, des paramètres distinctifs qui peuvent servir à notre travail ; la reconnaissance des émotions.

Concrètement :

- Sur ce corpus DES, nous avons pu constater la contribution importante des composants de la prosodie dans l'expression émotionnelle, particulièrement F_0 et l'intensité, ainsi que la contribution importante des composants dans le domaine fréquentiel montré par des coefficients tels que les MFCCs, LFCCs, etc ;
- le troisième composant - le débit phonétique - ne se montre pas encore très efficace avec notre méthode de calcul et avec le corpus étudié ; nous verrons dans la section suivante

portant sur les paramètres locaux que les tests avec ce même débit phonétique nous amènent à la même conclusion ;

- toujours sur le corpus DES, les deux autres paramètres du domaine temporel ont été également étudiés, le nombre de sommets se montre plus efficace par rapport au nombre de passages par zéro (ZCR). Cela peut s'expliquer par la dépendance très forte de ZCR de l'articulation phonémique que la normalisation neutre ne peut pas enlever : le ZCR est très élevé chez des phonèmes non-voisés (la plupart des consonnes) mais il est très bas chez des phonèmes voisés (des voyelles).

En conclusion, à travers ces études sur les paramètres globaux, bien que les résultats obtenus ne soient pas très élevés, on peut dire que dans le contexte de ce corpus, la construction d'un système de reconnaissance automatique de l'émotion et indépendant du locuteur se basant sur des caractéristiques globales est faisable. Cependant, il faudrait encore beaucoup d'études pour pouvoir sortir un ensemble de paramètres représentatifs, efficaces et indépendants du locuteur, du contexte, des dispositifs utilisés etc.

Notre normalisation neutre pourrait être l'une des solutions, mais elle nécessite encore des améliorations pour qu'elle soit efficace, et pour que les rapports RP_E soient plus exacts. Un exemple d'amélioration qui serait intéressant pour notre travail dans le futur est qu'au lieu de calculer des rapports entre des énoncés comme nous l'avons fait, nous pouvons calculer des rapports entre des unités phonémiques, donc à ce niveau, la dépendance de l'articulation phonémique est maximalement enlevée.

A partir des résultats obtenus de cette partie, dans la partie suivante, nous allons étudier des paramètres par une autre approche que nous appelons des paramètres locaux qui est plus réalisable et qui s'accorde plus avec notre objectif de détection automatique des émotions dans des énoncés audio.

5.3.2. Paramètres locaux

5.3.2.1 Fusion et Interpolation des paramètres

Tous les paramètres ne sont pas extraits avec la même fréquence d'échantillonnage (paramètres spectraux et débit phonétique par exemple). Afin de pouvoir les combiner, il est nécessaire d'en ré-échantillonner certains de façon à les ramener tous à la même fréquence d'échantillonnage. Nous utilisons pour cela une interpolation linéaire (Figure 27).

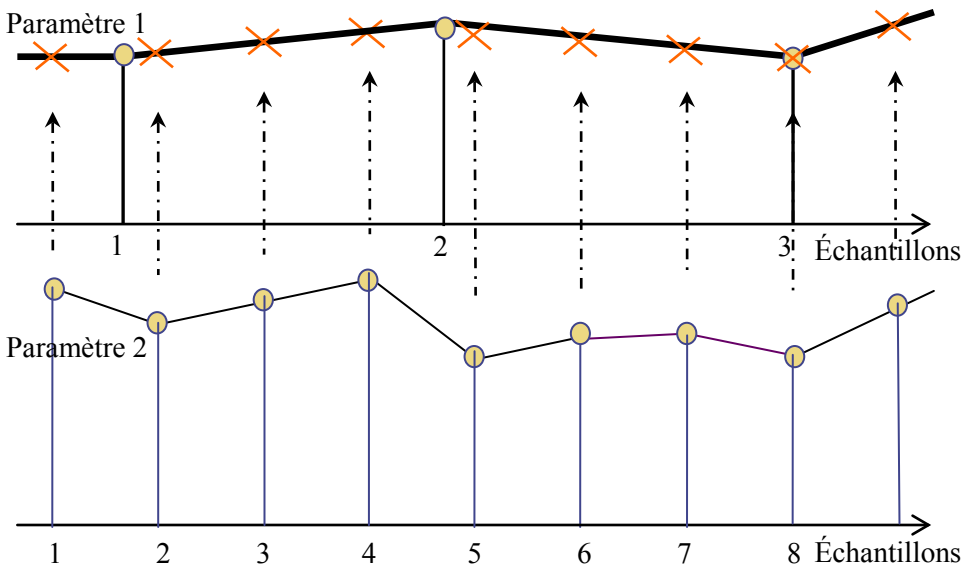

Figure 27 : Interpolation pour la fusion des paramètres.

5.3.2.2 Normalisation des paramètres

Les différents paramètres sont calculés dans des unités différentes et ne sont pas naturellement comparables entre eux. Pour compenser cela et éviter que certains, ayant une dynamique plus importante, ne dominent les autres, nous effectuons une normalisation en appliquant une transformation linéaire sur chacun d'entre eux de façon à normaliser leur moyenne (ramenée à 0) et leur variance (ramenée à 1).

5.3.2.3 Sélection des paramètres

5.3.2.3.1 Choix du modèle pour la validation

Un modèle fiable que nous appelons le modèle de validation doit être choisi pour évaluer la performance des caractéristiques d'entrée. Nous avons retenu le modèle de mélange des gaussiennes (GMM) car, par rapport aux autres modèles, celui-ci est assez simple en termes de nombre de paramètres à optimiser et son efficacité est déjà démontrée dans plusieurs systèmes de reconnaissance, comme ceux de [Lefort & al 02], [Moraru et al, 04] et [Aronowitz & Irony 05] par exemple.

La recherche du modèle le plus efficace pour la reconnaissance de l'émotion (parmi plusieurs autres modèles : VQ, SVM, particulièrement le modèle HMM qui peut capturer aussi l'information évolution temporelle des paramètres) sera effectuée après avoir trouvé l'ensemble de paramètres le plus efficace.

Dans le cas de la reconnaissance dépendante du locuteur avec les données du corpus DES, le nombre de 16 gaussiennes a été choisi en se basant sur les résultats expérimentaux obtenus avec les 12MFCCs+12ΔMFCCs+12ΔΔMFCCs comme dans le Tableau 36.

N. de gaussiennes	***1***	***2***	***4***	***8***	***16***	***32***	***64***	***128***
Taux de reconnaissance (%)	56,9	65,1	70,8	64,5	70,7	69,2	70,8	61,5

Tableau 36 : Taux de reconnaissance de 5 émotions du corpus DES en changeant le nombre de gaussiennes

où le taux de reconnaissance est définie par la formule suivante :

$$TR = \frac{\text{nombre d'échantillons correctement reconnus}}{\text{nombre d'échantillons à reconnaitre}} \qquad (2.9)$$

et les modèles GMMs sont entrainés par une partie d'apprentissage. Le taux TR est mesurée par (2.9) sur l'autre partie de test du corpus DES. La méthode « *Leave One Out* » est utilisé dans cette étape pour partitionner ces deux parties (voir la section suivante pour le détail).

Similairement, en se basant sur les résultats de reconnaissance des 5 émotions du corpus DES en utilisant 12MFCCs+12ΔMFCCs+12ΔΔMFCCs dans les deux autres cas : reconnaissance multi locuteurs et reconnaissance indépendante du locuteur, nous avons choisi le modèle GMM de 64 gaussiennes pour le modèle de validation dans le cas de la reconnaissance multi locuteurs et le GMM de 128 gaussiennes pour le cas de reconnaissance indépendante du locuteur.

5.3.2.3.2 Protocole de test

Nous poursuivons les trois cas d'étude : reconnaissance dépendante du locuteur, reconnaissance multi-locuteurs et reconnaissance indépendante du locuteur. Dans les deux premiers cas, le corpus DES a été choisi en raison de son équilibre en termes de nombre d'échantillons et de nombre d'émotions pour tous les locuteurs. Pour le troisième cas, le corpus BES a aussi été introduit car il contient plus de locuteurs que le corpus DES (10).

Dans le cas de la reconnaissance dépendante du locuteur, les données de chaque locuteur sont d'abord séparément étudiées pour une vue du comportement de système sur chaque locuteur. Les résultats moyens seront ensuite utilisés pour la comparaison avec d'autres approches. Au total nous avons 65 échantillons d'expression émotionnelle en 5 états émotionnels de 13 phases de texte pour chaque locuteur. Donc nous proposons de profiter au maximum de ces 65 fichiers pour l'apprentissage, ainsi que pour le test en utilisant la validation croisée avec le cas extrême de la validation croisée « *Leave One Out* ».

5.3.2.3.3 Sélection par le critère Fisher (FDR)

L'utilisation d'un nombre important de paramètres induit une charge de calcul importante. En fait, tous les paramètres ne sont pas utiles pour la reconnaissance. Certains peuvent même nuire en introduisant du bruit s'ils ne sont pas ou trop faiblement corrélés avec la classe à reconnaître. Pour extraire l'ensemble de paramètres les plus efficaces, pour éliminer la redondance ou le bruit, nous proposons tout d'abord de nous baser sur le critère de Fisher (FDR), une méthode permettant d'estimer la capacité discriminative de chaque paramètre pour les différentes classes en mesurant le chevauchement de leurs fonctions de densité de probabilité.

$$FDR = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K}(\overline{x[i]} - \overline{x[j]})^2}{\sum_{i=1}^{K} Var(x)[i]} \quad (5.10)$$

où $x[i]$ désigne la moyenne du paramètre x pour la classe i et $Var(x)[i]$ désigne la variance du paramètre x pour la classe i.

Pour chaque dimension du paramètre, il représente le rapport entre la distance qui sépare deux classes i, j et leurs variances. Dans cette formule, les contributions de toutes les K classes sont cumulées. Ainsi, ce paramètre peut être interprété comme le rapport de la variabilité interclasse du paramètre par la variabilité intraclasse du même paramètre. Donc, avec ce critère, les paramètres avec les meilleurs potentiels de pertinence peuvent être rapidement sélectionnés. Le désavantage de ce critère est qu'il n'intègre pas les relations de corrélation entre les paramètres, et que la valeur FDR dépend de la dynamique du paramètre considéré. C'est la raison pour laquelle, l'ensemble des meilleurs paramètres classifiés par ce critère ne donne pas généralement le meilleur résultat.

Dans la littérature, le critère de Fisher (FDR) est une des méthodes nous permettant l'estimation de la capacité discriminatives des paramètres. Et ce critère a été étudié dans notre cas de la reconnaissance de l'émotion. Cependant, les résultats obtenus avec ce critère ne sont pas très efficaces pour notre cas pour deux raisons suivantes :

- ce critère est censé être fiable pour un problème binaire (2 classes) mais il est notoirement inadaptable pour des problèmes multi-classes, en particulier pour des distributions non-convexes, asymétriques de données [Pachet & Roy 2007] ;
- ce critère ne prend pas en compte la corrélation entre des paramètres, donc l'utilisation de ce critère donne un jeu de paramètres « sous-optimal » [Istrate 2003].

C'est la raison pour laquelle, les tests ont été effectués pour tous les paramètres et nous ne basons nos évaluations des paramètres que sur les résultats expérimentaux. La comparaison entre les valeurs FDR et ces résultats expérimentaux nous montrera l'efficacité de cette approche FDR dans la reconnaissance de l'émotion.

Nous considérons tout d'abord les paramètres prosodiques avec les valeurs FDR montrées dans le Tableau 37. Dans ce tableau, Δ signifie la dérivée du 1er ordre, ΔΔ signifie la dérivée du 2ème ordre, F_0Rel, IntensitéRel, DébitRel sont les rapports de F_0, de l'intensité et du débit phonétique respectivement avec leurs moyennes de chaque énoncé.

F_0	25,19	**ΔF0**	6,71	**ΔΔF0**	7,53
F0Rel	7,24	**ΔF0Rel**	7,25	**ΔΔF0Rel**	4,98
Intensité	48,82	**ΔIntensité**	9,35	**ΔΔIntensité**	24,67
IntensitéRel	12,47	**ΔIntensitéRel**	2,14	**ΔΔIntensitéRel**	12,11
DébitP	4,72	**ΔDébitP**	0,03	**ΔΔDébitP**	0,22
DébitPRel	2,06	**ΔDébitPRel**	0,04	**ΔΔDebitPRel**	0,21

Tableau 37 : FDR des paramètres prosodiques

En se basant sur des valeurs FDR du Tableau 37, les mêmes résultats obtenus avec des paramètres globaux, l'intensité et F_0 sont aussi des paramètres locaux les plus discriminants (FDR les plus élevés). Et le débit phonétique est toujours le paramètre le moins efficace parmi

les paramètres de la prosodie. Pour comparaison, le Tableau 38 montre notre résultat expérimental.

Selon notre résultat sur le corpus DES, il y a en général une correspondance entre ces deux approches : les paramètres possédant les grandes valeurs de FDR sont en général des paramètres expérimentalement efficaces. Les cas de désaccords entre les résultats expérimentaux et les valeurs FDR expliquent la faiblesse de ce critère.

Paramètres	***Taux de reco. (%)***	***Param.***	***Taux de reco. (%)***	***Param.***	***Taux de reco. (%)***	***Combiner les trois param. (%)***
F0	34,7	ΔF_0	30,4	$\Delta\Delta F_0$	30,0	40,0
F0Rel	31,6	ΔF_0***Rel***	36,9	$\Delta\Delta F_0$***Rel***	30,8	38,5
Intensité	40,0	***ΔIntensité***	35,4	***ΔΔIntensité***	31,6	43,8
IntensitéRel	28,5	***ΔIntensitéRel***	25,4	***ΔΔIntensitéRel***	20,0	31,5
DébitP	26,9	***ΔDébitP***	19,2	***ΔΔDébit***	19,3	23,9
DébitPRel	22,3	***ΔDébitPRel***	17,0	***ΔΔDebitRel***	20,0	26,9

Tableau 38 : Taux de classification des paramètres prosodiques

Selon le résultat expérimental obtenu dans le Tableau 38 :

- 40,0 % est le taux assez élevé de la classification des 5 états émotionnels en utilisant seulement l'intensité (en comparaison avec 43,3 % qui est le taux de classification correcte le plus élevé des paramètres globaux pour les 4 émotions). Nous obtiendrons 43,8 % si la fusion les trois aspects de l'intensité est effectuée.

- les mauvais résultats ainsi que les basses valeurs FDR du débit phonétique dans ce cas nous raffirme la difficulté de l'utilisation du débit phonétique pour reconnaitre les émotions en mode isolé.

Le nombre de passages par zéro et le nombre de sommets sont aussi vérifiés par le Tableau 39 et par les résultats expérimentaux dans le Tableau 40. Au contraire aux résultats obtenus avec des operateurs globaux dans la partie précédente, les paramètres locaux du nombre de passages par zéro (ZCR) se montrent beaucoup plus performants que les paramètres du nombre de sommets. Avec la fusion de tous les trois aspects de ZCR, nous obtiendrons un taux de reconnaissance considérable : 44,9 % en moyenne pour les 4 locuteurs et en travaillant avec les 5 émotions. L'amélioration importante obtenue par cette fusion nous donne le constat que la corrélation est moins forte entre les trois aspects de ZCR que celle entre les trois aspects du nombre de sommets.

N. Sommet	***13,04***	***ΔN. Sommet***	***3,22***	***ΔΔN. Sommet***	***1,57***
ZCR	23,96	ΔZCR	4,03	ΔΔZCR	0,98

Tableau 39 : FDR du nombre de sommets et du nombre de passages par zéros

Paramètres	***Taux de reco. (%)***	***Param.***	***Taux de reco. (%)***	***Paramètres***	***Taux de reco. (%)***	***Combiner les trois param. (%)***
N. Sommet	29,3	ΔN. Sommet	32,3	ΔΔN. Sommet	23,1	33,1
ZCR	37,7	ΔZCR	27,7	ΔΔZCR	22,3	44,9

Tableau 40 : FDR des paramètres prosodiques

La Figure 28 montre les valeurs FDR des ensembles de 17 coefficients de MFCC : $MFCC_0$, $MFCC_1$, ..., $MFCC_{16}$, 17 coefficients de la dérivée en première ordre : $\mathbf{\Delta}MFCC_0$,… $\mathbf{\Delta}MFCC_{16}$, 17 coefficients de la dérivée en deuxième ordre : $\mathbf{\Delta\Delta}MFCC_0$,… $\mathbf{\Delta\Delta}MFCC_{16}$.

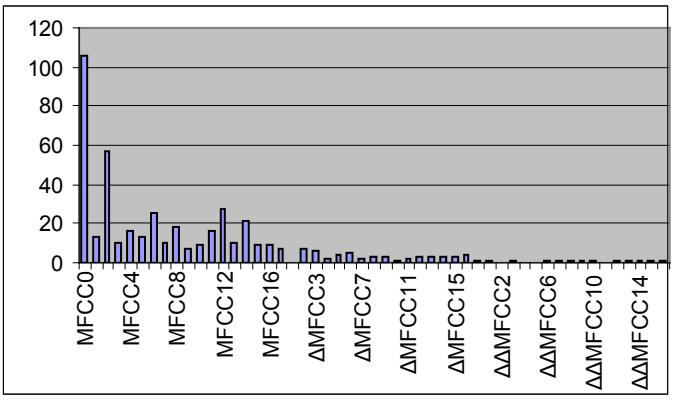

Figure 28 : Valeurs FDR des MFCCs, des ΔMFCCs et des ΔΔMFCCs

Selon la Figure 28, nous constatons que, si on se base sur des valeurs FDR, les dérivées sont les moins significatives surtout la dérivée en deuxième ordre. Cette remarque correspond généralement aux résultats expérimentaux du Tableau 41 ci-dessous, cependant ces valeurs FDRs ne sont pas toujours fiables en comparaison inter-ensemble (entre les trois types de paramètres). Par exemple, la valeur FDR de $\Delta MFCC_0$ est beaucoup moins importante que celle de $MFCC_{12}$ mais expérimentalement, il $\Delta MFCC_0$ se montre plus efficace que $MFCC_{12}$. Cela peut s'expliquer par des raisons que nous avons mentionnées ci-dessus. C'est la raison pour laquelle, nous ne basons les évaluations de la performance de chaque paramètre que sur nos résultats expérimentaux.

Le Tableau 41 montre l'efficacité expérimentale de chaque paramètre MFCC avec le modèle de validation GMM 16 gaussiennes. Dans ce tableau, la police en gros marque les 5 expérimentalement meilleurs résultats de chaque type de paramètres, les numéros (de 1 à 15) accompagnant ces résultats sont leur mise en ordre en fonction de la performance expérimentale. Les carrés foncés contiennent des paramètres dont leurs valeurs FDRs sont les plus élevées parmi chaque ensemble de paramètres.

MFCC0	46,2(1)	ΔMFCC0	45,1(3)	ΔΔMFCC0	27,7
MFCC1	40,5(5)	ΔMFCC1	32,8	ΔΔMFCC1	30,8
MFCC2	45,6(2)	ΔMFCC2	36,9	ΔΔMFCC2	23,1
MFCC3	32,8	ΔMFCC3	40,0(7)	ΔΔMFCC3	25,6
MFCC4	36,4(10)	ΔMFCC4	25,1	ΔΔMFCC4	25,6
MFCC5	40,0(6)	ΔMFCC5	37,9(8)	ΔΔMFCC5	23,6
MFCC6	35,9(12)	ΔMFCC6	31,3	ΔΔMFCC6	27,2
MFCC7	32,3	ΔMFCC7	31,3	ΔΔMFCC7	33,3
MFCC8	33,3	ΔMFCC8	34,9	ΔΔMFCC8	25,6
MFCC9	29,2	ΔMFCC9	32,8	ΔΔMFCC9	23,6
MFCC10	33,8	ΔMFCC10	21,0	ΔΔMFCC10	25,1
MFCC11	25,1	ΔMFCC11	32,3	ΔΔMFCC11	20,5
MFCC12	44,6(4)	ΔMFCC12	24,6	ΔΔMFCC12	20,5
MFCC13	28,2	ΔMFCC13	36,9(9)	ΔΔMFCC13	26,7
MFCC14	34,9	ΔMFCC14	33,8	ΔΔMFCC14	28,7
MFCC15	27,2	ΔMFCC15	20,5	ΔΔMFCC15	23,6
MFCC16	30,8	ΔMFCC16	36,4(11)	ΔΔMFCC16	23,1

Tableau 41 : Taux de reconnaissance expérimentale avec DES 16 mélanges de la première étape

D'après le Tableau 41 pour le corpus étudié : DES, pour chaque type de paramètres (paramètres originaux, dérivées du premier ordre et dérivées du deuxième ordre), il y a également une correspondance entre les valeurs FDRs et les résultats expérimentaux : les paramètres possédant les valeurs FDR élevées donnent normalement de bons résultats expérimentaux.

En comparant le Tableau 41 et la Figure 28 nous pouvons faire dans le cadre du corpus étudié les autres remarques suivantes :

- parmi les coefficients, $MFCC_0$ et $MFCC_2$ sont les deux paramètres les plus efficaces dans tous les deux cas de mesure : la meilleur valeur FDR et le taux le plus élevé de la reconnaissance expérimentale ;
- 46,2 % avec un seul coefficient $MFCC_0$ est un taux de classification élevé pour les cinq émotions par rapport au taux aléatoire (20 %) ; ce taux de classification est même plus élevé que ceux obtenus avec des paramètres prosodiques.

Cela peut s'expliquer par le fait que la bande de fréquences du filtre 2 correspond bien avec l'intervalle des fréquences fondamentales de ce corpus (de 74 Hz à 247 Hz). Le Tableau 42 montre les trois paramètres : f_b (basse fréquence), f_c (fréquence centrale) et f_h (haute fréquence) de 24 filtres triangulaires dans le domaine fréquentiel que nous avons utilisés pour ces 17 coefficients MFCCs.

La même explication peut être utilisée pour l'efficacité des coefficients $MFCC_1$, $MFCC_4$, $MFCC_5$, $MFCC_6$ parce que les filtres correspondants couvrent partiellement ou entièrement des harmoniques de la fréquence fondamentale. L'inefficacité de $MFCC_3$ peut s'expliquer par sa haute f_b (156) qui ne fait passer que la fréquence fondamentale des locutrices (~180 Hz pour ce corpus) et exclut entièrement les fréquences fondamentales des locuteurs (~110 Hz pour ce corpus). A notre avis, la correspondance des filtres avec les formants peut aussi être une des

raisons pour l'efficacité des coefficients MFCC, notamment pour les coefficients élevés comme $MFCC_6$, $\Delta MFCC_6$, $MFCC_{12}$, $MFCC_{14}$, $\Delta MFCC_{13}$, $\Delta MFCC_{16}$. Cependant, nous ne faisons pas une étude sur cet aspect.

	f_b (Hz)	f_c (Hz)	f_h (Hz)
Filtre 1	0	74	156
Filtre 2	74	156	247
Filtre 3	156	247	348
Filtre 4	247	348	459
Filtre 5	348	459	582
Filtre 6	459	582	718
Filtre 7	582	718	868
Filtre 8	718	868	1034
Filtre 9	868	1034	1218
Filtre 10	1034	1218	1422
Filtre 11	1218	1422	1646
Filtre 12	1422	1646	1895
Filtre 13	1646	1895	2171
Filtre 14	1895	2171	2475
Filtre 15	2171	2475	2812
Filtre 16	2475	2812	3184
Filtre 17	2812	3184	3596
Filtre 18	3184	3596	4052
Filtre 19	3596	4052	4556
Filtre 20	4052	4556	5113
Filtre 21	4556	5113	5730
Filtre 22	5113	5730	6412
Filtre 23	5730	6412	7166
Filtre 24	6412	7166	8000

Tableau 42 : Filtres triangulaires

Parmi ces 51 coefficients MFCC, nous choisissons les douze coefficients les plus efficaces : $MFCC_0$, $MFCC_1$, $MFCC_2$, $MFCC_4$, $MFCC_5$, $MFCC_6$, $MFCC_{12}$, $\Delta MFCC_0$, $\Delta MFCC_3$, $\Delta MFCC_5$, $\Delta MFCC_{13}$, et $\Delta MFCC_{16}$ pour tester la performance.

La raison de ce choix est de faciliter la comparaison avec les douze paramètres sélectionnés en utilisant la sélection forcée séquentielle avant. Le Tableau 43 présente le résultat obtenu de cette combinaison, en moyenne, nous avons 70,0 % de classification correcte. Ce taux est 73,8 % si nous choisissions les 12 paramètres dont les valeurs FDRs sont les plus élevées (Tableau 44). Cependant, en comparaison avec le résultat obtenu 78,7 % par la sélection forcée séquentielle avant dans le Tableau 45 (le processus de sélection forcée séquentielle avant sera présenté dans la section suivante), nous trouverons immédiatement que ces deux approches ne seront pas notre choix. L'inefficacité de ces deux approches peut s'expliquer par l'omission de la prise en compte de la corrélation entre les paramètres.

	colère (%)	***joie (%)***	***neutre (%)***	***tristesse (%)***	***surprise (%)***
colère	63,5	15,4	1,9	7,7	11,5
joie	17,3	63,5	0,0	1,9	17,3
neutre	19,2	5,8	67,3	1,9	5,8
tristesse	1,9	0,0	13,5	80,8	3,8
surprise	9,6	15,4	0,0	0,0	75,0
moyenne	70,0				

Tableau 43 : Taux de reconnaissance de la combinaison des coefficients qui sont isolement les plus efficaces : $MFCC_0$, $MFCC_1$, $MFCC_2$, $MFCC_4$, $MFCC_5$, $MFCC_6$, $MFCC_{12}$, $\Delta MFCC_0$, $\Delta MFCC_3$, $\Delta MFCC_5$, $\Delta MFCC_{13}$, et $\Delta MFCC_{16}$

	Colère (%)	***Joie (%)***	***Neutre (%)***	***Tristesse (%)***	***Surprise(%)***
Colère	67,3	17,3	0,0	3,8	11,5
Joie	3,8	71,2	0,0	1,9	23,1
Neutre	15,4	1,9	78,8	3,8	0,0
Tristesse	2,0	5,9	11,8	76,9	2,0
Surprise	5,8	17,3	1,9	0,0	75,0
Moyenne	73,8				

Tableau 44 : Taux de reconnaissance de la combinaison des 12 coefficients dont les valeurs FDR sont les plus élevées : $MFCC_0$, $MFCC_1$, $MFCC_2$, $MFCC_3$, $MFCC_4$, $MFCC_5$, $MFCC_6$, $MFCC_8$, $MFCC_{11}$, $MFCC_{12}$, $MFCC_{13}$, $MFCC_{14}$

	Colère (%)	***Joie (%)***	***Neutre (%)***	***Tristesse (%)***	***Surprise(%)***
Colère	81,4	8,5	0,0	1,7	8,5
Joie	7,7	78,8	0,0	1,9	11,5
Neutre	13,5	1,9	73,1	5,8	5,8
Tristesse	0,0	1,9	3,8	90,4	3,8
Surprise	11,5	19,2	0,0	0,0	69,2
Moyenne	78,8				

Tableau 45 : Taux de reconnaissance de la sélection forcée séquentielle avant avec la combinaison finale $MFCC_0$, $MFCC_2$, $MFCC_5$, $MFCC_6$, $MFCC_9$, $MFCC_{11}$, $MFCC_{12}$, $MFCC_{16}$, $\Delta MFCC_0$, $\Delta MFCC_{15}$, $\Delta\Delta MFCC_1$, et $\Delta\Delta MFCC2$

En conclusion, bien que les valeurs FDRs donne une performance acceptable pour choisir les meilleurs paramètres isolement, comme nous avons remarqué, à côté du problème de la corrélation entre des paramètres, nous devons aussi tenir en compte de la dépendance des valeurs FDRs au niveau de la dynamique des paramètres, et l'amplitude des paramètres est un aspect qui influe le plus sur cette mesure. Cela explique aussi pourquoi les douze coefficients choisis en se basant seulement sur ce critère sont tous les coefficients MFCCs malgré l'efficacité des coefficients ΔMFCCs. La différence entre 78,8 % (par la sélection forcée séquentielle en

avant) contre 73,8 % (par les valeurs FDR) démontre bien notre conclusion. Ce fait montre également l'efficacité de l'approche de la sélection forcée séquentielle en avant.

5.3.2.3.4Compresser par l'Analyse des Composantes Principales (PCA)

Des techniques comme l'Analyse des Composants Principaux (ACP) ou l'Analyse Discriminante Linéaire (ADL) permettent d'obtenir un ensemble de paramètres que l'on peut dire « optimal » à partir des paramètres « bruts ». Cette optimisation se fait en termes de diminution de la charge de calcul en diminuant la dimensionnalité de l'espace des caractéristiques mais en gardant en même temps presque la même quantité de l'information.

Cette capacité de compression des paramètres est non seulement profitable pour éliminer la redondance, mais elle est aussi une stratégie de recherche d'une combinaison optimale à partir d'un ensemble de paramètres bruts, au lieu des tests exhaustifs de toutes les combinaisons.

C'est la raison pour laquelle, nous avons choisi l'Analyse en Composantes Principales (ACP) un moyen pour compresser l'ensemble des paramètres. Le nombre de composantes est choisi en se basant sur leurs valeurs propres obtenues par le processus d'analyse de PCA. Effectivement, plus la valeur propre est importante, plus l'information portée par cette composante est significative ou, autrement dit, un composant avec une valeur propre nulle ne fournit aucune information et l'élimination de cette composante n'influe pas sur la richesse de l'information de l'espace. La Figure 29 montre un exemple des valeurs propres de l'espace de 51 dimensions corresponds aux 51 caractéristiques basées sur MFCC.

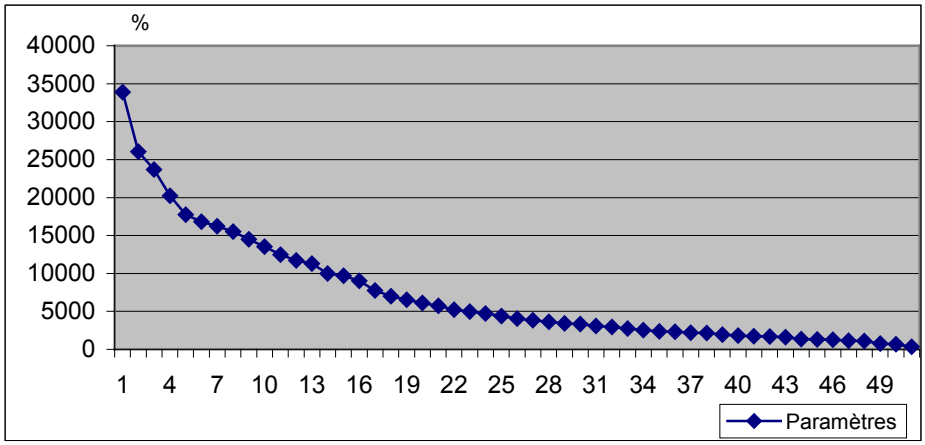

Figure 29 : Valeurs propres des vecteurs des 51 caractéristiques basées sur MFCCs

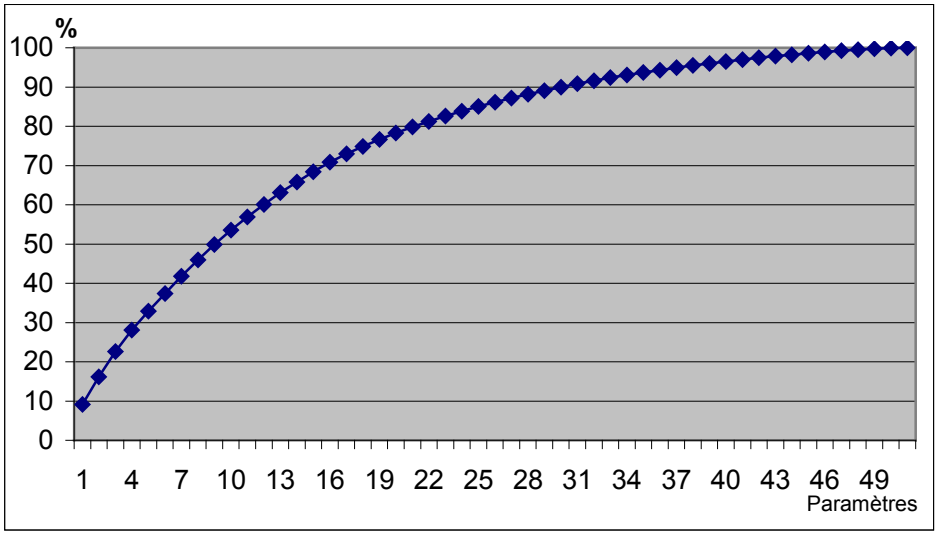

Figure 30 : Courbe sur la quantité d'information (%) des 51 caractéristiques basées sur MFCCs

Comme dans la Figure 30, nous constatons qu'en gardant 39 premiers composants (~ 80 % des paramètres) pour notre l'espace des caractéristiques, nous avons déjà 96 % d'information de tous les 51 composants. La section suivante donnera des résultats de nos expérimentations dans la reconnaissance de l'émotion avant et après la compression.

Dans les figures 30, 31, et 32, la méthode ACP a été utilisée pour compresser l'ensemble initial de paramètres en un nouvel ensemble moins chargé mais en gardant en même temps le maximum possible d'information de l'ensemble initial. Les ensembles initiaux de MFCC, de LFCC et de LPC se composent respectivement de 51, 48 et 48 paramètres comme nous l'avons présenté précédemment. Nous pouvons constater à partir des résultats que dans les deux cas extrêmes de l'ACP (compression maximale ou pas de compression), on obtient presque les mêmes performances par rapport aux résultats obtenus sans utiliser l'ACP comme le montre le Tableau 46.

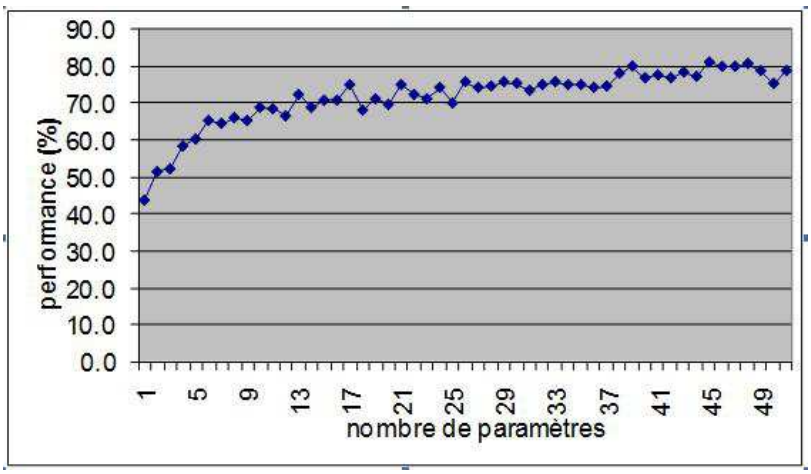

Figure 31 : PCA avec MFCCs

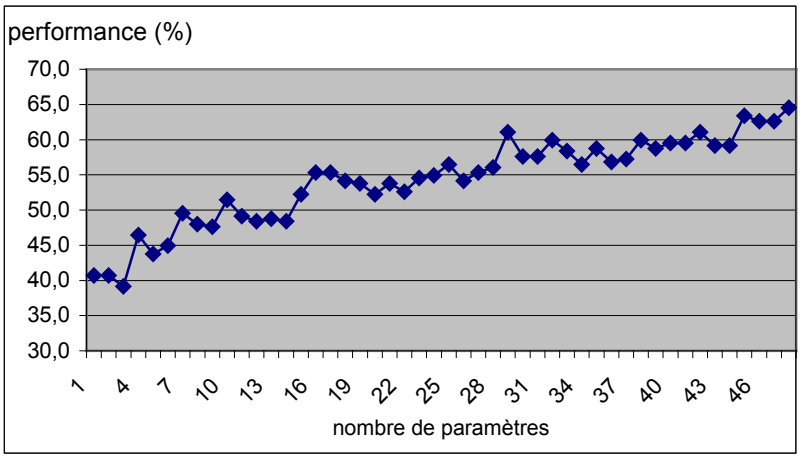

Figure 32 : PCA avec des LPCs

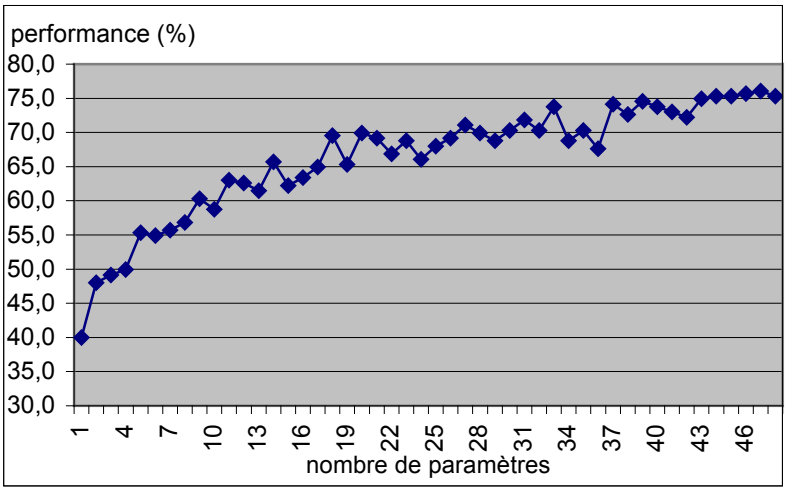

Figure 33 : PCA avec des LFCCs

Aussi selon ces résultats, nous obtiendrons la performance presque maximale avec environ 40 coefficients (78 % compression) des MFCCs, 37 coefficients (77 % compression) des LFCCs et 45 coefficients (94 % compression) des LPCs. La différence de la performance entre l'ensemble de MFCCs et l'ensemble de LFCCs peut s'expliquer par le fait que les coefficients MFCCs contiennent plus d'informations de la partie de basse fréquence où les informations de la prosodie se trouvent.

Cependant, pour obtenir le même niveau de performance obtenu par la méthode de sélection forcée séquentielle en avant, nous devrions compresser l'espace de paramètres en au moins 38 paramètres de MFCCs qui donne 78,2 % ou 29 paramètres de LPCs qui donne 61,1 % ou 33 paramètres de LFCCs qui donne 73,8 %.

En conclusion pour la partie d'étude des méthodologies pour filtrer des paramètres pour le corpus utilisé, nous constatons que parmi les trois approches : la sélection se basant sur des valeurs FDR, la sélection forcée séquentielle en avant et la compression utilisant la méthode

ACP. La première approche par FDR est la moins efficace, l'utilisation de la méthode ACP peut réduire la dimension de l'espace des caractéristiques mais nous ne donne pas le même avantage que la sélection forcée séquentielle en avant : la simplicité ou la rapidité et le nombre de paramètres sont beaucoup réduits. Donc, nous proposons également dans le cas de reconnaissance dépendante du locuteur d'utiliser les douze paramètres de MFCCs : $MFCC_0$, $MFCC_2$, $MFCC_{11}$, $MFCC_9$, $\Delta MFCC_0$, $\Delta MFCC_{15}$, $MFCC_{12}$, $\Delta\Delta MFCC_1$, $\Delta\Delta MFCC_2$, $MFCC_5$, $MFCC_6$, $MFCC_{16}$ qui donnent 78,8 % au lieu d'utiliser tout ensemble de 51 paramètres qui ne donne qu'une amélioration de 1,3 %. Dans la partie suivante, nous étudions la performance des autres paramètres en combinaison avec les MFCCs sélectionnés.

5.3.2.3.5 Sélection forcée séquentielle en avant

La méthode optimale pour sélectionner la meilleure combinaison de paramètres acoustiques est de tester expérimentalement toutes les combinaisons possibles et d'évaluer les taux de bonne classification qu'elles donnent afin de choisir la meilleure mais cette méthode est très coûteuse en temps de calcul et peu utilisée. Nous proposons donc un protocole appelé « sélection forcée séquentielle avant » SFSA qui est une amélioration de la méthode « sélection séquentielle avant » (Sequential Forward Selection - SSA) [Ververidis et al, 2005-2]. Nous effectuons aussi une comparaison entre des résultats obtenus par le SFSA et ceux obtenus par le critère de Fisher. La comparaison entre ces approches n'est effectuée que dans le cas de reconnaissance dépendante du locuteur pour vérifier et pour montrer que la méthode SFSA est toujours plus efficace. Dans les autres cas de reconnaissance multi-locuteur et indépendante du locuteur, nous ne nous servons que de la méthode SFSA.

La sélection séquentielle en avant (SSA) est effectuée comme suit : étant donné Z l'ensemble de tous les paramètres, N le nombre de paramètres à étudier. A partir d'un ensemble initialement vide des paramètres Z_0, à chaque étape d'avancement (inclusion) au niveau l, nous cherchons le paramètre z appartenant à $(Z\text{-}Z_l\text{-}1)$ tels que la performance obtenue avec $Z_l = Z_{l-1} \cup z$ est maximisée. Le processus s'arrête si la performance ne peut pas être améliorée.

L'inconvénient de cet algorithme est que nous pouvons tomber dans un maximum local et le résultat obtenu ne sera pas aussi bon que prévu. Pour que la sélection des paramètres soit plus efficace et plus robuste nous introduisons un autre critère dans le processus : le processus s'arrête seulement si la performance ne peut pas être améliorée et le résultat obtenu est supérieur à un seuil donné. Le seuil est connu avant, dans la plupart des cas nous choisissons le taux de reconnaissance T obtenu en utilisant tous les N paramètres pour le seuil. Cela assure que le processus de notre approche ne s'arrête pas à un mauvais point maximum local et que le résultat obtenu par notre approche est toujours supérieur à celui obtenu par la méthode SSA et au moins égal à T. Effectivement,

+ Si le processus s'arrêt au point où le nombre de paramètres choisis est inferieur à N et le résultat est supérieur à T. Nous avons trouvé une bonne combinaison de paramètres, (c'est le cas de reconnaissance indépendante du locuteur).

