ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

THÈSE PRÉSENTÉE À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

COMME EXIGENCE PARTIELLE
À L'OBTENTION DU
DOCTORAT EN GÉNIE
Ph. D.

PAR
Yazid ATTABI

RECONNAISSANCE AUTOMATIQUE DES ÉMOTIONS SPONTANÉES À PARTIR DU SIGNAL DE PAROLE

MONTRÉAL, LE 30 NOVEMBRE 2015

PRÉSENTATION DU JURY

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE :

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

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

M. Patrick Cardinal, membre du jury
Département de génie logiciel et des technologies de l'information à l'École de technologie supérieure

M. Sid-Ahmed Selouani, examinateur externe
Université de Moncton, Campus de Shippagan

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

LE 1er OCTOBRE 2015

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEUR

REMERCIEMENTS

Au terme de ce travail, je tiens à remercier mon directeur de recherche M. Pierre Dumouchel, Directeur général de l'École de technologie supérieure pour m'avoir donné l'opportunité de faire cette recherche de doctorat, après celle de la maitrise, pour ses précieux conseils et aussi pour sa confiance de me laisser libre d'explorer de nouvelles idées.

J'aimerai remercier M. Robert Sabourin, M. Patrick Cardinal et M. Sid-Ahmed Selouani qui ont accepté de participer au jury d'évaluation de cette thèse. Mes remerciements vont également à Mme Narjes Boufaden et encore une fois à M. Robert Sabourin pour leurs pertinentes remarques et suggestions émises lors de mon examen doctoral.

Je remercie également le centre de recherche informatique de Montréal (CRIM) et l'équipe de reconnaissance de la parole (RECO) où mes recherches de doctorat ont été effectuées. Je remercie en particulier le directeur de l'équipe RECO, Gilles Boulianne ainsi que les chercheurs principaux Patrick Kenny et Vishwa Cupta pour les riches et intéressantes discussions que j'avais eu avec eux et pour leur généreuse disponibilité. Je n'oublierai pas mes amis et collègues présents ou anciens membres de l'équipe RECO : Senoussaoui Mohamed, Jahangir Mohamed, Frédéric Osterrath, Themos Stafylakis, Walid Ziani, Pierre Ouellet et Najim Dehak.

J'adresse mes remerciements également au conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) ainsi qu'aux laboratoires universitaires Bell (LUB) pour avoir financé cette recherche.

Enfin, je ne remercierai jamais assez ma grande famille, et en particulier ma femme et mes enfants, pour leurs encouragements et leur patience inlassable durant ces années d'études.

À ma *mère* et à la mémoire de mon *père*, envers qui je resterai toujours redevable, je dédie ce travail.

RECONNAISSANCE AUTOMATIQUE DES ÉMOTIONS SPONTANÉES À PARTIR DU SIGNAL DE PAROLE

Yazid ATTABI

RÉSUMÉ

La reconnaissance automatique des émotions (RAE) à partir de la parole est une tâche difficile particulièrement lorsqu'il s'agit de classer des expressions spontanées issues du monde réel. Les émotions spontanées sont souvent subtiles, parfois mixtes, de courtes durées, caractérisées par une grande variabilité intraclasse en plus d'avoir une distribution de classes sévèrement biaisée. C'est dans ce contexte que s'inscrit notre objectif de proposer une méthodologie capable d'améliorer les performances des systèmes de RAE actuels.

La méthodologie proposée est motivée par les connaissances a priori sur les modèles théoriques des émotions en psychologie. L'idée est d'intégrer les concepts du modèle d'émotion dimensionnel dans la conception de classificateurs d'émotions discrètes. Deux concepts ont été dégagés du modèle dimensionnel : l'existence d'un espace dimensionnel dans lequel les émotions catégoriques peuvent être projetées et l'existence d'une relation de proximité entre ces catégories d'émotion relativement à chacune de ces dimensions. Le premier concept s'est traduit par l'extraction de traits de haut niveau destinés à jouer un rôle similaire à celui incarné par les dimensions du modèle théorique. Le second a motivé l'adoption d'une approche basée sur la similarité pour la représentation et la classification des émotions. Nous avons montré que les scores de vraisemblances générés par les modèles GMM constituent de puissants traits de similarité pour la RAE et répond bien à la contrainte relative à la taille limitée des énoncés.

Nous avons proposé une première méthode de classification, intitulée *le plus proche patron de similarité pondéré*. Cette méthode est bâtie autour d'un nouveau vecteur caractérisant un énoncé à travers la description de son patron de classes voisines. Les classes au sein d'un patron sont ordonnées selon leurs degrés de proximité estimés sur la base des scores de vraisemblance. Contrairement à la règle de décision Bayes, les rangs de tous les scores influencent la décision de classification. Deux types de modèles ont été proposés et expérimentés : linéaire et non linéaire.

Nous avons proposé également les modèles d'ancrage comme méthode de classification des émotions mais aussi comme outils d'aide à l'analyse du contenu émotionnel utile en psychologie. Les énoncés sont projetés dans un espace continu où chaque dimension est engendrée par un modèle de classe d'émotion qui mesure le degré de similarité d'un énoncé avec cette classe. Nous avons montré qu'il était possible d'appliquer avec succès les modèles d'ancrage aussi bien dans un contexte d'un problème multi-classe que celui d'une classification binaire- à travers une extension de l'espace d'ancrage avec de nouveaux modèles externes. Nous avons analysé et comparé les performances des modèles d'ancrage basés sur la distance euclidienne et cosinus en se basant sur les propriétés géométriques de leurs frontières de décision. Par ailleurs, nous avons montré que les modèles d'ancrage

peuvent servir aussi comme méthode puissante de combinaison de classificateurs moyennant une normalisation des scores plus adaptée au contexte de fusion. Leurs bonnes performances et leurs propriétés intéressantes (ex., insensibilité à la distribution biaisée des classes) font des modèles d'ancrage des solutions très adéquates au problème de RAE comparés à d'autres systèmes plus complexes.

Enfin, sur le plan des descripteurs acoustiques, de nouveaux traits plus discriminatifs ont été proposés. La combinaison de ces traits au moyen des modèles d'ancrage a permis de dépasser les résultats de l'état de l'art quand testés sur FAU AIBO Emotion, un corpus d'émotions spontanées commun à la communauté de recherche en RAE.

Mots clés : Reconnaissance des émotions, similarité, patron de voisinage pondéré, modèles de référence, combinaison de classificateurs, AMCC, Spectrum multitaper.

AUTOMATIC EMOTION RECOGNITION IN SPONTANEOUS SPEECH

Yazid ATTABI

ABSTRACT

Automatic emotions recognition (AER) from speech is a challenging task especially when dealing with real-life affective expressions. Spontaneous emotions are often subtle, sometimes mixed, of short periods, with large intra-class variability, in addition to have a skewed class distribution. It is in this context that our objective to propose a methodology capable of improving the performance of current AER systems is inscribed.

The proposed methodology is motivated by prior knowledge on theoretical models of emotion in psychology. The idea is to integrate the concepts of dimensional emotion model in the design of discrete emotions classifiers. Two concepts were identified from the dimensional model: the existence of a dimensional space in which categorical emotions can be projected and the existence of a similarity relationship between these categories of emotion with respect to each of these dimensions. The first concept leads to the extraction of high-level features that are intended to play a role similar to that played by the dimensions of the theoretical model. The second concept has motivated the adoption of a similarity-based approach for emotions representation and classification. We have shown that the likelihood scores generated by GMM models are powerful similarity-based features for AER task and responds well to the issue of the short duration length of utterances.

We have proposed a first method of classification, entitled *weighted ordered class-nearest neighbors*. This method is built around a new feature vector describing an utterance by its pattern of neighboring emotion classes. The classes inside the pattern are ordered according to their proximities and estimated on the likelihood scores basis. Unlike the Bayes decision rule, the ranks of all scores influence the classification decision. Two types of models have been proposed and tested: linear and nonlinear.

We also proposed anchor models as emotion classification method but which can be also used as a tool for emotional content analysis in psychology studies. The utterances are projected in a continuous space where each dimension is spanned by an emotion class model that measures the similarity level of an utterance with respect to this class. We have shown that it is also possible to successfully apply anchor models for multi-class problem context as for a binary classification one by expanding the anchor space with new external models. We analyzed and compared Euclidean- and cosine-based anchor models performances based on geometric properties of their decision boundaries. Furthermore, we showed that the anchor models can also be used as a powerful method of combining classifiers subject to scores normalization more suited for the fusion context. Their performances and their properties (e.g., insensitivity to skewed class distribution) make of anchor models very suitable solutions for AER task compared to more complex systems.

Finally, in terms of acoustic descriptors, new and more discriminative features have been proposed. The results achieved by fusion of these features using the anchor models outperformed the state-of-the-art when tested on FAU Emotions AIBO, a benchmark spontaneous emotion corpus for the AER research community.

Keywords: Emotion recognition, similarity, weighted neighborhood pattern, anchor models, classifiers combination, AMCC, multitaper spectrum.

TABLE DES MATIÈRES

Page

CHAPITRE 1 INTRODUCTION 1
1.1 Problématique 3
1.1.1 Incertitudes relatives à la définition de l'émotion en psychologie 4
1.1.2 Nature dynamique de l'émotion 5
1.1.3 Bruit au niveau des corpus des émotions 5
1. Erreurs induites par l'opération d'annotation : 5
2. Conditions d'enregistrement et de transmission du signal : 6
1.1.4 Chevauchement entre classes d'émotions dans l'espace des traits acoustiques 6
1.1.5 Propriétés statistiques des corpus de données 7
1.2 Objectif 8
1.3 Applications 8
1.4 Organisation de cette thèse 11

CHAPITRE 2 THÉORIES DES ÉMOTIONS 13
2.1 Définition des émotions 13
2.2 Expressions émotionnelles entre les effets *pousser* et *tirer* 14
2.3 Modèles psychologiques des émotions 15
2.3.1 Théorie de l'émotion discrète 15
2.3.2 Théorie dimensionnelle 16
2.3.3 Théorie de l'évaluation cognitive 17
2.3.4 Modèle d'émotion à composantes 17
2.4 Corpus de parole émotionnelle 18
2.4.1 Type de corpus des émotions 19
2.4.1.1 Émotions naturelles 19
2.4.1.2 Émotions simulées 19
2.4.1.3 Émotions induites 20
2.4.2 Constitution d'un corpus de parole émotionnelle 22
2.4.2.1 Collection des enregistrements 22
2.4.2.2 Annotation du corpus 23
2.4.2.3 Validation du corpus 24
2.4.3 Corpus de données d'émotion dans la revue de littérature 25
2.5 Conclusion 26