+ Si le processus ne peut pas améliorer la performance ou autrement dit il s'arrêt au point où le nombre de paramètres choisis est égal à N et le résultat obtenu est égal à T. Nous allons considérer les points maxima locaux pour choisir celui qui donne une bonne performance mais utilise moins de paramètres.

Nous constatons aussi que l'ajout continu des paramètres dans l'ensemble Z_l fait grandir rapidement Z_l. Et le volume important du Z_l causera naturellement la redondance de

l'information en raison de la corrélation entre des caractéristiques. La redondance devient une charge si le nombre de caractéristiques est important. La compression de l'espace des caractéristiques, ainsi que l'enlèvement de la redondance est donc une étape nécessaire.

Il faut noter que le résultat obtenu par ce protocole n'est pas le résultat optimal car plusieurs cas de combinaison ne sont pas tenus en compte, cependant ses avantages de la simplicité et de la performance assez élevée nous permettent obtenir une combinaison satisfaisante.

En utilisant ce protocole, en raison de nombre important de paramètres, nous nous intéressons tout d'abord aux trois ensembles principaux de paramètres : des MFCCs, des LFCCs et des LPCs. La sélection de la combinaison la plus efficace sera premièrement effectuée à l'intérieur de chaque ensemble. La combinaison qui donne le meilleur résultat sera utilisée pour l'étude de la fusion avec les autres caractéristiques.

Le Tableau 46 montre les résultats obtenus de notre processus de sélection forcée séquentielle en avant avec les trois ensembles de paramètres. Le Tableau 47 nous montre bien l'efficacité de la méthode de sélection en avant en termes de nombre de paramètres utilisés, mais aussi en termes de la performance obtenue. Cela peut s'expliquer par la capacité de détecter l`ensemble de paramètres les moins corrélés et l'utilisation de moins de paramètres réduit en même temps le bruit causé par des paramètres.

Nombre de paramètres	***Combinaison la plus efficace des MFCCs***	***Taux de class.*** *(%)*	***Combinaison la plus efficace des LPC***	***Taux de class.*** *(%)*	***Combinaison la plus efficace des LFCC***	***Taux de class.*** *(%)*
1	$Z_1=MFCC_0$	46,2	$Z_1=lpc_3$	41,5	$Z_1=LFCC_1$	40,4
2	$Z_2=Z_1+MFCC_2$	56,2	$Z_2=Z_1+lpc_{13}$	42,7	$Z_2=Z_1+LFCC_{10}$	41,2
3	$Z_3=Z_2+MFCC_{11}$	61,9	$Z_3=Z_2+\Delta lpc_6$	43,5	$Z_3=Z_2+LFCC_6$	45,8
4	$Z_4=Z_3+MFCC_9$	65,0	$Z_4=Z_3+lpc_1$	51,2	$Z_4=Z_3+LFCC_{13}$	51,9
5	$Z_5= Z_4+\Delta MFCC_0$	65,0	$Z_5=Z_4+\Delta\Delta lpc_7$	52,3	$Z_5=Z_4+\Delta LFCC_{11}$	51,9
6	$Z_6= Z_5+\Delta MFCC_{15}$	68,5	$Z_6=Z_5+lpc_5$	55	$Z_6=Z_5+LFCC_2$	53,8
7	$Z_7= Z_6+MFCC_{12}$	72,3	$Z_7=Z_6+lpc_9$	58,2	$Z_7=Z_6+LFCC_3$	59,2
8	$Z_8= Z_7+\Delta\Delta MFCC_1$	72,6	$Z_8=Z_7+\Delta lpc_1$	58,8	$Z_8=Z_7+LFCC_{11}$	64,2
9	$Z_9=Z_8+\Delta\Delta MFCC_2$	73,8	$Z_9=Z_8+lpc_{11}$	60,8	$Z_9=Z_8+\Delta LFCC_1$	65,4
10	$Z_{10}=Z_9+MFCC_5$	76,2	$Z_{10}=Z_9+lpc_{12}$	61,5	$Z_{10}=Z_9+LFCC_8$	67,7
11	$Z_{11}=Z_{10}+MFCC_6$	77,3	$Z_{11}=Z_{10}+\Delta lpc_2$	61,7	$Z_{11}=Z_{10}+LFCC_{15}$	68,5
12	$Z_{12}=Z_{11}+MFCC_{16}$	78,7			$Z_{12}=Z_{11}+\Delta LFCC_5$	72,3

Tableau 46 : Les paramètres obtenus en utilisant la sélection forcée séquentielle en avant

Type de paramètres	*Taux de classification (%)*	*Type de paramètres*	*Taux de classification (%)*	*Type de paramètres*	*Taux de classification (%)*
17 paramètres originaux de MFCC	77,7	*16 paramètres originaux de LPC*	61,9	*16 paramètres originaux de LPC*	73,5
17 dérivées premières de MFCC	55,0	*16 dérivées premières de LPC*	45,4	*16 dérivées premières de MFCC*	57,7
17 dérivées deuxièmes de MFCC	51,1	*16 dérivées deuxièmes de LPC*	40,4	*16 dérivées deuxièmes de LPC*	48,5
51 paramètres de MFCC	80,0	*48 paramètres de LPC*	65,0	*48 paramètres de LPC*	75,9
12 paramètres MFCC choisis par SFSA	78,7	*11 paramètres LPC choisis par SSA*	61,7	*12 paramètres LFCC choisis par SFSA*	72,3

Tableau 47 : Les résultats obtenus en comparaison

En conclusion, dans le contexte de nos expérimentation avec le corpus en danois DES, la sélection forcée séquentielle en avant est une approche simple mais efficace, elle nous permet d'obtenir une bonne performance avec un nombre de paramètres beaucoup moins important que le nombre initial de paramètres. (78,7 % avec 12 paramètres contre 77,7 % avec 17 paramètres MFCC originaux et contre 80,0 % avec 51 paramètres incluant les dérivées premières et secondes).

La méthode de sélection forcée séquentielle en avant s'arrête lorsque l'ajout d'un nouveau paramètre seul ne conduit plus à aucune amélioration. Elle ne permet pas toujours d'atteindre la performance maximale que l'on pourrait atteindre en ajoutant plus d'un paramètre mais elle atteint tout de même un point proche de l'optimum (78,7 % avec quatre fois moins de paramètres contre 80,0 % dans le cas optimal). Cette caractéristique sera très utile pour les systèmes qui travaillent en temps réels.

La méthode de sélection forcée séquentielle en avant fonctionne mieux que la méthode par analyse en composantes principales pour la réduction du nombre de dimensions. Dans le cas des MFCC, l'optimum est atteint avec 12 paramètres au lieu de 38 sur 51 pour une performance comparable.

Comme montré dans le Tableau 46, nous obtenons la performance optimale de 78,7 % avec 12 paramètres MFCC, 61,7 % avec 11 paramètres LPC et 72,3 % avec 12 paramètres LFCCs. Jusqu'ici nous pouvons conclure à une moins bonne efficacité des LPC pour la discrimination des émotions par rapport aux MFCC et aux LFCC et c'est la raison pour laquelle nous ne les considèrerons plus dans la suite de notre étude.

Une autre remarque est que pour tous les trois ensembles de paramètres, les paramètres originaux occupent la majorité de la combinaison. Les dérivées en deuxième ordre contribuent le moins à l'amélioration de la performance du système. Nous pouvons l'expliquer par le fait que les émotions sont des variations lentes, les Δ et les $\Delta\Delta$ n'ont pas donc beaucoup d'influence.

La Figure 34 qui nous montre le processus d'amélioration au long de la sélection forcée séquentielle en avant des trois ensembles de paramètres.

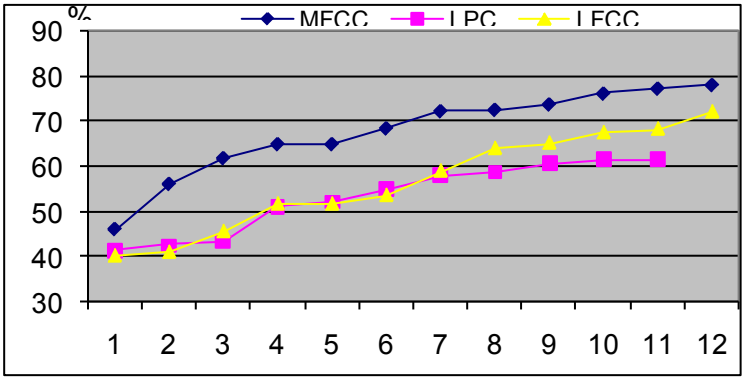

Figure 34 : Sélection forcée séquentielle en avant

5.3.2.4 Fusion des paramètres

La combinaison entre les douze MFCCs sélectionnés ou les douze LFCCs sélectionnés et les autres paramètres comme les paramètres prosodiques, le nombre d'extrémités, le nombre de passages par zéro et leurs dérivées ont été effectués afin de tester la contribution possible de ces paramètres dans l'amélioration de la performance du système en mode des paramètres locaux. Comme les résultats expérimentaux obtenus dans le Tableau 48, en comparaison avec 78,7 % obtenus avec seulement 12 MFCCs du Tableau 46 et 72,3 % obtenus avec 12 LFCCs, bien qu'il existe les deux cas de combinaison dont le taux moyen de classification sont un peu plus élevés que les 12 MFCCs originaux, nous ne pouvons pas confirmer l'utilité de ΔΔF0Rel et DebitPRel en tenant en compte l'intervalle de confiance de ces résultats comme montrés dans le Tableau 48 (l'intervalle de confiance à 95 %).

Paramètres	***Taux de classification** (%)*
12 MFCCs	78,7 ± 3,44
12 MFCCs + ΔΔF_0Rel	80,4 ± 4,10
12 MFCCs + DebitPRel	79,6 ± 3,84

Tableau 48 : La combinaison des MFCCs avec d'autres paramètres

En conclusion, dans cette partie, nous avons travaillé avec deux types de paramètres : globaux et locaux. Afin d'étudier le rôle des paramètres en aspect émotionnel, les paramètres globaux ont été utilisés et nous avons appliqué une comparaison relative entre des états émotifs avec l'état neutre pour enlever le plus possible les éléments qui dépendent du locuteur. En raison de l'instabilité de ces paramètres globaux, nous avons essentiellement exploité des paramètres locaux pour notre étude et pour l'amélioration de la performance du système. 80,4% est un taux assez élevé que nous avons obtenu avec des paramètres locaux. Nous constatons également l'efficacité de l'ensemble de MFCCs par rapport les deux autres ensembles voisins : LFCCs, LPC. C'est la raison pour laquelle, l'ensemble de LFCCs et de LPC seront éliminés de notre ensemble de paramètres à étudier.

En parallèle, en étudiant les trois approches de sélection des paramètres : par le critère FDR, par la sélection forcée séquentielle en avant et par la compression par la méthode PCA, nous montrons que la sélection séquentielle se montre une méthode simple mais efficace. Nous proposons donc d'utiliser cet ensemble de paramètres dans nos études des parties suivantes.

5.4. Performance des paramètres multi-locuteur

L'étude précédente a été faite dans le cas de la reconnaissance dépendante du locuteur dans lequel les données d'apprentissage et les données de test appartiennent à un même locuteur dans nos corpus. Dans le cas de reconnaissance multi-locuteur, nous utilisons les données d'apprentissage et de test qui appartient à des locuteurs différents. Les modèles entrainés sont plus généraux mais la performance de classification est aussi atténuée. La Figure 35 présente l'efficacité de chaque coefficient MFCC ; les tests sont faits avec le corpus DES ; le principe du « *Leave One Out* » est appliqué au niveau des énoncés. Nous rappelons que le modèle GMM 16 gaussiennes est toujours utilisé comme le modèle de validation pour les 5 émotions ; le choix au hasard est donc de 20 %.

De même que dans le cas de la reconnaissance dépendante du locuteur, $MFCC_0$ et $MFCC_2$ sont les coefficients les plus efficaces. Nous constatons aussi l'inefficacité des dérivées deuxièmes.

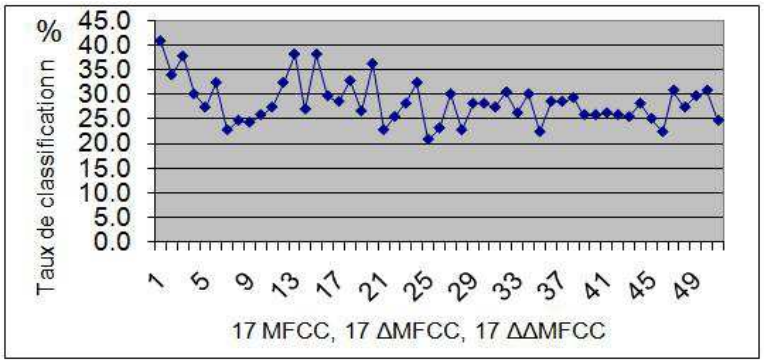

Figure 35 : Performance des MFCCs dans le cas de reconnaissance multi-locuteur

Les paramètres prosodiques et les autres paramètres comme le nombre de passages par zéro, le nombre d'extrémités ne sont plus aussi efficaces que dans le cas de reconnaissance dépendante du locuteur comme on peut le voir dans le Tableau 49, le Tableau 38.

Paramètres	***Taux de classification (%)***	***Paramètres***	***Taux de classification (%)***
F_0	31,9	***F_0Rel***	33,5
ΔF_0	34,2	***ΔF_0Rel***	38,5
$\Delta\Delta F_0$	23,8	***$\Delta\Delta F_0$Rel***	30,8
Intensité	36,5	***IntensitéRel***	19,6
ΔIntensité	26,9	***ΔintensitéRel***	23,1
ΔΔIntensité	24,2	***ΔΔIntensitéRel***	20,8
Debit	26,9	***DebitRel***	25,4
ΔDebit	21,5	***ΔDebitRel***	18,8
ΔΔDebit	25,0	***ΔΔDebitRel***	19,6
ZCR	26,9	***N. d'extrémités***	31,2
ΔZCR	29,6	***ΔN. d'extrémités***	29,2
ΔΔZCR	23,5	***ΔΔN. d'extrémités***	28,5

Tableau 49 : Performance de chaque paramètre

Cependant, nous remarquons que les coefficients relatifs de la fréquence fondamentale F_0 sont plus efficaces que les coefficients F_0 originaux.

5.4.1.1 Fusion des paramètres

Selon les résultats du Tableau 51, le meilleur taux de classification que nous pouvons obtenir est de 68,8 % en utilisant 51 paramètres MFCC. Le processus de la sélection séquentielle nous permet de raffiner cet ensemble de paramètres en gardant maximalement la performance du système. Les 7 paramètres $MFCC_0$, $MFCC_2$, $MFCC_{11}$, $\Delta\Delta MFCC_0$, $\Delta MFCC_0$, $MFCC_{12}$, $MFCC_5$ obtenus par cette sélection nous donnent 68,1 % et les 10 paramètres $LFCC_1$, $LFCC_{13}$, $LFCC_3$, $LFCC_5$, $LFCC_{15}$, $\Delta LFCC_4$, $\Delta LFCC_{13}$, $LFCC_{10}$, $\Delta LFCC_1$, $LFCC_2$ nous donnent 56,5 %.

Les paramètres dans le Tableau 50 appartiennent à l'ensemble des paramètres que nous avons sélectionnés pour le cas de reconnaissance dépendante du locuteur.

Nombre de paramètres	***Combinaison la plus efficaces des MFCCs***	***Taux de classification (%)***
1	$Z_1=MFCC_0$	40,8
2	$Z_2=Z_1+MFCC_2$	52,3
3	$Z_3=Z_2+MFCC_{11}$	58,1
4	$Z_4=Z_3+\Delta\Delta MFCC_0$	61,5
5	$Z_5= Z_4+\Delta MFCC_0$	62,7
6	$Z_6= Z_5+MFCC_{12}$	63,5
7	$Z_7= Z_6+MFCC_5$	68,1

Tableau 50 : Les paramètres obtenus en utilisant la sélection forcée séquentielle en avant

Type de paramètres	***Taux de classification (%)***
12 Paramètres trouvés dans le cas mono-locuteur	65,9
17 paramètres originaux de MFCC	66,5
17 dérivées premières de MFCC	49,6
17 dérivées deuxièmes de MFCC	35,4
51 paramètres de MFCC	68,8
7 paramètres obtenus par SFSA[6] pour le cas multi-locuteur	68,1

Tableau 51 : Comparaison de l'efficacité des ensembles de paramètres appliqués pour le cas de reconnaissance multi-locuteur

Jusqu'ici, avec les données du corpus DES, les paramètres MFCC se montrent toujours plus performants que les paramètres LFCCs, donc nous pouvons éliminer les LFCCs de l'ensemble de paramètres à étudier.

[6] voir la section 5.3.2.3.5

En général, nous trouvons que par rapport au cas de reconnaissance dépendante du locuteur, le taux de reconnaissance multi-locuteur est considérablement atténué en raison de la variété interlocuteur, cette atténuation est montrée par la différence significative entre les taux de classification correcte 68,8 % (en utilisant 51 paramètres MFCCs – multi-locuteur) contre 80,0 % (en utilisant 51 paramètres MFCCs – mono locuteur), 66,5 % (avec 17 paramètres MFCCs - multi locuteur) contre 77,7 % (avec 17 paramètres MFCCs - mono locuteur) et 68,1 % (avec 7 paramètres sélectionnés par la sélection forcée séquentielle en avant - multi locuteur) contre 78,7 % (avec 12 paramètres sélectionnés par la sélection forcée séquentielle en avant - mono locuteur).

Pour augmenter la performance du système, la combinaison de ces 7 paramètres MFCCs avec les autres paramètres prosodiques, le nombre d'extrémités et le nombre de passages par zéro ont aussi été effectués. Le Tableau 52 ci-dessous présente des résultats obtenus où la fusion de 7 paramètres MFCCs avec $\Delta F_0 Rel$ nous donne le meilleur résultat (69,6 %). Les autres tentatives pour combiner ces huit paramètres (7MFCCs + $\Delta F_0 Rel$) par la sélection forcée séquentielle en avant n'améliorent pas ce résultat.

Paramètre	Taux de classification (%)	***Paramètre***	Taux de classification (%)
F_0	64,2	$F_0 Rel$	68,8
ΔF_0	63,8	$\Delta F_0 Rel$	**69,6**
$\Delta\Delta F_0$	63,1	$\Delta\Delta F_0 Rel$	65,0
Intensité	66,2	***IntensiteRel***	59,2
Δintensite	67,3	***ΔintensiteRel***	68,1
ΔΔIntensite	65,4	***ΔΔIntensiteRel***	63,5
Débit	64,6	***DebitRel***	56,9
ΔDebit	64,6	***ΔDebitRel***	66,2
ΔΔDebit	67,3	***ΔΔDebitRel***	65,8
N, d'extremités	66,9	***ZCR***	68,1
ΔN, d'extremités	68,8	***ΔZCR***	67,3
ΔΔN, d'extremités	66,5	***ΔΔZCR***	66,5

Tableau 52 : Performance de chaque paramètre en combinaison avec les 7 MFCCs sélectionnés

Donc, en conclusion, dans le cas de reconnaissance multi-locuteur des 5 émotions, nous pouvons obtenir le taux de reconnaissance 69,6 % avec l'ensemble de huit paramètres 7MFCCs + $\Delta F_0 Rel$. Ce taux est la moyenne des résultats obtenus avec les 4 locuteurs (2 femmes et 2 homme) du corpus DES. Dans la partie suivante, nous étudierons la reconnaissance indépendante du locuteur, et la variété interlocuteur influence encore plus fortement la performance du système car le locuteur dont les données sont utilisées pour les tests n'est pas connu par le système lors de l'entrainement.

5.5. Performance des paramètres indépendante des locuteurs

Comme nous avons mentionné ci-dessus, la reconnaissance indépendante du locuteur est le cas le plus général mais aussi le plus difficile de tous les systèmes de reconnaissance de la parole. La raison principale est l'absence de données du sujet (pour le test) dans l'ensemble de données utilisés pour l'apprentissage.

Notre système d'indexation de l'émotion a le même problème car, dans la plupart des cas, le sujet est inconnu pour le système. Pour effectuer ces expérimentations, le principe du « *leave one out* » est appliqué au niveau des locuteurs. La dégradation importante de la performance est visible dans la Figure 36 et le Tableau 53.

Dans la Figure 36, nous présentons le taux de classification du système en utilisant isolement chaque paramètre de MFCCs, cela a pour but de faciliter la comparaison de l'efficacité de chaque paramètre isolé entre les cas de reconnaissance : mono et multi et indépendance du locuteur. Le Tableau 53 présente les résultats des 9 étapes de la sélection forcée séquentielle en avant qui trie les 9 meilleurs paramètres, ceux qui donne 52,7 % de classification correcte des 5 émotions. Bien que ce taux 52,7 % soit beaucoup plus élevé que le taux de 44,2 % obtenue en utilisant seulement les 17 paramètres MFCCs orignaux, en comparaison avec la reconnaissance multi locuteur du Tableau 50, nous constatons une différence considérable 15,4 % entre ce taux 52,7 % et 68,1 % du cas de reconnaissance multi-locuteur.

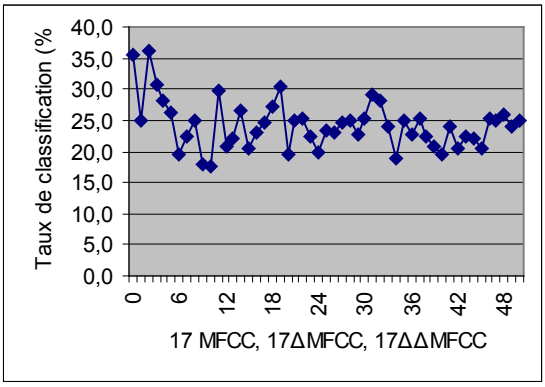

Figure 36 : Performance des MFCCs dans le cas de reconnaissance indépendante du locuteur

Nombre de paramètres	***Combinaison la plus efficaces des MFCCs***	***Taux de class.***
1	Z_1=$MFCC_2$	36,2
2	Z_2=Z_1+$MFCC_0$	40,8
3	Z_3=Z_2+$MFCC_3$	44,2
4	Z_4=Z_3+$MFCC_{11}$	47,3
5	Z_5= Z_4+$\Delta MFCC_2$	48,8
6	Z_6= Z_5+$\Delta MFCC_{12}$	51,2
7	Z_7= Z_6+$MFCC_7$	51,2
8	Z_8= Z_7+$\Delta MFCC_{14}$	50,4
9	Z_9= Z_7+$\Delta MFCC_{11}$	52,7

Tableau 53 : Les paramètres sélectionnés après la sélection forcée séquentielle en avant

Type de paramètres	*Taux de classification (%)*
7 paramètres trouvés dans le cas multi-locuteur	41,9
12 paramètres trouvés dans le cas mono locuteur	39,6
17 paramètres originaux de MFCC	44,2
17 dérivées premières de MFCC	41,5
17 dérivées deuxièmes de MFCC	31,9
51 paramètres de MFCC	43,1
9 paramètres choisis par SFSA pour le cas de reconnaissance indépendante du locuteur	52,7

Tableau 54 : Les paramètres sélectionnés après la sélection forcée séquentielle en avant

En faisant la comparaison dans le Tableau 54, une remarque dont nous voudrions parler ici est la meilleure performance en utilisant seulement les 9 paramètres (qui est de 52,7 % contre 43,1 % en utilisant tous les 51 paramètres). Effectivement, l'utilisation de toutes les informations ne donne pas toujours la meilleure performance. Pour expliquer ce problème, il nous faut être d'accord avec le fait que : plus le modèle est spécifique, plus le modèle est bon pour les cas spécifiques, mais moins le modèle est bon pour les cas généraux ; et c'est facile à constater que dans notre cas de reconnaissance indépendante du locuteur : l'ajout des paramètres veut aussi dire l'ajout des descriptions plus spécifiques pour les données des locuteurs que nous utilisons afin d'entrainer les modèles, autrement dit, les modèles entrainés par ces données sont plus spécifiques pour les locuteurs connus et le test effectué avec les données d'un locuteur inconnu atténue surement significativement la performance (-9,6 % dans notre cas). A partir de cette remarque, encore une fois, nous pouvons affirmer l'avantage de la méthode de sélection forcée séquentielle en avant pour le cas de reconnaissance indépendante du locuteur car elle nous aide à enlever des paramètres trop spécifiques pour les locuteurs.

A partir des résultats du Tableau 53, nous trouvons également que la différence importante 8,5 % entre le taux de classification 44,2 % en utilisant 17 paramètres MFCC originaux et 52,7 % en utilisant la sélection forcée séquentielle en avant, en plus, le nombre de paramètre est significativement réduit par cette méthode (9 paramètres contre 17 paramètres).

Dans cet ensemble de 9 paramètres sélectionnés, nous trouvons aussi que la contribution des premières dérivées devient importante (4/9 paramètres).

	Taux de classification en utilisant 17 paramètres originaux de MFCCs	***Taux de classification en utilisant tous les 51 paramètres de MFCCs***	***Taux de classification en utilisant les paramètres sélectionnés par SFSA***
Mono-locuteur	77,7 %	80,0 %	78,7 %
Muli-locuteur	66,5 %	68,8 %	68,1 %
Indépendance du locuteur	44,2 %	43,1 %	52,7 %

Tableau 55 : Synthèse des résultats de trois cas d'étude de la reconnaissance mono-locuteur, multi-locuteur et indépendante du locuteur

Le Tableau 55 ci-dessus fait une synthèse des résultats obtenus pour les trois cas d'étude : la reconnaissance mono-locuteur, multi-locuteur et indépendante du locuteur. Pour la comparaison, nous mettons à côte ces résultats les taux de reconnaissance obtenus en utilisant 17 paramètres originaux de MFCCs et en utilisant tous les 51 paramètres de MFCCs.

Parmi ces trois cas de reconnaissance, l'utilisation de tous les 51 paramètres donnera les taux de reconnaissance les plus élevés pour les deux premiers cas. Cependant, c'est différent pour le dernier cas, la reconnaissance indépendante du locuteur. Effectivement 43,1 % obtenu avec 51 paramètres de la reconnaissance indépendante du locuteur est même inférieur à 44,2 % obtenu avec 17 paramètres originaux. Comme nous avons expliqué ci-dessus, cela peut être causé par le fait que l'ensemble de 51 paramètres contient trop de paramètres dépendants du locuteur, et la dégradation causée par ces paramètres est plus forte que l'amélioration que ces paramètres peuvent y contribuer.

Et enfin, une remarque la plus importante que nous voulons souligner ici est que malgré l'avantage de la méthode SSA qui peut sortir les paramètres les moins dépendants du locuteur, les 9 paramètres sélectionnés juste ci-dessus sont encore « bruts », la performance du système peut être encore beaucoup améliorée parce que nous n'avons pas résolu un problème aussi important que le problème de sélection des paramètres dans le système de reconnaissance indépendante du locuteur : la normalisation.

Effectivement, les inconvénients de ces paramètres « bruts » sont :

- plus ou moins, leur valeur (y compris le minimum, le maximum, la moyenne, la dynamique etc.) dépend du locuteur ;
- en les mettant en ensemble dans un vecteur de paramètres pour la modélisation, l'influence de chaque paramètre sur le modèle est décidée fortement par sa dynamique : plus la dynamique de paramètre est importante, plus le paramètre a une influence ou autrement dit le paramètre joue un rôle plus important.

La partie suivante portera sur une proposition d'une méthode de normalisation qui est pour but de l'intégrer avec la méthode de sélection forcée séquentielle en avant afin d'obtenir l'ensemble de paramètres les plus indépendants du locuteur mais aussi les plus discriminatifs.

5.6. Normalisation symbolique

Dans le domaine de reconnaissance de la parole en général et pour la reconnaissance de l'émotion en particulier, le fait que toutes les valeurs absolues mesurées pour les paramètres dépendent plus ou moins du locuteur, influe fortement sur la performance des systèmes de reconnaissance automatique. Pour limiter cet effet, ces valeurs absolues sont soit remplacées par les valeurs relatives, soit modifiées par un processus de normalisation.

La normalisation par la moyenne et par l'écart type sont des approches bien connues les plus populairement utilisées par des études comme [Furui, 1981], [Kwon et al, 2003], [Blouin et al, 2005], [Viikki 1998], [Hsu et al, 2006]. Cependant cette approche a l'inconvénient d'utiliser encore les valeurs réelles des paramètres après la normalisation, la dépendance du locuteur existe donc encore et atténue encore la performance du système. La Figure 37 montre un exemple sur les deux distributions d'un paramètre (des deux locuteurs par exemple) qui a la même moyenne, le même écart-type (parce que $s_1 = s_2$) mais leurs champs dynamiques sont très différents.

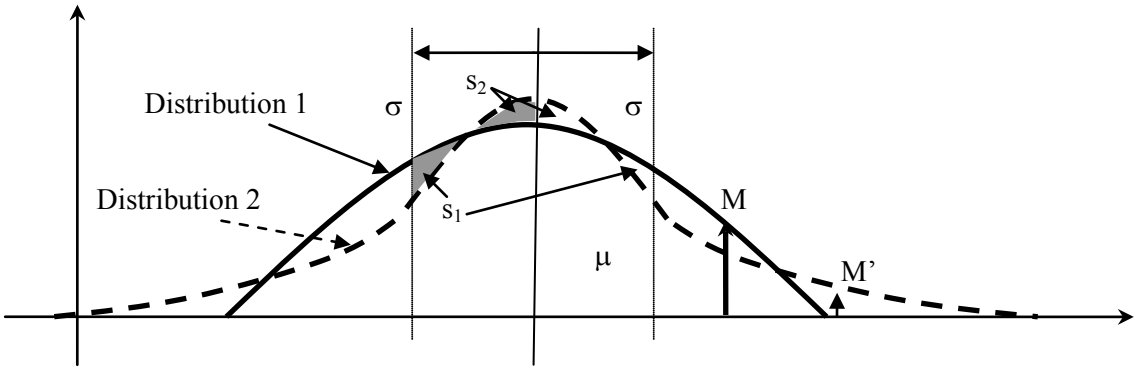

Figure 37 : Deux distributions ayant la même moyenne et le même écart-type

Dans le cas de la Figure 37, comment peut-on faire la correspondance entre la valeur M de la distribution 1 et la valeur M' de la distribution 2 ? Nous proposons donc d'utiliser une normalisation que nous appelons la normalisation symbolique. L'idée de cette approche est la suivante :

Considérons 2D et 2D' qui sont respectivement les champs dynamiques des valeurs de la distribution 1 et des valeurs de la distribution 2 comme dans la Figure 38.

Si on divise ce champ dynamique (2D et 2D') en N régions et remplace les valeurs dans chaque région par un symbole sémantique, à la place des valeurs absolues, pour tous les paramètres, nous aurons des valeurs représentés par un système de symboles unitaires.

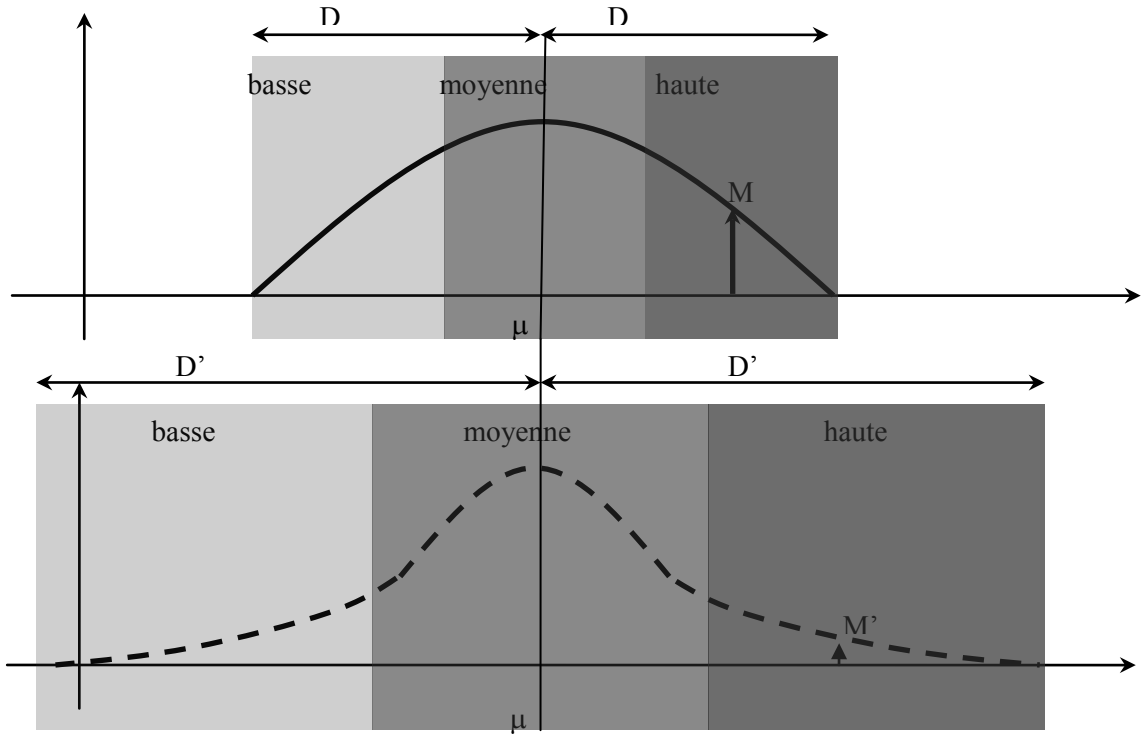

Figure 38 : Symbolisation des paramètres

La Figure 38 nous montre le cas où N = 3 avec les trois symboles sémantiques : basse, moyenne et haute. Il est alors clair que les deux points M et M' correspondent l'un à l'autre dans notre nouvelle échelle de représentation par les symboles.

Le nombre de régions peut varier de 1 à +∞. Si N=1 tous les paramètres sont identiques car ils sont tous représentés par un seul symbole. Théoriquement, plus N est grand, plus la nouvelle échelle des symboles est fine et plus la performance est améliorée mais il faut compter aussi plus de temps de traitement et, d'après nos résultats dans la reconnaissance de l'émotion, il est inutile de prendre une très grande valeur de N. La Figure 39 nous montre les résultats de notre test avec le paramètre $MFCC_0$ dans le cas de reconnaissance indépendante du locuteur (corpus DES, modèle GMM 16 gaussiennes) en fonction du nombre de régions symboliques N.

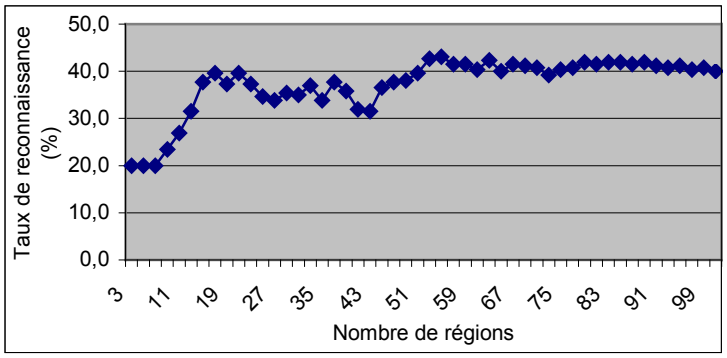

Figure 39 : Performance de $MFCC_0$ en fonction de nombre de régions symboliques

Nous pouvons constater que si le nombre de régions N est trop petit, la performance sera dégradée. Mais si N atteint une suffisamment grande valeur, la performance est saturée car l'amélioration est aussi limitée par la performance réelle du paramètre, pour notre cas dans la Figure 39, par l'observation nous avons choisi 53 pour le nombre de régions de saturation pour $MFCC_0$. Avec la même façon, le Tableau 56 présente des nombres de régions de saturation correspondant choisis pour 51 paramètres MFCCs.

Paramètres	***N. de régions symboliques***	***Paramètres***	***N. de régions symboliques***	***Paramètres***	***N. de régions symboliques***
MFCC0	53	ΔMFCC0	49	ΔΔMFCC0	33
MFCC1	25	ΔMFCC1	17	ΔΔMFCC1	47
MFCC2	7	ΔMFCC2	15	ΔΔMFCC2	9
MFCC3	9	ΔMFCC3	29	ΔΔMFCC3	15
MFCC4	17	ΔMFCC4	15	ΔΔMFCC4	21
MFCC5	11	ΔMFCC5	7	ΔΔMFCC5	9
MFCC6	3	ΔMFCC6	7	ΔΔMFCC6	11
MFCC7	5	ΔMFCC7	19	ΔΔMFCC7	29
MFCC8	15	ΔMFCC8	13	ΔΔMFCC8	10
MFCC9	17	ΔMFCC9	29	ΔΔMFCC9	19
MFCC10	25	ΔMFCC10	25	ΔΔMFCC10	7
MFCC11	23	ΔMFCC11	27	ΔΔMFCC11	31
MFCC12	9	ΔMFCC12	23	ΔΔMFCC12	12
MFCC13	9	ΔMFCC13	17	ΔΔMFCC13	11
MFCC14	17	ΔMFCC14	49	ΔΔMFCC14	21
MFCC15	7	ΔMFCC15	21	ΔΔMFCC15	35
MFCC16	13	ΔMFCC16	12	ΔΔMFCC16	9

Tableau 56 : Nombre de régions de saturation des 51 paramètres MFCCs

Selon les résultats du Tableau 41, comme 53 régions est le nombre le plus petit possible pour pouvoir maximiser l'efficacité de la transformation de tous paramètres en symbole, c'est la raison pour laquelle, nous avons choisi ces 53 régions pour notre processus de normalisation symbolique.

La Figure 40 nous donne les résultats obtenus avec les améliorations significatives en appliquant le processus de normalisation symbolique par rapport aux résultats obtenus sans normalisation. En moyenne nous obtenons respectivement les améliorations de 4,1 % ; de 2,0 % et de 0,95 % pour les 17 MFCCs originaux, pour leurs 17 dérivées premières et pour leurs 17 dérivées deuxièmes.

On peut donc dire que la normalisation ne fonctionne pas très bien avec les dérivées en premier et en deuxième ordre parce que ces dérivées eux-mêmes ne contiennent pas beaucoup d'informations de l'émotion comme nos résultats dans la partie des paramètres globaux l'ont montré.

Nous remarquons également que $MFCC_0$ se montre toujours le meilleur paramètre.

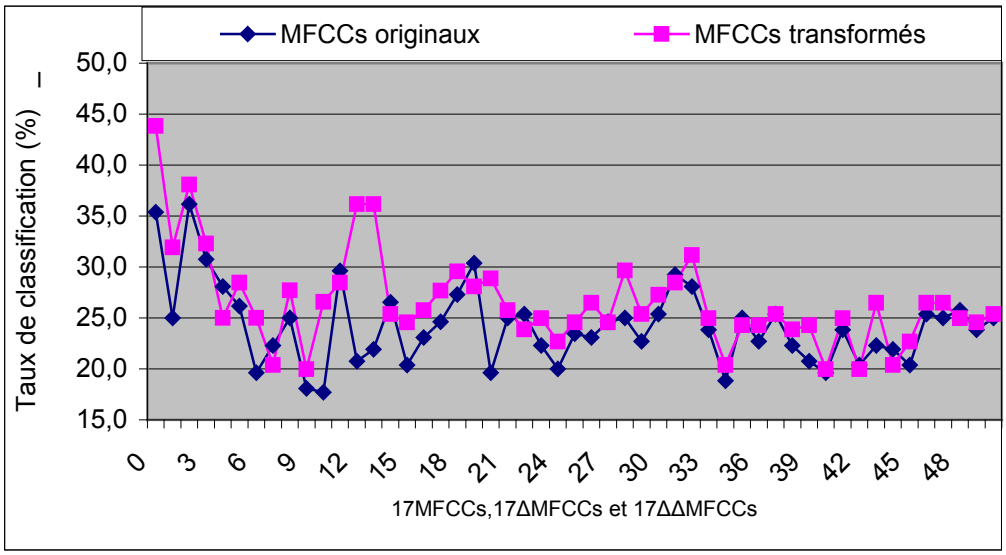

Figure 40 : Comparaison des performances de chaque paramètre MFCC avant et après la normalisation symbolique.

Comme nous avons mentionné dans la section précédente, la normalisation enlèvera la dépendance des paramètres du locuteur, et la transformation par cette normalisation uniformise la pondération des paramètres dans le vecteur de caractéristiques car tous les paramètres sont mesurés par un ensemble de symboles unitaires. Cela veut dire que, les paramètres qui ont une faible influence sur la performance avant la normalisation en raison de leur petite amplitude en termes des valeurs absolues, auront la même influence que les paramètres « forts ». Ce fait peut améliorer la performance du système si les paramètres renforcés sont efficaces (selon nos expérimentations, 17 MFCCs originaux se trouvent dans ce cas), sinon la performance sera dégradée (par exemple, le cas de 9 paramètres MFCCs sélectionnés pour la reconnaissance indépendante du locuteur, effectivement il y trop de paramètres dérivées dans cet ensemble, ceux qui sont prouvés par nos résultats moins efficaces que les paramètres originaux). Donc une étude du rôle des paramètres pour une pondération raisonnable serait intéressante dans le futur.

Pour notre étude, nous cherchons simplement une nouvelle combinaison pour montrer l'efficacité de cette méthode de normalisation dans la reconnaissance de l'émotion. C'est la raison pour laquelle nous utilisons simplement l'écart-type des paramètres avant la normalisation pour pondérer l'ensemble de paramètres après la normalisation. Les résultats de la recherche sont présentés dans le Tableau 57.

Type de paramètres	***Taux de classification*** *(%)*
9 MFCC_INDP : 9 paramètres sélectionnés par SFSA pour la reconnaissance indépendante du locuteur	52,7
9 MFCC_INDP normalisés sans pondération	51, 2
9 MFCC_INDP normalisés et pondérés	57,3

Tableau 57 : Comparaison de l'efficacité des trois ensembles de paramètres.

Comme nous avons expliqué, nous ne sommes pas surpris du fait que le taux de reconnaissance après la normalisation est inferieur à celui avant la normalisation car le rapport entre des paramètres ont été changé par la normalisation. Cependant, la restitution de ce rapport en utilisant la pondération (le troisième ensemble de paramètres) montre que nous allons dans le bon chemin et la normalisation prouve une amélioration significative pour la performance du système.

Pour la comparaison, nous avons aussi effectué la normalisation par la moyenne et par l'écart-type pour l'ensemble de 9 paramètres sélectionnés (9 MFCC_INDP), le taux de reconnaissance est de 49,2 %. Avec ce résultat, nous pouvons confirmer l'avantage de notre approche : la normalisation symbolique accompagnée par la sélection forcée séquentielle en avant.

Nous n'avons pas encore appliqué cette méthode de normalisation avec la sélection séquentielle pour les deux autres cas : la reconnaissance mono-locuteur et la reconnaissance multi-locuteur, mais théoriquement, on peut prévoir que cette méthode n'est pas utile pour le cas de reconnaissance mono-locuteur et elle est de plus ou moins utile pour l'amélioration de la performance dans le cas de la reconnaissance multi-locuteur.

Par nos expérimentations, nous n'avons pas trouvé d'autres combinaisons entre ces 9 MFCC_INDP avec des paramètres prosodiques ou des autres paramètres du domaine temporel qui donnent une meilleure performance. Nous proposons donc d'utiliser d'ici ces 9 paramètres $MFCC_2$, $MFCC_0$, $MFCC_3$, $MFCC_{11}$, $\Delta MFCC_2$, $\Delta MFCC_{12}$, $MFCC_7$, $\Delta MFCC_{14}$, $\Delta MFCC_{11}$ avec la normalisation symbolique et la pondération par l'écart-type comme l'ensemble de paramètres principal pour notre système de détection automatique des émotions.

5.7. Conclusion

Dans ce chapitre, nous avons analysé les caractéristiques statistiques des paramètres de l'émotion : les paramètres dans le domaine temporel et des paramètres du domaine fréquentiel.

Comme le corpus DES est un corpus construit dans un environement de simulation, les expressions émotionnelles de ce corpus ne sont pas spontanées (les émotions peuvent être exagérées). De plus, le corpus possède un nombre assez faible de locuteurs (2 locutrices et deux locuteurs) Nous voulons donc souligner avant de conclure ce chapitre que les conclusions ainsi

que les résultats que nous avons obtenus sur ce corpus DES pourraient ne pas se généraliser dans un contexte plus général, notamment à d'autres applications ou à d'autres types de données réelles où plusieurs autres problèmes peuvent se poser comme : la co-existence de plusieurs émotions, la transition entre des états émotionels et les cas où l'intensité de l'état émotionnel n'est pas élevée.

Nous ne nous sommes concentrés que sur les paramètres locaux (à court-terme) car selon nos expérimentations, en comparaison avec des paramètres globaux (à long-terme), ils donnent souvent les meilleurs résultats. En particulier, les paramètres locaux dépendent moins du contexte, ils fonctionnent donc mieux dans les deux cas de reconnaissance multi-locuteur et indépendante du locuteur.

Pour améliorer la performance de la reconnaissance, nous avons également proposé d'utiliser une méthode de normalisation symbolique qui se base sur la signification sémantique des régions de valeur des paramètres. Expérimentalement sur le corpus DES, cette méthode s'est montrée très efficace pour la reconnaissance de l'émotion en combinaison avec la sélection forcée séquentielle en avant.

Les combinaisons d'autres caractéristiques restantes avec ces 9 MFCCs normalisés sélectionnés par SFSA ont aussi été étudiées ; cependant le fait que ces combinaisons avec les 9 MFCCs ne donnent pas les meilleurs résultats dans le cas de reconnaissance indépendante du locuteur ne veut pas dire que la combinaison est inutile ou que les autres paramètres sont inefficaces, nous croyons que la raison de l'échec de la combinaison se situe au niveau de la pondération des paramètres après la normalisation. Dans le futur, l'étude sur la pondération est aussi un travail qui nous intéresse.

Et enfin, ce taux de 57,3 % de classification correcte obtenu par notre approche est comparable avec les autres résultats de classification correcte de 53 % obtenu par [Huang et al, 2006], de 42,3 % obtenu par [Kwon et al, 2003] et de 51,2 % obtenu par [Fernandez et al, 2003] aussi sur les cinq états émotionnels.

Chapitre 6. **Etudes des modèles**

Dans le chapitre précédent, nous avons analysé et précisément étudié l'optimalisation de la phase d'extraction des paramètres de notre système de détection de l'émotion sur les énoncés audio sur un corpus en Danois. Nous avons utilisé par défaut le modèle de mélange de 16 gaussiennes pour tester la performance des paramètres. Dans ce chapitre, dans le contexte du même corpus, nous étudierons quelle est la technique de classification la plus efficace pour la détection de l'état émotionnel des locuteurs, avec quel modèle et avec quelle configuration. Pour notre étude, nous distinguons dans ce chapitre les trois techniques de classification pour les trois grandes sections. La première technique vise à modéliser la distribution des paramètres et parmi les modèles de notre étude, nous testerons le modèle de quantification vectorielle (VQ) et le modèle de mélange de gaussiennes (GMM). La deuxième technique essaie de séparer explicitement l'espace des paramètres pour trouver les frontières des classes de l'émotion, comme par exemple le modèle SVM. Enfin, la troisième technique capable est de capturer des informations de l'évolution temporelle des vecteurs de caractéristiques pour renforcer la classification dont le modèle de Markov caché est le plus représentatif. Dans ce chapitre, les modèles sont aussi étudiés et sélectionnés dans trois cas de reconnaissance : la reconnaissance mono-locuteur, la reconnaissance multi-locuteur et la reconnaissance indépendante du locuteur. Durant l'expérimentation avec les modèles, nous effectuons également des tests de l'efficacité de la normalisation symbolique qui a donné de bons resultats sur le corpus DES.

6.1. Introduction

Par définition : « la reconnaissance est une action par laquelle on retrouve dans sa mémoire l'idée, l'image d'une chose ou d'une personne quand on vient à la revoir »[7]. Cela veut dire que, tous les processus de reconnaissance exigent un pré-processus pour collectionner, synthétiser et

[7] Définition de l'Académie française (8e édition 1932-1935)

mémoriser des connaissances de l'objet ou du phénomène que l'on doit reconnaître. Ce pré-processus de collection pour l'ordinateur est appelé le processus de l'apprentissage.

Il y en a différentes stratégies pour l'apprentissage : l'apprentissage non-supervisé et l'apprentissage supervisé. L'apprentissage non-supervisé est la stratégie qu'on emploie lorsqu'on ne connaît pas de sortie a priori pour le problème considéré. Le but de ce type d'apprentissage est de faire en sorte que le modèle identifie lui-même les motifs liés aux vecteurs d'entrée. Inversement, l'apprentissage supervisé (ou l'apprentissage à partir d'exemples) est adapté à tous les problèmes pour lesquels on peut déjà associer des sorties à certains vecteurs d'entrée. Donc, dans les cas où les vecteurs d'entrée sont connus, l'apprentissage supervisé fonctionne théoriquement mieux que l'apprentissage non-supervisé et c'est aussi la raison pour laquelle nous nous avons choisi la stratégie de apprentissage supervisé comme stratégie pour notre étude.

Apres l'apprentissage, simplement, la reconnaissance est la comparaison entre les modèles entrainés avec un échantillon d'entrée inconnu afin de pouvoir décider à quel modèle (ou à quelle classe, nous supposons qu'une classe est modélisée par un modèle) cet échantillon appartient. Généralement, ou dans notre cas en particulier, l'état émotionnel du locuteur d'un énoncé doit être la classe qui est la plus proche de cet énoncé. Chacun de modèles a sa propre unité de « distance », qui sont la distance (par exemple euclidienne) ou bien des valeurs de probabilité. Dans les parties suivantes, nous étudions brièvement chaque algorithme.

6.2. Modèle de quantification vectorielle

La quantification vectorielle (VQ en anglais) est une technique simple mais efficace et a été proposée par [Gray 1984]. Cette technique permet de compresser l'espace de données en représentant ces données par un certain nombre de composants représentatifs appelés centroïdes, L'essence de la représentation n'est rien qu'une approximation vers des vecteurs centraux (des centroïdes) qui est de même nature avec l'idée de l'arrondissement d'un nombre réel en entier.

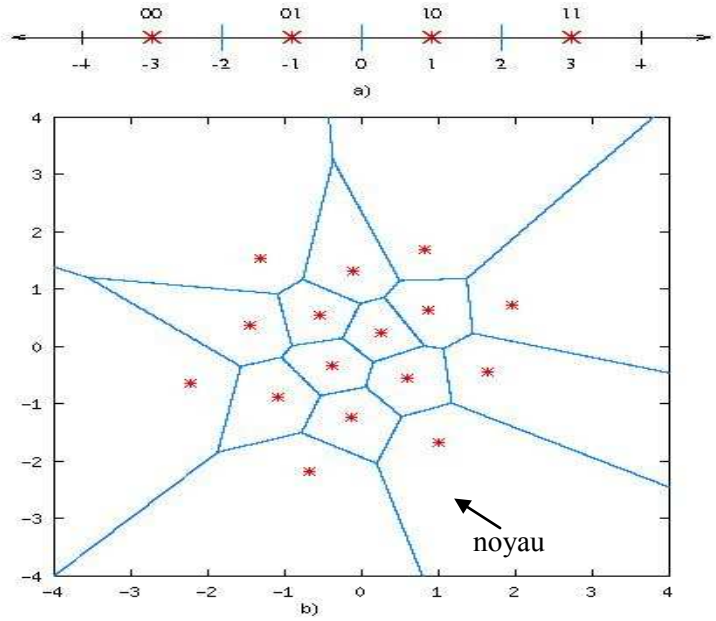

Figure 41 : Exemples de la quantification vectorielle dans l'espace 1 dimension (a) et l'espace 2 dimensions (b).

La Figure 41 montre les deux exemples de la quantification dans l'espace à une dimension et l'espace à deux dimensions. Dans le premier exemple avec l'espace à une dimension (sur la figure a), au lieu de stocker tous les valeurs de –4 à 4, on n'utilise que les 4 centroïdes étiquetés par 00, 01, 10, 11. Similairement, dans le deuxième exemple les vecteurs sont regroupés en régions où le noyau de chaque région sera le représentant pour toute la région.

Etant donnée une classe de l'émotion avec l'ensemble de N vecteurs de paramètres $X = \{x_i\}$ $1 \le i \le N$, le processus d'apprentissage de ce modèle est le processus de recherche des K centroïdes ($K << N$) de sorte que ces K centroïdes représentent le mieux l'ensemble des N vecteurs originaux. L'algorithme K-moyennes est une des techniques permettant la quantification vectorielle. La description plus précise de cet algorithme de quantification vectorielle peut se trouver dans [Gray 1984] où [Lien 1998].

Algorithme des *K*-moyennes L'algorithme des K-moyennes cherche à regrouper l'ensemble des vecteurs x_j d'apprentissage d'une classe en K sous-ensembles disjoints [Rabiner et al, 1993]. Cet ensemble de K sous-ensembles sera appelé le dictionnaire. Chaque sous-ensemble est caractérisé par son centroïde. Cet algorithme n'est que localement optimal, il est donc influencé par ses conditions initiales. La variante LBG (Linde-Buzo-Gray) [Linde et al, 1980] de cet algorithme comporte 4 étapes et elle sera décrite ensuite. L'optimisation qui est cherchée par l'algorithme, consiste en la réduction de la distance euclidienne entre chaque élément d'un sous-ensemble et le centroïde du sous-ensemble. Le nombre de sous-ensembles est la taille K du dictionnaire et ce sera une puissance entière de 2 (K=2p). Les étapes de cet algorithme sont les suivantes:

1. **Initialisation.** Le dictionnaire est constitué d'un seul sous-ensemble contenant tous les vecteurs d'apprentissage, son centroïde est la moyenne des vecteurs d'apprentissage. A cette étape transitoire, $k = 1$ (ce n'est pas encore une puissance de 2).

2. **Éclatement du dictionnaire.** Chaque sous-ensemble du dictionnaire va être éclaté en remplaçant chaque centroïde de coordonnées y_i par 2 nouveaux centroïdes de coordonnées respectives $y_i(1 + \varepsilon)$ et $y_i(1 - \varepsilon)$, avec $\varepsilon << 1$. La valeur de k est doublée par rapport à la valeur précédente.

3. **Optimisation du dictionnaire.** Chaque vecteur x_j sera examiné à tour de rôle. Dans une première étape la distance euclidienne séparant x_j et chacun des centroïdes est calculée ; le vecteur x_j étant alors affecté au sous-ensemble pour lequel cette distance est la plus faible. Lorsque chaque vecteur a été affecté à un sous-ensemble, la moyenne des vecteurs de chaque sous-ensemble est recalculée pour obtenir le centroïde correspondant. Cette étape doit être itérée plusieurs fois avant de passer à l'étape 4.

4. **Test d'arrêt.** Tant que $k < 2p$, le dictionnaire est à nouveau éclaté et optimisé en réitérant les étapes 2 et 3. Sinon, le dictionnaire a atteint la taille désirée et pourra alors s'arrêter.

Par la simplicité et grâce à la capacité de généraliser l'espace de données, nous trouvons que cette technique est bien adaptée pour résoudre notre problème de reconnaissance de l'émotion en travaillant avec l'espace de paramètres extraits à partir de la parole. Effectivement, chaque état émotionnel pourrait posséder un ensemble de centroïdes spécifiques, et par rapport aux autres ensembles de centroïdes, cet ensemble de centroïdes serait plus proche des vecteurs de paramètres des échantillons d'entrée qui sont dans le même état émotionnel. En perspectives, nous pensons aussi à utiliser cette distance comme une unité pour mesurer le dégrée d'émotion de l'échantillon.

6.2.1. Modèle de mélange de gaussiennes GMM

La reconnaissance de l'émotion à l'aide d'un modèle GMM comprend 2 phases : une phase d'apprentissage du système sur un ensemble de fichiers supposés représentatifs d'une classe et, une deuxième phase, de vérification de l'appartenance de l'émotion quelconque à cette classe.

Pendant la phase d'apprentissage, la modélisation statistique des paramètres acoustiques est effectuée. La répartition des paramètres acoustiques d'une classe dans l'espace est modélisée par une somme de fonctions de densités de probabilités *(pdf*), dans ce cas, ce sont des fonctions *pdf* gaussiennes.

Avec l'approche des paramètres locaux, à chaque instant d'échantillonnage (par exemple toutes les 10 ms), le système évalue les D paramètres acoustiques correspondants au signal de la parole. L'ensemble de ces D paramètres constitue le vecteur de caractéristiques.

Pour estimer le rapport de vraisemblance d'un vecteur de caractéristiques, dans ce modèle, la base de distributions multi-gaussiennes a été utilisée, c'est-à-dire que la distribution d'observations appartenant à une même classe d'émotion est modélisée par une somme pondérée de M distributions gaussiennes. Cela revient à considérer que les vecteurs observés, x_i, sont des réalisations de variables aléatoires mutuellement indépendantes dont la densité de probabilité $f_m(x_i)$ est de type gaussienne. La formule (5.11) donne la modélisation de la distribution d'observations $f(x_i)$ par une somme pondérée par les coefficients πm des distributions gaussiennes $f_m(x_i)$.

$$f(x_i) = \sum_{m=1}^{M} \pi_m . f_m(x_i) \qquad (5.11)$$

$$\text{où} \quad \pi_m \geq 0,\ \pi_m \geq 0, \forall m \in [1, M] \quad \text{et} \quad \sum_{m=1}^{M} \pi_m = 1$$

La densité de probabilité gaussienne sous la forme de l'équation (3.2) qui est réécrite comme suivante :

$$f_m(x_i) = \frac{e^{\left[-\frac{1}{2}(\vec{x}-\vec{\mu}_m)^T \Sigma_m^{-1}(\vec{x}-\vec{\mu}_m)\right]}}{(2\pi)^{D/2} \mid \det(\Sigma_m) \mid^{1/2}} \qquad (5.12)$$

où :

- $\vec{x}$ est le vecteur de caractéristiques à modéliser
- μ_m est le vecteur moyen des vecteurs $\vec{x}$
- Σ_m est la matrice de covariance des vecteurs $\vec{x}$
- D est le nombre de caractéristiques et aussi la dimension des vecteurs $\vec{x}$

Lors de la phase d'apprentissage, tous les vecteurs $\vec{x}$ d'une même classe de l'émotion sont utilisés pour déterminer le poids correspondant à chacune des N gaussiennes, le vecteur moyen μ_m et la matrice de covariance Σ_m de chacune des gaussiennes. Le vecteur moyen μ_m et la

matrice de covariance Σ_m se réduisent respectivement à la moyenne et à l'écart type dans le cas d'une distribution gaussienne mono-dimensionnelle (m=1).

Chacune des gaussiennes ($1 \leq m \leq M$) d'une classe Ωc est caractérisée par :

- la dimension D des vecteurs de paramètres $\vec{x}_c$
- les poids de chaque gaussienne $\pi_{c,m}$ qui respecte la condition : $\sum_{m=1}^{M} \pi_m = 1$
- les vecteurs moyens : $\mu_{c,m}$
- les matrices de covariance $\Sigma_{c,m}$

6.2.1.1 Apprentissage

L'apprentissage a pour but d'estimer les paramètres des gaussiennes qui composent le modèle à partir des vecteurs de paramètres de chaque classe de l'émotion. Le processus de l'apprentissage se décompose en deux étapes successives :

- approximation des paramètres des gaussiennes de la classe par l'algorithme des *K*-moyennes (ou «Kmeans ») (voir le modèle de quantification vectorielle),
- optimisation des valeurs de ces paramètres par un algorithme de type EM (Expectation Maximisation).

Algorithme EM L'algorithme EM fait intervenir des variables latentes que l'on ne peut observer directement. Dans ce cas, chaque vecteur x_j est décrit non seulement par les D paramètres acoustiques mais aussi par le sous-ensemble S_i (défini par un centroïde) auquel il se rattache. Nous avons noté que dans le cas de l'algorithme LBG, l'hypothèse avait été faite que chaque x se rattachait réellement à un sous-ensemble. Dans le cas de l'algorithme EM ce ne sera plus le cas. Celui-ci va maximiser la vraisemblance de façon itérative, mais le vecteur x sera maintenant rattaché aux M sous-ensembles Si avec une probabilité particulière, sans que l'on puisse déterminer à quel sous-ensemble Si il appartient réellement. C'est ce paramètre que l'on qualifie de donnée cachée ou latente, voir [Boite et al., 2000].

L'idée de base de l'algorithme EM consiste à raisonner sur les données observées et latentes, tout en prenant en compte le fait que l'information disponible sur les données latentes provient des données observées. A chaque étape k de l'algorithme, le calcul de la variable latente, pour chaque x_i et chaque gaussienne $\Theta_{k,m}$ des coefficients est effectué conformément à l'équation (5.13).

$$\gamma_i^{(k)}(m) = P(S_i = m \mid x_i; \Theta_{k,m}) = \frac{\pi_m . \left|\Sigma_m\right|^{-1/2} . e^{[-\frac{1}{2}(x_i-\mu_m)^T . \Sigma_m^{-1} . (x_i-\mu_m)]}}{\sum_{j=1}^{M} \pi_j . \left|\Sigma_j\right|^{-1/2} . e^{[-\frac{1}{2}(x_i-\mu_m)^T . \Sigma_m^{-1} . (x_i-\mu_m)]}} \quad (5.13)$$

Ce calcule utilise les paramètres $\Theta_{k,m}$ des gaussiennes déterminées à l'étape précédente ($k-1$) et permet le calcul de la quantité intermédiaire $\Psi_{\Theta k}(\Theta)$ que l'on doit maximiser avec l'équation (5.14).

$$\Psi_{\Theta_k}(\Theta) = \sum_{t=1}^{T}\sum_{i=1}^{M} \log(\pi_i f_i(x_t)).\gamma_t^{(k)}(i) \tag{5.14}$$

La ré-estimation des paramètres $\Theta_{k+1,m} = (\pi_m^{(k+1)}, \mu_m^{(k+1)}, \Sigma_m^{(k+1)})$, à partir des paramètres $\Theta_{k,m}$ constitue la deuxième étape de l'algorithme EM. La maximisation de (5.14) par rapport aux paramètres π_m, μ_m et Σ_m fournit les nouvelles valeurs estimées des paramètres [Cappé, 2000] pour l'itération (k+1). Les formules de calcul des paramètres à l'itération (k+1) sont données en (5.15).

$$\begin{cases} \pi_m^{(k+1)} = \dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\gamma_t^{(k)}(m) \\ \mu_m^{(k+1)} = \dfrac{\displaystyle\sum_{t=1}^{T}\gamma_t^{(k)}(m).x_t}{\displaystyle\sum_{t=1}^{T}\gamma_t^{(k)}(m)} \\ \Sigma_m^{(k+1)} = \dfrac{\displaystyle\sum_{t=1}^{T}\gamma_t^{(k)}(m).(x_t-\mu_m^{(k+1)}).(x_t-\mu_m^{(k+1)})^T}{\displaystyle\sum_{t=1}^{T}\gamma_t^{(k)}(m)} \end{cases} \tag{5.15}$$

6.2.1.2 Classification

Pendant la phase de classification, on doit déterminer la classe Ω_{c*} la plus probable à partir du calcul de la vraisemblance [Boite et al., 2000], pour le vecteur x obtenu à l'instant t, et pour chacune des classes de l'émotion Ω_c ($1 \le c \le C$) :

$$p(x \mid \Omega_c) = \sum_{m=1}^{M} \pi_{c,m} . \frac{1}{(2\pi)^{\frac{D}{2}} \left|\Sigma_{c,m}\right|^{\frac{1}{2}}} . e^{\left[-\frac{1}{2}(x-\mu_{c,m})^T \Sigma_{c,m}^{-1}(x-\mu_{c,m})\right]} \tag{5.16}$$

En pratique, il est nécessaire de déterminer la vraisemblance d'un son constitué d'une suite temporelle de N vecteurs $X = x_i$ $1 \le i \le N$. Elle peut être obtenue à partir de la vraisemblance de chacun des vecteurs x_i comme dans l'équation (5.17)

$$p(X \mid \Omega_c) = \prod_{i=1}^{N} p(x_i \mid \Omega_c) \tag{5.17}$$

Un signal à tester est transformé dans une suite de N vecteurs acoustiques X qui ont D paramètres. Il appartiendra avec le maximum de vraisemblance à la classe Ω_{c*} pour laquelle $p(X|\Omega_{c*})$ est maximale, conformément à l'équation (5.18).

$$p(X \mid \Omega_{c*}) = \max_{c=1}^{C}(p(X \mid \Omega_c)) \tag{5.18}$$

6.2.2. Modèle de machine à vecteur de support

Différemment des modèles VQ et GMM qui cherchent à modéliser l'espace des paramètres, la machine à vecteurs de support (SVM) est un classifieur à vecteurs de support qui sépare deux classes avec une marge maximale. La marge χ est définie par la distance du point le plus proche à l'hyperplan ou la frontière de décision. Les vecteurs de caractéristiques qui sont aux frontières des marges sont appelés des vecteurs de support. Le classifieur à vecteurs de support a été à l'origine conçu pour le problème de la classification à deux classes, mais il peut être étendu pour plusieurs classes.

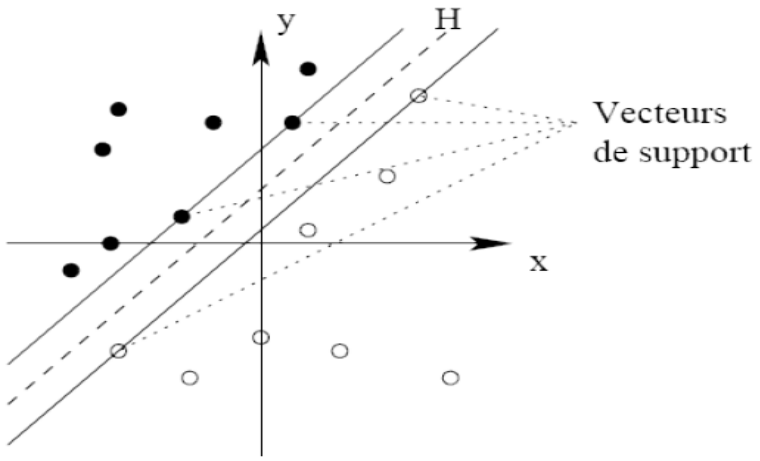

Figure 42 : Le modèle SVM

Considérons un ensemble d'échantillons d'apprentissage dénoté par $\{e_i\}_{i=1}^{N} = \{(y_i, l_i)\}_{i=1}^{N}$ où l_i=$\{-1,+1\}$ est la classe correspondante de chaque échantillon. Le classificateur est un hyperplan :

$$g(y) = \omega^T y + b \tag{6.1}$$

avec ω est le vecteur de gradient qui est perpendiculaire à l'hyperplan et b est le déplacement de l'hyperplan de l'origine. On peut démontrer que la marge est inversement proportionnelle à $\|\omega\|^2 / 2$. La valeur de $l_i g(y_i)$ peut être utilisée pour indiquer à quelle côté de l'hyperplan l'échantillon appartient, $l_i g(y_i)$ doit être plus grand que 1 si l_i = +1 et plus petit que –1 si l_i=–1. Ainsi, le choix de l'hyperplan dans le cas séparable peut être reformulé comme le problème de l'optimisation suivant :

minimiser $\frac{1}{2}\omega^T\omega$

avec

$$l_i(\omega^T y + b) \geq 1, i = 1, 2, ..., N \qquad (6.2)$$

L'optimum global pour les paramètres ω, b est trouvé en employant des multiplicateurs de Lagrange.

L'avantage de cette technique de classification est qu'elle peut être étendue à la séparation non-linéaire par la technique de la fonction noyau ; en plus, elle peut rester robuste sous haute dimensions de l'espace de données ; le détail de la raison de ce fait est complexe et il est donc hors de travail de cette thèse, cependant c'est essentiellement parce que la complexité de l'espace d'hypothèse ne se mesure pas en termes de nombre de dimensions, mais en termes de marges utilisées pour séparer l'hyperplan à partir des vecteurs de support. Comme le modèle de quantification vectorielle, la sortie du modèle SVM est aussi une valeur réelle représentant la distance positive ou négative avec l'hyperplan, cette valeur peut être utilisée avec profit pour représenter l'état émotionnel reconnu par rapport aux autres émotions et aussi pour conclure de quel dégrée est cet état émotionnel.

Avec tant d'avantages, SVM est aussi un modèle intéressant pour notre étude.

Comme la pratique connexionniste de l'apprentissage l'a prouvé, le passage des valeurs de données en échelle fixe (normalement entre -1 et 1 et entre 0 et 1) peut considérablement améliorer à la fois la précision et l'efficacité de SVM. Effectivement, la précision est affectée si certains paramètres possèdent une grande dynamique alors que les autres ont des petites dynamiques, cette différence est susceptible d'avoir différentes pondérations sur ces paramètres et d'avoir potentiellement différents effets dans l'apprentissage et donc pour la performance. De même, le passage en échelle fixe diminue également la taille du noyau de calculs parce que les arguments seront plus petits. Dans cette thèse, en utilisant le modèle SVM, tous les paramètres sont mis à l'échelle de valeurs entre 0 et 1.

Un problème de l'utilisation SVM est de sélectionner la fonction noyau (*kernel* en anglais). Nous expérimentons la fonction noyau gaussienne ou la fonction *rbf* (*radial-basis function*) qui est formulée par l'équation (6.3) pour les raisons suivantes :

- *rbf* est non-linéaire, donc elle nous permet de séparer l'espace de paramètres étudiés dans cette thèse ce qui est impossible à séparer linéairement ;
- cette fonction noyau possède un seul paramètre γ alors que les fonctions polynomiales ou sigmoïdes en ont 2, donc l'optimisation sur cette fonction noyau n'exige que l'étude de γ.

$$K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|}, \gamma \geq 0 \qquad (6.3)$$

où x_i et x_j sont des exemples et γ est un paramètre précisé par l'utilisateur.

En conclusion, en termes de paramètres du modèle, en travaillant avec SVM, nous chercherons à optimiser le couple de paramètres γ de la fonction noyau et C de coût du modèle SVM. Théoriquement, dans un intervalle de (C, γ) la performance du modèle SVM est distribuée comme des contours dans la Figure 43 [Hsu et al, 2003]. La zone la plus performante est entourée par le contour a, la zone moins performante est la zone entourée par le contour a et le contour b, et ainsi de suite.

Pour pouvoir dessiner cette distribution mais en même temps pour gagner du temps de calcul dans la recherche de ces deux paramètres, nous avons suivi une stratégie de recherche de propagation dans un certain intervalle de deux paramètres (C, γ) qui est montrée dans la Figure 43. Les résultats sont obtenus par la validation croisée par la division en 10 plis :

- étape 1 : notre processus de recherche commence par la flèche marquée 1 où nous testons verticalement toutes les possibilités des paires (C, γ) en gardant la valeur de C et en changeant la valeur de γ par des petits pas, nous obtiendrons à la fin les deux meilleurs résultats ;
- étape k : à partir des positions des deux meilleurs résultats obtenus dans l'étape k-1, nous lançons l'autre recherche horizontale (si celle de l'étape précédente est verticale et vice versa si celle de l'étape précédente est horizontale) pour toutes les possibilités des paires (C, γ) en gardant la valeur de γ et en changeant la valeur de C par des petits pas. Si un de ces deux résultats obtenus est meilleur que celui de l'étape k-1, on passe à l'étape k+1 et continue la recherche du meilleur résultat. Si aucun résultat obtenu est meilleur que celui de l'étape k-1, on continue la recherche avec les deux nouvelles positions des deux meilleurs nouveaux résultats qui viennent d'être obtenus.
- fin : l'algorithme se termine si on peut entourer la zone la plus performante où toutes les paires sont testées.

Dans cette thèse, $C = \{2^{-5},2^{-4}, \ldots,2^{15}\}$ et $\gamma = \{2^{-15},2^{-14}, \ldots, 2^{0}\}$ sont les valeurs de grille de notre recherche.

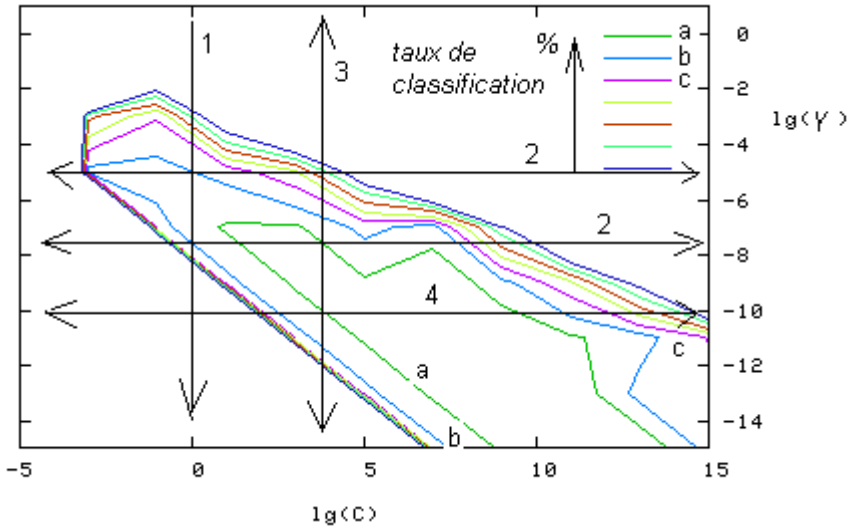

Figure 43 : Recherche des valeurs optimales (C, γ) du modèle SVM [Hsu et al, 2003]

Comme le modèle SVM est naturellement un classifieur binaire alors que nous travaillons avec cinq états émotionnels du corpus DES et 8 états émotionnels du corpus BES, une méthode permettant la classification multi-classes est exigée et nous avons choisi l'approche un face à un pour modéliser tous mes états émotionnels ; cela veut dire que (supposons que nous avons K émotions) nous aurons K.(K-1)/2 modèles correspondants avec tous les paires possibles d'émotions.

Pour classifier un énoncé d'entrée, son état émotionnel est ce qui correspond à la sortie la plus grande de ces $K.(K\text{-}1)/2$ modèles.

6.2.3. Modèle de Markov caché

Similairement au modèle de mélange de gaussiennes, le modèle de Markov caché est encore renforcé par sa capacité de capture des informations variant en fonction de temps grâce aux états de transition.

Effectivement, on considère un système de Markov qui est décrit par un ensemble de N états distincts S_1, S_2,..., S_N. Par exemple, le système avec $N = 5$ est illustré par la figure suivante [Rabiner, 1989] :

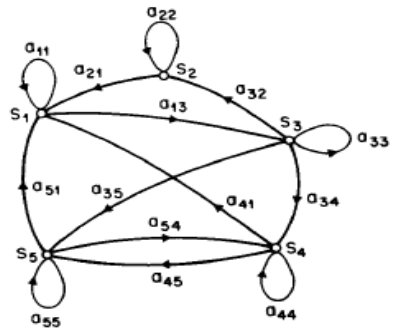

Figure 44 : Une chaîne de Markov de 5 états [Rabiner, 1989]

A chaque moment, le système peut changer d'état selon un ensemble de probabilités. Pour un HMM d'ordre 1, chaque probabilité ne dépend de que l'état en cour q_t et l'état précédant q_{t-1} :

$$P\,[q_t = S_j \mid q_{t-1} = S_i,\, q_{t-2} = S_k \ldots] = P\,[q_t = S_j \mid q_{t-1} = S_i]. \qquad (6.4)$$

Si on dénote a_{ij} la probabilité de transition d'état, telle que :

$$a_{ij} = P[q_t = S_j \mid q_{t-1} = S_i] \qquad \text{avec } 1 \le i, j \le N \qquad (6.5)$$

$$a_{ij} \ge 0$$

On peut appeler ce processus un Modèle de Markov observable dont chaque état du processus correspond à un élément (observable) physique.

Ici, on considère les modèles de Markov dont chaque état correspond à un événement observable (physique). Ceci est très restrictif si l'on veut modéliser des problèmes plus complexes. Le concept de modèle de Markov est donc étendu pour que l'observation soit une fonction de probabilité d'état. Le modèle est un processus aléatoire qui n'est pas observable (caché) mais qui peut être étudié à travers un autre ensemble de processus aléatoires qui produisent une série d'observations.

Un modèle Markov Caché peut être caractérisé par les éléments suivants :

- N, le nombre d'état du système : $S = \{S_1, S_2,\ldots,S_N\}$, et qt comme l'état du système au temps t.
- M, le nombre de symboles d'observations distingués par état. Les symboles d'observations correspondent à chaque sortie physique du système réel qu'on veut

modéliser. On peut noter ici l'ensemble des symboles d'observations du modèle : $V = \{V_1, V_2, ..., V_M\}$.

La matrice de probabilité de transition d'état $A = \{a_{ij}\}$ telle que :

$$a_{ij} = P\,[q_{t+1} = S_j \mid q_t = S_i] \qquad \text{avec } 1 \le i, j \le N \text{ et } a_{ij} \ge 0.$$

La distribution à l'état j de probabilité des symboles d'observation $B = \{b_j(k)\}$ dont :

$$b_j(k) = P\,[V_k \text{ à } t \mid q_t = S_i], \qquad \text{avec } 1 \le j \le N \text{ et } 1 \le k \le M.$$

La distribution d'état initial $\pi = \{\pi_i\}$ dont :

$$\pi_i = P[q_1 = S_i], \qquad \text{avec } 1 \le i \le N.$$

Etant donné les valeurs de N, M, A, B et π_i le modèle de Markov Caché peut générer la série d'observations $O = O_1, O_2,..., O_T$ avec $O_t \in V$ et T le nombre d'observations.

On appelle $\lambda = (A, B, \pi)$ les paramètres complets du modèle HMM.

Les trois problèmes principaux à résoudre pour un modèle de Markov caché sont : l'évaluation, le décodage et l'apprentissage d'un HMM.

- Problème 1 : le problème de l'évaluation : étant donné une série d'observation $O = O_1, O_2,..., O_T$ et le modèle $\lambda = (A, B, \pi)$, on doit calculer $P(O|\lambda)$, la probabilité de la série d'observation suivant le modèle donné. La solution à ce problème est donnée par l'algorithme Forward-Backward.
- Problème 2 : le problème du décodage : étant donné une série d'observation $O = O_1, O_2,..., O_T$ et le modèle $\lambda = (A, B, \pi)$, comment peut-on choisir les états optimaux q = q1, q2,..., qT correspondant à la série d'observation. Ce problème peut être résolu en utilisant l'algorithme de Viterbi.
- Problème 3 : le problème de l'apprentissage : Comment peut-on déterminer les paramètres $\lambda = (A, B, \pi)$, étant données des observations O, en optimisant $P(O|\lambda)$. La solution est l'algorithme de Baum-Welch et l'algorithme EM (Expectation Maximization en anglais)

Tous ces algorithmes : Forward-Backward, Viterbi et Baum-Welch ou EM peuvent être consultés dans [Rabiner 1989] [Rabiner 1993]

Une des éléments qui influence également sur la qualité de reconnaissance quand on utilise le modèle de Markov Caché est leur structure. En effet, dans la plupart des applications sur la reconnaissance de la parole, en raison de la continuité de la parole, on utilise souvent le modèle Gauche-Droite, celui dont la matrice de transition A possède des contraints suivantes :

$$a_{ij} = 0 \quad \text{avec} \quad j < i \tag{3.45}$$

Cela veut dire que dans le modèle Gauche-Droite, les transitions de l'état dont l'indice est inférieur à celui de l'état actuel sont interdites. Si n'importe quel état peut être joint (par un seul pas) à partir de n'importe quel autre état, on l'appelle le modèle ergodique (voir Figure 45)

La condition $a_{ij} = 0 \quad j > i + \Delta_i$ avec $\Delta_i = 1$ ou $\Delta_i = 2$ est ajoutée afin d'empêcher les changements brutaux. La matrice (6.6) montre le cas où $\Delta_i = 2$.

$$\pi_i = \begin{cases} 0, & i \neq 1 \\ 1, & i = 1 \end{cases} \tag{3.47}$$

Cette expression signifie que la séquence des états doit commencer à partir du 1er état (et se terminer au $N^{ème}$ état).

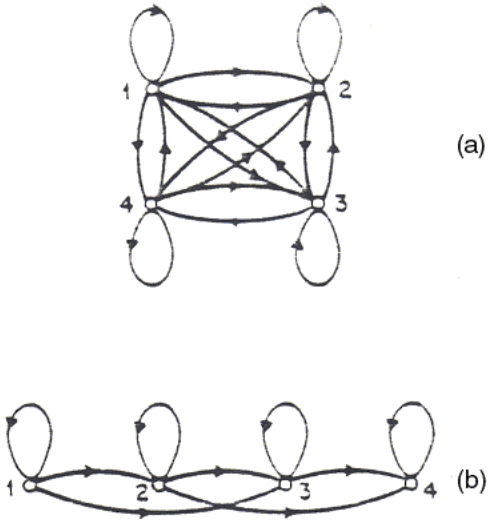

Figure 45 : a) Modèle de 4 états ergodiques b) Modèle de 4 états Gauche-Droite [Rabiner 1989]

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{bmatrix} \tag{6.6}$$

outre ces deux types de modèle, il y a encore beaucoup d'autres possibilités et d'autres combinaisons. En effet, le nombre de combinaisons différentes peut atteindre $2N$ où N est le nombre d'états.

Donc, quelle est la configuration la plus efficace ? Est-ce que le modèle Gauche-Droite, celui qui est largement utilisé dans la reconnaissance de la parole, est aussi le plus efficace pour la reconnaissance des états émotionnels ? Notre partie d'expérimentation suivante donnera ces réponses.

6.3. Expériences avec les modèles

6.3.1. Reconnaissance mono-locuteur

Bien que notre but ultime soit d'étudier le potentiel de la reconnaissance indépendante de l'émotion du locuteur, il est toujours instructif d'étudier de la capacité de modélisation et de reconnaissance de l'émotion dans les deux cas de reconnaissance mono et multi locuteur.