CHAPITRE 3 REVUE DE LITTÉRATURE SUR LES SYSTÈMES DE RECONNAISSANCE AUTOMATIQUE DES ÉMOTIONS 29
3.1 Introduction 29
3.2 Travaux basés sur des classificateurs simples 30
3.2.1 Travaux selon le type d'unité d'analyse 31
3.2.2 Travaux selon le type des traits caractéristiques 34
3.2.2.1 Prosodie 34

3.2.2.2 Traits spectraux 36
3.2.2.3 Traits de la qualité de la voix 37
3.2.2.4 Éclats affectifs 39
3.2.2.5 Information linguistique 40
3.2.2.6 Sélection des traits caractéristiques 41
3.2.3 Travaux selon la portée des traits caractéristiques 42
3.2.3.1 Information à court terme 43
3.2.3.2 Information à long terme 43
3.2.4 Travaux selon l'approche de classification 45
3.2.4.1 Approche dynamique 45
3.2.4.2 Approche statique 46
3.2.4.3 Approche logique floue 47
3.2.4.4 Approche basée sur la similarité 47
3.3 Combinaison de classificateurs 49
3.3.1 Combinaison en cascade 50
3.3.2 Combinaison hiérarchique 53
3.3.3 Combinaison parallèle 55
3.3.3.1 Diversification dans les types de traits 55
3.3.3.2 Diversification dans la portée temporelle de l'information acoustique 55
3.3.3.3 Diversification des unités d'analyse 56
3.3.3.4 Diversification des unités d'analyses et des types de descripteurs 56
3.3.3.5 Diversification des modèles de classification 56
3.3.3.6 Diversification des types et portées de traits, d'unités d'analyse et de modèles de classification 57
3.3.4 Combinaison série 58
3.4 Techniques d'amélioration des performances des systèmes de RAE 58
3.4.1 Techniques basées sur l'exploitation de l'information sur le mode opératoire 59
3.4.1.1 Mode dépendant- versus indépendant du locuteur 59
3.4.1.2 Mode dépendant versus indépendant du genre 59
3.4.2 Techniques basées sur le traitement du problème de rareté des données d'apprentissage 60
3.4.2.1 Combinaison de plusieurs corpus 61
3.4.2.2 Co-apprentissage 61
3.4.2.3 Étiquetage actif 62
3.5 Conclusion 63

CHAPITRE 4 MÉTHODOLOGIE ET APPROCHE BASÉE SUR LA SIMILARITÉ POUR LA CLASSIFICATION DES ÉMOTIONS 65
4.1 Introduction 65
4.2 Approche basée sur la similarité 66
4.2.1 Motivation 66
4.2.2 Traits basés sur la similarité 67

4.2.3 Méthodes de classification 68
4.3 Corpus de parole émotionnelle FAU AIBO Emotion 70
4.4 Protocole d'expérimentation 73
4.5 Choix des descripteurs de haut niveau 74
4.5.1 Supervecteurs et dérivées 75
4.5.1.1 Modélisation par mélange de gaussiennes 76
4.5.1.2 Méthode d'estimation du maximum de vraisemblance 77
4.5.1.3 Adaptation MAP 78
4.5.1.4 Adaptation MLLR 78
4.5.1.5 Combinaison de MLLR et MAP 79
4.5.1.6 Expérimentations 81
4.5.1.7 Analyse discriminante linéaire probabiliste (PLDA) 83
4.5.2 Scores de vraisemblance comme traits de haut niveau 85
4.5.2.1 Motivation 85
4.5.2.2 Scores de vraisemblance et mesure de similarité 87
4.5.2.3 Vecteur de traits VCE et l'analyse des émotions 88
4.6 Conclusion 89

CHAPITRE 5 MÉTHODE DU PLUS PROCHE PATRON DE SIMILARITÉ PONDÉRÉ 91
5.1 Introduction 91
5.2 Vue d'ensemble du système WOC-NN 92
5.3 Patron de proximité 94
5.4 Métrique de mesure de similarité 95
5.4.1 Pondération des rangs de classe 96
5.4.2 Modèle de régression logistique 98
5.4.3 Génération des données d'entrainement 100
5.4.4 Réduction de la dimensionnalité 101
5.4.5 Normalisation de la pondération 102
5.5 Interaction entre les classes dans un patron de proximité 103
5.6 Résultats expérimentaux 106
5.6.1 Patrons de proximité des classes d'émotion du corpus FAU AIBO Emotion 106
5.6.2 Résultats de la classification 109
5.6.3 Résultats du modèle non linéaire 110
5.7 Conclusion 112

CHAPITRE 6 MODÈLES D'ANCRAGE POUR LA RECONNAISSANCE MUTICLASSES D'ÉMOTION 115
6.1 Introduction 115
6.2 Modèles d'ancrage 116
6.2.1 Construction de l'espace d'ancrage 116
6.2.2 Mappage dans l'espace d'ancrage 117
6.2.3 Classification des énoncés émotionnels 118
6.3 Configuration expérimentale 119

6.4 Problème des données bruitées avec la métrique euclidienne 123
6.5 Normalisation des scores 127
6.5.1 Normalisation de la covariance intraclasse..... 127
6.5.2 Résultats et discussion 129
6.6 Vecteurs représentative des classes 131
6.6.1 Représentant unique versus représentants multiples..... 132
6.6.2 Résultats expérimentaux 133
6.7 Comparaison avec des systèmes dorsaux plus complexes.....137
6.7.1 Traitement du problème de distribution biaisée des classes de données 138
6.7.2 Résultats expérimentaux 139
6.8 Conclusion 143

CHAPITRE 7 MODÈLES D'ANCRAGE : PROPRIÉTÉS ET APPLICATION À UNE CLASSIFICATION BINAIRE.....145
7.1 Introduction.....145
7.2 Analyse géométrique des modèles d'ancrage dans espace bidimensionnel 147
7.2.1 Métrique euclidienne 147
7.2.2 Similarité cosinus..... 149
7.2.3 Relation entre les vecteurs représentatifs de classe et la métrique de similarité 150
7.2.4 Propriétés des vecteurs représentatifs de classe..... 152
7.3 Comparaison entre des règles décision Bayes et modèles d'ancrage 156
7.4 Expérimentation des modelés à ancrage à espace bidimensionnel.....160
7.4.1 Configuration expérimentale 160
7.4.2 Résultats avant la normalisation WCCN 161
7.4.3 Effet géométrique de la normalisation WCCN..... 164
7.5 Espace d'ancrage multidimensionnel 166
7.5.1 La distance euclidienne..... 167
7.5.2 La mesure cosinus..... 168
7.5.3 Résultats expérimentaux dans un espace d'ancrage à cinq dimensions.. 169
7.6 Comparaison de la complexité algorithmique et optimisation 172
7.7 Conclusion 173

CHAPITRE 8 LES MODÈLES D'ANCRAGE POUR LA COMBINAISON DE CLASSIFICATEURS.....175
8.1 Introduction.....175
8.2 Nouveaux traits spectraux pour la reconnaissance des émotions 175
8.2.1 Estimation du Spectrum multitaper 177
8.2.2 Extraction des MFCC et PLP multitaper 179
8.2.3 Extraction des traits AMCC..... 181
8.2.4 Évaluation des performances individuelles des traits proposés..... 183
8.2.4.1 Résultats des traits multitaper et des traits AMCC 183
8.3 Complémentarité des traits 184
8.3.1 Analyse des matrices de confusion..... 184
8.3.2 Combinaison des traits..... 185

8.4 Fusion avec les modèles d'ancrage....186
8.4.1 Définition de l'espace d'ancrage de fusion.... 186
8.4.2 Normalisation des scores 187
8.4.3 Résultats expérimentaux des modèles d'ancrage de fusion.... 188
8.4.4 Modèles d'ancrage versus autres méthodes de combinaison 189
8.5 Conclusion 191

CONCLUSION GÉNÉRALE....193

RECOMMANDATIONS 197

ANNEXE I CORPUS DE PAROLE ÉMOTIONNELLE....199

ANNEXE II LE CORPUS FAU AIBO EMOTION....203

ANNEXE III MÉTHODES D'ESTIMATION DES PARAMÈTRES....211

LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES....217

LISTE DES TABLEAUX

Page

Tableau 4.1 Matrice de confusion d'étiquetage (en %) des cinq annotateurs du corpus FAU AIBO Emotion après regroupement des petites catégories d'émotions (compilé à partir des valeurs du Tableau-A II-3 de l'ANNEXE II). 72

Tableau 4.3 Performances des supervecteurs en fonction de la méthode d'adaptation utilisée pour entrainer l'extracteur. Ces résultats sont obtenus sur les données de test du corpus d'émotion FAU AIBO. Le nombre (*D*) représente la dimension optimisée du supervecteur..... 82

Tableau 5.1 Patron de proximité du modèle linéaire de chacune des classes du corpus FAU AIBO Emotion ainsi que le poids associé à chaque rang appris à partir des données d'entrainement. 107

Tableau 5.2 Pondération des rangs du modèle non-linéaire des patrons de proximité à double interaction de chacune des classes du corpus FAU AIBO Emotion apprise à partir des données d'entrainement 109

Tableau 5.3 Résultats de classification du système WOC-NN obtenus sur les données de test en fonction des différentes configurations du vecteur de pondération du modèle linéaire 110

Tableau 5.4 Effets de la normalisation des poids et de l'interaction entre les rangs de classes sur les performances des systèmes WOC-NN testés sur le corpus d'émotion FAU AIBO. 111

Tableau 5.5 Comparaison des résultats du système proposé avec les meilleurs systèmes de la compétition *INTERSPEECH 2009 Emotion Challenge*, testés sur le corpus FAU AIBO Emotion 112

Tableau 6.1 Résultats des différents modèles d'ancrage évalués avec les données de test du corpus FAU AIBO Emotion. 131

Tableau 6.2 Comparaison des modèles d'ancrage de type *multifold* et *unifold* évalués sur les données de test du corpus FAU AIBO Emotion..... 136

Tableau 6.3 Comparaison des trois différents types d'architectures évaluées sur les données de de test du corpus FAU AIBO Emotion..... 140

Tableau 7.1 Pertinence des métriques euclidienne et cosinus pour les modèles d'ancrage en fonction de la valeur de la pente du segment reliant les vecteurs représentatifs des classes. 156

Tableau 7.2 Résultats UAR et WAR pour les dix expériences à deux classes évaluées sur des données de test du corpus FAU AIBO Emotion. Les performances des modèles d'ancrage sont évaluées dans un espace d'ancrage à cinq dimensions. 171

Tableau 8.1 Résultats des traits MFCC et PLP multitaper ainsi que les traits AMCC comparés aux traits MFCC et PLP conventionnels..... 184

Tableau 8.2 Résultats UAR de la combinaison de classificateurs avec les modèles d'ancrage en fonction des méthodes de normalisation. 189