Dans le chapitre précédent, nous avons montré que la combinaison de douze paramètres MFCCs : $MFCC_0$, $MFCC_2$, $MFCC_{11}$, $MFCC_9$, $\Delta MFCC_4$, $\Delta MFCC_{15}$, $MFCC_{12}$, $\Delta\Delta MFCC_1$, $\Delta\Delta MFCC_2$, $MFCC_5$, $MFCC_6$, $MFCC_{16}$ donne la meilleur performance pour la classification les 5 états émotionnels dans le cas de reconnaissance mono-locuteur (78,7 %), la combinaison de 7 paramètres MFCCs avec ΔF_0Rel : $MFCC_0$, $MFCC_2$, $MFCC_{11}$, $\Delta\Delta MFCC_0$, $\Delta MFCC_0$, $MFCC_{12}$, $MFCC_5$ donne le meilleur résultat pour le cas de reconnaissance multi-locuteur (69,6 %) et l'ensemble de 9 paramètres MFCCs : $MFCC_2$, $MFCC_0$, $MFCC_3$, $MFCC_{11}$, $\Delta MFCC_2$, $\Delta MFCC_{12}$, $MFCC_7$, $\Delta MFCC_{14}$, $\Delta MFCC_{11}$ nous donne le meilleur résultat dans le cas de reconnaissance indépendante du locuteur. La question posée est : est-ce que le nombre de 16 gaussiennes est suffisant pour modéliser la distribution des paramètres pour chaque cas ? Nous voudrions rappeler que nous nous sommes basés sur les résultats obtenus avec 36 paramètres de MFCCs (12MFCCs+12ΔMFCCs+12ΔΔMFCCs) pour choisir ce modèle de validation de 16 gaussiennes et parce que cet ensemble de 36 paramètres n'est pas l'ensemble le plus efficace selon nos résultats, nous devons vérifier les autres configurations de ce modèle GMM. En même temps, les autres modèles comme le modèle de quantification vectorielle, le modèle de machine à vecteurs de support, le modèle Markov caché seront aussi étudiés pour estimer la performance de ces modèles vers la modélisation des émotions. Les ensembles de paramètres correspondants listés ci-dessus seront utilisés lors de ces études.

6.3.1.1 Modèle à mélange de gaussiennes (GMM)

Avec le modèle GMM, le taux de reconnaissance en changeant le nombre de gaussiennes est montré dans le Tableau 58 pour le cas de reconnaissance dépendante du locuteur. 87,7 % est le meilleur résultat que nous pouvons obtenir avec une mélange de 64 gaussiennes. La matrice de confusion entre des émotions est montrée dans le Tableau 59

N. de gaussiennes	***Taux de reconnaissance*** *(%)*
1	81,1
2	75,7
4	81,2
8	81,2
16	78,7
32	83,1
64	**87,7**
128	85,2
256	86,6
512	86,1
1024	87,3

Tableau 58 : Recherche la configuration optimale du modèle GMM en fonction du nombre de gaussiennes pour les 12 MFCCs sélectionnés

	colère (%)	***joie (%)***	***neutre (%)***	***tristesse (%)***	***surprise (%)***
colère	88,5	5,8	0,5	0,0	4,8
joie	6,3	82,7	0,5	0,0	10,6
neutre	6,3	1,9	87,0	4,3	0,0
tristesse	1,4	2,4	3,4	92,8	0,0
surprise	2,9	9,1	0,0	0,0	88,0
moyenne	**87,7**				

Tableau 59 : Matrice de confusion du meilleur cas du modèle de mélange de 64 gaussiennes dans le cas de reconnaissance dépendante du locuteur

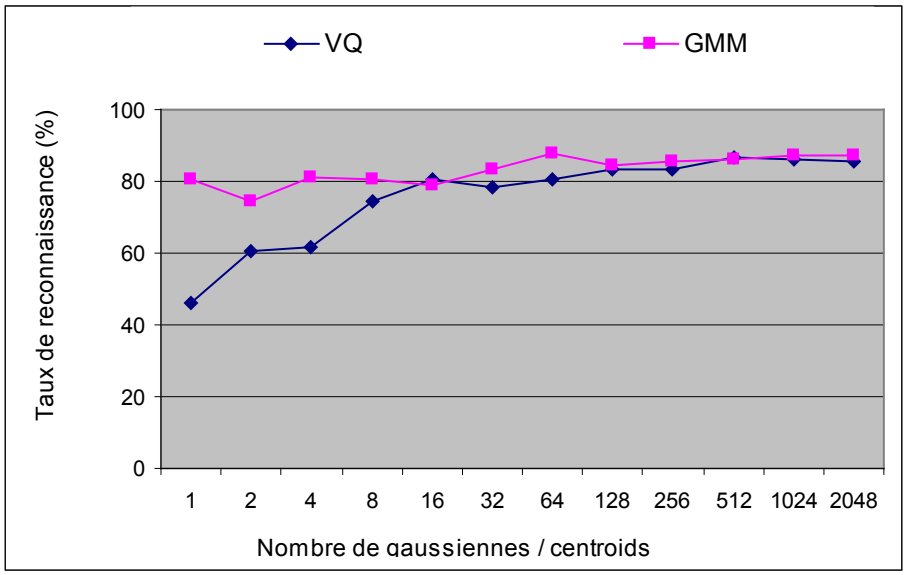

Figure 46 : Comparaison des performances entre le modèle GMM et le modèle VQ

6.3.1.2 Modèle de quantification vectorielle (VQ)

La Figure 46 contient la représentation des résultats du Tableau 58 et les résultats obtenus avec le modèle de quantification vectorielle dont le nombre de centroïdes varie également en exposant de 2 comme le nombre de gaussiennes dans le Tableau 58.

Il est visible que la performance du modèle GMM est supérieure à celle de modèle VQ dans presque tous les cas. Cependant, selon notre expérimentation avec le corpus DES, avec un nombre suffisamment grand de centroïdes (≥ 1024 centroïdes), les deux modèles convergent l'un vers l'autre en termes de performance. L'avantage de modèle VQ est alors sa simplicité par rapport au modèle GMM.

6.3.1.3 Modèle de Markov cachés continus (CHMM)

Pour vérifier les caractéristiques de l'évolution temporelle des émotions, nous avons utilisé le modèle de Markov cachés continus (CHMM) pour capturer ces évolutions si elles existent. Le nombre de 64 gaussiennes utilisés dans ce modèle de Markov caché est invariablement

maintenu parce qu'il correspond à la configuration la plus efficace du modèle GMM. Par contre, le nombre d'états de ce modèle CHMM varie de 1 à 11 et la matrice de connexion entre les états est pleine. On expérimente avec des modèles ergodiques (cela veut dire la matrice d'émission du modèle est uniforme – voir la section 6.2.3).

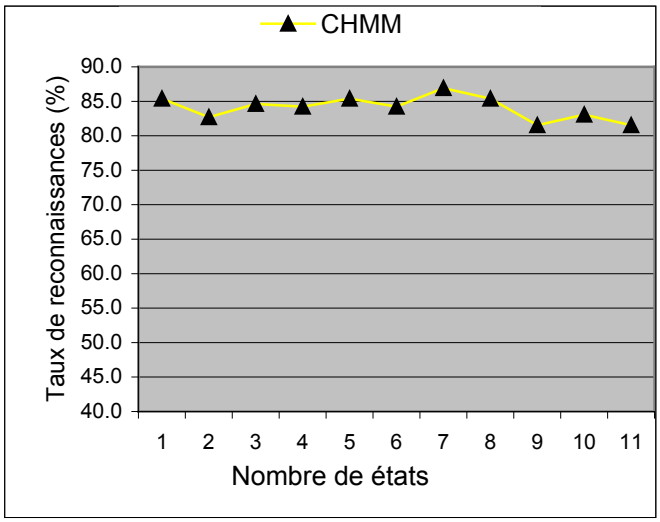

Figure 47 : Comparaison des performances avec le modèle CHMM

Nous pouvons constater dans la Figure 47 que l'ajout de nouveaux états n'améliore pas la performance du système et même que l'ajout d'un trop grand nombre d'états aux modèles CHMM diminue le taux de reconnaissance. Cela peut s'expliquer par le bruit introduit dans le système en raison des états inutiles pour la modélisation.

Il est aussi visible sur la Figure 47 que nous n'avons pas trouvé une amélioration significative avec ce modèle, cela veut dire que malgré les informations utiles de l'évolution locale que nous pouvons capturer par la mesure des paramètres Δ et $\Delta\Delta$, nous n'avons pas encore pu capturer les informations de l'évolution globale avec le modèle de HMM. Cela peut s'expliquer par le fait que cette information de l'évolution globale est différente aux niveaux différents (mot isolé, mots dans une phrase, phrase isolée, phrase dans un paragraphe etc.) et que le nombre d'états du modèle HMM de la Figure 47 n'est pas encore suffisant pour généraliser toutes ces variétés. Il faut rappeler que, ces résultats ont obtenus sur le corpus DES, les tests plus profonds nous demandent d'autres corpus plus grands.

Le même résultat a été obtenu quand nous expérimentons avec les différentes configurations de la matrice d'émission du modèle HMM. En effet, par nos expérimentations sur le corpus DES, les modèles HMM ergodiques (qui permettent tous les types d'évolutions) donnent de meilleurs résultats que le modèles gauche-droit (qui ne permet que l'évolution causale) [Le & al 2004].

Une autre remarque que nous constatons est que les résultats deviennent meilleurs avec les modèles dont le nombre d'états correspond avec le nombre de phonèmes des énoncés du corpus, cela nous suggère qu'une étude de l'émotion au niveau phonémique dans le futur pourrait aider à améliorer la performance du système.

6.3.1.4 Machines à vecteurs de support (SVM)

Toujours dans le contexte du corpus DES, avec les modèles VQ et GMM, nous avons testé la performance de la première approche de classification en nous basant sur la modélisation de l'espace des paramètres. Avec le modèle HMM, nous avons essayé de vérifier si nous pouvons capturer l'évolution temporelle pour améliorer la performance du système. Les résultats expérimentaux suivants dans le cas de la reconnaissance mono-locuteur avec le modèle SVM nous permettra tester la performance de la deuxième technique dans la classification de l'émotion en s'appuyant sur une séparation explicite de l'espace de données.

Le processus de recherche des paires optimales (C,γ) a été effectué sur le modèle SVM (voir section 6.2.2). La Figure 48 nous présente les résultats de 106 paires testées de (C,γ) qui nous permettent de déterminer la zone la plus performante Z de ces deux paramètres (la zone foncée sur la figure qui ne contient que des cas de reconnaissance avec un taux de 80 % à 90 %), le centre de cette zone se trouve être la paire (C,γ) = (2^9,2^{-6}).

Les paires (C,γ) de cette zone nous donnent en moyenne 83,1 % de taux de classification correcte. Ce taux est clairement moins élevé que ceux (87,7 %) obtenus avec des modèles de la première technique dont les deux modèles VQ et GMM sont des représentants. C'est la raison pour laquelle, nous n'allons pas plus loin avec ce modèle dans les deux autres cas de la reconnaissance multi-locuteur et de la reconnaissance indépendante du locuteur. Dans le futur, une recherche plus fine des valeurs de (C,γ) dans la région Z nous intéressera parce qu'il se peut que des valeurs plus fines de (C,γ) de cette région donnerait de meilleurs résultats que 83,1 %.

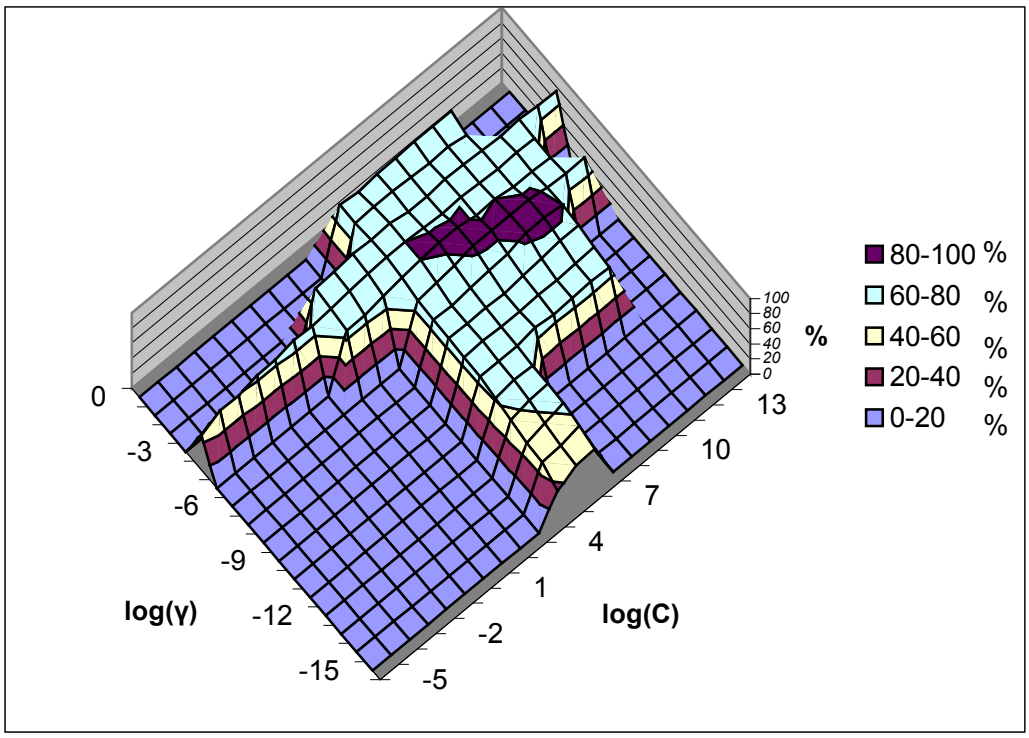

Figure 48 : Résultats correspondants avec 106 paires de (C,γ)

On peut être surpris par le fait que ce taux de reconnaissance automatique de l'émotion dépendante du locuteur sur le corpus DES soit nettement meilleur que celui obtenu par des être-humaines 87,7 % contre 67 % (voir le Chapitre 4). Ceci peut être expliqué par le fait que ces

deux cas de reconnaissance ne sont pas tout-à-fait identiques pour l'homme et pour la machine. Premièrement, la reconnaissance humaine ne se déroule pas de locuteur à locuteur comme le cas de la machine et, deuxièmement, l'homme n'a pas la possibilité de traiter presque tous les exemples d'un locuteur avant d'optimiser son choix comme la machine peut faire. [Womack et al, 1996] a fait le même constat.

6.3.2. Reconnaissance multi-locuteur

6.3.2.1 Modèle de mélange de gaussiennes

Comme nous l'avons mentionné, la reconnaissance mono-locuteur peut être considérée comme un cas particulier de la reconnaissance multi-locuteur où les données pour l'apprentissage et pour la reconnaissance appartiennent à une seule personne. Donc, théoriquement, la meilleure configuration pour le cas de reconnaissance mono-locuteur ne sera pas la meilleure configuration pour les cas de reconnaissance multi-locuteur parce que l'espace des valeurs des paramètres de plusieurs locuteurs est beaucoup plus large que celui d'un seul locuteur. Selon notre expérimentation sur le corpus DES, avec un nombre suffisamment grand de gaussiennes, la performance en reconnaissance du système multi-locuteur approche celle du système mono-locuteur. Cela est montré dans le Tableau 60. Ces résultats sont obtenus avec le corpus DES (4 locuteurs, et 5 états émotionnels) et le modèle GMM dont le nombre de gaussiennes varie de 1 à 1024 en puissances de 2. L'ensemble de 7 paramètres MFCCs et ΔF_0Rel a été utilisé dans ce cas.

Le Tableau 61 nous montre la matrice de confusion du meilleur cas avec GMM 256 gaussiennes. Comme les analyses dans la partie des paramètres globaux, nous trouvons aussi dans ce tableau la confusion forte entre la joie et la surprise. Selon les analyses que nous avons faites sur les paramètres globaux, il y a des indicateurs comme l'amplitude de F_0, *RisingFallingCountRatio* et *RisingFallingSumRatio* de F_0 qui peuvent nous permettre de distinguer la surprise et la joie, même la colère, cela sera donc une étude intéressante pour nous dans le futur sur la combinaison de ces deux ensembles de paramètres : globaux et locaux afin d'améliorer la qualité du système.

N. de gaussiennes	***Taux de reconnaissance*** (%)
1	67,1
2	62,3
4	62,5
8	66,0
16	68,1
32	81,1
64	84,5
128	85,8
256	**86,5**
512	86,2
1024	84,5

Tableau 60 : Recherche la configuration optimale du modèle GMM en fonction du nombre de gaussiennes pour les 7 MFCCs et ΔF_0Rel sélectionnés

	colère (%)	*joie (%)*	*neutre (%)*	*tristesse (%)*	*surprise (%)*
colère	84.6	7.7	1.9	0.0	5.8
joie	3.8	84.6	1.0	0.0	10.6
neutre	6.7	1.0	87.5	4.8	0.0
tristesse	1.0	1.9	2.9	94.2	0.0
surprise	4.8	13.5	0.0	0.0	81.7
moyenne	**86,5**				

Tableau 61 : Matrice de confusion du meilleur cas du modèle de mélange de gaussiennes dans le cas de reconnaissance multi-locuteur

Les résultats du Tableau 60 nous montrent que, au lieu de 64 gaussiennes, nous devons utiliser 256 gaussiennes pour le système le plus performent de reconnaissance de l'émotion dans le cas multi-locuteur. Effectivement, le modèle GMM du cas multi-locuteur a besoin plus d'espace pour pouvoir modéliser tous les locuteurs rencontrés.

6.3.2.2 Modele de quantification vectorielle

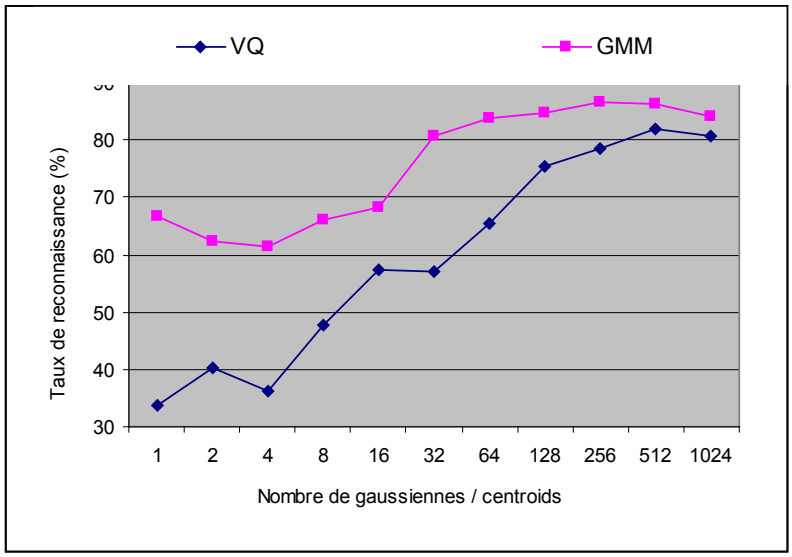

Figure 49 : Comparaison des performances entre le modèle GMM et le modèle VQ

La Figure 49 présente les résultats obtenus avec ces deux modèles en changeant le nombre de gaussiennes et le nombre de centroïdes. Similairement au cas de reconnaissance mono-locuteur, le modèle de GMM se montre toujours plus performant, cependant selon nos résultats avec un suffisamment grand nombre de centroïdes, la différence de performance de ces deux types de modèle n'est plus significative. Effectivement, nous pouvons obtenir un taux de classification de 82 % à partir de 512 centroïdes.

Malgré la simplicité du modèle de quantification vectorielle, parce que le modèle de mélange de gaussiennes le surpasse en termes de performance dans presque tous les cas, nous avons choisi le modèle GMM pour nos études en reconnaissance de l'émotion indépendante du locuteur dans la section suivante.

6.3.3. Comparaison relative avec d'autres travaux

Bien que la reconnaissance de l'émotion mono-locuteur et multi-locuteur ne soient pas notre objectif principal, afin nous placer dans les mêmes conditions que les autres travaux pour la comparaison avec ceux-ci, en plus d'utiliser le protocole « *leave one out* », nous avons également effectué des expérimentations avec le protocole de validation croisée par la division en 10 plis (90% pour l'apprentissage et 10% pour le test), par la division en 5 plis (80% apprentissage et 20% pour le test) et par la division en 3 plis (67% apprentissage et 33% pour le test), le Tableau 62 et le Tableau 63 montrent nos résultats ainsi que ceux de quelques autres équipes sur les mêmes tâches.

	Corpus utilisé	***Paramètres***	***Modèles***	***Proportion Apprentissage (%) / Test (%)***	***Taux de reconnaissance***
[New et al, 2001]	*6 émotions 2 locuteurs en birman*	*MFCCs*	*HMM discrète*	*60 / 40*	*72,2 %*
[Noble 2003]	*LDC 13 émotions 7 locuteurs en anglais*	*3000 paramètres de prosodie et qualité de voix*	*SVM*	*90 / 10*	*~55 %*
[Fernandez et al, 2003]	*4 émotions 4 locuteurs*	*20 paramètres obtenus par TEO (Teager Energy Operator)*	*Mélange de HMMs*	*80 / 20*	*61,2 %*
Notre système	*DES 5 émotions 4 locuteurs en danois*	*12 MFCCs*	*GMM*	*« Leave one out » 90 / 10 80 / 20 67 / 33*	*87,7 % 80,0 % 69,2 % 63,1 %*

Tableau 62 : Comparaison avec d'autres études dans le domaine pour le cas de reconnaissance mono-locuteur

A partir de ces deux tableaux, nous pouvons constater une dégradation du taux de reconnaissance lorsque nous passons du protocole « *leave one out* » au protocole de validation croisée les deux cas de reconnaissance : mono-locuteur et multi-locuteur. Cet effet peut s'expliquer par la perte des points forts du protocole « *leave one out* » dans le protocole de validation croisée :

+ Dans le cas de reconnaissance mono-locuteur, contrairement à la validation croisée, le protocole « *leave one out* » élimine bien des influences de l'aspect phonétique des échantillons lors de la séparation des données pour l'apprentissage et pour le test. Effectivement, lors du choix des données pour le test avec « *leave one out* », si l'énoncé d'une émotion est choisi, tous les énoncés des autres émotions ayant la même structure phonétique sont également choisis et mis avec pour le test. Cela nous permet d'éviter la perturbation causée par l'existence d'une structure phonétique dans la partie de l'apprentissage de quelques modèles mais pas dans tous

les modèles et par conséquent, lors de la reconnaissance, les modèles qui « connaissent » la structure phonétique de l'énoncé ont tendance à se baser sur la structure phonétique plutôt que sur l'émotion.

	Corpus utilisé	***Paramètres***	***Modèles***	***Proportion Apprentissage (%) / Test (%)***	***Taux de reconnaissance***
[Xiao et al, 2007]	*BES 7 émotions 10 locuteurs (5H, 5F) en allemand, données actées, 700 enoncés.*	*68 paramètres de F_0, formants, énergie, harmonies, zipf*	*Réseaux de neurones*	*80 / 20*	*78,3 %*
[Xiao, 2008]	*DES 5 émotions 4 locuteurs (2H, 2F) en danois, données actées, 260 énoncés.*	*Sélection des paramètres parmi 226 paramètres*	*Reseaux de neurones*	*90 / 10*	*81.2 %*
[Ververidis et al, 2005-1]	*DES 5 émotions 4 locuteurs (2H, 2F) en danois, données actées, 260 énoncés.*	*87 paramètres de formant, pitch, énergie*	*GMM*	*90 / 10*	*66%*
[Vidrascu & Devillers, 2005-1]	*2 émotions 6241 tours de locuteurs en français données « real-life » collectées dans un centre d'appels.*	*10 paramètres de la prosodie et du spectre*	*ADTree*	*80 / 20*	*73 %*
[Vidrascu & Devillers, 2007]	*5 émotions, 784 locuteurs français (271H, 513F) données « real-life » collectées dans un centre d'appels.*	*25 paramètres*	*SVM*	*90 / 10*	*56 %*
Notre système	*DES 5 émotions 4 locuteurs (2H, 2F) en danois, données actées, 260 énoncés.*	*7 MFCCs + ΔF0Rel*	*GMM*	*Leave one out 90 / 10 80 / 20 66 / 33*	*86,5 % 79,1 % 76,5% 70,5%*

Tableau 63 : Comparaison avec d'autres études dans le domaine pour le cas de reconnaissance multi-locuteur

+ Dans le cas de reconnaissance multi-locuteur, le fait de mettre ensemble tous les énoncés ayant la même structure phonétique pour toutes les émotions et pour tous les locuteurs avec le protocole « *leave one out* » nous permet non seulement d'éliminer fortement l'influence des aspects phonétiques mais aussi d'éliminer beaucoup d'influences de la dépendance au locuteur des modèles entraînés grâce à l'équilibre des échantillons pour chaque locuteur contenu dans la partie d'apprentissage ainsi que dans la partie de test.

+ « *leave one out* » nous permet encore de profiter au maximum des échantillons du corpus pour l'entraînement et aussi pour le test avec la proportion apprentissage / test la plus élevée : 240/20 ~ 92/8 pour tous les deux cas de reconnaissance mono-locuteur et multi-locuteur.

Dans le contexte du corpus DES, nous constatons aussi que la performance de la reconnaissance mono-locuteur a tendance de se dégrader plus vite que celle de la reconnaissance multi-locuteur si on diminue le nombre de plis appliqués. Cela peut s'expliquer par la petite taille du corpus DES qui est la cause de la différence importante de la taille des données utilisées pour l'entrainement des modèles des deux cas : lors de reconnaissance mono-locuteur, nous n'avons au total que 13 énoncés par modèle (pour chaque locuteur) contre 52 énoncés par modèle (pour tous les 4 locuteurs) dans le cas de reconnaissance mutli-locuteur.

En conlusion, bien que nos résultats, obtenus en utilisant le protocole de validations croisées à k plis, soient inférieurs à ceux de quelques études, par exemple celui de [Xiao et al, 2007], nous les trouvons raisonable car les paramètres utilisés dans ces expérimentations sont des paramètres qui ont été sélectionnés pour optimiser la reconnaissance avec le protocole « *leave one out* » et non pas la reconnaissance en validation croisée 80/20 ou 90/10. Nous trouvons que « *leave one out* » se révèle être une bonne approche pour les corpus équilibrés comme DES car elle permet de limiter l'influence de facteurs comme la structure phonétique ou la dépendance au locuteur. Nous avons rectenu ce protocole « *leave one out* » pour la suite de notre étude : la reconnaissance de l'émotion indépendante du locuteur.

6.3.4. Reconnaissance indépendante du locuteur

La partie centrale de cette thèse est d'analyser la capacité de reconnaissance des états émotionnels indépendante du locuteur en vue de détecter automatiquement les émotions dans la bande son des documents audiovisuel. Dans cette section, nous présenterons nos résultats de reconnaissance indépendante du locuteur obtenus avec les cinq émotions primaires du corpus DES.

Nous avons vu que les taux de reconnaissance des systèmes se dégradent en faisant la reconnaissance dans un environnement multi-locuteur dans la partie précédente. Cependant, la performance du système peut être rétablie si les modèles utilisés sont suffisamment « forts » afin de pouvoir modéliser tous les cas de tous les locuteurs ; cela se montre par l'augmentation des taux de reconnaissance en augmentant le nombre de gaussiennes ou le nombre de centroïdes dans les deux modèles GMM et VQ que nous avons présentés dans la section précédente.

Cette amélioration en changeant de configuration du modèle ne marchera pas très bien dans le cas de la reconnaissance indépendante du locuteur. En effet, comme les données du locuteur que nous devons traiter n'existent pas dans l'ensemble de données utilisées pour l'entrainement des modèles, le renforcement que nous pouvons effectuer sur les modèles sera inutile si les paramètres que nous extrairons à partir des données dépendent fortement du locuteur. Effectivement, la ligne en carrés dans la Figure 50 représentant les résultats obtenus avec les 9 paramètres originaux de MFCCs ne montre pas d'amélioration en fonction du nombre de gaussiennes du modèle GMM.

Le taux de classification obtenu avec cet ensemble de paramètres originaux est aussi très bas (46,2%) par rapport aux 87,7 % et 86,5 % des deux cas : mono et multi-locuteur.

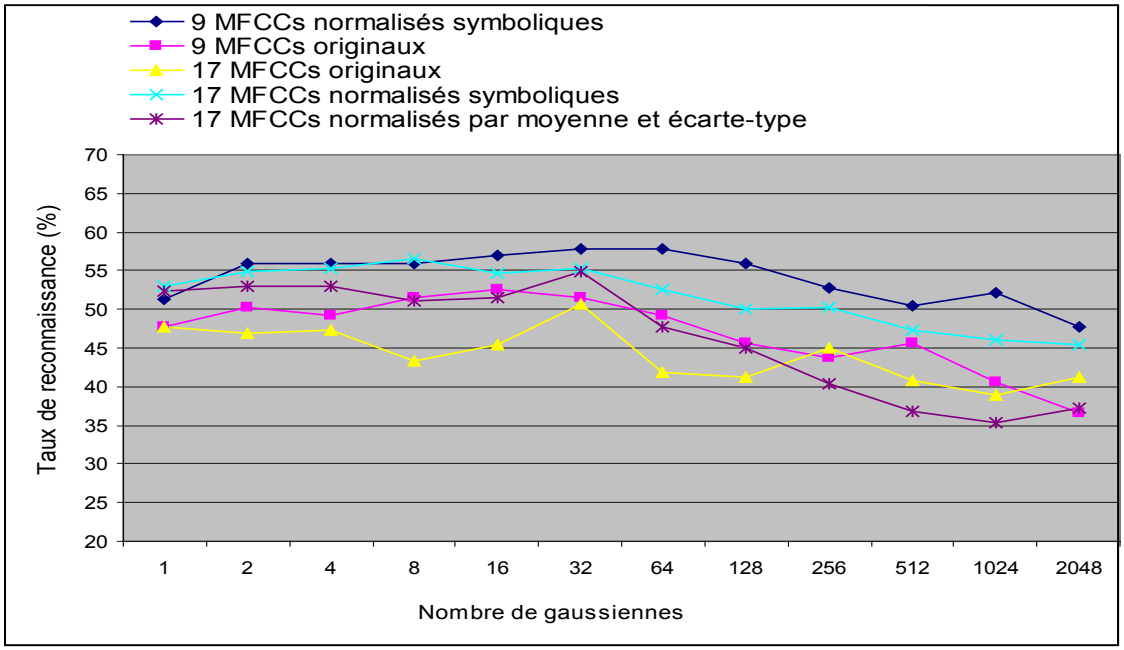

Figure 50 : Comparaison des performances des ensembles de paramètres MFCC avant et après la normalisation symbolique avec le modèle de mélange de gaussiennes.

	colère (%)	***joie (%)***	***neutre (%)***	***tristesse (%)***	***surprise (%)***
colère	54.8	25.5	0.0	1.9	18.8
joie	19.7	58.2	0.5	1.0	20.7
neutre	13.9	8.7	46.2	13.0	16.3
tristesse	2.4	4.3	19.7	63.0	10.6
surprise	14.4	17.3	0.5	3.4	64.4
moyenne	**57,3**				

Tableau 64 : Matrice de confusion du meilleur cas de 9MFCCs normalisés symboliques avec le modèle GMM de 64 gaussiennes

La ligne en fuseaux de Figure 50 présente le taux de classification obtenu avec les mêmes 9 paramètres de MFCC mais normalisés. Il est visible que l'ensemble de paramètres normalisés de MFCC nous donne une amélioration considérable par rapport aux autres ensembles de paramètres. En moyenne, la différence entre les ensembles de résultats est d'environ 7 %. Le taux maximum de classification est aussi monté jusqu'au 57,3 % grâce à cette normalisation.

Similairement aux deux cas de reconnaissance mono et multi-locuteur, la capacité de capture des caractéristiques de l'évolution temporelle du modèle de Markov caché ne conduit pas à une amélioration. L'augmentation de nombre d'états dans ce modèle cause même une dégradation de la performance du système.

La Figure 51 nous montre le taux de reconnaissance en augmentant le nombre d'états du modèle de HMM pour les deux ensembles de 9 paramètres sélectionnés de MFCC avant et après la normalisation symbolique.

Cependant, nous remarquons les deux régions où le taux de classification est plus élevé par rapport aux autres régions en changeant le nombre d'états du modèle de HMM ; ce sont la région de 2 à 3 états et la région de 10 et 11 états. Selon notre analyse, cette amélioration obtenue n'est pas en raison de la capture de la caractéristique de l'évolution temporelle des émotions, mais c'est en raison de la structure de corpus DES.

Effectivement, les longueurs des échantillons du corpus se concentrent autour de 2 ou 3 syllabes (avec des échantillons en mots) et autour de 6 à 7, et de 11 à 13 syllabes (avec des échantillons en phrases). Nous expliquons ce fait de ces deux cas comme suit : les états du modèle HMM capturent bien la structure phonétique des échantillons (cela veut dire que la transition entre états correspond à la transition phonétique), donc chaque état du modèle HMM modéliserait un phonème. Grâce à ce fait, l'alignement entre les mêmes phonèmes avec les mêmes états sera automatiquement effectué si les deux échantillons ont la même structure phonétique, en conséquence, le taux de reconnaissance devient meilleur.

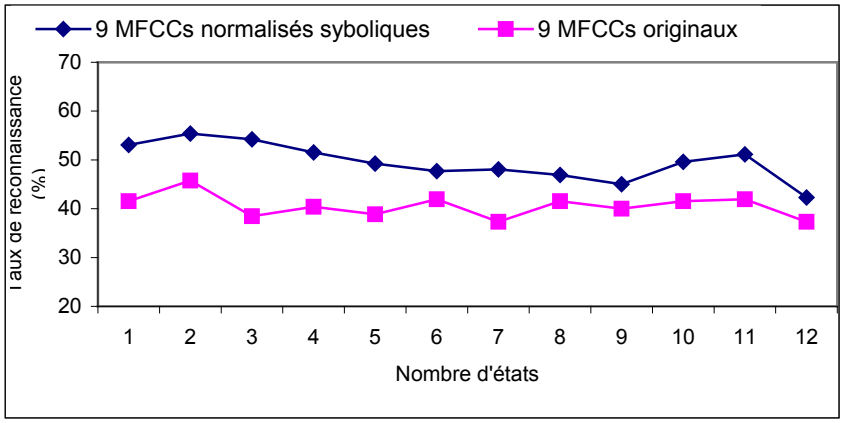

Figure 51 : La performance du modèle de Markov caché pour les ensembles de paramètres de MFCC avant et après la normalisation symbolique en fonction de nombre d'états.

6.4. Conclusion

Comme dans le chapitre précédent, avant de données des conlusions, nous voulons rappeler ici que nos expérimentations sont essentiellement réalisées sur le corpus DES dont le point faible est la taille et le niveau de naturel. C'est la raison pour laquelle on ne peut pas en déduire que ces résultats se généraliseraient dans d'autres conditions plus générales. C'est bien le cas pour le jeu précis de meilleurs paramètres que nous avons présentés dans le chapitre précédent, mais dans ce chapitre, nous pensons que c'est aussi le cas pour la recherche des méthodes de l'optimalisation avec des modèles : les modèles peuvent avoir les comportements différents dans des contextes différents. Effectivement, par exemple en raison de manque de données, nous n'avons pas trouvé l'amélioration de la performance du modèle HMM par rapport le modèle GMM.

En étudiant les trois branches des techniques de classification et de reconnaissance du domaine, nous constatons que dans le contexte de corpus étudié (le corpus DES : 4 locuteurs, 65 énoncés

en 5 états émotionnels pour chaque locuteur) l'approche de classification par la modélisation de l'espace de paramètres nous donne non seulement les meilleurs résultats, mais elle est aussi l'approche la plus efficace en raison de sa simplicité, ce qui est aussi important pour un système de détection automatique en temps réels. Effectivement, selon nos résultats sur le corpus DES, un modèle à 64 gaussiennes est déjà suffisant pour modéliser un état émotionnel indépendamment du locuteur. Pour la troisième branche des techniques de classification, la capacité à capturer l'information d'évolution temporelle des paramètres n'amène pas encore d'amélioration de performance du système, comme par exemple avec le modèle de Markov caché que nous avons étudié.

La comparaison relative avec les résultats des autres études dans le cas de reconnaissance dépendante du locuteur et dans le cas indépendante du locuteur nous permet une évaluation positive de notre approche dans tous les deux cas. La validation de cette approche sera effectuée avec le corpus BES (de plus grande taille) dans le chapitre suivant.

Chapitre 7. **Expérimentation inter-langue**

Nous avons utilisé jusqu'ici uniquement le corpus DES pour notre étude car celui-ci est bien équilibré pour le nombre d'échantillons entre les émotions, le nombre d'échantillons pour chaque locuteur, et le nombre d'hommes/femmes. Cependant, en raison de sa taille qui n'est pas suffisamment importante, particulièrement le nombre de locuteurs (4 locuteurs), l'étude et les résultats obtenus avec ce corpus doivent être vérifiés en utilisant un autre corpus. Le corpus BES avec 10 locuteurs a été choisi pour cet objectif.

7.1. Expérimentation avec le corpus BES

Le Tableau 65 montre l'ensemble de paramètres les plus efficaces sélectionnés par la méthode SFSA[8] pour le corpus BES. Les expérimentations sont faites dans les mêmes conditions que pour le corpus DES : reconnaissance indépendante du locuteur avec la méthode « *leave one out* ».

[8] voir la section 5.3.2.3.5

Nom. de paramètres sélectionnés par SFSA	***paramètres***	***résultats sur BES (%)***
1	$Z=MFCC_2$	41,0
2	$Z=Z+MFCC_9$	47,7
3	$Z=Z+\Delta MFCC_0$	51,1
4	$Z=Z+MFCC_5$	52,1
5	$Z=Z+\Delta MFCC_7$	54,1
6	$Z=Z+\Delta MFCC_{14}$	55,8
7	$Z=Z+\Delta MFCC_8$	57,7
8	$Z=Z+\Delta MFCC_9$	58,3
9	$Z=Z+\Delta MFCC_1$	58,8
10	$Z=Z+\Delta MFCC_{12}$	59,0
11	$Z=Z+\Delta\Delta MFCC_0$	59,5
12	$Z=Z+\Delta MFCC_5$	60,3
13	$Z=Z+\Delta\Delta MFCC_4$	61,3
14	$Z=Z+\Delta MFCC_{10}$	61,8
15	$Z=Z+\Delta\Delta MFCC_1$	61,8
16	$Z=Z+\Delta\Delta MFCC_{16}$	**62,1**

Tableau 65 : Ensemble de paramètres les plus efficaces obtenus par SFSA

A partir de ces résultats, et en comparaison avec les résultats obtenus avec le corpus DES dans le Tableau 53, les constats peuvent être retenus sur ces deux corpus :

- les paramètres les plus efficaces sélectionnés par SFSA pour les deux langues de ces deux corpus ne sont pas identiques, [Xiao, 2008] a également trouvé les deux ensembles différents pour ces deux langues, mais dans le cas de la reconnaissance dépendante du locuteur. D'après nous, cela est raisonnable parce que :
  - premièrement, bien qu'il y ait des points communs en termes d'expression émotionnelle entre des cultures différentes (voir Chapitre 2), il y a aussi beaucoup de points différents entre des langues comme l'accent, la prosodie, l'intonation etc. Par contre dans ce cas, nous avons cherché à optimiser séparément la performance de la reconnaissance pour chaque langue, c'est donc facile à comprendre que l'ensemble de paramètres trouvés va s'adapter aux caractéristiques spécifiques de chaque langue et par conséquence, ces deux ensembles de paramètres trouvés de chaque langue seront différents ;
  - deuxièmement, la différence dans la sélection des paramètres optimaux peut aussi s'expliquer par le fait que l'ensemble des émotions considérées ne sont pas identiques (5 émotions : *la colère, la joie, le neutre, la tristesse, la surprise* pour DES et 7 émotions : *neutre, la colère, la peur, la joie, la tristesse, le dégoût et l'ennui* pour BES) ;
- les émotions en danois du corpus DES ont tendance à être plus disjointes avec les coefficients MFCCs originaux tandis que les ΔMFCCs et ΔΔMFCCs (ceux qui signifient respectivement la vitesse et l'accélérateur des paramètres originaux) sont plus efficaces que les paramètres originaux pour discriminer les états émotionnels en allemand ; ceci pourrait aussi être dû aux caractéristiques spécifiques de chaque langue ;

- la présence des coefficients $MFCC_2$ dans les ensembles de paramètres sélectionnées par SFSA des deux corpus BES et DES (dans tous les trois cas de reconnaissance mono-locuteur, multi-locuteur et indépendante du locuteur) prouve l'efficacité/robustesse des coefficients $MFCC_2$.

Tenant compte des possibilités d'application et aussi pour la comparaison avec d'autres études, nous effectuons également des expérimentations sur la reconnaissance indépendante du locuteur du type émotif / neutre pour chaque émotion du corpus BES. Le Tableau 66 nous montre les résultats en utilisant la configuration optimale ci-dessus (16 MFCCs symboliquement normalisés et sélectionnés par SFSA et le modèle GMM 128 gaussiennes).

Emotions	***colère/ neutre***	***joie/ neutre***	***tristesse/ neutre***	***peur/ neutre***	***ennui/ neutre***	***dégoût/ neutre***
notre systèmee (%)	99,0	96,7	90,1	83,0	82,5	82,3

Tableau 66 : Classification par rapport au neutre

Parmi les six états du corpus DES, la colère est l'état le plus facile à détecter avec le taux de reconnaissance de 99%, contrairement, le dégoût est l'état le plus difficile (82,3%). Pourtant, nous trouvons que c'est tout-à-fait comparable avec des autres systèmes qui font aussi la classification émotif / neutre, par exemple [Yacoub et al, 2003] ont obtenu le taux de 94 % pour la classification colère / neutre ; [Yoon et al, 2007] ont obtenu 86,5 % pour la même classification colère / neutre.