Tableau 8.3 Tableau récapitulatif des meilleurs résultats obtenus pour chacune des méthodes de combinaison expérimentées sur les données de test du corpus FAU AIBO Emotion. 191

LISTE DES FIGURES

Page

Figure 1.1 Distribution des scores de décision pour les deux classes positives et négatives du corpus FAU AIBO Emotion. 3

Figure 1.2 Émotions dans l'espace Activation-évaluation. 7

Figure 4.1 Répartition des classes du corpus FAU AIBO Emotion pour les partitions d'entrainement et de test. 73

Figure 5.1 Exemple d'un patron de voisinage calculé pour la classe *colère* à partir des données d'apprentissage du corpus FAU AIBO Emotion. 93

Figure 5.2 La partie supérieure de cette figure illustre le processus de génération de patron proximité d'un énoncé **X**. La partie inférieure montre comment le vecteur de distance entre ce modèle et le patron de proximité de la classe d'émotion *k* est calculé. 96

Figure 5.3 Exemple de patron de distance calculé pour un énoncé **X** et le patron de proximité de la classe *colère*. 97

Figure 5.4 Schéma bloc de la méthode de génération des données d'entrainement utilisées l'apprentissage des coefficients de pondération de la classe *k* en utilisant la régression logistique. 101

Figure 5.5 Exemple de patron de proximité à interaction double entre les rangs de classe calculé pour la classe *colère*. 104

Figure 6.1 Exemple d'espace d'ancrage tridimensionnel engendré par les modèles des classes d'émotion *colère*, *emphatique* et *positive*. 117

Figure 6.2 Résultats UAR obtenus en utilisant la validation croisée à 9 plis s ur les données d'entrainement du corpus FAU AIBO Emotion en fonction du nombre de gaussiennes du GMM. Deux systèmes de modèles d'ancrage sont comparés : un basé la distance euclidienne et le second sur la similarité cosinus 120

Figure 6.3 Dans chaque graphique sont tracées les valeurs de la moyenne et de la variance des scores de vraisemblance des données des classes d'émotion en fonction de chacune des dimensions de l'espace d'ancrage (modèle d'émotion composant les vecteurs VCE). Dans cette figure, les valeurs tracées représentent les valeurs statistiques des coordonnées cartésiennes des vecteurs VCE (voir l'équation (6.1)). 121

Figure 6.4 Dans chaque graphique sont tracées la moyenne et la variance des scores de vraisemblance des données d'une classe d'émotion à l'égard de chaque modèle d'émotion (composantes des vecteurs VCE). Dans cette figure, les valeurs tracées représentent les valeurs statistiques des valeurs angulaires des vecteurs VCE par rapport à la base standard (voir l'équation (6.7)). 123

Figure 6.5 Exemple où la distance euclidienne séparant un point **x** d'un point **a** est plus petite que la distance séparant **x** de **b** sur une échelle linéaire alors que sur l'échelle logarithmique, le point **x** devient plus proche du point **b** que du point **a** 124

Figure 6.6 Graphique 3D montrant l'évolution des valeurs de distances entre deux variables unidimensionnelles *x* et *y* pour toutes les valeurs de probabilité possibles *x* et *y* comprises entre 0 et 1. L'allure de cette évolution est illustrée pour les valeurs *x* et ***y*** avant (graphique du haut) et après (graphique du bas) l'application du logarithme 126

Figure 6.7 Effet de la normalisation WCCN sur les performances UAR des modèles d'ancrage en fonction du nombre de gaussiennes des GMMs. Les résultats sont obtenus en utilisant la validation croisée à neuf plis sur les données d'apprentissage du corpus FAU AIBO Emotion.....128

Figure 6.8 Diagramme à surfaces de la distribution des scores des énoncés de la classe *colère* (A) pour les cinq modèles d'émotion avant (en haut) et après (en bas) la normalisation WCCN. Sur chaque boîte, la marque centrale représente la médiane, les bords les 25e et 75e percentiles, les moustaches (*whiskers*) indiquent les données les plus extrêmes, et les valeurs aberrantes sont représentées individuellement. 130

Figure 6.9 Résultats UAR moyenne de 50 exécutions des modèles d'ancrage en fonction du nombre de vecteurs VCE sélectionnés comme vecteurs représentatifs de classe. À chaque itération, un nouveau sous-ensemble des données d'entraînement est aléatoirement sélectionné comme classe vecteurs représentatifs. Les performances sont évaluées sur les données d'entrainement en utilisant la validation croisée à neuf plis.....134

Figure 6.10 Résultats UAR des modèles d'ancrage en fonction du nombre de grappes par classe. Les centres de grappes sont utilisés classe vecteurs représentatifs. Les performances sont évaluées sur les données d'apprentissage en utilisant la validation croisée à neuf plis.....135

Figure 6.11 Résultats UAR des modèles d'ancrage en fonction du nombre de pôles par classe utilisées comme vecteurs représentatifs. Les grappes sont pondérées par une valeur proportionnelle à la taille de la classe. Les performances sont évaluées sur les données d'entrainement en utilisant la validation croisée à neuf plis. 136

Figure 6.12 Types d'architectures de systèmes basés sur les scores du logarithme de probabilité de vraisemblance..... 137

Figure 7.1 Résultats de 10 expériences de classification binaire réalisées sur les données de test du corpus FAU AIBO Emotion. La figure présente les résultats UAR des systèmes GMM-Bayes et les modèles d'ancrage basés sur la similarité cosinus. 146

Figure 7.2 Exemples de pentes des segments reliant les vecteurs représentatifs de deux classes dans l'espace d'ancrage. Le vecteur représentant de chaque émotion est calculée en utilisant les données d'entrainement du corpus FAU AIBO Emotion. 155

Figure 7.3 Exemples de frontières de décision linéaires obtenues avec les classificateurs à base de la règle de décision *Bayes* (bleu) et le modèle d'ancrage basé sur la distance euclidienne (rouge). Les deux classificateurs prédisent les mêmes classes pour les données situées dans les régions (A) et (B) et des classes différentes pour (C) et (D). 158

Figure 7.4 Gain/perte relatif en UAR pour les modèles d'ancrage à base des mesures cosinus et euclidienne par rapport au système GMM-Bayes. Les systèmes sont évalués sur les données de test du corpus FAU AIBO Emotion. 161

Figure 7.5 Les régions de décision sont établies pour chacun des dix problèmes à 2-classes avant normalisation. La frontière de décision est représentée en bleu pour la règle Bayes, en rouge pour la distance euclidienne et en vert pour le cosinus. Les points représentatifs des classes sont reliés par une ligne continue. 163

Figure 7.6 Gain relatif en performance UAR pour les modèles d'ancrage à base des mesures cosinus et euclidienne par rapport au système GMM-Bayes après la normalisation WCCN. Les systèmes sont évalués sur les données de test du corpus FAU AIBO Emotion..... 164

Figure 7.7 Les régions de décision sont établies pour chacun des dix problèmes à 2-classes après la normalisation WCCN. La frontière de décision est dessinée en bleu pour le système GMM-*Bayes*, en vert pour le

cosinus et rouge pour la distance euclidienne. Les points représentatifs des classes sont reliés par une ligne continue.....165

Figure 7.8 Comparaison du gain relatif en par rapport aux systèmes GMM-*Bayes* pour les modèles d'ancrage à base des mesures euclidienne et cosinus dans l'espace à cinq-dimensions avant et après la normalisation WCCN. Les systèmes ont été évalués sur des données de test du corpus FAU AIBO Emotion. 169

Figure 7.9 Comparaison du gain relatif en UAR par rapport aux systèmes GMM-*Bayes* pour les modèles d'ancrage basés sur un espace à deux versus à cinq dimensions. Les performances des modèles d'ancrage sont évaluées après la normalisation WCCN en utilisant les données de test du corpus FAU AIBO Emotion.....170

Figure 8.1 Multitapers Thomson pour N = 256, $M = 6$ (a) du temps et (b) des domaines de fréquence.....179

Figure 8.2 Schéma illustrant l'extraction des traits MFCC et PLP basée sur l'estimation du spectre unique et multitaper. 180

Figure 8.3 Bloc diagramme du processus d'extraction des traits AMCC. 181

Figure 8.4 Résultats de classification par classe d'émotion en fonction du type de traits utilisé.....185

Figure 8.5 Résultats de classification au moyen de méthodes impliquant un apprentissage, en fonction de la méthode d'échantillonnage. Les résultats UAR sont obtenus en utilisant les données de test du corpus FAU AIBO Emotion. 190

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AAAC	Association for the Advancement of Affective Computing
AM-FM	Amplitude Modulation- Frequency Modulation
ACP	Analyse en Composantes Principales
AER	Automatic Emotion Recognition
AIBO	Chien robot de compagnie fabriqué par *Sony*
AMCC	Amplitude Modulation Cepstral Coefficients
BLSTM	Bidirectionnel Long Short-Term Memory
CMN	Cepstral Mean Normalization
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
EM	Espérance-Maximisation (Expectation-Maximization)
EP	Emotion Profile
F0	Fréquence fondamentale
FAU	Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
GMM	Gaussian Mixture Model
GMAVR	Gaussian Mixture Vector Autoregressive Models
GVV	Glottal Volume Velocity
HNR	Harmonics to Noise Ratio
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit (boîte à outils de reconnaissance vocale)
IDFT	Inverse Discrete Fourier Transform

INTERSPEECH	International Speech Communication Association
KNN	*k* plus proches voisins (k-*Nearest Neighbor*)
LDA	Analyse discriminante linéaire (Linear discriminant analysis)
LDC	Linguistic Data Consortium
LFPC	Log Frequency Power Coefficients
LP	Linear Prediction
LPC	Linear Predictive Coding coefficients
LPCC	Linear Prediction Cepstral Coefficients
LSTM	Long Short-Term Memory
MA	Modèles d'ancrage
MA-COS	Modèles d'ancrage basé sur la mesure cosinus
MA-EUC	Modèles d'ancrage basé sur la mesure euclidienne
MAP	Maximum A Posteriori
MFCC	Mel-Frequency Cepstral Coefficients
MLLR	Maximum Likelihood Linear Regression
ML	Maximum Likelihood estimation
MLP	Multilayer Perceptron
MONT	Montessori-Schule (école dans la ville Erlangen)
NEO	Nonlinear Energy Operator
NMC	Nearest Mean Classifier
OC-NN	WOC-NN sans pondération des rangs des patrons de voisinage
OHM	Ohm-Gymnasium (école dans la ville Erlangen)

PLDA	Probabilistic Linear Discriminant Analysis
PLP	Perceptual Linear Prediction
RAE	Reconnaissance automatique des émotions
RASTA PLP	Relative Spectral Perceptual Linear Predictive
RBF	Radial basis function
RBM	Restricted Boltzmann Machine
SAL	Sensitive Artificial Listener
SFFS	Sequential Floating Forward Selection algorithm
SFS	Sequential Forward Selection
SMOTE	Synthetic Minority Oversampling Technique
SNEO	Smoothed Nonlinear Energy Operator
SVM	Support Vector Machine
SW-OC-NN	WOC-NN basé sur un vecteur de pondération unique partagé
SWCE	Sinusoidal Weighted Cepstrum Estimator
TEO	*Teager* Energy Operator
UAR	Unweighted Average Recall
UBM	Universal Background Model
VCE	Vecteur caractéristique d'émotion
WAR	Weighted Average Recall
WCCN	Within-Class Covariance Normalization
WOC-NN	Weighted Ordered Classes-Nearest Neighbors
WOC-NN-2int	Modèle du système WOC-NN basé sur l'interaction double des rangs
WOC-NN-3int	Modèle du système WOC-NN basé sur l'interaction triple des rangs

WOC-NN-WFS	WOC-NN sans opération de sélection de traits (rangs)
WOZ	Wizard of OZ

LISTE DES SYMBOLES ET UNITÉS DE MESURE

$[.]^T$	Opérateur de transposé
$\otimes$	Opérateur de convolution
$\langle \mathbf{SV}_1, \mathbf{SV}_2 \rangle$	Produit scalaire des vecteurs $\mathbf{SV}_1$ et $\mathbf{SV}_2$
$\|\mathbf{SV}\|$	La norme euclidienne du vecteur **SV**
$\binom{n}{k}$	Nombre de combinaisons possibles de *k* parmi *n*
!	Factorielle d'un nombre entier
$\|\hat{a}(c,n)\|$	Composante AM du *c*-ième canal
$\#E$	Représente le cardinal de l'ensemble *E*
O	**M**atrice remplie de zéro
A	Matrice de décomposition de *Cholesky*
arccos	Fonction mathématique Arc cosinus, réciproque de la fonction *cosinus*
$\underset{i=1,\ldots,C}{\operatorname{argmin}}(\cdot)$	Argument du minimum
$b_i(\mathbf{x})$	Fonction de densités normales multidimensionnelles
Bark	Échelle psychoacoustique (bande de fréquence)
C	Nombre de classes d'émotion
cm^2	Centimètre carré
D	Dimension des vecteurs caractéristiques
$d(\cdot,\cdot)$	Fonction de calcul de distance
$\mathcal{D}_{wH}(,)$	Distance de *Hamming* pondérée
dB	Décibel
Diag()	Matrice diagonale
$(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_C)$	Base standard
exp	Fonction exponentielle
E_i	Classe d'émotion *i*
Hz	Hertz (cycles par seconde)
K	Kilo
kHz	Kilo Hertz
L(**X**)	Vecteur de logarithme des probabilités de vraisemblance d'un énoncé **X**
log	Logarithme
Logit	Fonction *log odds*
Mel	Échelle psycho acoustique de hauteurs des sons
(mc_{ij})	Nombre de données de la classe *i* classifiés dans la classe *j*
ms	Milliseconde
n_i	Nombre de données d'apprentissage appartenant à la classe *i*
n_i^j	Nombre de données d'apprentissage de la grappe *j* de la classe *i*
$O()$	Notation grand O
$p(i)$	Probabilité a priori de la classe *i*
$p(\mathbf{x}\|\lambda)$	Densité de probabilité d'un GMM
r	Vecteur de type *patron de proximité*

$\mathbf{S}_i$	Matrice de covariance de la classe *i*
$\hat{S}_d(m,k)$	Estimation du périodogramme fenêtré pour m^{e} trame et k^{e} série de fréquence
SV	Supervecteur
$s(m,j)$	Signal de parole dans le domaine temporel
$\hat{S}_{MT}(\cdot)$	Estimation spectrale multitaper
T	Taille d'une séquence de trames correspondant à un énoncé de parole
v	Vecteur de type de patron de distance entre deux patrons de proximité
W	Matrice de covariance intraclasse
Wald	Test statistique de *Chi-deux*
Watt	Unité de quantification de la puissance
$\mathbf{w}_i$	Vecteur de pondération associé à un patron de proximité de classe *i*
$w(j)$	Fonction de fenêtrage dans le domaine temporel
X	Une séquence de vecteurs acoustiques associée à un énoncé de parole
x	Vecteur de trait acoustique associé à une trame de parole
$x(c,n)$	*n*-ième trame vocale du *c*-ième canal
z-score	Normalisation de la moyenne et de la variance
μ_k	Moyenne de la classe *k*
λ	Modèle GMM d'une classe d'émotion
$\lambda(p)$	Coefficient de pondération associé au *p-ième* périodogramme fenêtré
$\mathbf{\Sigma}_i$	Matrice de covariance de la *i*-ième gaussienne
ξ_j	Variable d'écart (*slack variable*)
γ_j	Valeur de pondération associée au point de données *j*
$\Psi(x(c,n))$	Le NEO standard de $x(c,n)$

CHAPITRE 1

INTRODUCTION

Par nos discours, et au-delà du message sémantique que nous voulons communiquer, nous transmettons à notre interlocuteur, sciemment ou inconsciemment, des informations sur nos traits personnels tels que notre identité, âge, genre, état de santé, personnalité et surtout notre émotion. D'ailleurs, d'après Mehrabian and Wiener (1967), quand il s'agit de communiquer ses sentiments et ses états d'esprits, le message verbal compte pour 7 % de l'information transmise alors que la communication non-verbale et visuelle compte pour 38 % et 55 % respectivement.

En effet les **signes non-verbaux** tels que l'intonation, le rythme, l'interruption de la phonation, le cri, le silence, les interjections ou les éclats affectifs sont des indices véhiculés par notre voix qui renseignent sur notre état affectif. Outre la voix, l'émotion peut également être perçue à travers : (i**)** nos **expressions faciales** telles que le froncement de sourcils, l'orientation du regard, diamètre de la pupille, larmes, rougissement, rides, étirement de la bouche, bâillement; (ii) nos **signes physiologiques** tels que la salivation, la respiration (changement du rythme, soupir, une inspiration soudaine), la transpiration (de Melo and Gratch, 2009), la conductance de la peau (Kapoor *et al.* 2007) et la température; (iii) nos **postures corporelles** telles que l'orientation et la position de la tête, le type du geste (tête, mains); et enfin (iv) notre **manière d'accomplir une tâche** (dynamisme, fluidité, impulsivité et vitesse) (Castellano *et al.* 2010), comme par exemple la manière de frapper à la porte (Pollick *et al.* 2001), de conduire une voiture (McMahon *et al.* 2008) ou le degré de pression que nous exerçons sur une souris (Kapoor *et al.* 2007). Ce reflet s'explique par l'effet induit par le changement d'émotion sur nos actions- concentration, jugement d'un risque et sur notre niveau d'énergie (Cowie, *et al.* 2010a).

Parce que ces expressions audibles ou visibles extériorisent un état interne, elles constituent par conséquent un aspect majeur de la communication sociale qui informent les autres sur nos sentiments, nos préférences (prédicteur de satisfaction d'un service), nos réactions à des évènements et sur nos intentions aux actes (Scherer *et al.* 2010a). Les intentions sociales de ces messages émotionnels sont décodées par notre interlocuteur, qui lui aussi peut être affecté à son tour de façon consciente ou inconsciente par ces messages (Bänziger *et al.* 2010b).

Par ailleurs, face à l'ubiquité informatique (omniprésence de systèmes informatiques) croissante dans notre vie quotidienne, le contexte d'interaction humain-humain se voit de plus en plus remplacé par celui d'une interaction humain-machine. Dans ce contexte, les messages affectifs envoyés par un interlocuteur humain resteront lettres mortes tant et aussi longtemps que la machine est *affectivement* « sourde ». Nous pouvons penser par exemple au cas d'un client confus ou en colère ayant des difficultés à interagir avec un système de réponse automatisé. Dans cette perspective, doter les machines d'une compétence émotionnelle permettra de révolutionner ce nouveau mode d'interaction. Munir les machines d'une telle compétence revient à développer des systèmes capables de reconnaitre et de synthétiser automatiquement les émotions. Cette thèse s'intéressera particulièrement à la partie reconnaissance des émotions de l'informatique affective à partir du signal de parole. Bien que le domaine de la recherche sur la détection automatique des émotions véhiculées par la voix est relativement nouveau, ce domaine connait un intérêt croissant et un engouement particulier vu l'étendue des domaines applicatifs pouvant bénéficier de cette technologie. Cependant, détecter l'émotion d'un locuteur dans des conditions de la vie réelle demeure un défi. Nous montrerons dans la section 1 pourquoi dans de telles conditions la reconnaissance des émotions est un problème difficile (problématique de recherche). La section 2 décrira l'objectif que nous avons fixé dans cette thèse. Nous exposerons dans la section 3 des exemples sur les domaines d'application potentiels qui pourraient bénéficier de cette technologie avant de donner un aperçu sur l'organisation de cette thèse.

1.1 Problématique

Le grand d'intérêt affiché ces dernières années pour la reconnaissance automatique des émotions (RAE) à partir du signal de parole s'est traduit par l'extraction et l'exploration de milliers de traits caractéristiques et l'expérimentation de multitudes d'approches de classification. Pourtant, les performances des systèmes de RAE demeurent relativement basses pour un déploiement effectif. Les faibles taux de reconnaissance obtenus reflètent un niveau de confusion élevé entre les différentes classes d'émotion. Pour illustrer cette confusion, nous avons tracé dans la Figure 1.1 la distribution des scores de décision issue de la classification automatiques des deux classes d'émotion *positive* et *négative* du corpus FAU AIBO Emotion en utilisant les modèles GMM (Gaussian Mixture Model). Le tracé des deux distributions met clairement en évidence l'important chevauchement des scores de décision des deux classes d'émotion.