Quelques autres cas intéressants et applicables sont aussi expérimentés pour le but de tester notre approche de la reconnaissance indépendante du locuteur. Les résultats sont présentés dans le Tableau 67. Ces résultats sont obtenus en utilisant le corpus BES et les 16 coefficients symboliquement normalisés et sélectionnés par SFSA pour ce corpus et le modèle GMM 128 gaussiennes. D'après ces résultats, nous constatons que, dans le context du corpus BES la tristesse / le neutre et la colère qui sont très disjoints par notre approche et qui peuvent nous donner le taux de reconnaissance très élevé de 94 % ; la joie et la colère sont les deux états assez confondus par la machine.

	taux de classification de notre approche sur BES
joie / neutre / colère	79,7%
neutre / colère / tristesse	94,0%
joie / neutre / colère / tristesse	79,3%
joie / neutre / colère / tristesse / peur	70,7%
joie / neutre /colère / tristesse / peur / ennui	68,6%

Tableau 67 : Les ensembles d'émotions souvent rencontrées dans la littérature

Pour la comparaison, nous citons ici deux autres systèmes : celui de [Yacoub et al, 2003] qui a réussi à classifier les trois groupes : joie / neutre + tristesse / colère avec 79,7 % et celui de [Lee et al, 2004-2] qui a réussi à classifier les quatre groupes joie / neutre / colère / tristesse avec 76,12 %. En général, les taux de reconnaissance indépendante du locuteur que nous avons

obtenus avec le corpus BES sont assez élevés en comparaison avec ces études. Le Tableau 68 donne une autre comparaison entre nos résultats obtenus sur les deux corpus DES et BES avec les quatre autres systèmes. Nous voulons souligner que cette comparaison n'est qu'une comparaison relative en raison de la variété des corpus utilisés, de la langue.

	Corpus utilisé	***Paramètres***	***Modèles***	***Proportion Apprentissage (%) / Test (%)***	***Taux de reconnaissance***
[Huang et al, 2006]	*1087 énoncés en 5 émotions actées par 7 locuteurs*	*ZEPS*	*HMM*	*« Leave one speaker out » (86 % / 14 %)*	*53 %*
[Noble 2003]	*LDC 13 émotions actées en anglais par 7 locuteurs*	*3000 paramètres de prosodie et qualité de voix*	*SVM*	*« Leave one speaker out » (86 % / 14 %)*	*~23,60 %*
[Yacoub et al, 2003]	*2433 énoncés en 15 émotions actées en anglais par 8 loctuteurs*	*37 paramètres de F_0, énergie, durée*	*Réseaux de neurones*	*« Leave 2 speakers out » (75 % / 25 %)*	*8,7 % (contre 6,7% pour le cas de classification aléatoire)*
Notre système sur DES	*DES 260 énoncés en 5 émotions actées en danois par 4 locuteurs*	*9 MFCCs*	*GMM*	*« Leave one speaker out » (75 % / 25 %)*	*57,30 %*
Notre système	*BES 7 émotions actées en allemand par 10 locuteurs*	*16 MFCCs*	*GMM*	*« Leave one speaker out » (90 % / 10 %)*	*62,1 %*

Tableau 68 : Résultats de la reconnaissance indépendante du locuteur sur les deux corpus DES et BES en comparaison avec d'autres systèmes

7.2. Expérimentations inter-langue, inter-culture

7.2.1. Croisement de la sélection des paramètres

La première expérimentation que nous effectuons sur ces deux corpus simultanément est de tester la performance des ensembles de paramètres les plus efficaces trouvés dans une langue sur l'autre langue. Bien que les différents résultats sélectionnés par SFSA dans la section précédente nous permettent déjà d'affirmer théoriquement que les paramètres optimaux pour une langue ne seront pas ceux optimaux dans une autre langue, cette expérimentation donne une comparaison précise.

Le Tableau 69 montre les résultats obtenus pour les deux corpus DES et BES en utilisant l'ensemble de 17 paramètres MFCCs originaux, l'ensemble de 17 paramètres MFCCs normalisés par la moyenne/l'écart-type, l'ensemble de 17 paramètres MFCCs symboliquement normalisés, l'ensemble de 9 paramètres MFCCs symboliquement normalisés et sélectionnés par SFSA pour le corpus DES et 16 paramètres MFCCs symboliquement normalisés et sélectionnés par SFSA pour le corpus BES.

	17 originaux	*17 originaux normalisés par moyenne et par écart-type*	*17 originaux normalisés symbolique*	*9 MFCCs normalisés symboliques sélectionnés par SFSA pour DES*	*16 MFCCs normalisés symboliques sélectionnés par SFSA pour BES*
DES ***5 émotions, 4 locuteurs***	46,2 %	48,5 %	50,0 %	57,3 %	**40,4 %**
BES ***7 émotions, 10 locuteurs***	54,3 %	56,6 %	57,7 %	**53,4 %**	62,1 %

Tableau 69 : Le taux de reconnaissance obtenus avec les deux corpus DES et BES

D'après les résultats montrés par le Tableau 69, les conclusions suivantes peuvent être retenues dans notre contexte :

- l'ensemble de 16 paramètres et l'ensemble de 9 paramètres les plus efficaces respectivement pour la reconnaissance de l'émotion indépendante du locuteur sur le corpus BES en allemand et sur le corpus DES en danois ne sont plus les ensembles optimaux s'ils sont appliqués dans une autre langue ;
- l'utilisation directe des coefficients sans normaliser n'obtient pas non plus de bons résultats dans les deux langues ;
- la normalisation classique par la moyenne et par l'écart-type est moins efficace que la normalisation symbolique proposée.

7.2.2. Croisement des modèles

La deuxième expérimentation que nous avons effectuée sur ces deux corpus est la reconnaissance croisée entre ces deux langues. Les énoncées de 4 émotions communes de ces deux corpus (la colère, le neutre, la joie et la tristesse) sont utilisés pour entrainer les 4 modèles GMM 128 gaussiennes dans une langue et reconnaitre les émotions contenus dans les énoncées de l'autre langue et vice versa.

Bien que les résultats précédents montrent les différences des ensembles de paramètres entre ces deux langues, cette expérimentation a pour le but d'étudier s'il existe des points communs en termes de paramètres entre les deux langues pour l'expression des émotions. Les résultats du Tableau 70 nous permet der confirmer cette existence, les points communs sont les 9 paramètres MFCCs symboliquement normalisés et sélectionnés par la méthode SFSA ; ce sont MFCC5, MFCC2, MFCC1, MFCC10, ΔMFCC4, ΔMFCC3, ΔΔMFCC10, ΔMFCC11 et ΔΔMFCC3.

Ces 9 paramètres nous donne les taux de classification assez élevés 62,1 % dans le cas de l'entrainement sur le corpus DES et le test sur le corpus BES (A) et 54,8 % dans le cas de l'entrainement sur le corpus BES et le test sur DES (B). La raison pour laquelle le taux de classification du cas (B) est même plus faible que celui obtenu par le cas (A) peut s'expliquer par la petite taille, le petit nombre de locuteurs du corpus DES qui ne permet pas une bonne généralisation sur les modèles entrainés en utilisant les données du corpus DES ;

	17 MFCCs originaux	17 MFCCs normalisés par moyenne et par écart-type	17 MFCCs normalisés symbolique	9 MFCCs normalisés symboliques sélectionnés par SFSA pour DES	16 MFCCs normalisés symboliques sélectionnés par SFSA pour BES	9 MFCCs normalisés symboliques sélectionnés par SFSA pour la reconnaissance croisée
Entrainement sur BES Test sur DES	28,3	28.3	29,32	29,3%	44,7%	**62,1%**
Entrainement sur DES Test sur BES	34.3%	28,1%	37,6%	27,2%	25,0%	**54,8%**

Tableau 70 : Résultats de la reconnaissance croisée entre les deux corpus DES et BES

Les taux de reconnaissances semblent meilleurs dans ce cas-ci que dans le cas précédent (croisement des paramètres seulement) parce qu'ici, la reconnaissance ne s'effectue que sur quatre états émotionnels au lieu de cinq ou sept.

En comparaison avec un autre système de [Xiao, 2008] qui utilise les deux mêmes corpus pour la reconnaissance croisée entre les deux langues, nous constatons que notre approche (la combinaison entre la normalisation symbolique et la sélection forcée séquentielle en avant SFSA) peut améliorer la performance du système. Effectivement, l'auteur a obtenu pratiquement le taux de classification correcte (59,35 %) mais pour un cas plus strict : la reconnaissance avec le genre spécifique du locuteur ; les résultats de cet auteur fait en moyenne environ 53 % de classification correcte.

	Colère (%)	Neutre (%)	Joie (%)	Tristesse (%)
Colère	**62,0**	0,0	38,0	0,0
Neutre	11,4	**88,6**	0,0	0,0
Joie	19,8	0,0	**80,2**	0,0
Tristesse	4,8	77,4	0,0	**17,7**
Moyenne	62,1%			

Tableau 71 : Matrice de confusion de la reconnaissance croisée entre les deux corpus DES/BES

Le Tableau 71 présente la matrice de confusion de notre meilleur cas. Comme nous pouvons constater, la tristesse est très souvent confondue avec le neutre, la colère est souvent reconnue comme la joie, une étude dans le futur de la combinaison de cet ensemble de 9 paramètres avec des autres paramètres plus discriminatifs pour ces émotions confondues ou une approche de classification hiérarchique pourrait améliorer la qualité du système.

7.3. Expérimentation avec corpus Orator

Comme présenté dans la section 4.4 ; parmi les trois corpus que nous avons utilisés, l'Orator est le seul corpus qui est construit en utilisant l'approche continue (évaluation en échelle continue) et dimensionnelle (évaluation par plusieurs aspects en même temps). Cette approche a son avantage d'être plus proche des expressions émotionnelles naturelles mais elle possède aussi son inconvénient de ne pas avoir suffisamment d'échantillons typiques et représentatifs pour les études théoriques ainsi que pour pouvoir entrainer des modèles des émotions. Nous avons donc utilisé les modèles entrainés par les expressions émotionnelles « pures » du corpus BES et testé avec les énoncés du corpus Orator. Parmi les 7 états émotionnels du corpus BES et les 7

dimensions affectifs du corpus Orator, nous constatons qu'il n'y a que la colère et la joie qui sont les états émotionnels communs.

Pour sortir les échantillons ayant la joie ou la colère du corpus Orator, nous nous basons sur les évaluations de 20 évaluateurs en échelle de -2 à 2 qui ont été faites par l'auteur de ce corpus. Cependant, avant de pouvoir utiliser les valeurs des évaluations, il nous faut un processus pour les normaliser auprès chaque évaluateur car premièrement l'état émotionnel de l'évaluateur au moment de jugement n'est pas toujours en neutre, il influence donc de plus ou moins sur les décisions (ceci influence sur la moyenne des valeurs choisies) ; deuxièmement, les habitudes différentes des évaluateurs entrainent également la différence de la dynamique des valeurs choisies (ceci influence sur l'écart-type des évaluations). La Figure 52 montre la distribution des évaluations d'un locuteur avant et après la normalisation par moyenne et par l'écart-type. Après la normalisation, la moyenne et l'écart-type des évaluations de chaque locuteur sont respectivement égaux à 0 et à 1.

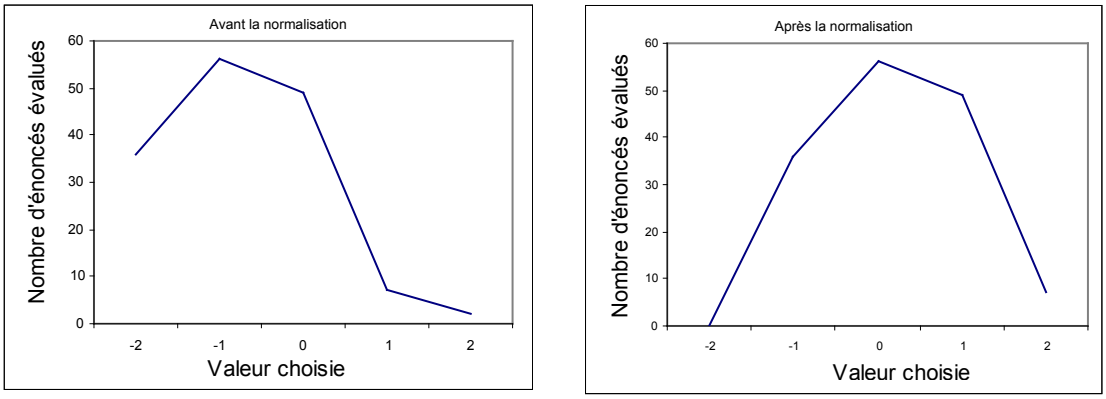

Figure 52 : Evaluations d'un locuteur avant et après la normalisation par moyenne et par l'écart-type.

Ensuite, la moyenne des évaluations de tous les évaluateurs pour chaque énoncé a été calculée. Un énoncé est donc considéré comme une expression d'une émotion quelconque si sa moyenne des évaluations de tous les évaluateurs pour cette émotion est supérieure à un seuil S1 = 0,5.

Selon ce critère, nous avons obtenu 31 énoncés et 26 énoncés respectivement pour la joie et pour la colère. En utilisant les deux modèles GMM 128 gaussiennes entrainés par les échantillons en colère et en joie du corpus BES pour reconnaître cet ensemble de 57 énoncés du corpus Orator, nous obtenons en moyenne le taux de reconnaissance de 79,9 %. Le serrement de notre critère en augmentant le seuil S1, ou autrement dit l'augmentation de l'intensité des émotions, nous donne les meilleurs résultats ; la Figure 53 nous montre les résultats obtenus en fonction du seuil S1 (en fonction de l'intensité des émotions continues dans les énoncés).

Ce résultat nous suggère qu'une étude dans le futur qui portera ou s'appuie sur l'estimation du degré des émotions contenues dans un énoncé sera intéressante et réalisable.

	joie (%)	colère (%)
Joie	**71,0**	29,0
Colère	11,5	**88,5**
moyenne	**79,9** %	

Tableau 72 : Matrice de confusion de la reconnaissance croisée entre les deux corpus BES/Orator

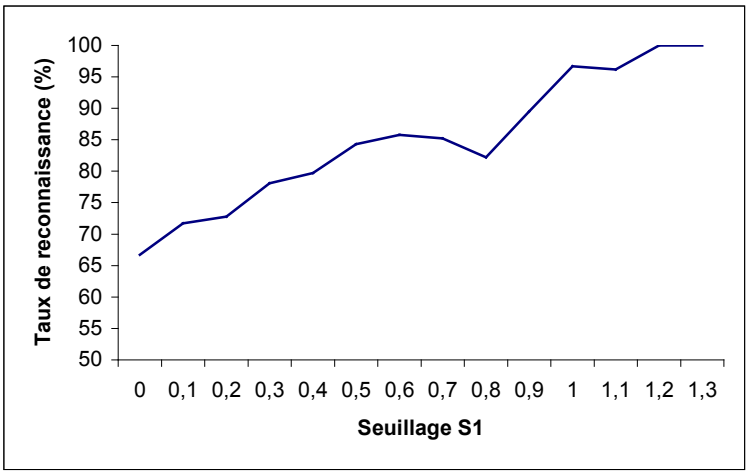

Figure 53 : Taux de reconnaissance sur Orator en fonction de seuillage S1.

Nous nous rappelons que les deux corpus Orator et BES sont tous en allemands, cependant selon notre expérimentation du Tableau 73, ce taux assez élevé de 79,7% de reconnaissance correcte est obtenu avec les 9 coefficients symboliquement normalisés et sélectionnés par la méthode SFSA pour le cas de reconnaissance inter-langue (allemande/danois), par contre, l'utilisation des modèles GMM 128 gaussiennes entrainés par 16 coefficients qui ont été sélectionnés dans le même langue (sur le corpus BES) pour tester ces 57 énoncés du corpus Orator ne nous donne que le taux de 59,8% de connaissance correcte dont il y a une confusion importante de la colère pour la joie. Cela peut s'expliquer par la différence de monologue de chaque corpus et ce fait réaffirme également la robustesse de 9 paramètres trouvés avec la normalisation symbolique.

	17 MFCCs originaux	17 MFCCs normalisés par moyenne et par écart-type	17 MFCCs normalisés symbolique	16 MFCCs normalisés symboliques sélectionnés par SFSA pour BES	9 MFCCs normalisés symboliques sélectionnés par SFSA pour le cas de reconnaissance inter-langue
Entrainement sur BES Test sur Orator	53,4 %	63,8 %	67,1 %	59,8 %	**79,9 %**

Tableau 73 : Classification joie / colère sur le corpus Orator en utilisant les modèles entrainés par les données du corpus BES

Jusqu'ici, nous avons étudié et travaillé avec trois corpus. Nous voudrions rappeler que les résultats sont obtenus dans des conditions précises de ces trois corpus et qu'on ne peut pas en

déduire que ces résultats se généraliseraient dans un contexte plus général. Cependant, nous voudrions tester nos résultats sur un système d'indexation des documents réels.

7.4. Indexation sur un corpus réel

7.4.1. Approche de segmentation

Le premier problème auquel nous devons faire face est de segmenter raisonnablement la parole continue en segments de sorte qu'il n'y ait pas de changement de l'état émotionnel du locuteur dans chaque segment. Autrement dit, la question porte sur le niveau (de phonème, de mot, de group de mots, de phase, etc.) sur lequel nous devons nous baser pour segmenter ?

Dans la littérature, nous n'avons pas trouvé d'études portant sur cet aspect. La plupart des recherches sur la reconnaissance émotionnelle néglige ce problème en utilisant des corpus construits par simulation (voir chapitre 4). Avec ces corpus, plusieurs types d'unités comme les mots, les phrases ou les paragraphes ont été préparés, et les expressions émotionnelles des acteurs les lisant sont toujours considérées comme uniformément présentes sur toute la longueur de chaque énoncé.

Ceci n'est pas toujours correct pour la parole spontanée dans laquelle l'état émotionnel peut changer n'importe quand. Effectivement, il est en effet facile de constater que l'état émotionnel ne dure pas toujours toute longueur d'une phrase ou d'un parcours de la parole, alors *quand le changement de l'état émotionnel se produit-il* ?

De notre côté, nous proposons de nous baser sur l'unité de mot pour segmenter pour les deux raisons suivantes :

- dans le cadre d'un système d'indexation automatique de l'émotion, nous ne nous intéressons pas aux niveaux acoustiques inférieurs comme le niveau syllabique ou le niveau phonétique ;
- en prenant le mot comme unité de base de l'expression émotionnelle, nous avons un moyen pour déterminer la frontière entre des états émotionnels différents apparaissant dans une phrase ou dans un paragraphe.

En se basant sur l'unité de mot, afin de déterminer le changement de l'état émotionnel dans un paragraphe ou monologue continu, nous proposons d'utiliser la distance temporelle relative entre les mots consécutifs de ce monologue. L'idée est basée sur notre observation du fait que l'homme effectue souvent des « pauses » (silence) avant le changement d'état émotionnel.

En pratique, la « pause » peut aussi correspondre au changement de tour des deux locuteurs dans un dialogue ou à la vraie pause d'un seul locuteur pour marquer la fin d'une phrase, d'une idée complète. Dans ces deux cas, nous constatons que la proposition de la « pause » est raisonnable. Effectivement :

- si la pause marque un changement de tour des locuteurs (changement de locuteur), de toute façon, un changement de l'état émotionnel doit être mis à ce point, car ce sont clairement des segments de parole des deux locuteurs différents ;
- si la pause marque la fin d'une phrase, d'une idée, il est également raisonnable de mettre une marque de changement de l'état émotionnel à ce point car la probabilité du changement d'état émotionnel à ce point est élevée en raison du changement de contenu sémantique ; même dans le cas où il n'y a pas de changement d'état émotionnel à ce

point, cette segmentation n'affecte pas la qualité de l'indexation car nous pouvons facilement regrouper les segments consécutifs qui contiennent les mêmes états émotionnels d'un locuteur.

L'algorithme suivant explique notre approche pour détecter les points de segmentation.

Etant donné un paragraphe d'un seul ou de plusieurs locuteurs et qui se compose d'une suite des mots m_1, m_2, …, m_n avec c_i est le temps où le mot m_i commence à être prononcé, et t_i est le repère du le temps de la fin de la prononciation du mot m_i. Alors, $t_i<t_{i+1}$; $c_i<c_{i+1}$ et $t_i<c_{i+1}$. La distance entre les deux mots est définie par $d_i=c_{i+1}-t_i$. Le Tableau 74 donne un exemple du calcul de ces distances.

Mot	m_0	m_1	…	m_{n-1}	m_n
Exemple d'une phrase	Bonjour	tout		bonne	soirée
Le point de commencement	c_0	c_1	…	c_{n-1}	c_n
Le point de terminaison	t_0	t_1	…	t_{n-1}	t_n
La distance entre les mots consécutifs d'une phrase	$d_0=c_1-t_0$	$d_1=c_2-t_1$		$d_{n-1}=c_n-t_{n-1}$	$d_n = \infty$ (pour imposer une pause)

Tableau 74 : Calcul de distances des mots

La détermination des « pauses » est réalisée par la détermination du d_i dont la valeur est anormalement changée par rapport avec la valeur du d_{i-1} précédent sur la chaine des valeurs consécutives d_0, d_1, …, d_i, …, d_n.

Le changement est défini comme anormal s'il est supérieur à un seuil $(1+\alpha).d_m$ (la distance est anormalement grande) ou inferieur à un seuil $(1-\alpha).d_m$ (la distance est anormalement petite) où d_m est la distance moyenne et α est un coefficient. Dans notre cas, par l'observation et par des tests, nous avons choisi $\alpha = 0,7$.

1. Mettre le pointeur $i = 0$; $d_m = 0$;
2. Calculer $d_m = d_i$ et passer à l'étape 3 ;
3. i = i +1 ;

   si $d_i \geq (1+\alpha).d_m$ → Marquer d_i comme la pause dont t_i, c_{i+1} sont respectivement les points de commencement et de fin de la pause ; remettre $i = i+1$; refaire l'étape 2 ;

   si $d_i \leq (1-\alpha).d_m$ → Marquer d_{i-1} comme la pause dont t_{i-1}, c_i sont respectivement les points de commencement et de fin de la pause ; refaire l'étape 2 ;

   si $(1-\alpha).d_m < d_i < (1+\alpha).d_m$ → Mettre à jour $d_m = \frac{1}{M+1}\sum_{k=i-M}^{i} d_k$ où M est le nombre de distances consécutives sans pause avant i.

La Figure 54 illustre les deux cas de pause dans l'étape 3.

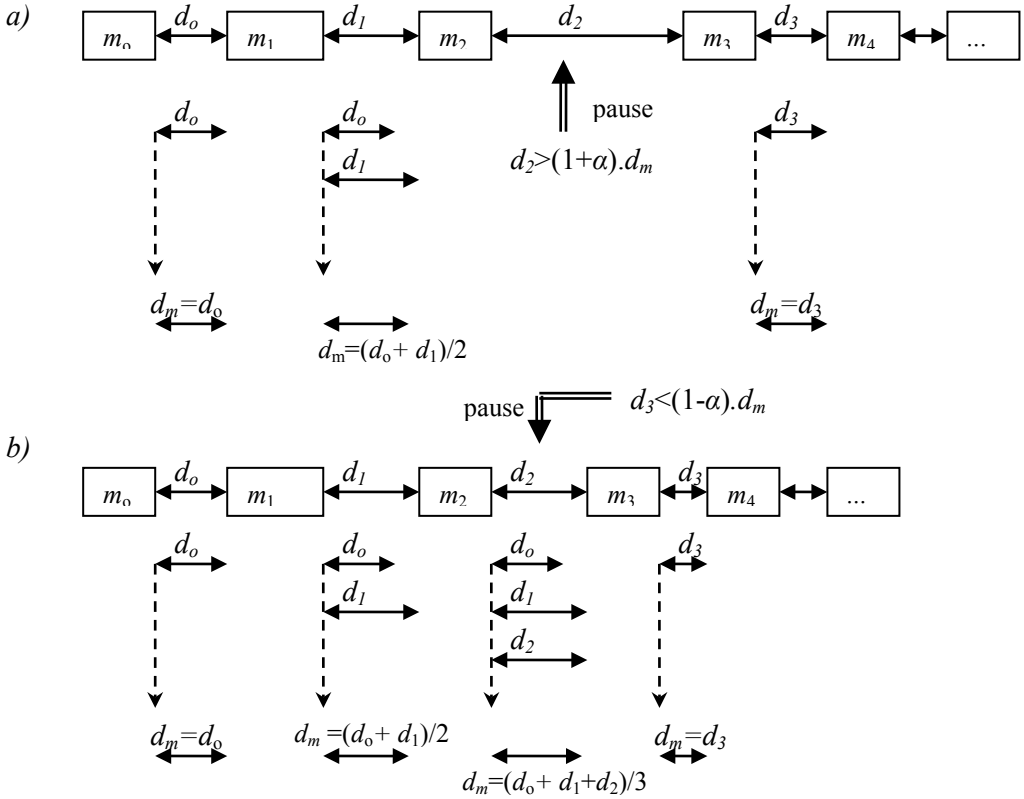

Figure 54 : Dépistage des pauses
a) la distance est anormalement grande; b) la distance est anormalement petite

En appliquant cette approche sur 254 fichiers contenant la bande son d'environ 12h de journal télévisé de CNN et ABC, nous avons obtenu au total 105178 segments dont la longueur en fonction du nombre de mots est présenté dans le Tableau 75 et la Figure 55.

Nombre de mots	Nombre de segments	Durée moyenne des segments (s)	Durée moyenne des mots (s)
1	13561	0,42	0,42
2	10097	0,74	0,37
3	9353	1,02	0,34
4	9022	1,29	0,32
5	7843	1,56	0,31
6	7532	1,86	0,31
7	6809	2,13	0,30
8	6285	2,45	0,31
9	5367	2,70	0,30
10	4634	2,99	0,30
11	4048	3,27	0,30

12	3413	3,56	0,30
13	3088	3,83	0,29
14	2570	4,14	0,30
15	2146	4,42	0,29
16	1747	4,69	0,29
17	1521	4,92	0,29
18	1203	5,19	0,29
19	1052	5,50	0,29
20	868	5,71	0,29
21	662	5,97	0,28
22	506	6,17	0,28
23	402	6,53	0,28
24	318	6,94	0,29
25	289	7,17	0,29
26	179	7,35	0,28
27	154	7,46	0,28
28	125	7,96	0,28
29	98	8,30	0,29
30	64	8,22	0,27
31	47	8,46	0,27
32	41	8,82	0,28
33	31	9,10	0,28
34	18	9,68	0,28
35	18	9,31	0,27
36	18	10,14	0,28
37	10	9,90	0,27
38	11	10,63	0,28
39	2	11,23	0,29
40	3	9,22	0,23
41	6	10,35	0,25
42	5	12,54	0,30
43	3	12,06	0,28
44	4	12,60	0,29
45	1	10,16	0,23
49	1	11,31	0,23
50	2	13,64	0,27
60	1	15,69	0,26

Tableau 75 : Statistique du nombre de segments et la durée moyenne de mots en fonction de la longueur du segment.

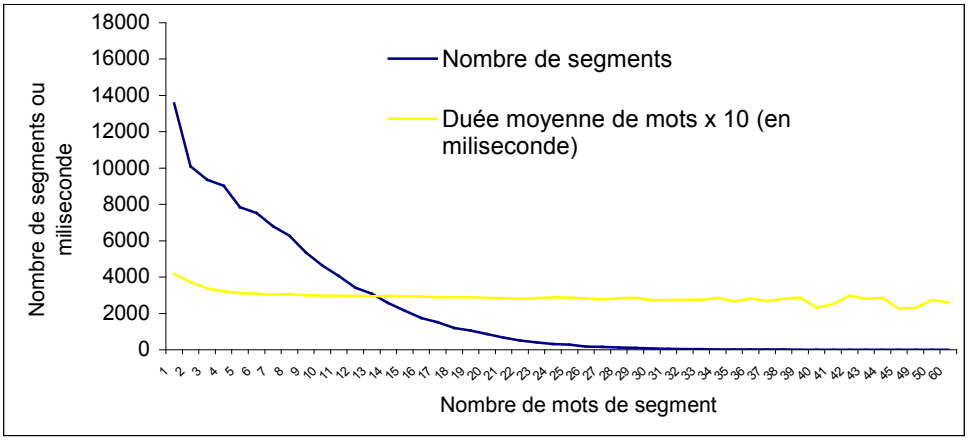

Figure 55 : Distribution des segments et des longueurs moyennes de mots en fonction de la longueur de segments

Le Tableau 76 montre les résultats obtenus du processus de détection des émotions contenus dans ce corpus TREC 2003 de 12h de journal.

Etat émotionnel	***Nombre de segments***
Colère	9103
Joie	24536
Neutre	66077
Tristesse	5462

Tableau 76 : Distribution en 4 états émotionnels

En conclusion, dans le contexte du corpus TREC 2003, premièrement cette approche nous permet d'extraire des segments de la parole où l'ensemble de mots qui sont naturellement attachés. Deuxièmement, avec cette approche, nous pouvons nous adapter avec la vitesse actuelle du locuteur. Cependant, l'inconvénient de cette approche est qu'il exige une annotation au niveau des mots correspondante avec le signal acoustique, donc il ne peut s'appliquer que dans les systèmes qui possèdent une transcription du signal ou un moteur de reconnaissance automatique de la parole. Pour notre cas, nous utilisons cette approche pour segmenter des segments du corpus TREC 2003, celui qui est fourni pour des campagnes de compétition dans la recherche d'information et celui qui possède les deux types de données nécessaires : le signal acoustique et la transcription. Comme nous avons mentionné, le corpus TREC 2003 contient 12h de journal télévisé CNN et ABC, et une partie importante de ce document contient le signal de parole des présentateurs et aussi de la publicité qui sont normalement en neutre et en joie. D'après nous, c'est la raison principale pour laquelle il y a beaucoup de segments reconnus comme neutre et joie dans le Tableau 76. Nous voudrions rappeler que le modèle que nous avons utilisé pour l'indexation sur le corpus TREC 2003 est le modèle entrainé par le corpus BES et la meilleure performance de la reconnaissance des émotions inter-langues entre l'allemand et le danois que nous pouvons obtenir est de 62,1 % avec les quatre émotions.

7.5. Conclusion

Encore une fois, bien que dans ce chapitre nous travaillions aussi avec le corpus BES dont la taille est un peu plus grande que celle du corpus DES, les expressions émotionnelles de ces deux corpus sont toujours du type de simulation. Cela veut dire que l'état émotionnel dans chaque échantillon de ces deux corpus est normalement « pur » et avec une intensité assez élevée. Ce n'est plus le cas si on travaille avec les corpus de données réelles où on peut trouver toujours la co-existence de plusieurs états émotionnels et normalement avec une intensité plus faible ainsi que d'autres influences du côté de locuteur comme son âge, son sexe, son milieu social, etc. Avant de conclure ce chapitre et de passer dans un autre chapitre où nous construisons et travaillons avec un corpus plus naturel, nous voulons rappeler encore une fois que les résultats obtenus doivent être considérés dans notre contexte précis des corpus utilisés.

Pour conclure, dans ce chapitre, avec les corpus DES et BES, et afin de tester notre approche de la reconnaissance de l'émotion indépendante du locuteur par la normalisation symbolique en combinaison avec la sélection forcée séquentielle en avant et avec le modèle de mélange des gaussiennes qui est expérimentalement montré le plus efficace par les données de corpus DES, nous avons utilisé les résultats obtenus avec ce corpus DES pour expérimenter sur le corpus BES qui possède un plus grand nombre de locuteurs (voir Chapitre 4).

Bien que le nombre d'émotions du corpus BES soit plus important que celui du corpus DES, le résultat obtenu avec le corpus BES est sensiblement meilleur que celui obtenu sur le corpus DES (62,1 % contre 57,3 %). Cela s'explique peut-être par le fait que les énoncés du corpus BES sont plus exagérés par les locuteurs comme décrit dans le Chapitre 4 (des amateurs ont aussi été utilisés pour interpréter les énoncés émotionnels du corpus BES) mais il nous semble que c'est la taille du corpus qui a le plus d'influence sur la performance du système. Effectivement, malgré la normalisation, avec les 4 locuteurs (2 hommes/2 femmes) du corpus DES, en appliquant la méthode « *Leave One Out* » au niveau du locuteur, le petit nombre de locuteurs restants et le déséquilibre hommes/femmes dégradera la qualité des modèles entrainés. Cela se montre également par les résultats obtenus par [Xiao, 2008] et nos résultats dans le cas de reconnaissance croisée entre les deux langues : nous obtenons 59,3 % en entrainant les modèles par les données du corpus BES (grand nombre de locuteurs), et 54,3 % en entrainant les modèles les modèles par les données du corpus DES (petit nombre de locuteurs).

Chapitre 8. **Corpus vietnamien**

Durant ces dernières années, des études sur la reconnaissance et la synthèse de l'émotion dans la parole ont été effectuées dans plusieurs langues : l'allemand [Quast 2002], le danois [Engberg et al, 1996], le chinois [Jiang et al, 2004], le coréen [Chung 2000], l'anglais [Chung 2000], le français [Devillers et al, 2004] [De Abreu et al, 2006] le hongrois [Fek et al, 2004], le japonais [Iida 2002]… mais on ne trouve pas encore de telles études sur la langue vietnamienne. Celle-ci est une langue tonale dans laquelle la prosodie de la phrase est fortement influencée par le ton des mots composants. Par ailleurs, les trois corpus que nous avons présentés ci-dessus sont construits en utilisant des acteurs et des actrices : bien que les processus de construction soient très rigoureux, la parole produite n'est pas aussi naturelle que celle obtenue à partir de conversations quotidiennes ou à partir de films. Le corpus VnEm est donc construit d'une part pour étudier comment l'état émotionnel du locuteur est exprimé dans la parole spontanée et comment l'auditeur identifie les émotions à travers les traits prosodiques et d'autre part pour des travaux futurs pour savoir si le ton joue un rôle important pour l'expression émotionnelle dans une langue tonale. Les caractéristiques du ton et de la prosodie seraient intéressantes pour améliorer la qualité de la parole synthétique, pour améliorer la performance des systèmes de reconnaissance de la parole ou pour améliorer la qualité de l'indexation en vietnamien.

8.1. Acquisition des données.

Le corpus VnEm est construit par extraction des segments à partir des trois films en DVD : « Mùa len trâu », « Nụ hôn thần chết », et « Cú và chim se sẻ ».

Ces trois films sont de type drame. Ces films étaient compressés sous la forme DivX (l'outil se trouve sur http://www.divx.com/) pour faciliter le travail de l'extraction sur l'ordinateur. Le logiciel FadeToBlack (http://www.thoughtman.com/) d'édition des vidéos en DivX avec des fonctions permettant un positionnement au niveau de la trame a été utilisé.

Les clips extraits sont ensuite réencodés et enregistrés sous la forme wmv, vidéo : 640 x 480 @ 30 fps, 1150 kbps et audio : 48000 Hz stéréo 16 bits, 192 kbps. La vidéo est uniquement utilisée pour aider les annotateurs dans leur travail.

Le signal audio est ensuite extrait à partir de la vidéo et enregistré en format PCM 16000Hz mono 16 bits pour les expérimentations.

La sélection des segments dans les films pour l'extraction est réalisée par une seule personne qui regarde le film, écoute le son et choisit des repères des segments. Cette personne doit compléter la segmentation (l'extraction des segments) d'un film avant de pouvoir travailler avec un autre. Avant l'extraction, elle a du entièrement revoir ce film afin d'avoir une vue globale et ainsi pour bien positionner des scènes du film.

Les segments extraits peuvent contenir un seul mot comme « có » hoặc « không » (« oui » ou « non » en français) ou être plus longs comme une phrase ou un passage continu. Le Tableau 77 fournit des statistiques sur la longueur des segments.

Films	*Nombre de segments*	*Longueur moyenne en seconde*
Mùa len trâu	782	1.93
Nụ hôn thần chết	1021	2.18
Cú và chim se sẻ	489	2.63
Total	2292	~ 5021 secondes ~ 84 minutes

Tableau 77 : Nombre et longueur moyenne des segments du corpus VnEm

Les données video et audio de ce corpus peuvent être accédées par l'adresse : http://mrim.imag.fr/corpus/VnEm

Chaque segment peut être retrouvé dans le film grâce à un fichier TimeCode.xml contenant les repères de la position de ce segment dans le film.

8.1.1. Format du fichier TimeCode

La séquence suivant illustre le format du fichier TimeCode :

<file annot="" start="00:04:54.40" end="00:04:55.91" score="0001141613">.Fem01001.wav</file>

où Fem01001.wav est le nom du fichier contenant le segment extrait ; les trois lettres Fem indiquent qu'il s'agit d'une locutrice ; les deux chiffres juste après : 01 indiquent que c'est la première locutrice apparaissant dans le film et enfin les trois derniers chiffres 001 indiquent que le segment Fem01001.wav est le premier segment extrait pour cette locutrice. Ces numéros ne sont pas forcément successifs car des segments ont pu être supprimés lors de la revue par l'auditeur en raison de leu faible qualité.

L'attribut « annot » est prévi pour contenir la transcription manuelle de ce qui est dit dans le segment extrait, cette transcription peut être remplie avec le logiciel utilisé (voir la section Annotation ci-dessous). Par manque de temps, cet attribut n'a pas encore été rempli pour la plupart des segments.

Les attributs « start » et « end » contiennent respectivement les points de début et le point de fin du segment dans le film complet selon le format : HH:MM:SS.CS avec HH pour heure, MM pour minute, SS pour seconde et CS pour centiseconde.

L'attribut « score » contient le score le plus élevé de nos modèles de reconnaissance et comme les autres attributs (voir le fichier TimeCode.xml dans le corpus), ces données sont utilisées seulement par l'auteur, elles peuvent donc être ignorées.

8.1.2. La structure du corpus

Le corpus se compose au total de trois grands répertoires correspondants aux trois films. Chaque grand répertoire contient 7 sous-répertoires.

Le répertoire « Annotations » contient les fichiers des évaluations brutes de 10 évaluateurs. Ces 10 fichiers peuvent être utilisés ou modifiés avec l'outil accompagnant dans le répertoire « EvaluationTool ». Le format de ces fichiers est maintenu compatible par l'outil : les deux premières lignes sont des informations qui servent à l'outil et qui ne doivent pas être modifiées manuellement ; le reste se compose de 8 colonnes correspondant aux 8 évaluations pour chaque segment pour chaque état émotionnel en ordre : Colère, Ennui, Dégoût, Peur, Joie, Neutre, Tristesse, Surprise. Les valeurs pour l'évaluation vont de 0 à 5 pour les 8 colonnes sauf pour la colonne du neutre qui ne comporte que les valeurs binaires 1 ou 0 pour « Neutre » ou « Non-neutre ».

Les répertoires « Audio16 » et « Audio48 » contiennent les mêmes signaux audio pour tous les segments extraits. Les différences sont la fréquence d'échantillonnage des signaux et le nombre de canaux : 16000 Hz mono et 48000 Hz stéréo (format original) respectivement.

Le quatrième répertoire « CompleteAudio » contient un fichier audio du film complet en 16000 Hz mono.

Le répertoire « EvaluationTool » contient un petit outil appelé AnnotEm (fichier Evaluation.exe) et un fichier « Input_FileList.xml ». Ce fichier liste tous les segments du film ainsi que l'emplacement de ces segments. Il est utilisé par AnnotEm et son nom ne doit pas être modifié.

Le fichier « TimeCode.xml » se trouve dans le répertoire « TimeCode ».

Le répertoire « Video » contient tous les clips video correspondants aux signaux audio présents dans les répertoires « Audio16 » et « Audio48 »

8.2. Locuteurs

Lors de la sélection manuelle des segments de films, nous avons constaté que les quantités de segments de voix masculines et féminines n'étaient pas équilibrées. Par exemple, le film « Mùa len trâu » contient essentiellement des voix masculines. La plupart de temps également, ce sont les personnages principaux qui apparaissent et qui parlent. En conséquence, une grande partie du corpus contiendra les voix de ces personnages. Le Tableau 78 indique le nombre de locuteurs et de segments dans chaque film et la part occupée par les personnages principaux.

Films	*Nombre de locuteurs / Nombre de segments*	*Nombre de locutrices / Nombre de segments*	*Nombre de personnages principaux / Nombre de segments*
Mùa len trâu	18 locuteurs pour 653 segments	3 locutrices pour 129 segments	1 locuteur pour 193 segments
Nụ hôn thần chết	17 locuteurs pour 546 segments	8 locutrices pour 475 segments	1 locuteur pour 244 segments 1 locutrice pour 310 segments
Cú và chim se sẻ	7 locuteurs pour 329 segments	9 locutrices pour 160 segments	1 locuteur pour 178 segments 1 locutrice pour 94 segments
Total	42 locuteurs pour 1528 segments (67%)	20 locutrices pour 764 segments (33%)	5 personnages principaux pour 1019 segments (44%)

Tableau 78 : Nombre de locuteurs et de locutrices du corpus VnEm

Au total, nous avons 62 locuteurs et locutrices pour 2292 segments dans le corpus. Les segments produits par des personnages principaux du corpus représentent 44%. Ce déséquilibre est un des inconvénients des corpus extraits et les études sur ce type de corpus doivent ternir compte de ce problème.

Tous les locuteurs de ces corpus sont des adultes. Les segments correspondant à des personnages mineurs ont été éliminés afin d'éviter des problèmes causés par la différence entre la voix des adultes et la voix des enfants. Le Tableau 79 donne quelques informations sur les acteurs (et les actrices) principaux dans les films de notre collection, elles sont retenues à partir de page de web : http://www.imdb.com

L'âge des acteurs et des actrices est compté à la date de sortie du film, pour ce corpus l'âge moyen des acteurs (et des actrices) principaux est environ de 27 ans.