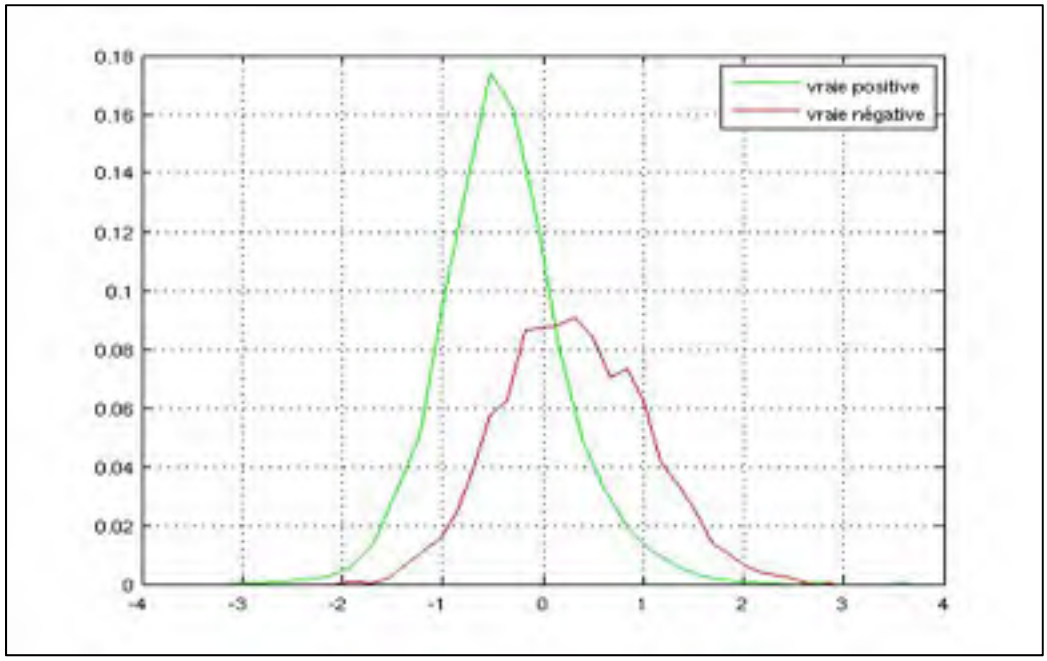

Figure 1.1 Distribution des scores de décision pour les deux classes positives et négatives du corpus FAU AIBO Emotion

D'ailleurs cette confusion n'est pas une caractéristique propre aux systèmes automatisés, mais touche aussi au même titre les personnes humaines. En effet, dans des expériences où il a été demandé à un certain nombre de personnes d'annoter des énoncés d'émotion actée, il a été relevé que le taux de reconnaissance obtenu varie entre 55% et 65% (5 à 6 fois supérieur à une prédiction par chance) (Scherer, 2003). Cette ambigüité est aussi relevée au cours des opérations d'annotation des corpus d'émotions spontanées. Afin d'atténuer ce problème, l'opération d'annotation est confiée à plusieurs annotateurs pour ne sélectionner que les énoncés ayant bénéficié d'un consensus ou le cas échéant, d'un vote majoritaire.
L'amélioration des performances des systèmes de RAE passera donc en premier par l'identification des sources de cette confusion afin de proposer une solution adéquate pour y remédier. L'ambigüité entre classes peut avoir plusieurs causes qui contribuent à rendre la discrimination entre classes un problème difficile. Nous identifierons dans ce qui suit, les sources qui nous semblent être les plus déterminantes et qui peuvent être d'ordre conceptuels ou opérationnels.

1.1.1 Incertitudes relatives à la définition de l'émotion en psychologie

La RAE fait appel à des notions et modèles théoriques qui relèvent du domaine de la psychologie et qui sont déterminants dans la motivation de certains choix de conception d'un système de RAE (i.e., modèles et/ou classes d'émotions, types de corpus de parole émotionnelle et méthodes de leurs collections et d'annotations). Cependant, l'étude des émotions en psychologie est caractérisée par un ensemble d'incertitudes qui sont reflétées par une absence de consensus entre théoriciens autour de certains éléments clés tels que : (i) la définition de l'émotion par rapport aux autres types d'états affectifs tels que l'humeur ou les traits de personnalité; (ii) quel modèle d'émotion est le plus approprié : discret ou dimensionnel; (iii) le nombre et le nom des différentes catégories d'émotion. Ainsi, pour le nombre de catégories d'émotion, certaines écoles de pensée recensent 15 classes alors que d'autres proposent six classes d'émotions universelles (*Big Six*), et même pour ceux qui proposent six classes d'émotion, il n'existe pas d'entente sur la nature des six classes. Nous pourrions contourner cette difficulté en partie dans le domaine des sciences du traitement de

l'information en proposant des solutions dépendantes du domaine applicatif. Selon le contexte applicatif, une certaine catégorisation des émotions est proposée qui mettra en évidence les émotions les plus pertinentes. Dans un contexte de détection d'appels problématiques des clients d'un centre d'appels par exemple, nous nous intéresserons à deux grandes classes d'émotions positive vs négative. La classe négative comprendra principalement la *colère* et la *frustration*. Dans le contexte d'un système d'apprentissage interactif, nous intéresserons particulièrement à la classe d'émotion *ennui* comme classe négative, alors qu'appliqué aux systèmes de sécurité pour la surveillance des lieux publics, c'est l'*angoisse* qui représentera la classe négative.

1.1.2 Nature dynamique de l'émotion

L'émotion est un épisode relativement bref qui est susceptible de changer rapidement (Scherer, 2000). Par conséquent un changement d'état émotionnel du locuteur peut se produire à l'intérieur d'un même énoncé. Ceci soulève la question de la segmentation des énoncés en unités de parole pertinentes pour l'analyse du contenu émotionnel de la voix (i.e., mot, tour de parole ou une unité intermédiaire,…). Cette unité devrait être suffisamment petite pour ne contenir qu'une seule émotion mais être suffisamment longue pour pouvoir en extraire des propriétés statistiques valides.

1.1.3 Bruit au niveau des corpus des émotions

Les performances d'un système de reconnaissance de formes basé sur une approche statistique guidée par les données, telle que nous allons utiliser dans cette thèse, dépendent de la taille et de la qualité des données d'apprentissage. La présence aiguë du bruit influencera négativement la précision des modèles de classification. Nous pouvons distinguer principalement deux sources de bruit dans les corpus d'émotion:

1. Erreurs induites par l'opération d'annotation : Une source importante de bruit qui peut altérer la qualité d'un corpus émotionnel, notamment spontané, provient des erreurs

induites par l'opération d'étiquetage des données. Ceci est particulièrement vrai quand il s'agit d'annoter un énoncé véhiculant des émotions subtiles ou mixtes. Des émotions mixtes peuvent être de types (i) mélangées : superposition de deux émotions simultanées qui sont conflictuelles (ex. : expression de la joie par des larmes) ou ambiguës (deux émotions dans une même catégorie générale); (ii) en succession rapide; (ii) ou masquées : la personne essaie de supprimer son émotion réelle ou de la masquer par une autre émotion (Devillers *et al.* 2005). On peut citer l'exemple de clients d'un centre de service de la bourse qui ressentiront de la colère car ils auront peur de perdre de l'argent (Vidrascu et Devillers, 2007). L'annotation de tels énoncés devient problématique quand il va falloir choisir une seule catégorie d'émotion.

2. Conditions d'enregistrement et de transmission du signal : Un autre facteur important qui influe sur la qualité des données, et par conséquent sur les modèles, est les conditions d'acquisition et de transmission du signal audio. En effet, un signal audio enregistré véhicule non seulement la parole et l'émotion, mais aussi un bruit additif occasionné par l'environnement sonore lors de la prise du son, des distorsions relatives au microphone et éventuellement au canal de transmission.

1.1.4 Chevauchement entre classes d'émotions dans l'espace des traits acoustiques

Les différentes classes d'émotions sont caractérisées par un chevauchement dans l'espace de réalisation (expression vocale) de ces émotions, c.-à-d. dans l'espace des traits acoustiques. Ce chevauchement se trouve particulièrement le long de la dimension valence (voir Figure 1.2, i.e., joie et colère) (Banse et Scherer, 1996; Lee et Narayanan, 2003; Ververidis et Kotropoulos, 2004; Yildrim *et al.* 2004). Ce chevauchement dans l'espace des traits acoustiques peut s'expliquer, en partie, par l'absence de frontières claires entre les différentes catégories d'émotion qui marque le passage d'une classe d'émotion vers une classe voisine. Ceci est particulièrement vrai pour les émotions appartenant à la même grande classe (famille) d'émotion telles que colère et irritation. Un autre facteur important qui accentue ce chevauchement est la diversité importante qui existe entre différents individus dans

l'expression d'une même émotion. Cette large variabilité peut avoir plusieurs origines telles que le milieu culturel de la personne, son âge, son genre (homme ou femme) et la langue parlée. Deux individus (ou le même individu à des moments différents) peuvent réagir avec différentes émotions à la même situation car les objectifs, les valeurs et les styles d'adaptation (ex. : l'externalisation ou l'internalisation) qui sont spécifiques à un individu peuvent conduire à des évaluations différentes (Scherer *et al.* 2010c). Le problème de chevauchement met également en évidence l'absence d'un ensemble de traits acoustiques optimal qui permettrait de discriminer convenablement les classes d'émotion.

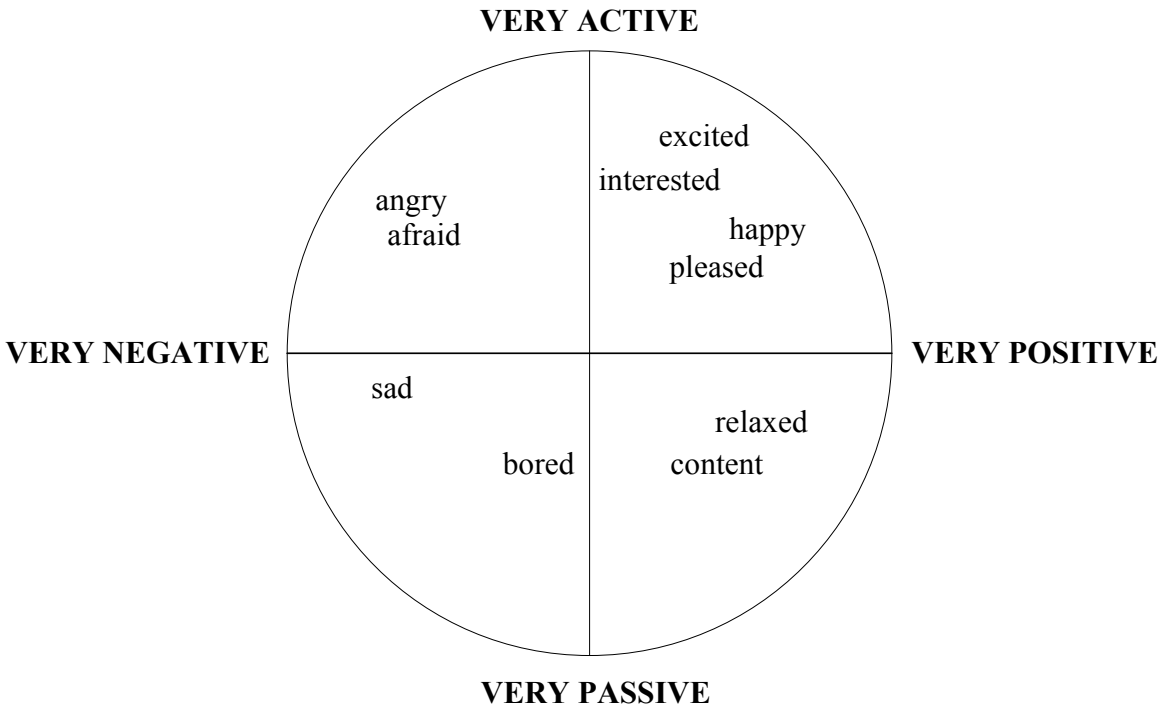

Figure 1.2 Émotions dans l'espace *activation-évaluation*

1.1.5 Propriétés statistiques des corpus de données

Dans une perspective d'apprentissage machine et de reconnaissance de formes, les corpus de parole émotionnelle présentent des propriétés statistiques assez contraignantes et

défavorables pour un apprentissage robuste de modèles et ce particulièrement quand il s'agit d'émotions réelles. Les données sont d'une taille assez limitée pour l'entrainement, et les énoncés de test sont de tailles courtes (empêchant l'extraction robuste de paramètres statistiques) alors que la distribution des classes est sévèrement biaisée.

1.2 Objectif

Notre objectif dans cette thèse est de trouver de nouveaux traits et approches de classification permettant d'améliorer les performances actuelles des systèmes de RAE à partir de la parole opérants dans des conditions réelles (*real-life settings*), ayant par conséquent un haut degré de complexité. Ainsi, la méthodologie proposée devrait répondre aux besoins de reconnaitre des émotions (i) naturelles et spontanées, donc plus subtiles, non complètes (*not full-blown*) et non primaires, (ii) véhiculant un vocabulaire non contrôlé, (iii) caractérisées par une large variabilité dans l'expression d'une même émotion à travers la participation d'un grand nombre de personnes, (iv) dans un contexte de classification multi-classes (donc caractérisé par une plus grande confusion en comparaison à un problème à deux classes), (v) où la distribution des classes est largement biaisée.

Afin de rendre les améliorations des performances des systèmes de RAE apportées dans cette thèse transférables à différents contextes d'utilisation, aucune connaissance a priori identifiable à une situation particulière ne sera exploitée. Ainsi la solution proposée sera indépendante du locuteur, de son genre et de la langue parlée (en excluant les traits linguistiques). Il sera donc par la suite facile d'adapter le modèle général proposé à un cas d'utilisation particulier pour en améliorer davantage les performances.

1.3 Applications

Nous avançons de plus en plus vers un monde où l'informatique affective deviendra une technologie omniprésente dans les systèmes à venir (robots, téléphones intelligents, machines informatiques,…). Un agent affectivement compétent offrira une interaction plus efficace, naturelle et plus sensible au comportement de l'utilisateur. Nous exposerons dans ce qui suit quelques contextes potentiels d'utilisation de cette technologie.

- Amélioration du service à la clientèle lorsque le système de RAE est intégré aux serveurs vocaux interactifs des centres d'appels commerciaux (Lee et Narayanan, 2003; Maganti *et al.* 2007; Petrushin, 2000; Yacoub *et al.* 2003). La RAE est nécessaire car les systèmes de la compréhension automatique de la parole ne sont pas exempts d'erreur et engendrent souvent des difficultés de communication et par conséquent peuvent provoquer la colère des appelants. La détection précoce d'un appel problématique à travers la détection d'un état émotionnel négatif d'un client permettra à la machine d'entreprendre plusieurs stratégies de gestion de l'échec de l'appel (i.e., restreindre et guider le dialogue; traiter l'appel en priorité par un opérateur humain en temps réel ou en différé s'il y a bris de communication).

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

- L'état émotionnel peut être utilisé comme indicateur d'aptitude à exercer correctement une tâche ayant des incidences sur la sécurité des personnes, telle que la conduite ou le pilotage afin d'activer les routines de sécurité. Ceci est motivé par le lien établi, dans plusieurs études, entre l'état émotionnel du conducteur et ses performances au volant (Jones et Jonsson, 2007). Ainsi, un support ou une rotation des travailleurs peut être prévus par exemple au cours d'une expérience de conduite pénible.

- Détection de la présence d'émotions extrêmes, principalement la peur, dans le cadre de la surveillance dans les lieux publics (Clavel *et al.* 2006) ou pour l'évaluation de l'urgence d'un appel pour prendre une décision dans le cadre d'un centre d'appel médical offrant un service de conseils médicaux aux patients (Devillers et Vidrascu, 2007); ou encore pour prioriser automatiquement les messages cumulés dans la boîte vocale selon

différents axes affectifs tels que l'urgence, valence (heureux vs triste) et l'excitation (calme vs excité) pour alerter le client et lui permettre d'écouter les messages importants en premier (Inanoglu et Caneel, 2005);

- Repérage ou indexation automatique d'émotion d'acteurs ou d'évènements émotionnels utiles pour la récapitulation (*summarization*) automatique des films (Malandrakis *et al.* 2011); ou encore l'évaluation de l'impact d'annonces publicitaires dans les médias sur les clients (McDuff *et al.* 2014);

- Utilisation des traits spéciaux véhiculés par les émotions pour le développement de systèmes de vérification automatique de locuteurs (Panat et Ingole, 2008) ou de systèmes de reconnaissance automatique de la parole (Womack et Hansen, 1999) plus robustes et plus précis.

- Des robots affectivement sensibles peuvent servir comme (i) assistants thérapeutiques pour personnes avec handicap cognitif ou social, les autistes en l'occurrence, afin d'améliorer leurs habilités de communication sociale et leurs apprendre à exprimer leurs émotions (El Kaliouby *et al.* 2006; Marchi *et al.* 2012); (ii) ou comme compagnons artificiels pour personnes âgées ou personnes dépendantes à la toxicomanie (Hill *et al.* 2013). Outre le domaine médical, les compétences affectives peuvent être intégrées dans des jouets interactifs pour des fins ludo-éducatives et de divertissement. Kismet (Breazeal, 2003) est l'exemple populaire de robot affectivement sensible.

L'importance de l'étendue du domaine applicatif de l'émotion computationnelle et de son impact sur les personnes est telle qu'on étudie d'ores et déjà les problèmes d'éthique que pourrait engendrer l'utilisation à mauvais escient de cette technologie dans le futur. On pourrait penser par exemple aux cas d'exploitation de l'information émotionnelle d'un utilisateur de la machine sans un consentement express de sa part, ou encore la mise en œuvre d'une stratégie de manipulation des émotions pour influencer les opinions (en

politique par exemple) ou du comportement (dans le domaine du marketing par exemple) (Scherer *et al.* 2011).

1.4 Organisation de cette thèse

Notre thèse est structurée comme suit. Dans le CHAPITRE 2 nous décrirons les émotions vues d'une perspective des domaines de la psychologie, de philosophie et de psychophysiologie. Nous mettrons en exergue la complexité et les incertitudes entourant la définition des émotions et de leurs modèles théoriques. Nous nous intéresserons également à la description des différentes méthodes de collecte des corpus, et aux consignes à suivre lors d'une opération d'annotation et de validation d'un corpus de parole émotionnelle.

Dans le CHAPITRE 3 nous allons présenter une revue de littérature détaillée sur les systèmes de RAE. Nous commencerons par présenter une taxonomie des systèmes basés sur des classificateurs simples selon la nature et la portée de l'information acoustique, type d'unité d'analyse émotionnelle et l'approche du classificateur. Nous analyserons par la suite les différentes stratégies de combinaison de classificateurs suivies pour améliorer la robustesse des systèmes de RAE.

Dans le CHAPITRE 4, nous allons présenter la méthodologie proposée dans cette thèse afin d'améliorer les performances des systèmes de RAE. Nous commencerons d'abord par présenter notre approche basée sur le concept de la similarité. Nous utiliserons la similarité comme alternative pour à la fois représenter et classifier les énoncés émotionnels. Nous déterminerons également dans ce chapitre le type de descripteurs de haut niveau qui serviront pour le calcul des traits basés sur la similarité.

Après avoir déterminé les traits basés sur la similarité, nous introduirons dans le CHAPITRE 5 une nouvelle méthode de classification des émotions basée sur la similarité que nous avons intitulée : le *plus proche patron de similarité pondéré* (WOC-NN, *Weighted Ordered Classes-Nearest Neighbors*). Cette méthode est basée sur l'extraction de descripteurs de haut

niveau appelés les *patrons de proximité pondérés* construits à partir des vecteurs de probabilité de vraisemblance.

Si dans le modèle WOC-NN, le calcul de la proximité a été basé sur une mesure qualitative (rangs des classes), nous proposons d'estimer dans les chapitres suivants la similarité entre énoncés et classes à travers des métriques quantitatives (euclidienne et cosinus). Ainsi, dans le CHAPITRE 6, les modèles d'ancrage sont proposés pour la classification d'émotions dans un contexte d'un problème multi-classe. Les propriétés des modèles d'ancrage déduites de leurs régions de décision sont étudiées et comparées avec les systèmes GMM basés sur le la règle de décision *Bayes* dans le CHAPITRE 7. Nous expérimenterons également dans le même chapitre les modèles d'ancrage dans un contexte d'une classification binaire.

De nouveaux types de descripteurs plus discriminants seront proposés dans le CHAPITRE 8 i) les coefficients cepstraux sur l'échelle *Mel* (*Mel-Frequency Cepstral Coefficients*, MFCC) et les coefficients de prédiction linéaire perceptuelle (*Perceptual Linear Prediction*, PLP) basés sur une estimation de spectre de type multifenêtre ii) les coefficients cepstraux de modulation d'amplitude. Nous montrerons également qu'il existe une complémentarité d'information entre les différents traits exploitables à travers une stratégie de combinaison de classificateurs. Enfin, nous montrerons que les modèles d'ancrage constituent aussi une méthode puissante de combinaison des classificateurs moyennant une normalisation préalable des scores plus adaptée au contexte de fusion.

CHAPITRE 2

THÉORIES DES ÉMOTIONS

2.1 Définition des émotions

> "*Because we -as human and as researchers-, all understand emotions, because we all produce emotional displays in our daily lives, we are tempted to believe that we are natural experts and easily can get the hang of such phenomena. Unfortunately, this is an illusion; emotional expressions are complex, fleeting, and difficult objects to study, and research in this field can be limited by preconceptions or by restrictive and sometimes implicit models*" (Bänziger et al. 2010b).