Films	*Nom du locuteur*	*Sexe*	*Age*
Mùa len trâu	The Lu Le	Homme	20
Nụ hôn thần chết	Tri Nguyen	Homme	33
	Thanh Hang	Femme	26
Cú và chim se sẻ	Cat Ly	Femme	~27
	The Lu Le	Homme	23
Moyenne			27

Tableau 79 : Ages des acteurs (et actrices) principaux du corpus VnEm

8.3. Annotation

Comme mentionné ci-dessus, pour valider l'état émotionnel des énoncés, nous avons utilisé l'annotation perceptive. L'annotation perceptive sur les états émotionnels est effectuée avec des choix obligatoires dans la majorité des études sur l'émotion, comme [Schlosberg 1952], [Engberg et al, 1996], [Burkhardt 2005] (voir Chapitre 2). Cela veut dire l'évaluateur a comme seule possibilité de répondre « Oui » ou « Non » aux questions posées sur l'état émotionnel de

locuteur ce qu'il peut ressentir en écoutant cette expression. Il s'agit d'une évaluation bipolaire. Ce type d'évaluation est bien adapté pour le deuxième groupe de corpus de simulation où chaque expression a été conçue et interprétée pour une seule émotion. L'avantage de ce type d'évaluation est son caractère explicite qui facilite beaucoup les travaux se basant sur ces données. Mais, comme mentionné précédemment, à cause de la possible coexistence de plusieurs émotions dans un même segment, et parce que l'intensité des émotions dans les énoncées n'est pas toujours la même, l'évaluation bipolaire des émotions dans les énoncés n'est pas très adaptée dans le cas présent.

Récemment, on peut trouver dans les travaux avec le corpus HUMAINE [Douglas-Cowie et al, 2007] les techniques permettant d'annoter les données multimodales ainsi que de tenir compte de plusieurs dimensions qui pourraient être utiles pour les études de l'émotion avec les données réelles. Les auteurs ont effectués les deux niveaux d'annotation : les descripteurs globaux et les descripteurs « locaux ». Les descripteurs globaux comprennent les informations qui ne varient pas rapidement comme l'information de contexte, l'état émotionnel lié, le type de combinaison des émotions, l'événement « clef », les mots émotifs fréquemmnent utilisés, et les catégories de l'évaluation perceptive (« appraisal » en anglais, [Devillers et al, 2006]). Les huit descripteurs locaux sont les estimations des évaluateurs de la personne en question sur : l'intensité exprimée, la manifestation des états émotionnels, la dissimulation des états émotionnels, le niveau actif, le niveau positif / négatif, le contrôle de la situation, la surprise de l'événement, et l'intensité de chaque émotion (voir [Douglas-Cowie et al, 2007] pour le détail).

Pour notre travail, afin d'obtenir des jugements plus adaptés, nous définissons une échelle à 6 niveaux pour toutes les émotions. On demande aux évaluateurs de juger si un segment contient une certaine émotion et de donner une note allant de 0 à 5 selon sa perception de celle-ci. Parce que l'état Neutre est binaire, nous avons demandé aux évaluateurs de ne pas choisir des autres états si le Neutre est coché et vice versa, ils ne doivent pas choisir Neutre si au moins un de 7 autres états est coché.

La Figure 56 montre l'interface de notre outil avec les choix selon l'échelle retenue pour chaque émotion. Elle permet à l'utilisateur de revenir en arrière ou d'avancer s'il veut modifier ses jugements. Comme mentionné précédement, la transcription de la parole dans les segments (l'attribute « annot » dans le fichier TimeCode.xml) peut éventuellement être manuellement insérée ici.

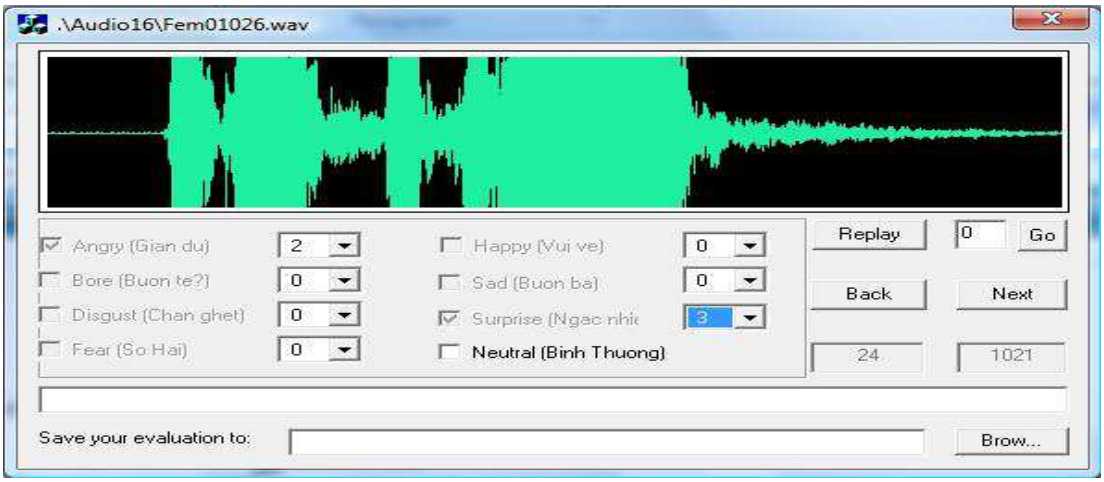

Figure 56 : AnnotEm, l'outil pour l'annotation

Pour chaque segment, le résultat du processus est une liste de jugements pour toutes les émotions. Ces évaluations seront utilisées pour classer et éliminer des segments ne possédant pas une intensité suffisante.

Le Tableau 80 montre un exemple des jugements d'un auditeur.

Expression	***Colère***	***Ennui***	***Dégoût***	***Peur***	***Joie***	***Neutre***	***Tristesse***	***Surprise***
Exp.1.	1	0	0	0	0	0	0	0
Exp 2.	0	0	0	0	3	0	0	2
….								
Exp 1391	0	0	0	0	0	1	0	0
Exp 1392	5	0	0	0	0	0	0	0

Tableau 80 : Evaluation d'un évaluateur

Pour ce corpus VnEm, dix personnes (7 hommes et 3 femmes) ont été invitées à évaluer l'état émotionnel des locuteurs dans les 2292 segments. Tous les auditeurs sont des étudiants vietnamiens de 20 à 30 ans, l'âge moyen étant de 23,7 ans (Tableau 81).

	Evaluateur	***Sexe***	***Age***
1	LE X.C.	Homme	23
2	DOAN T.N.H.	Femme	26
3	NGUYEN H.	Homme	20
4	NGUYEN V.S.	Homme	26
5	NGUYEN T.H.	Femme	21
6	HUYNH C.P.	Homme	30
7	NGA	Homme	23
8	LE X. H.	Homme	26
9	LE M. H	Homme	21
10	NGUYEN T.T.	Femme	21
	Moyen		23,7

Tableau 81 : Evaluateurs pour le corpus VnEm

L'objectif de la recherche et les étapes du processus d'évaluation ont préalablement été expliqués aux évaluateurs. Avant l'écoute, il a été demandé aux évaluateurs de juger de l'état émotionnel selon leurs sensations et non pas en s'appuyant sur le contenu sémantique de la parole, cependant les évaluations obtenues étaient influencées par le contenu sémantique auquel les auditeurs ont naturellement réagi, les segments de parole étant dans leur langue maternelle. Environ 10 évaluations d'entraînement ont été effectuées afin que l'auditeur s'habitue au processus et qu'il ait une idée de l'intervalle des valeurs de l'échelle d'évaluation. L'évaluateur a eu suffisamment de temps pour effectuer tous les jugements (il n'y avait aucune limite de temps pour l'évaluation et l'évaluateur décidait de son rythme). Les données audio du corpus ainsi que l'outil d'annotation AnnotEm ont été fournis aux évaluateurs pour qu'ils puissent effectuer les évaluations chez eux. Pour chaque auditeur, les évaluations ont duré en moyenne 4 jours pour ce corpus VnEm.

8.4. Expérimentations

8.4.1. Traitement des annotations brutes

Comme mentionné précédemment, les jugements par test perceptif dépendent beaucoup de l'auditeur : ils dépendent de sa sensibilité et de ses habitudes mais aussi de son état émotionnel. Ceci est démontré par le fait que l'amplitude de jugements produits varie d'un auditeur à un autre. Nous pouvons le constater dans la vie quotidienne : une personne joyeuse voit les événements d'une manière plus positive qu'une personne triste et elle donnera probablement plus de points pour l'évaluation des émotions positives.

Afin de limiter l'influence de la subjectivité sur l'évaluation ou, autrement dit, pour limiter la différence d'amplitude des jugements inter-auditeurs, nous effectuons une normalisation en moyenne et en écart type des jugements de chaque auditeur. Pour chaque émotion, la moyenne et l'écart type sont calculés pour chaque auditeur et pour l'ensemble des auditeurs (nous les appelons : la moyenne générale et l'écart type général). La moyenne et l'écart type de chaque auditeur sont ramenés à la moyenne et l'écart type global par l'application à ses jugements pour l'émotion considérée de la transformation affine :

$$JA'(i) = \mu(E) + (JA(i) - \mu A(E)).(\sigma(E)/ \sigma A(E)) \quad (6.7)$$

Avec :

$\mu(E)$: moyenne des notes attribuées pour l'émotion E par l'ensemble des auditeurs,

$\mu A(E)$: moyenne des notes attribuées pour l'émotion E par l'auditeur A,

$\sigma(E)$: moyenne des notes attribuées pour l'émotion E par l'ensemble des auditeurs,

$\sigma A(E)$: moyenne des notes attribuées pour l'émotion E par l'auditeur A,

JA(i) : jugement de l'auditeur A pour l'expression i de l'émotion E avant la normalisation, l'intervalle de JA(i) est clairement de 0 à 5.

JA'(i) : jugement de l'auditeur A pour l'expression i de l'émotion E après la normalisation, JA'(i) est entier, et son intervalle est aussi de 0 à 5. Cela veut dire qu'après la normalisation, toutes les valeurs de JA'(i) ont été arrondies, et celle inferieures à 0 sont considérées comme 0, celles plus supérieures à 5 prendront la valeur 5.

Après cette normalisation, tous les auditeurs ont presque la même moyenne et le même écart type pour leurs évaluations. En réalité, il existe effectivement une petite différence de la moyenne et de l'écart type entre des auditeurs, elle peut s'expliquer par l'arrondissement des JA'(i). La Figure 57 montre un exemple de la distribution des notes d'un auditeur pour la colère avant et après la normalisation. La moyenne des notes de cet auditeur pour la colère est supérieure à la moyenne générale (2,32 contre 2,21), et cet auditeur possède aussi une valeur de l'écart type plus large que celle générale (1,29 contre 1,09). Apparemment, cette personne était plus énervée que des autres au moment de jugement (valeur moyenne plus élevée), et elle est également plus sensible (valeur d'écart type plus large).

En comparant la Figure 57a et la Figure 57b, nous pouvons facilement constater l'effet de la normalisation : les notes de cet auditeur pour la colère sont devenues plus petites pour abaisser sa valeur moyenne. Les valeurs des notes ont été également devenues plus proches les unes des autres pour réduire l'écart type.

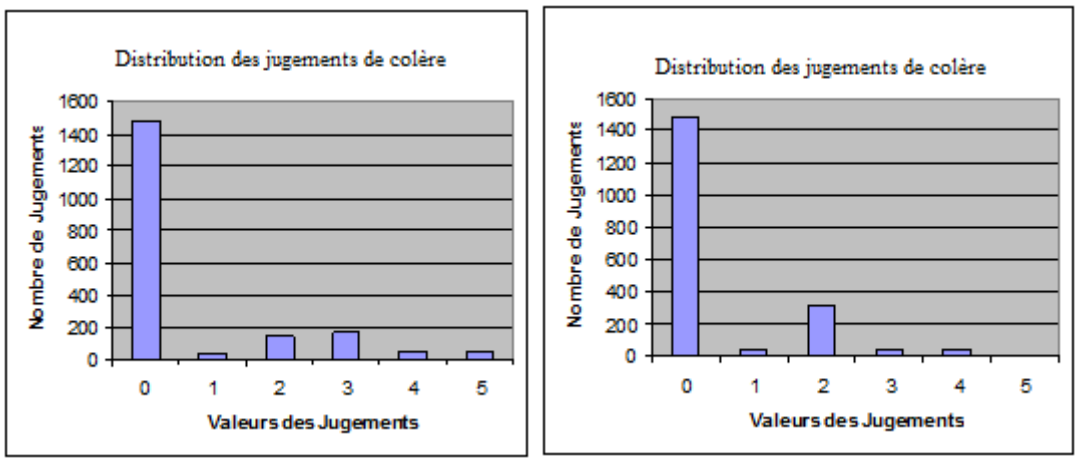

Figure 57 : La distribution des jugements avant (a) et après (b) la normalisation de valeur moyenne.

8.4.2. Seuillage des annotations

Comme précédemment mentionné, la sensation dépend beaucoup de chaque personne, de son caractère, de son état émotionnel actuel, si on ne veut pas lister ici plusieurs autres facteurs comme : l'âge, le sexe masculin ou féminin, le niveau intellectuel, la culture, etc. C'est la raison pour laquelle, la validation des états émotionnels dans les expressions doit prendre en compte l'accord entre les auditeurs. Jusqu'ici, nous avons des notes de jugements normalisés selon 6 niveaux qui s'étendent de 0 à 5. Dans la partie suivante, nous présenterons notre approche pour le seuillage et pour la sélection des expressions.

Clairement, les évaluations de l'existence d'une émotion dans un énoncé par plusieurs auditeurs ne sont pas toujours identiques : certains auditeurs sont capables de reconnaître cette émotion mais d'autres ne reconnaissent pas sa présence dans cet énoncé.

Nous pouvons appliquer une règle de majorité, soit par le nombre d'accords entre les évaluateurs, soit par la note moyenne des évaluateurs. Par exemple, nous présentons ici le cas où le choix est réalisé en se basant sur le nombre d'accords : parmi les notes données pour un énoncé i, si la majorité des notes sont supérieures à 0, nous considérons l'énoncé i comme une expression en cette émotion. La majorité est d'au moins 6 auditeurs en accord (6/10 auditeurs). Avec ce critère, et sans tenir en compte de la coexistence de plusieurs émotions dans un énoncé, nous avons obtenus des accords pour 1173 énoncés selon les 8 états émotionnels : 51 pour la joie, 653 pour le neutre, 101 pour la colère, 132 pour la tristesse, 25 pour la peur, 10 pour l'ennui, 98 pour dégoût et 103 pour la surprise.

Jusqu'ici, nous avons donc des résultats comparables avec ceux des systèmes utilisant l'évaluation bipolaire. En effet, avec cette règle, pour chaque émotion, nous avons deux classes : une pour des énoncés contenant cette émotion et l'autre pour des énoncés qui ne la contiennent pas. L'annotation à ce niveau peut servir aux études sur la discrimination : Emotion / Neutre comme la discrimination entre la Colère et le Neutre, la Surprise et le Neutre, etc.

8.4.3. Détection des états émotionnels avec les paramètres interlangues.

Nous avons expérimenté la détection des états émotionnels interlangue avec les trois corpus DES, BES et Orator. Avec ces corpus, nous avons trouvé que la détection croisée des émotions primaires entre le dannois et l'allemand est possible.

Pour la même raison, nous avons essayé d'utiliser les modèles GMM 128 gaussiennes entrainés par les corpus DES et BES et aussi d'utiliser les 9 paramètres obtenus à partir du chapitre 7 pour détecter l'état émotionnel dans les énoncés du corpus VnEm. Dans cette expérimentation, seules les données des personnages principaux (3 hommes et 2 femmes) du corpus VnEm sont utilisées car l'utilisation de la normalisation symbolique a besoin beaucoup de données de chaque locuteur pour extraire les informations statistiques avant la normalisation.

Les résultats que nous avons obtenus ne sont pas beaucoup plus élevé que ceux du choix aléatoire : 22,5% en moyenne pour la détection des 5 états émotionnels : la colère, la joie, le neutre, la surprise et la tristesse, en utilisant les modèles entrainés par le corpus DES pour faire le test sur le corpus VnEm et 16,4% en moyenne pour la détection des 7 états émotionnels : le neutre, la colère, la peur, la joie, la tristesse, le dégoût et l'ennui, en utilisant le modèle entrainé sur le corpus BES pour faire le test sur le corpus VnEm (voir le Tableau 82). En observant les résultats obtenus dans la dernière case, nous constatons que, la plupart des énoncés sont reconnus comme l'état Peur. A notre avis, c'est la différence entre une langue tonale et une langue non-tonale qui est une des causes de cette confusion et du faible taux de reconnaissance. Effectivement, dans le cas de la discrimination émotion/neutre et en travaillant avec les personnages principaux, pour le cas de la reconnaissance independante du locuteur et avec le protocole « Leave One Out », nous trouvons que le taux de reconnaissance est amélioré d'environ 9 % si nous utilisons des données dans la même langue. Le Tableau 83 et le Tableau 84 donnent les résultats pour ces deux cas de discrimination Emotion/Neutre.

Nous constatons que dans le cas où on utilise les données du corpus VnEm pour entrainer le modèle de reconnaissance, l'amélioration porte essentiellement sur la détection de l'état neutre, l'état qui n'est pas bien reconnu par les modèles de l'autre langue.

	Joie	***Neutre***	***Colère***	***Tristesse***	***Peur***	***Ennui***	***Dégoût***
Joie	5,3	13,2	2,6	0	50	0	28,9
Neutre	7,0	5,0	0	0,6	70,8	0	16,7
Colère	18,3	0	1,4	0	73,2	0	7,0
Tristesse	12,7	1,0	1,0	2,0	68,6	0	14,7
Peur	0	0	5,3	0	73,7	0	21,1
Ennui	0	0	0	0	100,0	0	0
Dégoût	8,2	0	1,4	0	63,0	0	27,4
Moyenne	16,4%						

Tableau 82 : Détection de l'état émotionnel des locuteurs vietnamiens en utilisant les modèles entrainés par le corpus BES en allemand.

	Toutes les émotions	***Neutre***
Toutes les emotions	87	13
Neutre	79	21
Moyenne	53 %	

Tableau 83 : Résultat de la discrimination de l'état Emotion / Neutre des énoncés en vietnamien en utilisant les modèles entrainés par le corpus BES en allemand.

	Toutes les emotions	***Neutre***
Toutes les emotions	85	15
Neutre	66	34
Moyenne	62 %	

Tableau 84 : Résultat de la discrimination de l'état Emotion / Neutre des énoncés du corpus VnEm

Comme il a déjà été observé dans le chapitre précédent, les paramètres les plus efficaces dans des langues différentes sont différents. De plus, le vietnamien est une langue tonale. Il est donc compréhensible que l'application de l'ensemble de 9 paramètres obtenus sur DES et BES donne le faible taux de 62% pour la détection émotion/neutre.

Dans la section suivante, nous allons chercher un ensemble de paramètres efficaces pour la détection des émotions en vietnamien.

Selon les modèles théoriques que nous avons présentés dans le chapitre 2, le degré de l'activité est une des deux dimensions la plus étudiées, ce sera donc interessant d'étudier les émotions du corpus VnEm dans cette dimension. De plus, nous voulons savoir si dans cette dimension le ton a aussi une influence.

Parmi les 8 états émotionnels du corpus VnEm, nous n'avons pas utilisé les données de la surprise car la surprise n'est pas encore certaine (voir chapitre 2). La joie est un état émotionnel primaire selon les modèles du chapitre 2, mais elle n'est pas utilisée dans ce cas car elle est le seul état positif et la fusion de la joie avec la colère + la peur pourrait causer la confusion. Nous n'avons pas non plus assez de données pour cet état pour le mettre seul dans un groupe. Les autres émotions sont regroupées en trois classes en raison de manque de données : la classe des émotions « fortes », le neutre et la classe des émotions « faibles », qui correspondent aux trois groupes : la colère + la peur, le neutre, et la tristesse + l'ennui + le dégoût.

8.4.4. Détection de trois classes émotions « fortes »/neutre/émotions « faibles »

La première étape pour cette expérimentation est la sélection des échantillons qui contiennent effectivement des états émotionnels intéressés.

Comme nous avons mentionné précédemment, notre sélection est basée sur une majorité, soit par le nombre d'accords entre des évaluateurs, soit par la moyenne des évaluations. Ces deux approches donnent les mêmes résultats pour le cas du neutre car les évaluations sur le neutre ne possèdent que la valeur binaire (0 ou 1) ; c'est pourquoi la moyenne obtenue correspond bien aux nombre d'accords.

Il y a une autre remarque pour la sélection que nous avons faite qui est que : en raison du manque de données pour quelques états émotionnels et pour obtenir un nombre d'énoncés assez

équilibré entre les trois classes étudiées, des critères différents ont été appliqués pour les états émotionnels différents.

Plusrécisément, pour le neutre, nous n'avons retenu que les énoncés qui ont l'accord d'au moins 9 évaluateurs. Ce critère nous permet d'obtenir 111 énoncés en état neutres pour les trois films. Pour la colère et la peur, nous avons choisi les énoncés dont la note moyenne des évaluations est supérieure à 2,8. Ce critère nous donne au total 46 échantillons pour la classe des émotions fortes (30 énoncés en colère et 16 énoncés en peur). Pour la tristesse, l'ennui et le dégoût, une moyenne minimum de 2,0 a été appliquée et nous avons obtenu 49 échantillons au total (35 pour la tristesse, 0 pour l'ennui et 14 pour le dégoût).

Une sélection aléatoire a été faite parmi les échantillons obtenus dans les deux classes : neutre et émotions faibles pour que chaque classe de ces deux classes ne contienne que 46 échantillons comme la classe d'émotions fortes.

Au total, nous avons les données de 23 locuteurs différents pour les trois classes, 15 locuteurs pour la classe des émotions fortes, 11 locuteurs pour le neutre, et 5 locuteurs pour la classe des émotions faibles. Il est facile de constater qu'il y a des locuteurs dont les données apparaissent dans au moins deux classes différentes.

L'approche que nous avons utilisée pour cette expérimentation est la reconnaissance multi-locuteur avec le protocole « Leave One Out » et le modèle GMM avec 128 gaussiennes.

Le Tableau 85 et le Tableau 86 montrent la performance des paramètres MFCCs, de l'intensité ainsi que la fréquence fondamentale en appliquant séparément chaque parmètre pour la discrimination de ces trois classes.

Paramètres	***Taux de reconnaissance***	***Paramètres***	***Taux de reconnaissance***	***Paramètres***	***Taux de reconnaissance***
MFCC0	*63,8*	ΔMFCC0	42,8	ΔΔMFCC0	47,1
MFCC1	52,9	ΔMFCC1	33,3	ΔΔMFCC1	35,5
MFCC2	49,2	ΔMFCC2	39,1	ΔΔMFCC2	36,2
MFCC3	23,9	ΔMFCC3	42,8	ΔΔMFCC3	35,5
MFCC4	46,4	ΔMFCC4	48,6	ΔΔMFCC4	29,7
MFCC5	42,8	ΔMFCC5	38,4	ΔΔMFCC5	38,4
MFCC6	50,7	ΔMFCC6	45,7	ΔΔMFCC6	41,3
MFCC7	48,6	ΔMFCC7	45,7	ΔΔMFCC7	40
MFCC8	44,9	ΔMFCC8	42	ΔΔMFCC8	42
MFCC9	46,4	ΔMFCC9	42	ΔΔMFCC9	37,6
MFCC10	42,8	ΔMFCC10	39,1	ΔΔMFCC10	39,9
MFCC11	45,7	ΔMFCC11	46,4	ΔΔMFCC11	39,9
MFCC12	43,5	ΔMFCC12	44,2	ΔΔMFCC12	39,1
MFCC13	41,3	ΔMFCC13	45,7	ΔΔMFCC13	37,6
MFCC14	32,6	ΔMFCC14	43,5	ΔΔMFCC14	37,6
MFCC15	34,1	ΔMFCC15	44,9	ΔΔMFCC15	33,3
MFCC16	42	ΔMFCC16	45,7	ΔΔMFCC16	31,9

Tableau 85 : Taux de reconnaissance les trois classes d'émotions en utilisant les coefficients MFCCs séparemment.

Paramètres	*Taux de reconnaissance*
F0	*58,7*
Δ F0	*43,5*
ΔΔ F0	*45,7*
Intensité	62,1
Δ Intensité	*25,4*
ΔΔ Intensité	*24,1*

Tableau 86 : Taux de reconnaissance les trois classes d'émotions en utilisant les coefficients F0 et l'intensité séparemment.

Comme nous pouvons le constater, l'intensité et MFCC0 donnent les meilleurs résultats, cela peut s'expliquer par le fait que nous avons choisi ces trois classes en se basant sur l'échelle « valence » des émotions. Pourtant, selon le Tableau 86, la fréquence fondamentale donne aussi un taux assez élevé (58,7% contre 33,3% pour la réponse aléatoire, voir la section 5.3.1.4.1 pour la même conclusion).

Le Tableau 85 montre aussi que les dérivées secondes de MFCC sont moins efficaces que les dérivées premières, les dérivées premières sont moins efficaces que les MFCC originaux.

La fusion de tous ces paramètres nous donne un taux de reconnaissance de 78,9 % comme montré dans le Tableau 87.

	Emotions « fortes » (%)	***Neutre (%)***	***Emotions « faibles » (%)***
Emotions « fortes »	78.3	8.7	13.0
Neutre	23.9	67.4	8.7
Emotions « faibles »	6.5	2.2	91.3
Moyenne	79.0		

Tableau 87 : Taux de reconnaissance les trois classes d'émotions en utilisant la fusion de tous les paramètres.

Comme dans les expérimentations avec les autres langues, nous recherchons la combinaison la plus efficace pour détecter ces trois classes d'émotions. L'algorithme SFSA a été utilisé pour cette recherche sur ces 57 paramètres (51 pour MFCCs, 3 pour F0 et 3 pour l'intensité). Le coefficient MFCC0 est choisi pour initialiser l'ensemble de paramètres de combinaison et le Tableau 88 nous montre les résultats des étapes de sélection.

Nombre de paramètres	*Combinaison la plus efficaces des MFCCs*	*Taux de class.*
1	Z_1=$MFCC_0$	63,8
2	Z_2=Z_1+$MFCC_1$	73,2
3	Z_3=Z_2+$MFCC_{14}$	75,4
4	Z_4=Z_3+$MFCC_9$	76,8
5	Z_5= Z_4+$MFCC_4$	79,0
6	Z_6= Z_5+$\Delta MFCC_2$	79,0
7	Z_7= Z_6+ $\Delta MFCC_3$	79,0
8	Z_8= Z_7+ $MFCC_{15}$	80,4
9	Z_9= Z_8+ $MFCC_{16}$	81,9
10	Z_{10}= Z_9+ $MFCC_{13}$	82,6
11	Z_{11}= Z_{10}+ $\Delta MFCC_1$	83,3

Tableau 88 : Taux de reconnaissance en appliquant l'algorithme SFSA.

Nous constatons que les paramètres sélectionés se trouvent essentiellement dans l'ensemble des MFCCs originaux. Les trois paramètres de dérivée première des MFCCs qui ont été choisis sont ΔMFCC2, ΔMFCC3 et ΔMFCC1. Si nous revoyons sur le Tableau 42 portant sur les filtres lors de l'extraction des coefficients MFCC, nous constatons que les deux coefficients ΔMFCC2, ΔMFCC3 correspondent avec les filtres dont la bande de fréquences sont de 74 Hz à 247 z et de 156 Hz à 348 Hz. Ce sont les bandes de fréquences contenant la fréquence fondamentale de l'homme et de femme. Et comme nous pouvons constater dans le processus de sélection par l'algorithme SFSA dans le Tableau 88, bien que l'ajout de ces deux paramètres n'améliore pas immédiatement la performance du système, il nous permet de continuer l'algorithme et de trouver la combinaison efficace entre ces deux paramètres avec les paramètres suivants. La dernière combinaison nous donne une amélioration significative (de 79,0 % à 83,3 %). Le Tableau 89 détaille ce résultat.

	Emotions « fortes »	***Neutre***	***Emotions « faibles »***
Emotions « fortes »	80.4	8.7	10.9
Neutre	15.2	76.1	8.7
Emotions « faibles »	2.2	4.3	93.5
Moyenne	83.3 %		

Tableau 89 : Le taux de détection des trois classes d'émotions en utilisant 11 coefficients MFCCs trouvés.

En observant les résultats donnés dans le Tableau 89, nous constatons qu'il y a encore des confusions entre le neutre et la classe d'émotions fortes. A notre avis, pour une langue tonale, il serait possible que le ton, qui doit varier beaucoup même dans l'état neutre (pour tenir le sens des mots), soit une des raisons de chevauchement entre le neutre et les autres émotions.

Pour le test de l'utilité de la normalisation symbolique dans le cas de reconnaissance indépendante du locuteur avec la langue vietnamienne, à partir de l'ensemble de données

trouvées (46 échantillons pour la classe des émotions fortes, 49 échantillons pour la classe des émotions faibles et 111 échantillons pour le neutre), nous gardons seulement les données des personnages principaux et éliminons toutes les autres. En faisant cette élimination, nous obtenons au total 107 échantillons pour les trois classes : 24 échantillons de 3 locuteurs principaux pour la classe des émotions fortes, 19 échantillons de 2 locuteurs pour la classe des émotions faibles et 64 échantillons de 4 locuteurs pour le neutre. Le protocole « Leave One Out » a été utilisé en prenant un locuteur pour le test et tous les autres pour l'apprentissage.

Pour le résultat (Tableau 90 et Tableau 91), nous n'avons pas trouvé d'amélioration significative en utilisant la normalisation symbolique sur l'ensemble de 11 paramètres trouvés ci-dessus par rapport à l'utilisation directe ces 11 paramètres sans la normalisation. Cela peut s'expliquer par le déséquilibre dans les données des personnages principaux et aussi par la taille des données de ces personnages qui ne sont pas encore assez grande pour couvrir tous les variations possibles ; cela ne nous permet pas d'obtenir des informations statistiques suffisament bonnes pour une normalisation symbolique correcte ; il est par exemple rare de trouver le personnage principal du film « Mùa len trâu » en peur ou en neutre.

	Emotions « fortes »	***Neutre***	***Emotions « faibles »***
Emotions « fortes »	56.0	32.0	12.0
Neutre	54.2	25.0	20.8
Emotions « faibles »	10.5	5.3	84.2
Moyenne	52,9 %		

Tableau 90 : Le taux de reconnaissance indépendante du locuteur en utilisant 11 paramètres MFCCs originaux trouvés

	Emotions « fortes »	***Neutre***	***Emotions « faibles »***
Emotions « fortes »	48.0	36.0	16.0
Neutre	29.2	45.8	25.0
Emotions « faibles »	15.8	10.5	73.7
Moyenne	54,4 %		

Tableau 91 : Le taux de reconnaissance indépendante du locuteur en utilisant 11 paramètres MFCCs trouvés normalisés par la normalisation symbolique.

8.4.5. Conclusion

Dans ce chapitre, nous avons présenté le corpus VnEm qui est construit par l'extraction manuelle de segments de parole émotionnellement chargée dans trois films vietnamiens. Nous avons aussi présenté une approche pour l'annotation de ce corpus qui est réalisée en effectuant des tests perceptifs avec dix auditeurs vietnamiens. Le corpus est construit pour l'étude de l'état émotionnel « spontané » dans la parole mais aussi il peut être utile pour des études portant sur

les expressions émotionnelles dans une laugue tonale par rapport aux autres langues non-tonales.

Nous avons aussi effectué des experimentations sur la détection de trois classes d'émotions : émotions fortes / neutre / émotions faibles sur ce corpus VnEm. Nous constatons que les paramètres qui fonctionneny bien pour les corpus DES et BES ne donnent plus de bons résultats sur le corpus VnEm. L'optimalisation intra-langue nous permet d'obtenir un ensemble de 11 paramètres MFCCs donnant un taux de discrimination de 83,3 % pour les trois classes. Cela peut s'expliquer par le fait que les deux ensembles de données sont très différents en termes d'acquisition, d'enregistrement, de niveau de bruit et de niveau naturel. Mais, à notre avis, le ton crée aussi beaucoup de différences dans les expressions émotionnelles dans les langues tonales par rapport aux langues non-tonales. Enfin, avec ce corpus, nous espérons fournir des données nécessaires pour les études sur cet aspect.

Chapitre 9. **Conclusions et perspectives**

9.1. Contribution

Notre travail se place dans le cadre de la reconnaissance automatique de l'émotion dans les documents audiovisuels.

9.1.1. Etude des émotions

Comme l'émotion est un domaine nouveau et difficile, une étude précise de l'émotion est nécessaire. Nous avons commencé par passer en revue les recherches dans ce domaine, qui s'organisent selon les deux principaux axes suivants :

- les études portant sur l'universalité de l'émotion, y compris l'innéité de l'émotion et son aspect interculturel. Bien qu'il y ait encore beaucoup de discussions, la plupart des auteurs acceptent la notion « big six » des émotions primaires et universelles qui sont : la peur, la colère, la joie, la tristesse, la surprise et le dégoût de [Ekman et al 1969] ; l'innéité de l'émotion est également rapportée dans plusieurs études. Quelques autres résultats ont aussi affirmé l'aspect interculturel des émotions mais à certains niveaux seulement [Haidt et al, 1999], [Scherer, 2000], [Xiao, 2008] ;
- les études portant sur la modélisation des émotions. Nous distinguons deux approches : la première considère les émotions comme des états discrets et la seconde représente les émotions dans un espace continu. La plupart des études s'accordent sur les deux dimensions principales suivantes :
  - l'activité qui indique le niveau actif/passif des états émotionnels ;
  - la valence qui indique le niveau positif/négatif des états émotionnels.

Comme l'approche continue sur l'émotion est encore en discussion et est un défi même pour la reconnaissance par l'être-humain, parce qu'elle nécessite encore beaucoup d'études, et dans l'objectif de la réalisation d'un système de détection automatique de l'émotion, nous n'avons retenu que l'approche discrète et seulement avec les émotions primaires qui sont largement considérées comme universelles.

9.1.2. Étude des corpus

La disponibilité de corpus annotés est un problème crucial pour l'entraînement et l'évaluation de tous les systèmes de reconnaissance. Dans le chapitre 4, nous avons discuté trois approches pour la collection d'un corpus de l'émotion :

- l'enregistrement des conversations naturelles : très peu d'études considèrent cette approche en raison de sa difficulté malgré le très haut niveau de spontanéité des échantillons obtenus ;
- l'utilisation des expressions émotionnelles interprétées par des professionnels ; contrairement à la première approche, la plupart des études cherchent à utiliser les corpus de ce type en raison de la facilité de leur production et de la possibilité de gestion de la structure du corpus par l'auteur ;
- la troisième approche s'appuie sur la collection des morceaux de l'expression émotionnelle à partir de journaux télévisés ou à partir des films. Les expressions obtenues par cette approche sont aussi assez spontanées mais la difficulté essentielle de cette approche se situe au niveau du fort bruit qui accompagne les scènes.

En raison de l'absence d'un corpus en anglais et aussi d'un corpus en vietnamien, nous avons construit nous-mêmes deux corpus, appelés VnEm, en vietnamien, et EnEm, en anglais :

- *VnEm* contient 2292 échantillons de 62 locuteurs collectés à partir de 3 films vietnamiens, ces échantillons sont évalués sur 8 états émotionnels par 10 auditeurs de 20 à 30 ans.
- *EnEm* contient 1392 échantillons de 58 locuteurs collectés à partir de 6 films américains, les échantillons de ce corpus ne sont pas encore validés par les tests de perception comme avec VnEm.

Parmi les corpus sur l'émotion étudiés et construits, les deux corpus DES et le BES ont été choisis en raison de la disponibilité des émotions primaires dans ceux-ci et de la distribution quasi uniforme de leurs échantillons pour toutes les émotions pour tous leurs locuteurs. Cela favorise effectivement l'objectif de notre étude : la reconnaissance indépendante du locuteur. Grâce au corpus VnEm en vietnamien, une langue tonale, quelques experimentations ont été effectués pour la vérification de la possibilité d'utiliser les resultats obtenus dans une langue non-tonale comme l'allemand vers une langue tonale comme le vietnamien.

9.1.3. Etude des paramètres

La performance d'un système de reconnaissance en général et de reconnaissance de l'émotion en particulier est la combinaison de la performance de deux processus : l'extraction des paramètres (analyse) et l'utilisation de ces paramètres (modélisation). Autrement dit, la dégradation de la performance d'un de ces deux processus entraîne la dégradation de la performance du système. C'est la raison pour laquelle, dans ce travail de thèse, nous nous sommes intéressés à ces deux étapes. Dans la première étape, nous avons effectué des analyses et des études sur les paramètres pour trouver la combinaison de paramètres la plus efficace, et

nous avons testé des modèles bien connus dans le domaine de classification pour trouver le modèle le plus adapté pour la reconnaissance de l'émotion.

Dans l'étape de l'extraction des paramètres, nous nous sommes intéressés aux deux aspects du signal de parole :

- le domaine temporel où la prosodie est la caractéristique la plus intéressante avec les paramètres de F_0, l'intensité et de débit phonétique ainsi que leurs dérivées premières et secondes ; deux autres paramètres : le nombre de passages par zéro et le nombre d'extrémités, ainsi que leurs dérivées premières et deuxièmes ont également été analysés ;
- le domaine fréquentiel avec les trois ensembles MFCCs, LFCCs et LPCs qui sont des candidats intéressants, particulièrement les coefficients MFCCs qui sont majoritairement utilisés dans les systèmes actuels de reconnaissance automatique de la parole : nous avons étudié 147 paramètres au total pour ce domaine.

Nous avons commencé par étudier les informations globales sur un énoncé pour une première vue des performances de chaque paramètre. L'information globale d'un paramètre est constituée des valeurs obtenues par les 8 opérateurs, Max, Min, Moyenne, Médiane, Variance, Range, RisingFallingCountRatio, RisingFallingSumRatio, qui donnent une vue générale de la dynamique, de la distribution et de la forme des paramètres analysés. Les résultats ont été obtenus en effectuant des tests en utilisant le modèle de l'arbre de décision et la validation croisée par la division en 10 plis (« folds »).

Comme nous l'avons mentionné, la prosodie est une caractéristique indispensable pour l'étude de l'émotion. Pour généraliser les résultats obtenus en évitant le plus possible la dépendance des paramètres par rapport au locuteur et à la structure phonémique, nous avons proposé une normalisation par rapport au neutre : au lieu d'utiliser les valeurs originales des paramètres (A_E) mesurées sur chaque échantillon de chaque émotion de chaque locuteur, nous utilisons le rapport A_E/A_N où A_N est le même paramètre mais mesuré avec le même échantillon du même locuteur mais en neutre. L'efficacité de la normalisation par rapport au neutre a été démontrée par le taux assez élevé (44 %) de classification correcte que nous avons obtenu en utilisant isolément chaque paramètre de la prosodie.

Parmi les conclusions obtenues par l'analyse des informations globale, les quelques conclusions ou réaffirmations suivantes sont remarquables sur la prosodie :

- les trois aspects de la prosodie ne contribuent pas seulement à l'expression émotionnelle, mais leur rôle est aussi important dans la discrimination des émotions ; cela se montre par un taux de classification correcte beaucoup plus élevé avec ces paramètres qu'avec une classification aléatoire ;
- parmi les trois aspects de la prosodie, l'intensité et F_0 se montrent les plus efficaces (> 40 %) ; par contre, avec les mesures que nous avons effectuées, nous ne trouvons pas encore de contribution efficace pour le débit phonétique, le troisième aspect de la prosodie dans la discrimination des émotions ;
- parmi les paramètres globaux analysés, quelques paramètres sont très discriminants pour des émotions spécifiques ; la médiane de l'intensité, par exemple, peut donner un taux de classification correcte de 81 % sur les deux ensembles tristesse / non-tristesse ; cette remarques sera intéressante pour les études de classification hiérarchiques où, pour éviter la confusion entre des émotions causée par l'utilisation d'un seul ensemble de paramètres, on utilise les différents paramètres pour séparer les groupes différents ; [Xiao

2008] arrive à la même conclusion en suivant cette approche mais elle n'utilise pas des caractéristiques de prosodie ;

- en utilisant les opérateurs globaux, nous constatons également l'existence d'une différence en termes de l'évolution temporelle des caractéristiques de la prosodie pour quelques émotions mais nous ne pouvons pas encore nous servir de cette caractéristique car elle ne se montre pas claire pour tous les états émotionnels ; il se peut que cette caractéristique soit un paramètre utile pour les essais de classification hiérarchique.

A côté de la prosodie, nous avons également vérifié le comportement global des 49 paramètres dans le domaine fréquentiel comme les MFCCs, les LFCCs et les LPCs et leurs premières et deuxièmes dérivées, ce qui fait au total 49×3 = 147 paramètres, en appliquant 8 opérateurs sur les valeurs instantanées de chaque paramètre. Les mêmes conclusions sont obtenues :

- les caractéristiques fréquentielles sont également des signes discriminatifs de l'émotion ;
- parmi les trois ensembles, les MFCCs se montrent les plus efficaces, suivis par les LFCCs et les LPCs qui sont les moins efficaces ;
- en étudiant les MFCCs, nous constatons le rôle important du paramètre $MFCC_0$, celui qui est souvent retiré par tous les systèmes de reconnaissance automatique de la parole, mais qui est très efficace pour la discrimination des émotions ; ce paramètre peut être un candidat pour remplacer l'intensité car il est plus stable.