Cette citation reflète bien les incertitudes et difficultés, non perceptibles d'emblée, mais qui pourtant transparaissent à mesure qu'on franchit les étapes de conception et de développement d'un agent affectivement compétent. La première incertitude liée au domaine de la psychologie à laquelle on pourrait faire face est l'absence d'un consensus autour de la définition de l'émotion et de ses différents types. L'absence du consensus apparaît déjà dans la distinction de l'émotion des autres types d'états affectifs que sont l'humeur (*mood*), attitudes interpersonnelles (*interpersonal stances*), attitudes et traits de personnalité affectifs. Afin de distinguer entre ces cinq types d'états, Scherer (Scherer, 2000) a suggéré une approche caractéristique basée sur un ensemble de critères distinctifs tels que l'intensité, la durée, le degré de synchronisation des sous-systèmes organismiques, le degré de dépendance avec un événement déclencheur et le degré d'impact de l'état affectif sur le comportement. Selon ces critères, les différents types d'états affectifs peuvent être caractérisés comme suit :

Émotion : Relativement bref épisode, d'une réponse synchronisée de l'ensemble ou la plupart des sous-systèmes organismiques, en réponse à l'évaluation d'un événement externe ou interne, d'une importance majeure (ex. colère, tristesse, joie, peur, honte, fierté, désespoir).

Humeur : État affectif diffus, plus prononcé comme changement de sentiment subjectif, mais de faible intensité, d'une durée relativement longue, souvent sans cause apparente (ex. joyeux, triste, irritable, apathique, déprimé, vif).

Stances interpersonnelles : Attitude affective envers une autre personne lors d'une interaction, caractérisant l'échange interpersonnel (ex. distant, froid, chaud, favorable, méprisant).

Attitudes : Relativement durable, conviction teintée d'affection, préférence et prédisposition envers des objets ou des personnes (ex. plaisant, affectueux, détestable, appréciable, désirant).

Traits de personnalité : Émotionnellement chargé, caractères de personnalité et tendances de comportement stables, typique pour une personne (ex. nerveux, anxieux, imprudent, morose, hostile, envieux, jaloux).

L'émotion se distingue donc, par rapport aux autres états affectifs, par une plus grande intensité, mais de plus courte durée, qui possède un grand impact sur le comportement de l'individu et qui est susceptible de changer plus rapidement.

2.2 Expressions émotionnelles entre les effets *pousser* et *tirer*

Scherer (Scherer *et al.* 2010c) explique que l'expression émotionnelle est façonnée à travers une combinaison d'effets *pousser* et *tirer* (*push and pull*). Dans le cas de l'effet *pousser*, les facteurs internes naturels de l'organisme, tels que les modifications physiologiques liées à l'activation d'émotions fortes, poussent le comportement moteur dans certaines directions. L'expression qui en résulte est très variable et peut changer rapidement dans le temps. Les grognements et cris des nourrissons, les éclats d'affect et les énoncés émotionnels soudains et incontrôlés sont des exemples de cas où l'effet *pousser* est dominant. En revanche, les facteurs *tirer* externes sont au service de visées spécifiques de communication, des attentes

qui nécessitent l'affichage de traits expressifs relativement clairs et compriss ou de normes culturellement définies. Par exemple, une personne est tenue socialement d'avoir un air heureux même si elle est en réalité déçue du cadeau reçu (Bänziger *et al.* 2010b). Dans une étude menée par Cowie (Cowie *et al.* 2010b) sur l'induction par observation de films, il a été montré que le type de signes visibles manifestés sur le visage (effet *tirer*) dépend non seulement de la présence ou non d'autrui, mais aussi de son identité (proche ou étranger) et du type d'émotion exprimée. L'encodage d'effet tirer se caractérise par un haut degré de symbolisation et de stylisation, on peut s'attendre, par conséquent, à ce que les différences individuelles soient relativement petites et peu nombreuses (Scherer *et al.* 2010c).

2.3 Modèles psychologiques des émotions

Tout système doté d'une compétence affective doit être bâti autour d'un modèle théorique d'émotion. Cependant, il existe plusieurs théories vouées à la modélisation des concepts de l'émotion. Deux théories ont traditionnellement fortement influencé le passé de la recherche dans le domaine des modèles émotionnels, à savoir la théorie de l'émotion discrète et la théorie dimensionnelle. Cependant, un récent modèle, appelé modèle à composants, basé sur la théorie de l'évaluation cognitive ne cesse de gagner en intérêt.

2.3.1 Théorie de l'émotion discrète

La théorie de l'émotion discrète se concentre particulièrement sur l'étude de l'expression motrice ou du schème de conduite adaptative. Les théoriciens de cette mouvance proposent l'existence d'un petit nombre, compris entre 9 et 14, d'émotions de base ou fondamentales caractérisées par des modèles de réponse très spécifiques (Scherer, 2003). Ces modèles de réponse sont produits par un programme neuro-moteur inné caractérisé par ses composants neurophysiologiques, expressif (faciales et vocales) et subjectif. La lutte en cas de colère et la fuite (vol) en cas de peur (*fight* or *flight*) sont deux exemples de conduites bien connus qui sont souvent cités pour illustrer la différentiation des réponses autonomes activées pour chacune des deux émotions. En cas de colère, une augmentation de la perfusion sanguine des

mains (signe d'une mobilisation pour une querelle violente) est prévue alors qu'une augmentation de la perfusion au visage est considérée comme un message d'excitation de colère. Par contre en cas de peur, une perfusion accrue des membres à mobiliser pour une course rapide est prédite. Cette distribution du sang vers les extrémités pourrait causer un visage pâle. D'ailleurs une augmentation de la transpiration est prévue de se produire et qui peut servir comme moyen échappatoire efficace à l'emprise du prédateur (Kreibig *et al.* 2010). Les premières études qui se sont intéressées à l'effet vocal des émotions ont utilisé ce modèle et ont choisi d'examiner particulièrement l'effet de la joie, la tristesse, la peur, la colère et la surprise. Sur le plan de production de la parole, d'après Williams and Stevens (1981), les émotions colère, peur et joie vont exciter le système nerveux sympathique qui va provoquer une augmentation de la fréquence cardiaque et de la pression artérielle, la bouche deviendra sèche et des tremblements musculaires occasionnels sont à prévoir. La parole devient par conséquent forte et rapide avec une forte énergie à haute fréquence. D'autre part, l'excitation du système nerveux parasympathique, par une émotion tristesse par exemple, causera une diminution de la fréquence cardiaque et de la pression artérielle et une augmentation de la salivation. Par conséquent la production de la parole est lente avec peu d'énergie dans les hautes fréquences (Nwe *et al.* 2003).

2.3.2 Théorie dimensionnelle

La théorie dimensionnelle s'intéresse principalement à la description verbale des sentiments subjectifs (*subjective feeling*). Dans cette théorie le couplage étroit entre certaines émotions et certains comportements, tel que proposé par les modèles d'émotions discrètes, est critiqué. L'absence d'une relation une-à-une entre les émotions et les comportements est avancée comme argument. La variété des réactions que pourrait avoir un animal stimulé par un choc électrique (réactions à la peur) est donnée comme exemple (Kreibig *et al.* 2010). Dans cette tradition, les différents états émotionnels sont représentés par des points dans un espace continu de deux ou trois dimensions. Les deux dimensions principales consistent en la dimension *valence* (agréable-désagréable) et la dimension intensité de l'activité (actif-passif), ou *arousal* en anglais. La dimension *valence* reflèterait l'existence dans le cerveau de deux

systèmes motivationnels : le système *appétitif* (un système activé en présence de stimuli appétitifs provoquant une tendance d'approche des stimuli) et le système *aversif* (ou système défensif déclenché par des stimuli déplaisants provoquant un comportement d'évitement). L'excitation (*arousal*) est considérée comme une représentation de l'activation métabolique et neurale soit du système appétitive ou du système aversive (Kreibig *et al.* 2010). La troisième dimension, si elle est utilisée, représente souvent le contrôle ou la puissance intellectuelle. Dans leurs travaux relatifs à l'effet de l'émotion sur la voix, les partisans de ce modèle se limitent souvent à l'étude des différences qui existent entre l'état émotionnel positif versus négatif et actif versus passif.

2.3.3 Théorie de l'évaluation cognitive

La théorie de l'évaluation cognitive (*appraisal theory* en anglais) postule que la plupart des émotions sont provoquées par une évaluation cognitive des situations et des évènements antécédents. Ainsi, le modèle de réaction dans les différents domaines de réponse est commandé par ce processus d'évaluation. Par exemple, l'évaluation subjective de l'importance d'un évènement pour un organisme et sa capacité à en faire face sont supposées déterminer la nature de l'émotion correspondante (Scherer, 2010b). L'évaluation se fait selon un processus continu ce qui permet de se réadapter progressivement à la situation. Cette tradition est à la base de la majorité des modèles computationnels des émotions (Marsella et al. 2010).

2.3.4 Modèle d'émotion à composantes

Le modèle d'émotion à composantes (*componential model of emotion*), proposé par Scherer, est un des modèles ayant élaboré la conceptualisation de la théorie de l'évaluation cognitive (Scherer, 2010b). Le champ d'intérêt de ce modèle ne se limite pas à l'étude des sentiments subjectifs (tel est le cas pour la théorie dimensionnelle) ni au nombre supposé d'émotions de base (comme c'est le cas avec la théorie discrète). Ce modèle met l'accent sur la variabilité des différents états émotionnels, tels que produits par différents types de patrons d'évaluation

(*appraisal patterns*) (Scherer, 2003). L'évaluation est basée sur quatre critères qui représentent les variables qui permettraient de prédire l'émotion et son intensité, à savoir i) la pertinence de l'évènement (affecte-il la personne ou le groupe social) ii) implications (impacts de cet évènement sur son bien-être et sur l'atteinte de ses objectifs) iii) capacité à faire face à ces conséquences et iv) significativité de cet évènement par rapport à ses convictions personnelles et ses valeurs sociales. Ils offrent également la possibilité de modéliser les différences qui existent entre les membres de la même famille d'émotion, telle que la colère forte, la colère froide et le mépris. D'après (Scherer, 2003), ces approches fournissent une base solide pour une élaboration théorique des mécanismes qui sont censés sous-tendre la relation émotion-voix et permettent de générer des hypothèses très concrètes qui peuvent être testées empiriquement.

Dans le cadre du modèle des composantes, Klaus Scherer (2010b) définit une émotion comme un épisode de changements d'état intervenant dans tous ou la plupart des cinq sous-systèmes organiques de manière interdépendante et synchronisée en réponse à l'évaluation d'un stimulus externe, ou interne, par rapport à un intérêt central pour l'individu. Les cinq sous-systèmes organiques touchés par les changements sont les composants: cognitif (activité du système nerveux central), psychophysiologique (réponses périphériques), motivationnel (tendance à répondre à l'événement), moteur (mouvement, expression faciale, vocalisation), sentiment subjectif.