Bien que les analyses des informations globales nous donnent des indices importants, elles ne permettent pas d'atteindre un niveau correct de performance pour un système de reconnaissance automatique de l'émotion. C'est la raison pour laquelle, notre travail se focalise sur les paramètres locaux.

Les paramètres locaux sont les valeurs instantanées mesurées tout au long du signal. Notre étude des paramètres locaux a été réalisée en trois étapes : pour la reconnaissance mono-locuteur, pour la reconnaissance multi-locuteur et pour la reconnaissance indépendante du locuteur.

Globalement, nous pouvons résumer que notre approche de l'étude des paramètres comme un processus de filtres successifs en fonction de l'indépendance des paramètres du locuteur. Effectivement, nous commençons par l'analyse globale des comportements des paramètres pour l'ensemble des états émotionnels. Grâce à cette étape, les premiers points de vue de la performance de chaque paramètre sont établis. La reconnaissance mono-locuteur est le premier cas, le plus spécifique mais aussi le plus tolérant (en termes de l'indépendance des paramètres du locuteur) pour tester l'efficacité des paramètres et aussi pour éliminer ceux qui sont inefficaces. Les paramètres LPC, par exemple, sont éliminés dès cette étape. La reconnaissance multi-locuteur est un cas moins spécifique mais aussi moins tolérant ; effectivement, il y a une dégradation significative du taux de reconnaissance dans les résultats obtenus dans ce cas en utilisant le même modèle que le cas mono-locuteur. Dans cette étape, nous commençons à constater que les paramètres relatifs deviennent plus efficaces que les paramètres originaux. La reconnaissance indépendante du locuteur est le cas le plus général mais aussi le plus sélectif pour les paramètres. Autrement dit, un paramètre dépendant trop du locuteur ne donne pas de bons résultat dans ce cas malgré l'utilisation des modèles très performants.

L'algorithme que nous utilisons pour sélectionner des paramètres dans chaque étape est la sélection forcée séquentielle en avant (SFSA). La proposition de forcer la continuité de bouclage (en ajoutant la caractéristique correspond au meilleur résultat) jusqu'à ce que les résultats obtenus soient supérieurs à un seuil nous permet d'atteindre au moins la performance de tout l'ensemble de paramètres (voir la section 5.3.2.3.5).

Quelques autres méthodes permettant d'évaluer l'efficacité des paramètres, comme le critère de Fisher, ou la compression de l'espace des paramètres, comme avec l'analyse en composantes principales, ont été étudiées. Expérimentalement, la méthode SFSA nous donne la meilleure performance avec un nombre de paramètres très réduit.

Effectivement, sur le corpus DES, l'ensemble de 12 paramètres MFCCs sélectionnés par SFSA, 7 paramètres MFCCs sélectionnés par SFSA et la première dérivée de la fréquence fondamentale relative ΔF_0Rel, et 9 paramètres MFCCs sélectionnés par SFSA nous donnent respectivement des taux de 87,7 %, de 86,5 % et de 52,7 % de classification correcte dans les cas de reconnaissance mono-locuteur, multi-locuteur et indépendante du locuteur.

L'optimisation de la performance du système dans le cas de reconnaissance indépendante du locuteur est un des objectifs de cette thèse. En étudiant les méthodes actuelles de normalisation, nous avons proposé d'utiliser, dans ce cas, une normalisation symbolique. L'idée principale de cette approche est de symboliser sémantiquement les intervalles des valeurs de chaque paramètre. Autrement dit, les paramètres sont tous représenté par des symboles d'un ensemble défini ; il n'existe donc plus de décalage entre les valeurs selon les locuteurs. Ceci veut dire que l'approche de normalisation symbolique permettra théoriquement l'utilisation de paramètres indépendamment du locuteur en général et pour notre cas en particulier.

La normalisation symbolique a aussi l'avantage de favoriser la fusion entre deux types de paramètres qui sont assez différents en termes d'amplitude et/ou de variance : les paramètres à moyen terme comme la prosodie et les paramètres à court terme, comme les MFCCs par exemple. Cependant, nous n'avons pas encore identifié de combinaison convenable entre ces deux types de paramètres pour améliorer encore la performance du système. Une étude dans le futur portant sur la pondération adéquate pour la combinaison des paramètres serait peut-être une bonne direction pour améliorer encore le système.

La combinaison entre la SFSA et la normalisation symbolique s'est révélée très efficace dans le cas de reconnaissance indépendante du locuteur en nous permettant d'obtenir une amélioration relative d'environ 15 % par rapport à l'utilisation de 17 paramètres MFCCs originaux, et d'environ 10 % par rapport à l'utilisation de ces 17 paramètres MFCCs normalisés par la moyenne et par l'écart-type. Le score de classification correcte le plus élevé que nous ayons obtenu est de 57,3 % ; bien qu'il soit encore très loin en comparaison avec les 87,7 % de classification correcte dans le cas mono-locuteur, c'est un bon score en comparaison avec le taux de classification aléatoire 20 % et en comparaison avec 53 % obtenu par l'étude de [Huang et al, 2006] et 23,6 % obtenu par l'étude de [Noble 2003].

1.1.1. Etudes sur le vietnamien

Pour pouvoir effectuer des expérimentations avec des données plus spontanées mais aussi pour vérifier les résultats obtenus à partir des deux corpus DES et BES (des langues tonales), un corpus en vietnamien, VnEm, a été créé. Pour la construction de ce corpus, nous avons proposé une approche de l'annotation qui permet de prendre en compte le problème des émotions complexes (l'existence de plusieurs émotions en même temps) et aussi l'intensité des émotions. Cette approche a pour but de permettre de mener dans le futur des travaux plus complexes portant sur la reconnaissance des émotions dans un contexte réel où les émotions peuvent être mélangées, où leur intensité peut être très faible et où l'ambigüité ou le conflit peuvent toujours se produire.

En travaillant avec ce corpus en vietnamien, nous avons effectué quelques premières expérimentations. Nous constatons que les taux de la reconnaissance des émotions inter-langue

(entre l'allemande BES ou le danois DES avec le vietnamien) sont faibles (à peine supérieurs à ceux du hasard). Cela veut dire que les indices que nous avons sélectionnés pour la reconnaissance des émotions dans l'allemand ou le danois ne sont pas appropriés pour le vietnamien. A notre avis, ce fait peut s'expliquer par la différence entre une langue non-tonale et une langue tonale où le ton participe fortement aux sens des mots. Un autre facteur est certainement lié au type de données : actées avec peu de locuteurs (DES, BES) contre parole plus spontanée (contexte de films) avec un plus grand nombre de locuteurs. Il n'a pas été possible d'évaluer séparément l'impact de chacun de ces deux facteurs.

9.1.4. Etudes des modèles

L'état émotionnel peut être identifié en analysant la distribution des paramètres. Dans cette partie du travail de thèse et dans le contexte des corpus utilisés, nous expérimentons trois approches ou groupes de techniques en apprentissage supervisé afin de rechercher la technique la plus efficace mais aussi de vérifier quelques autres caractéristiques de l'espace de paramètres comme l'évolution temporelle des paramètres :

- le premier groupe de techniques vise à représenter l'escape des paramètres sous une autre forme, plus simple, mais plus générale. Parmi les modèles que nous étudions, le modèle de mélange de Gaussiennes (GMM) est le plus représentatif pour ce groupe. Beaucoup plus simple que le modèle GMM, le modèle de Quantification vectorielle (VQ) appartient aussi à ce groupe.
- le deuxième groupe des techniques cherche à séparer explicitement l'espace des paramètres en zones indépendants correspondants aux classes spécifiques. Les machines à vecteurs de support appartiennent à ce groupe.
- le troisième groupe des techniques comprend des techniques qui sont en plus capable de capturer l'information de l'évolution temporelle et les modèles de Markov cachés appartiennent à ce groupe.

De même que pour l'étude sur les paramètres, les modèles moins adéquats pour la reconnaissance émotionnelle ont été éliminés successivement en allant de l'étape la plus tolérante (reconnaissance mono-locuteur) vers l'étape la moins tolérante (reconnaissance indépendante du locuteur). Concrètement, l'approche de classification par la séparation explicite de l'espace des paramètres est éliminée à partir du cas de reconnaissance mono-locuteur parce qu'elle ne se montre pas aussi efficace que les autres approches (83,8 % contre 87,8 %) ; dans le premier groupe, malgré sa simplicité, le modèle VQ a aussi été éliminé car il se montre souvent moins efficace qu'un autre membre, le modèle GMM.

Il nous restait les deux modèles HMM et GMM à étudier, sachant que le modèle GMM peut être vu comme un cas particulier du modèle HMM dans lequel on néglige l'information d'évolution temporelle des paramètres. Donc, en analysant les résultats obtenus par ces deux types de modèles, nous pouvons répondre à la question : au-delà des évolutions temporelles locales déjà capturées par notre mesure (dans les Δ, $\Delta\Delta$), le modèle de HMM peut-il exploiter les évolutions globales de ces paramètres pour la discrimination des émotions. Nos expérimentations, n'ont pas mis évidence de différence significative en termes de performance entre ces deux types de modèles. Cela peut s'expliquer par le fait que pour l'expression émotionnelle, l'évolution globale des paramètres aux différents niveaux sont différents (mots isolés, mots dans une phrase, phrase isolé, phrase dans un paragraphe, etc.). Effectivement, en travaillant avec le modèle de HMM et avec le corpus DES, nous avons fait les deux constatations suivantes (voir la section 6.3.1.3) :

- le modèle ergodique de HMM (qui permet tous les types d'évolution) donne les meilleurs résultats par rapport au modèle gauche-droit de HMM (qui ne permet que les évolutions causales) ;
- l'amélioration de la performance lors que le nombre d'états correspond aux nombre de phonèmes nous suggère qu'une étude de l'émotion au niveau phonémique pourrait aider à améliorer la performance du système.

Pour terminer la conclusion de cette thèse, nous voudrons rappeler que les résultats dans les chapitres 5, 6 et 7 sont obtenus en travaillant avec les corpus dont les données ont des limites de la taille et ainsi que du niveau naturel, ces résultats ne pourraient pas donc se généraliser ou être appliqués dans des autres contextes plus généraux sans vérifier sa vérité dans ces contextes.

9.2. Perspectives

Comme la reconnaissance de l'émotion est un domaine très nouveau et bien que les résultats obtenus (57,3 %) soient applicables et acceptables pour un système d'indexation automatique de l'émotion, nous avons identifié d'autres pistes pour améliorer la performance du système.

Premièrement, au niveau des corpus, bien que le nombre d'émotions (cela veut dire aussi le nombre de classes à identifier) du corpus BES soit plus important que celui du corpus DES (8 contre 5), le fait que le résultat de reconnaissance indépendante du locuteur obtenu avec BES (62,1 %) soit meilleur que celui obtenu avec DES (57,3 %) pourrait s'expliquer par la taille (le nombre de locuteurs) du corpus BES qui est beaucoup plus importante que celle du corpus DES. Une de nos premières perspectives est donc d'élargir les corpus ou de construire et de travailler sur d'autres corpus de grande taille, y compris le développement de nos deux corpus EnEm et VnEm.

Deuxièmement, au niveau de paramètres, dans ce travail, nous avons ciblé les paramètres locaux. Cependant, comme nous avons montré dans le chapitre 5, les paramètres globaux portent également des informations émotives, une étude profonde de cette approche avec plus de paramètres, ceux qui présentent mieux l'information émotive indépendante du locuteur (par exemple pour discriminer la tristesse et le neutre), avec des modèles plus performants que le modèle de l'arbre de décision que nous avons utilisé, promettrait des meilleurs résultats. Ainsi, la fusion entre ces deux approches (locale et globale) serait intéressante.

En analysant des paramètres, nous constatons aussi que certains paramètres sont très efficaces pour discriminer certaines émotions, mais ils ne sont pas très efficaces pour le reste, c'est la raison pour laquelle, nous envisageons également une classification hiérarchique pour optimiser le taux de reconnaissance. L'efficacité de cette approche a aussi été prouvée par [Xiao, 2008] avec d'autres ensembles paramètres.

Malgré l'avantage et l'efficacité de la normalisation symbolique proposée ; l'utilisation de cette normalisation unifie le rôle de tous les paramètres, et nous croyons qu'avec une autre étude dans le futur sur la richesse d'information émotive afin de pondérer adéquatement chaque paramètre donnera les meilleurs résultats de classification.

En limitant la normalisation sur un seule genre à l'aide d'un détecteur du genre de locuteur est aussi une approche prometteuse et réalisable par le fait que la détection automatique du sexe de locuteur n'est pas difficile à réaliser et que certaines améliorations possibles ont été prouvées par [Blouin et al, 2005] et [Hu et al, 2007] ; [Xiao, 2008] a aussi récemment prouvé une amélioration la plus grande d'environ 20 % du taux de reconnaissance dans le cas multi-locuteur en utilisant l'information complémentaire du genre.

Au niveau de la langue, dans cette thèse, nous avons étudié et obtenus des résultats avec les deux corpus DES en danois et BES en allemand. Bien que l'émotion soit interculturelle à certains niveaux comme l'ont montré nos résultats et par quelques autres études citées, l'application sur une langue répandue comme l'anglais et le français est une cible de notre système d'indexation automatique de l'émotion. Pour le faire, enrichir et compléter le corpus EnEm, tester et appliquer sur le corpus TRECVID est un des travaux réalisables dans le futur.

L'étude de l'émotion dans une langue tonale comme le vietnamien est aussi un travail vraiment intéressant, non seulement pour l'indexation automatique, mais aussi pour améliorer les systèmes de synthèse automatique du vietnamien que nous sommes aussi en train d'élaborer. Les travaux futurs pourront commencer par l'enrichissement de ce corpus et ensuite par la complétion de l'annotation au niveau phonémique. En effet, l'influence très forte des tons dans les langues tonales exigerait des études à ce niveau afin de pouvoir renforcer la reconnaissance des émotions.

Enfin, au niveau de la modalité, du fait que la parole n'est pas un seul moyen pour transmettre l'information émotionnelle, que le renforcement réciproque entre des modalités est prouvé par plusieurs études et que notre objectif définitif est aussi d'intégrer dans un système d'indexation automatique de l'émotion des documents audio-visuels, l'étude et l'utilisation de l'information des autres modalités est une tendance de notre travail et du domaine. Dans une première étape, nous pensons à la combinaison des deux aspects de la parole : l'information contenue dans le signal acoustique et l'information linguistique/sémantique contenue dans la transcription de la parole celle-ci ayant été montrée comme étant une bonne approche par plusieurs études de Devillers ([Devillers et al, 2002], [Devillers et al, 2003], [Devillers et al, 2004], [Devillers et al, 2006]). Intégrer la reconnaissance de l'émotion dans un système multimodal est un sujet encore très ouvert et il n'y a aucun rapport d'un effort de recherches qui vise à intégrer toutes les modalités non-verbales dans un seul système pour l'analyse affective du comportement humain. [Sebe et al, 2005].

BIBLIOGRAPHIE

[Abelin et al, 2000] Abelin, A., Allwood, J., *Cross-linguistic interpretation of emotional prosody*, In: Proceedings of the ISCA Workshop on Speech and Emotion, Belfast, Ireland, Textflow, Belfast, 2000, p. 1-18.

[Akbar et al, 1998] Mohammad Akbar, Jean Caelen, *Parole et traduction automatique: le module de reconnaissance RAPHAEL,* International Conference On Computational Linguistics Proceedings of the 17th international conference on Computational linguistics, Montreal, Quebec, Canada, 1998, vol. 1, p. 36-40.

[Ang et al, 2002] Ang J., Dhillon R., Krupski A., et al, *Prosody-based automatic detection of annoyance and frustration in human-computer dialog,* ICSLP, USA, 2002, vol. 3, p. 2037.

[Archer et al, 1996] Archer Dane, *The human face: Emotions, Identities and Masks*. Berkeley, CA: University of California Extension Center for Media and Independent Learning, 1996. 1 videocassette (30 min.): sd., col. with b&w sequences; 1/2 in. + 1 guide (12 p. : ill. ; 28 cm.)

[Aronowitz & Irony 05], Aronowitz H. Irony D., *Modeling intra-speaker variability for improved speaker recognition*, in SLSF' - Subspace, Latent Structure and Feature Selection techniques: Statistical and Optimisation perspectives Workshop, 2005, Bohinj.

[Banse et al, 1996], Banse, R. & Scherer, K. R., *Acoustic Profiles in Vocal Emotion Expression*, Journal of Personality and Social Psychology, 1996, vol. 70, No. 3, p. 614-636.

[Bänziger et al, 2006] Bänziger, T., Pirker, H., & Scherer, K. (2006), *GEMEP - GEneva Multimodal Emotion Portrayals: a corpus for the study of multimodal emotional expressions*, Proceedings of LREC'06 Workshop on Corpora for Research on Emotion and Affect, p. 15-19

[Barrett 2006] Barrett, L. F., *Are Emotions Natural Kinds?,* Perspectives on Psychological Science, 2006, vol. 1, p. 28-58.

[Batliner et al, 2006] Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., & Aharonson, V., *Combining efforts for improving automatic classification of emotional user states*, In Erjavec, T. and Gros, J. (Ed.), Language Technologies, *IS-LTC 2006* (pp. 240-245)

[Beritelli et al, 2005] Beritelli, F.; Casale, S.; Russo, A.; Serrano, S., *A Genetic Algorithm Feature Selection Approach to Robust Classification between Positive and Negative Emotional States in Speakers,* Signals, Systems and Computers, 2005, Conference Record of the Thirty-Ninth Asilomar Conference on Volume, Issue, October 28 - November 1, 2005 p. 550 - 553

[Black et al, 1995] M.J. Black and Y. Yacoob, *Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion*, In Proc. International Conf. On Computer Vision, 1995, p. 374–381.

[Blouin et al, 2005] Blouin C., Maffiolo V. (2005), *A study on the automatic detection and characterization of emotion in a voice service context*, In InterSpeech, Lisboa, 2005, p. 469-472.

[Boite et al., 2000] Boite, R., Bourlard, H., Dutoit, T., Hang, J., and Leich, H. (2000), *Traitement de la parole,* ISBN 2-88074-388-5. Presses Polytechniques et universitaires Romandes, Lausanne, Suisse.

[Boersma et Weenink 2005] P. Boersma & D. Weenink, *Praat: doing phonetics by computer (Version 4.3.14)*, [Computer program]. Retrieved May 26, 2005, from http://www.praat.org/

[Bosh 2000] Bosh, L.T. *Emotions: what is possible in the ASR framework*? In: Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, 2000, p. 189-194.

[Bosch 2003] Bosch, L.F.M. ten (2003), *Emotions*, *Speech and the ASR framework*, Speech Communication, 2003, vol. 40, Issue 1-2, p. 213-225.

[Boucouvalas et al, 2002] Boucouvalas, A. Ac, and Zhe, X. 2002, *Text-to-emotion engine for real time internet communication*, In Proceedings of the International Symposium on CSNDSP 2002, Staffordshire Univ., July 15-17, p. 164-168.

[Braun 2005] Braun, Angelika / Katerbow, Matthias, *Emotions in dubbed speech: an intercultural approach with respect to F0*, In InterSpeech, Lisboa, 2005, p. 521-524

[Breazeal 2000] Cynthia Breazeal, *Sociable Machines: Expressive Social Exchange Between Humans and Robots*, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, PhD Thesis, May 2000.

[Burkhardt et al, 2000] Burkhardt, F., Sendlmeier, W. *Verification of acoustical correlates of emotional speech using formant-synthesis*, In: Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, UK, 2000, p. 151-156.

[Burkhardt et al, 2005] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, *A Database of German Emotional Speech,* Interspeech, Lisboa, 2005, p. 1517-1520.

[Busso et al, 2004] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, *Analysis of emotion recognition using facial expressions, speech and multimodal information*, in Sixth International Conference on Multimodal Interfaces ICMI, State College, PA, 2004, p. 205-211.

[Calliope 1989] Calliope, *La parole et son traitement automatique*, Masson et CENT-ENST, Paris, 1989, ISBN 2-225-81516-X

[Campbell 2002] Campbell, W. N., *The Recording of Emotional speech, JST/CREST database research*, Proceeding LREC 2002, vol. 6, p. 2029-2032

[Cappé, 2000] Cappé, O. (2000), *Modèles de mélange et modèles de markov cachés pour le traitement automatique de la parole*, (ENST/Paris) [http://www.tsi.enst.fr/~cappe/cours/tap.pdf]

[Cauldwell 2000] Cauldwell, R. *Where did the anger go? The role of context in interpreting emotions in speech*, ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, UK, 2000, p. 127-131

[Chan et al, 1995] Chan, D., Fourcin, A., Gibbon, D., Granström, B., Hucvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Lindberg, B., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., Veld, C., Zeiliger, J. *EUROM- A Spoken Language Resource for the EU*, Proceedings of the 4th European Conference on Speech Communication and Speech Technology, Eurospeech, Madrid, 1995, vol. 1, p. 867-870.

[Chan et al, 1998] Chan, S.W.K. And Franklin, J. *Symbolic connectionism in natural language disambiguation,* IEEE Tran. Neural Network, 1998, vol. 9, p. 739-755.

[Chateau et al, 2004] Chateau N., Maffiolo V. and Blouin C., *Analysis of emotional speech in voice mail message: The influence of speaker's gender*, ICSLP, Jeju Island, Corea 2004, p. 885-888.

[Chen et al, 1998] L.S. Chen, H. Tao, T.S. Huang, T. Miyasato, and R. Nakatsu. *Emotion recognition from audiovisual information.* In Proc. IEEE Workshop on Multimedia Signal Processing, 1998, p. 83–88.

[Chen et al, 2000] L.S. Chen and T.S. Huang. *Emotional expressions in audiovisual human computer interaction*. In Proc. International Conference on Multimedia and Expo (ICME), 2000, p. 423–426.

[Chung 2000] Soo-Jin Chung, *L'expression et la perception de l.émotion extraite de la parole spontanée: évidences du coréen et de l'anglais,* thèse, Université de la Sorbonne Nouvelle (Paris III), 2000.

[Clore 1992] Clore, G.L.: *Cognitive phenomenology: Feelings and the construction of judgment*. In Martin, L., Tesser, A. (eds.): The Consruction of Social Judgments (Lawrence Erlbaum Associates, Hillsdale 1992)

[Cohen et al, 2003] I. Cohen, N. Sebe, A. Garg, L. Chen, and T.S. Huang, *Facial expression recognition from video sequences: Temporal and static modelling*, Computer Vision and Image Understanding, 2003, vol. 91, p. 160–187.

[Cohen et al, 2004] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T.S. Huang, *Semi-supervised learning of classifiers: Theory, algorithms, and applications to human-computer interaction*, IEEE Trans. on Pattern Analysis and Machine Intelligence, to appear, 2004, vol. 26, p. 1553-1566.

[Cowie et al, 2000] E. Douglas-Cowie, R. Cowie, and M. Schröder, *A new emotion database: Considerations, sources and scope.* Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Belfast 2000, p. 39-44.

[Curtis Roads 1996] Curtis Roads, *The Computer Music Tutorial,* MIT Press, Cambridge, 1996.

[Daly 1983] Daly, E. M., Lancee, W. J., Polivy, J. *A conical model for the taxonomy of emotional experience*, Journal of Personality and Social Psychology, 1983, vol. 45, p. 443-457.

[Davis et Mermelstein 1980] S.B. Davis & P. Mermelstein. *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Trans, 1980, vol 28, issue 4, p. 357-366.

[De Abreu et al, 2006] Sophie De Abreu, Catherine Mathon, Daniela Perekopska, *Perception de la colère dans un corpus de français spontané par des apprenants portugais et tchèques,* Journées d'Etude sur la Parole, Dinard, France, 12-16 juin 2006. (http://jep2006.irisa.fr/index.htm).

[De Cheveigné et al, 2002] Alain de Cheveigné and Hideki Kawahara. *Yin, a fundamental frequency estimator for speech and music*, Journal of the Acoustical Society of America, 2002, vol. 111, issue 4, p. 1917-1930.

[De Silva et al, 2000] L.C. De Silva and P.C Ng, *Bimodal emotion recognition,* In Proc. Automatic Face and Gesture Recognition, 2000, p. 332–335.

[Dellaert et al, 1996] Dellaert, F., Polzin, T.,Waibel, A., 1996, *Recognizing emotion in speech*, In: Proc. Int. Conf. Spoken Language Processing (ICSLP '96). Vol. 3. p. 1970–1973.

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

[Devillers et al, 2002] Devillers, L., Vasilescu, I., And Lamel, L. *Annotation and detection of emotion in a task-oriented human-human dialog corpus*, In Proceedings of the ISLE Workshop on Dialogue Tagging for Multi-Modal Human-Compute Interaction 15-17 Dec. 2002, Edinburgh.

[Devillers et al, 2003] Devillers, L., Luniel, L., And Vasilescu, I. 2003, *Emotion detection in task-oriented spoken dialogues*, In Proceedings of the International Conference on Multimedia and Expo, Baltimore, MD, July 6-9, vol. 3, p. 549-552.

[Devillers et al, 2004] Devillers, L. & Vasilescu, I. (2004), *Détection des émotions à partir d'indices lexicaux, dialogiques et prosodiques dans le dialogue oral*. Journées d'Etude sur la Parole (JEP 2004). Fès, Maroc, 19-22 avril 2004. (http://www.lpl.univ-aix.fr/jep-taln04/proceed/actes/jep2004/Devillers-Vasilescu.pdf)

[Devillers et al, 2005] Devillers, L., Vidrascu, L., & Lamel, L. (2005), *Challenges in real-life emotion annotation and machine learning based detection,* Neural Networks: 18, 407-422.

[Devillers et al, 2006] Devillers, Laurence / Vidrascu, Laurence (2006): *"Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs",* In InterSpeech-2006, paper 1636-Tue1A3O.3.

[Devillers et al, 2006] Devillers, L., Cowie, R., Martin, J.-C., Douglas-Cowie, E., Abrilian, S., Mc Rorie, M., *Real life emotions in French and English TV video clips : an integrated annotation protocol combining continuous and discrete approaches*, LREC 2006. Fifth International Conference on Language Resources and Evaluation (2006)

[Dijkstra et al, 1994] Dijkstra, K., Zwaan, R.A., Graesser, A.C., And Magliano, J.P. 1994, *Character and reader emotions in literary texts*, Poetics 23, p. 139-157.

[Dorken et al, 1994] Erkan Dorken and S. Hamid Nawab, *Improved musical pitch tracking using principal decomposition analysis,* In International Conference on Acoustics, Speech and Signal Processing, IEEE, 1994, vol. II, p. 217–220.

[Douglas-Cowie et al, 2007] Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., Martin, J., Devillers, L., Abrilian, S., Batliner, A., Amir, N., and Karpouzis, K., *The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In Proceedings of the 2nd international Conference on Affective Computing and intelligent interaction*, Lisbon, Portugal, September 12-14, 2007

[Doval et al, 1991] Boris Doval and Xavier Rodet, *Estimation of fundamental frequency of musical sound signals*, International Conference on Acoustics, Speech and Signal Processing, IEEE, 1991, p. 3657–3660.

[Doval et al, 1993] Boris Doval and Xavier Rodet, *Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs*, International Conference on Acoustics, Speech and Signal Processing, IEEE 1993, vol. I, p. 221–224.

[Dutoit et al, 1993] T. Dutoit, H. Leich, 1993, *MBR-PSOLA : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database*, Speech Communication, vol. 13, n° 3-4, p. 435-440.

[Eide et al, 1996] E. Eide, H. Gish, *A Parametric Approach to Vocal Tract Length Normalization*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Atlanta, GA, May 1996, vol. I, p. 346-349.

[Ekman et al, 1971] Ekman, P., *Universals and cultural differences in facial expressions of emotion,* In J. Cole (Ed.), Nebraska Symposium on Motivation 1971, Lincoln, NE: University of Nebraska Press, vol. 19, p. 207-283.

[Ekman et al, 1978] P. Ekman and W.V. Friesen, *Facial Action Coding System: Investigator's Guide*, Consulting Psychologists Press, 1978.

[Ekman 1982] Ekman, P., (Ed.), *Emotions in the Human Face*, London: Cambridge University Press, New York, 1982, p. 353-395.

[Ekman et al, 1986] Ekman P & Friesen WV, *A new pan-cultural facial expression of emotion. Motivation and Emotion*, 1986, vol. 10, n° 2, p. 159-168

[Ekman 1994] P. Ekman, *Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique*. Psychological Bulletin, 1994, vol. 115, n° 2, p. 268–287.

[Ekman 1999] Ekman P., *Basic Emotions*, In T. Dalgleish and M. Power (Eds.). *Handbook of Cognition and Emotion*, Sussex, U.K., John Wiley & Sons, Ltd., 1999.

[Engberg et al, 1996] I. S. Engberg and A. V. Hansen, *Documentation of the Danish Emotional Speech Database (DES*), Internal AAU report, Center for Person Kommunikation, Denmark, 1996.

[Engberg et al, 1997] I. S. Engberg, A. V. Hansen, O. Andersen and P. Dalsgaard, *Design, Recording and Verification of a Danish Emotional Speech Database*, EUROSPEECH'97 : 5th European Conference on Speech Communication and Technology, Rhodes, Greece, September, 1997, vol. 4, p.1695-1698

[Essa et al, 1997] I.A. Essa and A.P. Pentland, *Coding, analysis, interpretation, and recognition of facial expressions*, IEEE Trans. on Pattern Analysis and Machine Intelligence, 1997, vol. 19, n° 7, p. 757–763.

[Fek et al, 2004] Fék M. – Németh G. – Olaszy G. – Gordos G.: *Designe of a Hungarian Emotional Database for Speech Analysis and Synthesis*, Proc. of Workshop on Affektive Dialogue Systems, 2004, vol. 3068, p. 113-116.

[Fernandez et al, 2003] Fernandez, R., Picard, R., 2003, *Modeling drivers' speech under stress*, Speech Communication, 2003, vol. 40, n° 1, p. 145–159.

[Fletcher 1953] Fletcher, H. *Speech and hearing in communication*, Contributors: Harvey Fletcher - author. Publisher: D. Van Nostrand. Place of Publication: Princeton, NJ. Publication Year: 1953.

[France et al, 2000] France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., Wilkes, M. *Acoustical properties of speech as indicators of depression and suicidal risk*. IEEE Trans. Biomedical Engineering, 2000, vol.47, issue 7, p. 829–837.

[Fukunaga, 1990] K. Fukunaga, *Introduction to statistical pattern recognition*, Academic Press, Boston, 2nd edition, 1990.

[Furui, 1981] Furui S., *Cepstral Analysis Technique for Automatic Speaker Verification*, IEEE Trans, 1981, vol. 29, issue 2, p.254-272.

[Garau et al, 2005] G. Garau, S. Renals, and T. Hain., *Applying vocal tract length normalization to meeting recordings*, In Proceeding Interspeech, Lisbon, Portugal, September 2005, p. 265-268.

[Gobl et al, 2000] Gobl, C., Chasaide, A.N. *Testing affective correlates of voice quality through analysis and resynthesis*, In: Proceedings of the ISCA Workshop on Emotion and Speech, Northern Ireland, 2000, p. 178-183.

[Goleman 1995] D. Goleman, *Emotional Intelligence*, Bantam Books, 1995

[Geoffriois 1996] Edouard Geoffriois, *The multi-lag-window method for robust extended-range f0 determination*, In Fourth International Conference on Spoken Language Processing, 1996, vol. 4, p. 2239–2243.

[Gerhard 1999] David Gerhard, *Audio visualization in phase space.* In Bridges: Mathematical Connections in Art, Music and Science, August 1999, p. 137-144.

[Gerhard 2003] David Gerhard, *Pitch extraction and Fundamental Frequency: History and Current Techniques*, Technical Report TR-CS 2003-6, University of Regina Department of Computer Science, November 2003

[Gouyon et al, 2000] Gouyon, F. Pachet, F. Delerue, O. 2000, *On the use of zero-crossing rate for an application of classification of percussive sounds*' *Proceedings of* COST G6 Conference on Digital Audio Effects, Verona, Italy, 2000, p. 147-152

[Gray 1984] R.M. Gray, *Vector Quantization*, IEEE ASSP Magazine, April 1984, vol. 1, p. 4-29.

[Guyon et al, 2002] Guyon I., *Gene Selection for Cancer Classification using Support Vector Machines*, Journal of Machine Learning Research, 2002, vol. 46, p. 389-422.

[Guyon et al, 2003] Guyon I., Elisseeff A., *An introduction to feature and variable selection,* Journal of Machine Learning Research, vol. 3, 2003, p. 1157-1182.

[HARRIS 1978] Harris, F.J. *On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform,* Proc IEEE, 1978, vol. 66, p. 51-83.

[Hager & Ekman 1983] Hager, J.C., Ekman, P., *The Inner and Outer Meanings of Facial Expressions,* In: Cacioppo, J.T., Petty, R.E. (eds.) Social Psychophysiology: A Sourcebook, The Guilford Press, New York, 1983, p. 348-356.

[Haidt et al, 1999] Haidt J & Keltner D, *Culture and facial expression: Open-ended methods find more expressions and a gradient of recognition*, Cognition and Emotion, 1999, vol. 13, n° 3, p. 225-266.

[Hieronymus 1991] Hieronymus J.L., *Formant Normalisation for Speech Recognition and Vowel Studies*, Speech Communication, NL, Elsevier Science Publishers, Amsterdam, Dec. 1, 1991, vol. 10, No. 5/06, p. 471-478.

[Houtgast et al, 1985] Houtgast T. & Steeneken J. M. (1985*), A Review of the MTF Concept in Room Acoustics and its Use for Estimating Speech Intelligibility in Auditoria*, Journal of the Acoustical Society of America, vol. 77, n° 3, p.1069-1077.

[Hsu et al, 2003] C.-W. Hsu, C.-C. Chang, C.-J. Lin. *A practical guide to support vector classification, 2003,* http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[Hsu et al, 2006] Hsu, Chang-wen, Lee, Lin-shan, *Extension and further analysis of higher order cepstral moment normalization (HOCMN) for robust features in speech recognition*, In InterSpeech-2006, paper 1748-Mon1A2O.5.

[Hu et al, 2007] Hao Hu, Ming-Xing Xu, and Wei Wu, *GMM Supervector based SVM with Spectral Features for Speech Emotion Recognition*, IEEE Int. Conf. Acoustic, Speech and Signal Processing (ICASSP'07), vol. 4, p. 413-416.

[Huang et al, 2006] Huang, R.[Rongqing], Ma, C.[Changxue], *Toward A Speaker Independent Real-Time Affect Detection System*, International Conference on Pattern Recognition 2006, vol. I, p.1204-1207.

[Iida, 2002] Iida A, *A study on corpus-based Speech Synthesis with Emotion*, PhD thesis, Keio University, 2002.

[Istrate 2003] D. Istrate, *Détection et reconnaissance des sons pour la surveillance médicale*, thèse, Université de Grenoble, 2003

[Jiang et al, 2004-1] Jiang, Dan-Ning / Cai, Lian-Hong (2004), *Classifying emotion in Chinese speech by decomposing prosodic features*, In InterSpeech-2004, p. 1325-1328

[Jiang et al, 2004-2] Jiang, D. N., Cai, L. H. *Speech emotion classification with the combination of statistic features and temporal features*. In: Proc. Int. Conf. Multimedia and Expo (ICME '04). Taipei, vol. 3, issue 27-30, p. 1967-1970.

[Jones et al, 2000] Jones, S., Meddis, R., Lim, S.C., *Toward a digital neuromorphic pitch extraction system*, IEEE T Neural Networks, vol. 11, p. 978-987.

[Kadambe et al, 1990] Kadambe, S.; Boudreaux-Bartels, G.F, *A comparison of a wavelet transform event detection pitch detector with classical pitch detectors*, Signals, Systems

and Computers, 1990. Conference Record Twenty-Fourth Asilomar Conference on 5-7 Nov 1990, vol. 2, p. 1073

[Kaernbach et al, 1998] Kaernbach, C., and Demany, L. *Psychophysical evidence against the autocorrelation theory of pitch perception*, J. Acoust. Soc. Am., 1998, vol. 104, p. 2298-2306.

[Kanade et al, 2000] Kanade T, Cohn JF, Tian Y. *Comprehensive database for facial expression analysis*, in 4th IEEE International Conference on Automatic Face and Gesture Recognition, 2000, p. 46-53.

[Keller et al, 1997] Keller E. & Werner S. (1997), *Automatic Intonation Extraction and Generation for French*, Proc. 14th CALiCO Annual Symposium, West Point, NY, USA, September 1997, ISBN 1-890127-01-9.

[Kedem 1986] Kedem, B., *Spectral analysis and discrimination by zero-crossings*, Proceedings of the IEEE, Nov. 1986, vol. 74, issue 11, p. 1477-1493.

[Keltner et al, 1999] Keltner, D., & Haidt, J. (1999), *Social functions of emotions at four levels of analysis*, Cognition and Emotion, vol. 13, p. 505-522.

[Kienast et al, 1999] Kienast, M., Paeschke, A., & Sendlmeier, W. *Articulatory reduction in emotional speech*, Proceedings of Eurospeech, Budapest, Hungary, 1999, p. 117-120.

[Kira et al, 1992] Kira K., Rendell L., *A practical approach to feature selection*, Proceedings of the International Conference on Machine Learning, 1992, vol. 1, p. 249-256.

[Klasmeyer 1995] G. Klasmeyer, *Emotions in Speech,* Institut fur Kommunikationswissenschaft, Technical University of Berlin, 1995.

[Kohler 1995] Kohler, K. J. *Articulatory Reduction in Different Speaking Styles*, Proceedings ICPhS '95, Stockholm, 1995, Vol. 2, p. 12-19.

[Kwon et al, 2003] Kwon, O. W., Chan, K. L., Hao, J., Lee, T. W., 2003. *Emotion recognition by speech signals*. In: Proc. European Conf. Speech Communication and Technology, Eurospeech '03, vol. 1. pp. 125–128.

[Lanitis et al, 1995] Lanitis, A.; Taylor, C.J.; Cootes, T.F., *A unified approach to coding and interpreting face images*, Computer Vision, 1995. Proceedings., Fifth International Conference, 20-23 Jun 1995, p. 368-373.

[Larivey 2002] Larivey M., *La puissance des émotions,* Éditions de l'Homme, 2002 ISBN 2-7619-1702-2

[Le V. B. 2006] Le Viet Bac, Thèse 1 Juin 2006, *Reconnaissance automatique de la parole pour des langues peu dotées*, Université Joseph Fourier.

[Lee et al, 1996] L. Lee, R. Rose, *Speaker Normalization using Efficient Frequency Warping Procedures*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Atlanta, GA, May 1996, vol. I, pp. 353–356.

[Lee et al, 2004-1] Yoonjae Lee, Hanseok Ko, *Multi-Eigenspace Normalization for Robust Speech Recognition in Noisy Environments*, ICSLP, Oct, 2004, vol. 3, p. 2097-2100.

[Lee et al, 2004-2] Lee C. M., Yildirim S., Bulut M., Kazemzadeh A., Busso C., Deng Z., Lee S., Narayanan S.S, *Emotion Recognition based on Phoneme Classes*, in Proc. of ICSLP 2004, Korea, 2004.

[Lee et al, 2005] Lee, C. M., Narayanan, S. S. *Toward detecting emotions in spoken dialogs*, IEEE Trans. Speech and Audio Process, 2005, vol. 13, n° 2, p. 293-303.

[Lee et al, 2007] Lee Cheongjae, Gary Geunbae, *Emotion Recognition for Affective User Interfaces using Natural Language Dialogs*, The 16th IEEE International Symposium on Robot and Human interactive Communication, 2007, p. 798-801.

[Lefort & al 02] L. Lefort, T. Merlin, J.-F. Bonastre, P. Nocera, *Le projet MTM - Reconnaissance de la parole et du locuteur sur une plateforme embarquée*, XXIVème Journées d'Etude sur la Parole, Nancy, 24-27 juin 2002,

[Li et al, 2000] Yongxin Li, Yuqing Gao and Hakan Erdogan, *Weighted Pairwise Scatter to Improve Linear Discriminant Analysis*. Proc. ICSLP 4, 2000, p. 608-611.

[Lien 1998] J. Lien, *Automatic recognition of facial expressions using hidden Markov models and estimation of expression intensity,* PhD thesis, Carnegie Mellon University, 1998

[Lima et al, 2004] Lima, C. S. Tavares, Adriano Silva, Carlos A. Oliveira, Jorge F, *Spectral normalization MFCC derived features for robust speech recognition*, SPECOM, International Conference Speech and Computer, 9, Saint Petersburg, 2004.

[Linarès et al, 2007] G. Linarès, P. Nocéra, D. Massonié, D. Matrouf, *The LIA speech recognition system : from 10xRT to 1xRT*, International Conference of Text, Speech and Dialogue (TSD 2007), Pilsen, Czvech Republic, vol. 4629, p. 302-308.

[Lindberg, 1995] B. Lindberg & H. Christensen (1995), *Documentation of the Danish EUROM.1 Database*, Esprit project 2589 (SAM) Multi-lingual speech input/output assessment, methodology and standardisation. CPK Denmark.

[Linde et al, 1980] Y.Linde, A.Buzo, and R.M.Gray, *An algorithm for Vector Quantization Design,* IEEE Transaction on Communications, COM, January 1980, vol. 28, p. 84-95.

[LtiLib 2006] Peter Dörfler and José Pablo Alvarado Moya, *LTI-Lib - a C++ Open Source Computer Vision Library*, Advanced Man-Machine Interaction, K.-F. Kraiss (ed), Springer, http://ltilib.sourceforge.net/doc/html/index.shtml

[Lyons et al, 1998] Lyons MJ, Akamatsu S, Kamachi M, Gyoba J. *Coding facial expressions with gabor wavelets*. in Third IEEE International Conference on Automatic

Face and Gesture Recognition. 1998. Nara Japan: IEEE Computer Society, ISBN: 0-8186-8344-9, p. 200-205.

[Mase 1991] Kenji MASE, *Recognition of Facial Expression from Optical Flow*, IEICE TRANSACTIONS on Information and Systems, 1991, Vol.E74-D n°.10 p.3474-3483.

[Martin et al, 2001] A. Martin, D. Charlet, and L. Mauuary, *Robust speech/non-speech detection using LDA applied to MFCC,* in Proceedings of IEEE International Conference on Acoustics,Speech, and Signal Processing, 2001, vol.1, p. 237-240.

[Martin et al, 2003] A. Martin. L. Mauuary, *Voicing parameter and energy based speech/non-speech detection for speech recognition in adverse conditions,* in Proc. of Eurospeech, Geneva, Switzerland, Sept. 2003, p. 3069–3072.

[Matsumoto 1998] D. Matsumoto, *Cultural influences on judgments of facial expressions of emotion*, In Proc. ATR Symposium on Face and Object Recognition, 1998, p. 13-15.

[Mayer et al, 1993] Mayer, J.D., Salovey, P. *The intelligence of emotional intelligence.* Intelligence, 1993, vol. 17, p. 433-442.

[MBROLA project 1996] http://tcts.fpms.ac.be/synthesis/mbrola.html

[McGilloway et al, 2000] McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, C. C. A. M., Westerdijk, M. J. D., Stroeve, S. H. Approaching automatic recognition of emotion from voice: A rough benchmark. In: Proc. ISCA Workshop Speech and Emotion, 2000, vol. 1. p. 207–212.

[Ming et al, 1996] Ming, J. O'Boyle, P. McMahon, J. Smith, F.J. Sch. of Electr. Eng. & Comput. Sci., Queen's Univ., Belfast, *Speech recognition using a strong correlation assumption for theinstantaneous spectra*, ICSLP Proceedings., Fourth International Conference, 1996, vol. 2, p. 1061-1064.

[Moorer 1974] Moorer, J. A., *The optimum comb method of pitch period analysis of continuous digitized speech,* Acoustics, Speech, and Signal Processing, IEEE Transactions, Oct 1974, vol. 22, Issue 5, p. 330-338.

[Moorer 1977] Moorer, J. A., *On the transcription of musical sound by computer*, Computer Music Journal, November 1977, p. 32-38.

[Montero et al, 1999] Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., & Pardo, J. M., *Analysis and Modelling of Emotional Speech in Spanish*, ICPhS 99, p. 957-960.

[Moraru et al, 04] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, J.-F. Bonastre (2004), *Segmentation selon le locuteur : les activités du consortium ELISA dans le cadre de Nist RT03*, In: JEP 2004 (Maroc).

[Mozziconacci 2004] Mozziconacci, S. J. L., & Hermes, D. J., *Role of intonation patterns in conveying emotion in speech*, ICPhS 1999, p. 2001-2004.

[Murray et Arnott 1993] Murray, I. R. & Arnott, J. L. (1993), *Toward the simulation of emotion in synthesized speech: A review of the literature on human vocal emotion*, Journal of Acoustic Society of America, vol. 93, n°. 2, p. 1097-1108.

[Muray et Arnott, 2008] I. R. Murray, J. L. Arnott, *Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech*, Computer Speech and Language, 2007, Elsevier, p.107-129.

[Noble 2003] Noble, James (2003), *Spoken Emotion Recognition with Support Vector Machines*, Honours thesis, Department of Computer Science and Software Engineering, University of Melbourne.1

[Noll A. M. 1967] Noll A. Micheal, *Cepstrum Pitch Determination*, The Journal of the Acoustical Society of America, February 1967, vol. 41, Issue 2, p. 293-309.

[Nuttall 1981] Nuttall, A, *Some windows with very good sidelobe behaviour***,** Acoustics, Speech, and Signal Processing IEEE Transactions, Feb 1981, vol. 29, Issue 1, p. 84-91.

[New et al, 2001] Tin Lay Nwe, Say Wei Foo, and L. C. De Silva, *Speech Based Emotion Classification*. Proceedings of IEEE Region 10 International Conference on Volume 1, Issue, 2001, vol. 1, p. 297 - 301.

[O'Toole et al, 2005] O'Toole AJ, Harms J, Snow SL, Hurst DR, Pappas MR, et al., A *video database of moving faces and people*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, vol. 27, p. 812-816.

[Ortony et Turner 1990] Ortony, A. & Turner, T. J. (1990). *What's basic about basic emotions?,* Psychological Review, vol. 97, n°.3, p. 315-331.

[Otsuka et al, 1997] T. Otsuka and J. Ohya, *Recognizing multiple persons' facial expressions using HMM based on automatic extraction of significant frames from image sequences*, In Proc. International Conf. on Image Processing, 1997, p. 546-549.

[Oudeyer 2002] Oudeyer P-Y. (2002), *The production and recognition of emotions in speech: features and algorithms*, International Journal in Human-Computer Studies, vol. 59, n°1-2, p. 157-183, special issue on Affective Computing.

[Panagiotakis et al, 2005] Panagiotakis, C. Tziritas, G. *, A speech/music discriminator based on RMS and zero-crossings,* Multimedia, IEEE Transactions on, 2005, vol. 7, Issue 1, p. 155-166.

[Pantic et al, 2005] Pantic M, Valstar MF, Rademaker R, Maat L, *Web-based database for facial expression analysis*, in IEEE International Conference on Multimedia and Expo (ICME). 2005. Amsterdam, p. 5.

[Pachet & Roy 2007] Pachet, F. and Roy, P., *Exploring billions of audio features*, In Eurasip, editor, Proceedings of Content-Based Multimedia Indexing 2007, p. 227-235.

[Pelachaud et al, 1996] C. Pelachaud, N. Badler, and M. Steedman. *Generating facial expression for speech*. Cognitive Science, 1996, vol. 20, p. 1-46.

[Pellom et al, 1996] Pellom, B. L., Hansen, J. H. L., 1996. *Text-directed speech enhancement using phoneme classfication and feature map constrained vector quantization*. In: Proc. Inter. Conf. Acoustics, Speech, and Signal Processing (ICASSP'96), vol. 2, P. 645-648.

[Petrushin 1999] Petrushin, V. A., 1999. *Emotion in speech recognition and application to call centers*. In: Proc. Artificial Neural Networks in Engineering (ANNIE 99), vol. 1, p. 7–10.

[Pfitzinger et al, 1996] Pfitzinger, H.R. Burger, S. Heid, S., *Syllable detection in read and spontaneous speech, Spoken Language*, 1996. ICSLP 96. Proceedings., Fourth International Conference on, vol 2, P. 1261-1264.

[Picard 1997] R. W. Picard, *Affective Computing*, MIT Press, Cambridge, MA (1997).

[Picard et al, 2001] Picard, R.W., Vyzas, E., Healey, J. *Toward machine emotional intelligence: Analysis of affective physiological state*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001, vol. 23, issue 10, p. 1175–1191.

[Piszczalski et al, 1979] Martin Piszczalski and Bernard A. Galler, *Predicting musical pitch from component frequency ratios,* Journal of the Acoustical Society of America, September 1979, vol. 66, n° 3, p. 710-720.

[Pittam et Scherer 1993] Pittam, J., & Scherer, K. R. (1993), *Vocal expression and communication of emotion*, In M. Lewis, & J. Haviland (Eds.), The handbook of emotions. Guilford, New York, p. 185-197.

[Plutchik 1980] Plutchik, R. (1980), *A structural model of the emotions*, Emotion, A Psychoevolutionary Synthesis, Harper and Row, New York, p. 152-172.

[Plutchik 2003] Plutchik, R. (2003), *Emotions and life; perspectives from psychology, biology and evolution*, American Psychological Association, Washington, DC,

[Potamianos et al, 2003] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior. *Recent advances in the automatic recognition of audiovisual speech.* Proceedings of the IEEE, 2003, vol. 91, n° 9, p. 1306-1326.

[Praat] Praat is a free scientific software program for the analysis of speech in phonetics. It has been designed and continuously developed by Paul Boersma and David Weenink of the University of Amsterdam. It can can run on a wide range of operating systems, including various Unix platforms, Mac and Microsoft Windows (95, 98, NT4, ME, 2000, XP). The program also supports speech synthesis, including articulatory synthesis. http://www.fon.hum.uva.nl/praat/ (wikipedia).

[Press et al., 2002] Press,W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling,W. T. (2002). *Numerical Recipes in C; The Art of scientific Computing* ;The second Edition. ISBN 0-521-43108-5. Cambridge University Press.

[Quast 2002] Quast, H. (2002), *Automatic Recognition of Nonverbal Speech: An Approach to Model the Perception of Para- and Extralinguistic Vocal Communication with Neural Networks,* Machine Perception Lab Tech Report 2002/2. Institute for Neural Computation, UCSD (2002)

[Quinlan 1993] R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993, ISBN:1-55860-238-0.

[Rabiner 1978] L-R. Rabiner, R-W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[Rabiner et al, 1993] Rabiner, L. R., Juang, B. H., 1993, *Fundamentals of Speech Recognition*, NJ: Prentice-Hall.

[Rank et al, 1998] Rank, E., & Pirker, H. *Generating emotional speech with a concatenative synthesizer,* Proceedings of the 5th International Conference of Spoken Language Processing, Sydney, Australia, 1998, p. 671–674.

[Robnik-Sikonjaet al, 2003] Robnik-Sikonja M., Kononenko I., *Theoretical and Empirical Analysis of ReliefF and RReliefF*, Journal of Machine Learning Research, 2003, vol. 53, n° 1-2, p. 23-69.

[Rosenblum et al, 1996] M. Rosenblum, Y. Yacoob, and L.S. Davis, *Human expression recognition from motion using a radial basis function network architecture,* IEEE Trans. on Neural Network, 1996, vol. 7, n° 5, p.1121–1138.

[Rossignol et al, 1998] Stephane Rossignol, Xavier Rodet, J¨oel Soumagne, Jean-Luc Collette, and Philippe Depalle, *Features extraction and temporal segmentation of acoustic signals,* In International Computer Music Conference, 1998, p. 199–202.

[Rossignol 2000] Stéphane Rossignol*, Segmentation et indexation des signaux sonores musicaux*, Thèse de doctorat, Université Paris VI, Juillet 2000.

[Russell 1980] Russell, J. (1980), *A circumplex model of affect*, Journal of Personality and Social Psychology, vol. 39, p.1161-1178.

[Ryynänen et al, 2004] M. P. Ryynänen and A. Klapuri, *Modelling of Note Events for Singing Transcription*, in Proc. of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing, Oct. 2004, p. 6.

[Salovey, 1990] P. Salovey and J.D. Mayer., "*Emotional intelligence*", Imagination, Cognition, and Personality, 1990, vol. 9, p. 185-211.

[Schlosberg 1952] Schlosberg, H. (1952), *The description of facial expressions in terms of two dimensions,* Journal of Experimental Psychology, vol. 44, n°. 4, p. 229-237.

[Schlosberg 1954] Schlosberg, H. *Three dimensions of emotion*, The Psychological Review, 1954, vol. 61, n°. 2, p. 81-88.

[Schüller et al, 2004] Schüller, B., Rigoll, G., Lang, M., 2004. *Speech emotion recognition combining acoustic features and linguistic information in a hybrid support*

vector machine-belief network architecture. In: Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP '04). vol. 1, p. 557-560.

[Schüller et al, 2005] Schuller, R. Müller, M. Lang, G. Rigoll, *Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles*, INTERSPEECH 2005, Special Session: Emotional Speech Analysis and Synthesis: Towards a Multimodal Approach , Lisbon, Portugal, 04.-08.09.2005, p. 805-809.

[Scheirer et al, 1997] Eric Scheirer and Malcolm Slaney, *Construction and evaluation of a robust multifeature speech/music discriminator,* In International Conference on Acoustics, Speech and Signal Processing, IEEE, 1997, vol. II, p. 1331–1334.

[Scherer et al, 1981] Scherer, K. R., *Speech and Emotional States*, in: Darby, J. K. (ed.), The Evaluation of Speech in Psychiatry, New York: Grune & Stratton, 1981, p. 189-220.

[Scherer et al, 1991] Scherer, K.R., Banse, R., Wallbott, H.G., Goldbeck, T., *Vocal cues in Emotion Encoding and Decoding*, Motivation and Emotion, 1991, vol. 15, p. 123-148.

[Scherer et al, 2001] Scherer, K.R., Banse, R., Wallbott, H.G., 2001, *Emotion inferences from vocal expression correlate across languages and cultures,* J. Cross-Cult. Psychol, 2001, vol. 32, n° 1, p. 76–92.

[Scherer, 2002] Scherer, Klaus R. (2000), *A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology*, In ICSLP-2000, vol. 2, p. 379-382.

[Scherer, 2003] Scherer, Klaus R. (2003), *Vocal communication of emotion: A review of research paradigms*, Speech Communication, vol. 40, p. 227-256.

[Schröder 1999], Schröder, M. (1999), *Can emotions be synthesized without controlling voice quality?*, Phonus 4, Research Report of the Institute of Phonetics, University of the Saarland, p. 37-55. http://www.dfki.de/~schroed.

[Sebe et al, 2005] N. Sebe, I. Cohen, Th. Gevers, T. S. Huang, *Multimodal approaches for emotion recognition: a survey* (Invited Paper), SPIE, Internet Imaging, San Jose, 2005, vol. 5670, p. 56-57.

[Shannon et al, 2004] Shannon Benjamin J., Paliwal Kuldip K. (2004), *MFCC computation from magnitude spectrum of higher lag autocorrelation coefficients for robust speech recognition*, In INTERSPEECH-2004, p. 129-132.

[Shannon 1949] C. E. Shannon. *Communications in the presence of noise.* Proc IRE, Jan. 1949, vol. 37, p. 10-21.

[Shaver et al, 1987] Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987), *Emotion Knowledge: Further exploration of a prototype approach*, Journal of Personality and Social Psychology, vol. 52, n°. 6, p. 1061-1086.

[Shi et al, 2003] Shi, R. P., Adelhardt, J., Zeissler, V., Batliner, A., Frank, C., N¨oth, E., Niemann, H., 2003. *Using speech and gesture to explore user states in multimodal dialogue systems*. In: Proc. ISCA Tutorial and Research Workshop Audio Visual Speech Processing (AVSP '03), vol. 1, p. 151–156.

[Slaney et al, 1998] Slaney, M., McRoberts, G., 2003. *Babyears*: *A recognition system for affective vocalizations*. Speech Proceeding of the 1998 International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seattle, WA, May 12-15, 1998, IEEE, vol. 39, n° 3, p. 367-384.

[Söderström et al. 2004] U. Söderström and H. Li, *Emotion recognition and estimation from tracked lip features*, Swedish Symposium on Image Analysis (SSBA'04), pp Uppsala, March 11-12 2004, vol. 5.

[Subasic et al 2001] Subasic, P. And Huettner, A. 2001, *Affect analysis of text using fussy semantic typing*, IEEE Trans. Fuzzy Systems, vol. 9, p. 483-496.

[Tato 2002] Tato, R., 2002. *Emotional space improves emotion recognition*. In: Proc. Int. Conf. Spoken Language Processing (ICSLP '02), Colorado. Vol. 3, p. 2029-2032.

[Tickle 2000] Tickle, A., 2000, *English and Japanese speaker's emotion vocalizations and recognition: a comparison highlighting vowel quality*, ISCA Workshop on Speech and Emotion, Belfast, 2000

[Tide Project 1995] Tide Project: TP1174-VAESS, Technical Annex, 26-6-1995: http://www.speech.kth.se/speech/proj/vaess.html

[Tao et al, 2004] Tao, J. And Tan, T. 2004. *Emotional Chinese talking head system*. In Proceedings of the 6th International Conference on Multimodal Interface (Oct. 13-15), p. 273-280.

[Tran et al, 2005] Tran D.D., Castelli E., Serignat J.F., Trinh V.L. & Le X.H. (2005), *Influence of F0 on Vietnamese syllable perception*, Insterspeech - Eurospeech 2005, Lisbon, Portugal, September 4-8, p. 1697-1700.

[Tran 2007], Tran D. D, *Synthèse de la parole à partir du texte en langue vietnamienne. Application à la reproduction des emotions*, thèse cotutelle entre le laboratoire MICA et le laboratoire CLiPS-IMAG 2003-2007.

[Vacher et al, 2003] M. Vacher, D. Istrate, L. Besacier, E. Castelli, J.-F. Serignat, *Smart Audio Sensor for Telemedecine*, presented at Smart Objects Conference (SOC) 2003, Grenoble, 15-17 Mai, 2003, p. 222-225.

[Vaufreydaz 2002] Vaufreydaz D. *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*, thèse de l'Université J. Fourier – Grenoble I, Janvier 2002.

[Ververidis et al, 2004-1] Ververidis, D., Kotropoulos, C., 2004. *Automatic speech classification to five emotional states based on gender information*. In: Proc. European Signal Processing Conf. (EUSIPCO '04), vol. 1, p. 341–344.

[Ververidis et al, 2004-2] Ververidis, D., Kotropoulos, C., Pitas, I., 2004a. *Automatic emotional speech classification*. In: Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP '04), Montreal, vol. 1, p. 593–596.

[Ververidis et al, 2005-1] Ververidis, D., Kotropoulos, C. *Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm*, In: Proc. Int. Conf. Multimedia and Expo (ICME '05), p. 1500-1503.

[Ververidis et al, 2005-2] Ververidis, D., Kotropoulos, C.**,** *Sequential Forward Feature Selection With Low Computational Cost*, Aristotle University of Thessaloniki, Greece, http://www.ee.bilkent.edu.tr/~signal/defevent/papers/cr1411.pdf

[Ververidis et al, 2006] Ververidis Dimitrios, Kotropoulos Constantine 2006, *Emotional speech recognition: Resources, features, and methods. Speech communication (Speech commun.),* ISSN 0167-6393 CODEN SCOMDH, vol. 48, n°. 9, p. 1162-1181.

[Vidrascu & Devillers, 2005-1] Vidrascu, L., & Devillers, L. (2005), *Annotation and detection of blended emotions in real human-human dialogs recorded in a call center,* IEEE ICME 2005, p. 4.

[Vidrascu & Devillers, 2005-2] Vidrascu, L., & Devillers, L. (2005), *Real-Life Emotion Representation and Detection in Call Centers Data*, Affective Computing and Intelligent Interaction, vol. 3784/2005, p. 739-746.

[Vidrascu & Devillers, 2007] Vidrascu, L., & Devillers, L. (2007), *Five emotion classes detection in real-world call centre data: the use of various types of paralinguistic features*, workshop Paraling 2007 de ICPHs 2007.

[Vroomen et al, 1993] Vroomen, J., Collier, R., & Mozziconacci, S. J. L., Duration *and Intonation in Emotional Speech*, Eurospeech 1993, Vol. 1, p. 577-580.

[Vu 2007] Vu Minh Quang, *Exploitation de la prosodie pour la segmentation et l'analyse automatique de signaux de parole*, thèse INP de Grenoble, France, 2007.

[Vu et al, SFC2005] Vu M.Q., Castelli E., Boucher A. & Besacier L. (2005), *Classification de parole en Question et NonQuestion par arbre de décision*, SFC 05, 12èmes Rencontre de la Société Francophone de Classification - Montréal, 30 mai - 1er juin 2005

[Vu et al, MajeSTIC2005] Vu M.Q., Besacier L., Castelli E. & Pham Thi N. Y. (2005), *Extraction automatique de Questions dans les corpus de réunions et de dialogues,* MajeSTIC05, Manifestation des jeunes chercheurs francophones dans les domaines des STIC. Novembre 2005, Rennes, France.

[Xiao et al, 2007] Zhongzhe Xiao; E. Dellandrea; Weibei Dou; Liming Chen, *Two-stage Classification of Emotional Speech*, Zhongzhe Xiao; E. Dellandrea; Weibei Dou; Liming Chen Digital Telecommunications, 2006. ICDT apos; 06. International Conference on Volume , Issue , 2006, p. 32-32.

[Xiao, 2008] Zhongzhe Xiao, *Classification of Emotion in Audio Signals*, thèse de l'École centrale de Lyon, 2008.

[Xu Shaoal 2004] Xu Shao, Milner, B., *Acoustics, Pitch prediction from MFCC vectors for speech reconstruction*, Speech, and Signal Processing, ICASSP 2004. IEEE International Conference, 17-21 May 2004, vol. 1, p. 97-100.

[Yacoob et al, 1996] Y. Yacoob and L.S. Davis, *Recognizing human facial expressions from long image sequences using optical flow*, IEEE Trans. on Pattern Analysis and Machine Intelligence, 1996, vol. 18, n° 6, p. 636–642.

[Yacoub et al, 2003] Yacoub Sherif, Simske Steve, Lin Xiaofan, Burns John (2003), *Recognition of emotions in interactive voice response systems*, in EUROSPEECH-2003, p. 729-732.

[Yanaru 1995] Yanaru, T. 1995. *An emotion processing system based on fuzzy inference and subjective observations*. In Proceedings of the 2nd New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems (Dunedin, N Z., Nov. 20-23). IEEE Computer Society Press, New York, p. 15-20.

[Yoon et al, 2007] Won-Jung Yoon, and Kyu-Sik Park, *A Study of Emotion Recognition and Its Applications*, MDAI 2007, LNAI 4617, p.455-462.

[Yost 1996] Yost, W. A. (1996), *Pitch strength of iterated rippled noise*, The Journal of the Acoustical Society of America 100, p. 3329-3335.

[Young et al, 1996] Steve Young et al. (1996*), HTK - Hidden Markov Model Toolkit (2.0),* Entropic Cambridge Research Laboratory: http://htk.eng.cam.ac.uk/

[Zeng et al, 2009] Zeng, Z.H.[Zhi-Hong], Pantic, M.[Maja], Roisman, G.I.[Glenn I.], Huang, T.S.[Thomas S.], *A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions*, PAMI(31), No. 1, January 2009, pp. 39-58

[Zhang et al, 2005] Zhang, Y.; Li, Z.; Ren, F.; Kuroiwa, S., *Semi-automatic emotion recognition from textual input based on the constructed emotion thesaurus,* Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on 30 Oct.-1 Nov. 2005, p. 571-576.

[Zheng et al, 2001] Zheng Fang, Guoliang Zhang, Zhanjiang Song, *Comparison of Different Implementations of MFCC*, Journal of Computer Science and Technology 2001, vol. 16, issue 6, p. 582-589.

[Zhou et al, 2007] Xi Zhou Yun Fu Ming Liu Hasegawa-Johnson, M. Huang, T.S., *Robust Analysis and Weighting on MFCC Components for Speech Recognition and Speaker Identification*, Multimedia and Expo Conference, Beijing July 2007, p. 188-191.

[Zhu Xuan et al, 2002] Zhu Xuan; Chen Yining; Liu Jia; Liu Runsheng, *Feature selection in Mandarin large vocabulary continuous speech recognition***,** Signal Processing, 2002 6th International Conference on Volume 1, Issue , 26-30 Aug. 2002, vol. 1, p. 508-511.

[Zovato et al, 2004] Zovato, E., Pacchiotti, A., Quazza, S., & Sandri, S. (2004), *Towards emotional speech synthesis: a rule based approach*, Proc. 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA, USA, p. 219–220.

[Wakita 1977] H. Wakita, Normalization *of Vowels by Vocal-Tract Length and its Application to Vowel Identification*, IEEE Trans. on Acoustics, Speech, and Signal Processing, April 1977, vol. 25, p. 183–192.

[WaveEdit] http://www.aldostools.com/wavedit.html

[Wegmann et al, 1996] S. Wegmann, D. McAllaster, J. Orloff, B. Peskin, *Speaker Normalization on Conversational Telephone Speech*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Atlanta, GA, May 1996, vol. I, p. 339–342.

[Wery et al, 1989] B. R. Wery, A. Leroux, H. Ph. Delbrouck, J. Leclerc, *A new parametric speech analysis and synthesis technique in the frequency domain*, EUROSPEECH '89 First European Conference on Speech Communication and Technology, Paris, France September 27-29, 1989, vol. 6, n° 4, p. 277-289.

[Weston et al, 2003] Weston J., Elisseeff A., Scholkopf B., Tipping M., *Use of the Zero-Norm with Linear Models and Kernel Methods*, Journal of Machine Learning Research, 2003, vol. 3, p. 1439-1461.

[Wiegrebe et al, 1998] Wiegrebe, L., Patterson, R.D., Demany, L., & Carlyon, R.P. *Temporal dynamics of pitch strength in regular interval noises,* Journal of the Acoustical Society of America, 1998, vol. 104, p. 2307-2313.

[Williams et al, 1972] Williams C.E & Stevens K.L (1972), *Emotions and speech: Some acoustical correlates*, Journal of Acoustic Soc. Am. Vol. 52, issue 4B, p. 1238-1250.

[Womack et al, 1996] Womack, B. D., Hansen, J. H. L., 1996, *Classification of speech under stress using target driven features,* Speech Communication, vol. 20, p. 131-150.

[Womack et al, 1999] Womack, B. D., Hansen, J. H. L., 1999. *N-channel hidden Markov models for combined stressed speech classification and recognition*. IEEE Trans. Speech and Audio Processing, vol. 7, n° 6, p. 668–667.

[Wu et al, 2006] Wu, C.H., Z.J. Chuang, and Y.C. Lin, *Emotion Recognition from Text using Semantic Label and Separable Mixture Model*, ACM Transactions on Asian Language Information Processing, 2006, vol. 5, n° 2, p. 165-183.

Annexe A. Extraction de la F_0

Dans cette annexe, nous présenterons brièvement les méthodes utilisées pour détecter F_o dans les trois domaines généraux groupés par le type d'entrée et du traitement. Les méthodes dans le domaine temporel seront présentées d'abord, car elles sont habituellement simples au niveau du calcul. Les méthodes statistiques qui utilisent la théorie des probabilités pour faciliter la décision seront présentées à la fin. Nous discuterons aussi des améliorations qui peuvent être appliquées pour les algorithmes en général, et pour notre cas en particulier.

A.1. Les méthodes dans le domaine temporel.

Les approches les plus fondamentales et les plus concrètes de la détection de la fréquence fondamentale est de regarder la forme du signal acoustique qui représente la pression atmosphérique en fonction de temps et d'essayer de déterminer F_o à partir de cette forme. Il y a une branche des méthodes que cherchent à découvrir combien de fois l'onde répète entièrement sa forme et on appelle cette branche la détection se basant sur des événements temporels. Une autre branche qui est aussi dans le domaine temporel et que nous avons choisi pour notre travail est constituée des méthodes qui utilisent l'autocorrélation.

b.1.1. La détection se basant sur le taux d'événements temporels

L'idée principale de cette méthode est que si l'onde est périodique, il y a donc des événements extractibles répétés en fonction de temps qui peuvent être comptés. Le nombre de ces événements se produisant dans une unité de temps est inversement lié à la fréquence. Cependant, le signal contient naturellement plusieurs autres composantes fréquentielles qui peuvent causer les mêmes événements dans le signal. Par conséquent, des méthodes se basant sur ce type d'événement ne sont pas assez fiables. Un exemple de l'utilisation du ZCR (le nombre de passages par zéro) pour la détection de F_0 est l'étude de [Rossignol et al, 1998]. Le sommet peut aussi être un événement dans la détection de F_0.

b.1.2. L'autocorrélation

La corrélation entre les deux signaux est une mesure de leur similitude. La fonction d'autocorrélation est la corrélation d'un signal avec lui-même en fonction du temps de retard. On a toujours la similitude exacte au point de retard zéro. La dissimilitude augmente avec l'augmentation du temps de retards. La définition mathématique de la fonction de l'autocorrélation d'un signal discret infini $s[n]$ est donnée dans l'équation (6.8) qui présente la définition mathématique suivante si le signal est fini $s'[n]$ et d'une taille N.

$$R_s(k) = \sum_{n=-\infty}^{\infty} s[n]s[n+k] \tag{6.8}$$

$$R_{s'}(k) = \sum_{n=0}^{N-1-k} s'[n]s'[n+k] \tag{6.9}$$

La caractéristique la plus importante est que la fonction d'autocorrélation d'un signal périodique avec la période P est aussi périodique avec la même période : $R_s(k) = R_s(k+P)$. En effet, la valeur de la fonction $R(k)$ diminue à la valeur minimum quand le temps de retard k s'approche vers des points au milieu des périodes ($k \rightarrow nP/2$ avec n=1,2.3…) car la phase de la forme d'onde est inverse avec sa copie à ces points. Si k continue à augmenter vers la fin d'une période ($k \rightarrow nP$), le signal et la copie retardée auront la même phase et la valeur de la fonction d'autocorrélation rétablira un nouveau maximum. Alors, la détermination de la période fondamentale P du signal de la parole peut être remplacée par la détermination les sommets maxima de la fonction d'autocorrélation : le premier sommet maximum de la fonction d'autocorrélation (au point P) indique donc la période du signal.

Le même problème se produit quand le signal possède des composants harmoniques de haute fréquence. L'algorithme YIN essaye de résoudre ces problèmes par quelques manières.

Le détecteur YIN est une nouvelle technique développée par Alain de Cheveigné et Hideki Kawahara [De Cheveigné et al, 2002], elle est nommée selon le nom du principe philosophique oriental de l'équilibre Yin-Yang9 représentant la tentative de cet auteur d'équilibrer entre l'autocorrélation et l'annulation de cet algorithme. YIN est basé sur la fonction de différence, bien qu'il soit semblable à la fonction d'autocorrélation, il essaie de minimiser la différence entre la forme d'onde et sa reproduction retardée au lieu de maximiser le produit comme l'autocorrélation. Nous appelons les minima de cette fonction les sommets négatifs. L'équation (6.10) présente la fonction de différence, Figure 58 a) montre cette fonction :

$$d_m(k) = \sum_{n=1}^{N} (s_n - s_{n+k})^2 \tag{6.10}$$

L'idee de l'algorithme YIN est d'employer une fonction moyenne cumulative qui « désemphase » des hautes fréquences dans la fonction de différence :

[9] http://asiarecipe.com/yinyang.html

$$d'_m(k) = \begin{cases} 1, & k = 0 \\ \dfrac{d_m(k)}{\dfrac{1}{k}\sum_{j=k}^{j=1} d_m(j)}, & \text{sinon} \end{cases} \qquad (6.11)$$

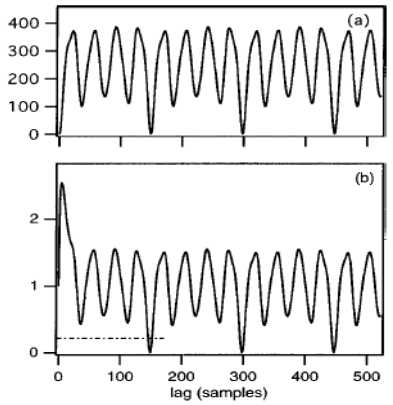

Figure 58 : a) fonction normale de différence b) fonction améliorée de la différence

D'autres intérêts de cet algorithme peuvent se trouver dans [De Cheveigné et al, 2002], y compris la possibilité de l'enlèvement de l'intervalle de recherche des sommets négatifs. Des intervalles se font en se basant sur les informations de la fréquence fondamentale maximum et minimum du locuteur. La possibilité de l'enlèvement de cet intervalle nous donne un grand intérêt pour appliquer dans le cas de la reconnaissance indépendante du locuteur car dans ce cas, l'extraction du pitch est normalement effectuée avec des personnes inconnues.

Pour d'autres améliorations du détecteur YIN et une discussion plus complète de cette méthode, y compris l'exécution et les résultats statistiques, voir le papier cité.

b.1.3. Espace de phase :

L'espace de phase est une manière relativement nouvelle de visualiser le son. [Gerhard 1999]. On peut se baser sur sa caractéristique de périodicité du diagramme de phase pour détecter la fréquence du signal original (voir [Gerhard 1999]). Cependant, cette méthode est apparentée au problème de la détection de taux de passage à zéro et elle a donc des mêmes problèmes associés.

A.2. Les méthodes dans le domaine fréquentiel

Il y a beaucoup d'informations dans le domaine fréquentiel qui peuvent être liées à F_o du signal. Il y en a les quatre axes principaux des approches pour l'extraction de F_o : l'approche basée sur le taux des partiels fréquentiels, l'approche utilisant des filtres, l'approche appliquant l'analyse de cepstre, et l'approche multi-résolution.

b.2.1) Approches basée sur le taux de partiels fréquentiels

Depuis 1979, [Piszczalski 1979] a cherché à extraire F_o du signal de la musique pour détecter les frontières des notes (en supposant qu'une note simple est présente à chaque point de temps). Dans le procédé original, Piszczalski a commencé par la transformation spectrale et

l'identification des composants dans le signal en utilisant la détection des maxima. Pour chaque paire de ces composants dont les fréquences sont fx et fy, l'algorithme cherche à trouver « les nombres harmoniques les plus petits » i et j qui correspondent aux harmonies de ces deux composants. Les critères pour déterminer i et j se trouvent dans [Piszczalski 1979]. Chacune paire de ces nombres harmoniques est alors employée comme hypothèse pour la fréquence fondamentale du signal. L'avantage de cette méthode est qu'elle n'exige pas la présence de la fréquence fondamentale.

[Dorken et al, 1994] ont présenté une amélioration pour la méthode de Piszczalski. Ils suggèrent de calculer « le spectre conditionné » en utilisant une méthode qu'ils avaient précédemment employée pour l'analyse du composant principal. La condition améliore l'exactitude de l'extraction des composants fréquentiels, et par conséquent, la transformation globale est plus exacte.

b.2.2 Méthodes par des filtres.

Avec cette approche, pour l'extraction de F_o, en général, plusieurs filtres sont utilisés avec des différentes fréquences de centre pour comparer leurs sorties. Si le sommet spectral aligne avec un filtre de passe-bande, la valeur de la sortie de ce filtre sera plus élevée que celles des autres. Le filtre optimum en peigne « optimum Comb Filter » est une de ces méthodes. Comme dans [Moorer 1974] et [Moorer 1977]. le détecteur de F_o utilisant le filtre optimum en peigne est un algorithme robuste mais coûte cher en calcul.

b.2.3 Analyse du cepstre

L'analyse de cepstre est une forme de l'analyse spectrale où on travaille avec la transformation inverse de Fourrier du logarithme de spectre au lieu de travailler directement avec le spectre. Le cepstre sert à séparer deux composants superposés : l'excitation glottale (corde vocale) et la résonance du conduit vocal. Donc, une manière simple de décrire le cepstre est : elle tend à séparer le composant de pitch du reste de spectre. [Curtis Roads 1996]. La théorie de cette méthode se fonde sur le fait que si nous supposons que la parole voisée est un produit de convolution entre une séquence de l'excitation glottal et la réponse discrète du conduit vocal. Dans le domaine de fréquence, la relation de convolution devient une relation de multiplication et en utilisant la propriété de la fonction logarithmique : log(AB) = log(A) + log(B), la relation de multiplication peut être transformée en relation de l'addition. Enfin, la transformation inverse de Fourier préserve la propriété de cette relation additives [Noll A. M. 1967]: IFT(log(AB)) = IFT(log(A)) + IFT(log(B)).

Comme la plupart des autres approches, cet algorithme possède aussi des désavantages [Kadambe et al, 1990]: 1) il n'est sensible aux variations du pitch puisqu'ils estiment la période moyenne du pitch travers un segment donné du signal de la parole. 2) Il ne marche pas bien non-plus pour tous les locuteurs de haute fréquence fondamentale et de bas fréquence fondamentale car la longueur de la trame est fixée. 3) il n'est pas robuste non-plus avec le bruit.

b4) Multi-résolution : une amélioration peut être appliquée à n'importe quelle méthode dans le domaine spectrale est d'employer la résolution multiple [Geoffriois 1996]. L'idée est relativement simple : si l'exactitude d'un certain algorithme à une certaine résolution est suspecte, nous pouvons confirmer ou annuler le résultat en utilisant le même algorithme à une résolution plus haute ou basse. Cela veut dire, nous pouvons utiliser une plus grande ou plus petite fenêtre pour recalculer le spectre. Si le sommet fréquentiel apparaît dans toutes ou presque toutes les fenêtres, ceci peut être considéré comme une confirmation de l'hypothèse de F_o.

Cependant, des nouvelles résolutions impliquent également des nouveaux calculs lourds. C'est la raison pour laquelle l'analyse de Fourier en résolution multiple est lente et couteuse.

A.3. Les méthodes statistiques

Le problème de l'évaluation automatique de F_o peut être considéré, par certains côtés, comme le problème statistiques. Plusieurs méthodes modernes ont été testées y comprises les deux méthodes : le réseau neurone et les estimateurs du maximum de vraisemblance.

L'approche de l'évaluation de F_o à l'aide des estimateurs du maximum de vraisemblance peut se trouver dans des articles de [Doval et al, 1991] [Doval et al, 1993]. L'idée principale de cette approche est de chercher la probabilité maximum entre une observation et des fréquences fondamentales de candidat.

Différemment, l'approche par le réseau neurone a pour le but d'employer des modèles connexionnistes pour l'évaluation du pitch est de modeler le système auditif humain, comme dans [Jones et al, 2000] où un modèle de réseau neurologique basé sur les mécanismes cochléaires de l'oreille humaine a été présenté. Le désavantage du modèle connexionniste est que même si un bon modèle est trouvé, il ne fournit aucune information de la façon dont le problème est résolu, toutes ces informations sont stockées dans les poids des connections, et dans le cas des grands modèles avec des milliers ou des millions de connections, il est presque impossible de traduire ces poids en description ou en algorithme. Ils deviennent donc des « boîtes noires » faisant ce qu'il fait sans savoir pourquoi ou comment. Pour bien utiliser un modèle connexionniste, le domaine et les données de l'apprentissage doivent aussi être soigneusement choisis [Gerhard 2003]

Plusieurs algorithmes de détection du pitch ainsi que leurs points forts et leurs points faibles ont été discutés. Parmi ces techniques, l'autocorrélation avec les techniques de YIN dans le domaine temporel ont été choisies parce que dans notre cas, nous devons effectuer l'extraction du pitch et en temps réels de plusieurs locuteurs, y comprises des personnes inconnues, donc un algorithme robuste et assez rapide est préféré. YIN nous donne ces réponses : lors de l'utilisation avec les fréquences d'échantillonnage courantes aujourd'hui de la parole, cette routine de d'autocorrélation est tout à fait précise et assez robuste [Keller et al, 1997]. De plus, en utilisant YIN nous pouvons coordonner l'extraction du taux de voisement avec l'extraction de la fréquence fondamentale dans une phase. Notre processus suivant est utilisé pour extraire ces deux paramètres :

1. Calculer la fonction $d_m(k)$ où τ est le retard du temps.

$$d_m(k) = \sum_{n=m}^{m+N-1} (s_n - s_{n+k})^2 \tag{6.12}$$

où $0 \leq k < N$

2. Évaluer la fonction de la différence moyenne cumulative normalisée

$$d'_m(k) = \begin{cases} 1, & k = 0 \\ \dfrac{d_m(k)}{\dfrac{1}{k}\sum_{j=1}^{j=k} d_m(j)}, & \text{sinon} \end{cases} \quad (6.13)$$

3. Chercher le minimum local dans l'intervalle :

$$(1-\alpha)\frac{F_e}{Pitch_{\max}} < k < \alpha\frac{F_e}{Pitch_{\min}}$$

Dénoter cette valeur minimum $d'_m(k')$.

4. Pour notre système la fréquence fondamentale F_o et l'intensité de voisement vm sont obtenus par :

$$F_0 = \frac{F_e}{k'} \quad (6.14)$$

$$v_m = 1 - d'_m(k') \quad (6.15)$$

La valeur de v_m s'étend de 0 à 1 et plus elle est grande, plus l'intensité de voisement n'est forte. Une valeur de v_m supérieure à un seuil absolu ($v_m > C$) dénote que ce morceau de signal est voisé. La valeur de C est choisi par l'expérimentation, par exemple [Ryynänen et al, 2004] a choisi la valeur 0.85 en travaillant avec la musique, et 0.9 dans [De Cheveigné et al, 2002]. Pour notre cas nous avons choisi 0.2 pour *C*.

$$V_m = \begin{cases} 1, & v_m > C \\ 0 & \text{sinon} \end{cases} \quad (6.16)$$

Calculer le taux de voisement en durée TV_m par la moyenne des V_m pendant une unité de temps. Etant donné M est le nombre de trames dans une unité de temps. Nous avons choisi 200ms pour une unité de temps qui correspondante avec la durée moyenne un phonème. Avec 10ms de chevauchement entre des fenêtres d'analyse, *M* est donc égal à 20.

$$TV_m = \frac{1}{M}\sum_{j=1}^{M} V_m \quad (6.17)$$

Pour améliorer la performance de ce processus, le prétraitement : l'évidement central est appliqué sur le signal avant de calculer (6.12) :

$$s[n] = \begin{cases} s[n] & si\ s[n] \geq C_L \\ 0 & si\ s[n] < C_L \end{cases} \quad (6.18)$$

avec CL est égal à 30% de maximum de s[n] dans l'intervalle : $0 \leq n < N$