2.4 Corpus de parole émotionnelle

La constitution d'un corpus de données représente un élément fondamental dans le processus de construction d'un système de reconnaissance de formes en général et celui des émotions en particulier. La qualité et la quantité des données influent considérablement sur la capacité de généralisation et de prédiction du modèle appris. C'est pourquoi on accordera une attention particulière aux différentes étapes nécessaires à la constitution d'un corpus qui sont particulièrement délicates quand il s'agit d'un corpus de parole émotionnelle. En fait la constitution d'une base d'émotions résume bien, à elle seule, la plupart des défis et difficultés rencontrées lors de la conception et le développement d'un système de reconnaissance

d'émotions. Ces étapes comprennent, après le choix du modèle émotionnel, le choix du type du corpus, sa collection, son annotation et enfin sa validation.

2.4.1 Type de corpus des émotions

Nous distinguons essentiellement trois catégories de corpus émotionnel utilisées dans le domaine de la détection automatique des émotions : les émotions naturelles, simulées et induites.

2.4.1.1 Émotions naturelles

Les émotions naturelles sont des enregistrements d'états émotionnels vécus naturellement et spontanément. Ce corpus de données est caractérisé par une très haute validité écologique. L'inconvénient est que ces données sont très limitées en nombre de locuteurs, de courtes durées, souvent de piètre qualité, et en plus d'être très difficiles à collecter et à étiqueter en classes d'émotions (Scherer 2000, Scherer 2003).

2.4.1.2 Émotions simulées

Les émotions simulées sont des émotions produites par des acteurs professionnels ou semi-professionnels en se basant sur le nom de la classe d'émotion et/ou de scénarios typiques. Cette méthode représente le moyen préféré pour constituer les données dans ce domaine étant donné que les émotions naturelles et intenses surviennent de façon imprévisible et loin de l'observation du public (Bänziger *et al.* 2010a). Cependant, certains griefs sont adressés à cette méthode. Il est soupçonné, par exemple, que l'émotion simulée est stéréotypée et qu'elle soit caractérisée par une plus grande intensité que l'émotion naturelle. Les partisans des émotions simulées minimisent l'impact du caractère d'émotion stéréotypée en arguant que même les émotions naturelles fortes sont aussi fortement sujettes à un contrôle social ou à une autorégulation. Par ailleurs, afin de remédier au problème d'exagération en intensité dans l'expression des émotions simulées, Bänziger et Scherer ont proposé un nouveau

scénario de simulation d'émotions et qui a été appliqué pour la collection du corpus GEMEP (Bänziger *et al.* 2010a). Dans ce scénario, il a été consigné aux acteurs de simuler certaines émotions avec des intensités supérieures et inférieures à ce qui correspond à l'intensité habituelle pour une émotion donnée. L'hypothèse sous-jacente est que les interprétations (*portrayals*) produites avec moins d'intensité sont susceptibles de reproduire plus fidèlement les émotions vécues dans nos interactions quotidiennes. Les acteurs ont été également instruits de masquer partiellement quelques émotions afin de reproduire le phénomène d'autorégulation constaté dans certaines émotions réelles i.e. en simulant une tentative ratée de déception. Enfin, Scherer (Scherer *et al.* 2010c) souligne que le but de l'utilisation des interprétations d'acteurs n'est pas d'étudier les émotions qui se produisent spontanément ou de détecter les émotions ressenties par les acteurs mais plutôt l'identification et la représentation prototypique des émotions dans la communication sociale (code partagé d'expression des émotions).

2.4.1.3 Émotions induites

Les émotions induites ont été utilisées au départ dans le domaine de la psychologie afin de déterminer si la stimulation des états émotionnels du locuteur produit les changements acoustiques correspondants. Les émotions de cette catégorie sont induites expérimentalement dans des laboratoires en utilisant des techniques d'induction. Un nombre impressionnant de techniques d'induction ont été créées par les psychologues et parmi celles-ci on peut citer, les techniques de visualisation d'images ou de films émouvants, l'écoute de la musique, l'imagination et la remémoration, ou l'exposition des sujets à des tâches difficiles à accomplir sous la contrainte de délai.

Scherer (Scherer *et al.* 2010c), dresse les limitations et obstacles rencontrés avec cette méthode. Ainsi, certaines de ces techniques produisent des effets relativement fiables et l'intensité des états obtenus est généralement basse, avec des expressions peu observables de l'extérieure. Les contraintes d'ordre éthique, coût et pratique (un laboratoire ne permet pas d'action adaptative) empêchent souvent les chercheurs de confronter les participants avec des stimuli d'une importance assez élevée susceptibles de produire de véritables émotions. Par

ailleurs, les artefacts observés dans un contexte d'émotion naturelle ne sont pas à exclure également. Sous l'influence des règles d'apparence sociales ou personnelles, les participants peuvent inhiber ou modifier les expressions naturelles survenues qu'ils considèrent inappropriées ou encore simuler les émotions souhaitées par l'expérimentateur même si celles-ci ne sont en réalité pas ressenties (Scherer *et al.* 2010c).

Cowie et son équipe se sont intéressé à développer et à améliorer les techniques d'induction afin de fournir des modèles pour la génération des données selon les besoins théoriques et pratiques. Le paradigme nommé « données spaghetti » développé par cette équipe, a permis de produire des enregistrements d'émotions fortes et spontanées de diverses types. Toutes les émotions ont été induites en présence d'un observateur. L'analyse des variables sociales montre que le niveau d'émotion affiché dépend à la fois du genre de l'observateur et de l'intelligence émotionnelle du participant. Cowie et son équipe (2010b) se sont également intéressés à améliorer les méthodes d'induction dans un contexte d'interactions sociales en développant le paradigme SAL (*Sensitive Artificial Listener*). SAL est un agent artificiel doté de compétence affective et qui peut fonctionner en mode entièrement automatique contrairement au paradigme du magicien d'Oz (*Wizard-of-OZ*, WOZ). L'introduction du WOZ avait déjà permis d'induire des interactions émotionnelles très riches en comparaison avec une interaction avec un opérateur humain (les utilisateurs interagissent d'une manière plus relâchée face à une machine). Cependant, certains éléments conversationnels clés dans WOZ, contrairement à SAL, étaient limités car l'opérateur humain qui contrôle l'agent automatique à distance à l'insu de l'utilisateur passe énormément de temps à regarder le script et non l'utilisateur. Des paradigmes tel que SAL sont prometteurs dans le sens où ils permettent d'ouvrir de nouvelles perspectives de contrôle qui sont très difficiles à atteindre dans un cadre d'interaction humain-humain. Notons par ailleurs que les travaux préliminaires réalisés par Cowie et son équipe sur les enregistrements issues de l'interaction avec des agents soulèvent de nouveaux enjeux qui leurs sont propres tels que les signes de confusion ou désengagement, susceptibles d'être une caractéristique importante de l'interaction humain-machine.

2.4.2 Constitution d'un corpus de parole émotionnelle

La qualité du corpus des données revêt une importance primordiale dans la qualité du système de RAE ou dans toute analyse exploratoire. Par conséquent, la procédure de collection devrait être préparée minutieusement et les conditions permettant de minimiser le bruit et d'éviter que le contenu émotionnel du corpus soit biaisé doivent être réunies (Bänziger *et al.* 2010a). Notons que certains des critères que devraient remplir un corpus, du point de vue d'ingénierie, sont parfois presque inconciliables. Collecter un large corpus avec un enregistrement de qualité et une distribution de classes équilibrée sont des critères très difficilement conciliables avec l'exigence que ces émotions soient survenues naturellement et issues de la vraie vie et non contrôlées et interprétées dans un laboratoire (Cowie *et al.* 2010a).

2.4.2.1 Collection des enregistrements

Les critères que devrait remplir un corpus et la méthodologie suivie dans sa collecte dépendront des différentes fonctions qu'il devrait servir. Ainsi, pour un corpus destiné à la reconnaissance des émotions, il est primordial que le contexte (la source) utilisé pour collecter les enregistrements soit identique à celui utilisé pour le déploiement du système. Pour la synthèse de la voix, avoir la représentation de toutes les voyelles dans tous les contextes phonétiques pour chaque style émotionnel constituera une priorité (Cowie *et al.* 2010a). Par ailleurs, la construction d'un corpus d'émotions actées nécessite une planification minutieuse pour établir la liste des émotions à inclure et les instructions d'interprétation assignées aux acteurs afin de répondre aux objectifs des différents domaines de recherche auxquels le corpus est destiné (Bänziger *et al.* 2010a). Afin de s'assurer que les modèles d'expressions produites sont des répliques plausibles et crédibles des expressions spontanées réelles typiques de certaines émotions, il est recommandé de faire appel aux acteurs professionnels (Scherer *et al.*). Ces acteurs devraient être encadrés par un metteur en scène professionnel qui utilise une méthode basée sur des techniques dramatiques impliquant prise de rôle, souvenirs personnels et empathie afin d'augmenter la crédibilité des interprétations

(Scherer *et al.* 2010c). Le choix des énoncés prononcés par les acteurs peut parfois se faire parmi des textes ayant une sémantique neutre, tels que les dates et les nombres (Banse et Scherer, 1996), des éclats d'affect (Bänziger *et al.* 2010a) ou encore des séquences de phonèmes pseudo-linguistiques (combinaisons plausibles de phonèmes mais dénuées de sens) choisies par un phonéticien. À chaque émotion simulée, les acteurs sont libres de la signification sémantique à attribuer à ces phonèmes pseudo-linguistiques (Bänziger *et al.* 2010a).

2.4.2.2 Annotation du corpus

Dans une base de données d'émotion en sus des enregistrements, on retrouve également le système d'annotation. Le système d'annotation s'intéresse non seulement à la description du contenu émotionnel affectif mais s'étend également à la description du contexte (Cowie *et al.* 2010a).

Annotation du contenu émotionnel :

La description du contenu émotionnel consiste à associer à une unité d'analyse tel qu'un enregistrement, une étiquette d'état émotionnel qui peut être de type catégorique ou dimensionnel (active, passive) (Cowie *et al.* 2010a). Cette association est réalisée par un ensemble d'annotateurs. Généralement, les annotateurs sont invités à juger l'émotion évoquée par les signes externes affichés par une personne indépendamment de la véracité de cette perception. C'est avec ce type de directive qu'un corpus d'émotion simulée, par exemple, est annoté (Scherer *et al.* 2010c). Dans certaines situations, l'annotateur est tenu d'aller au-delà de l'émotion perçue et identifier l'émotion vécue (réelle) par la personne. Ce type d'annotation n'est requis que dans certaines applications où il est primordial que le système identifie les vraies émotions comme cela pourrait être le cas dans un contexte de diagnostic. Dans ce cas, l'annotation devrait être confiée à des observateurs experts, étant donné que l'émotion perçue et réelle peuvent diverger (Cowie *et al.* 2010a).

Description du contexte :

La deuxième étape d'une opération d'annotation consiste à fournir une description détaillée du corpus collecté ainsi que les méthodes utilisées lors de sa collecte. Ces informations serviront comme référence pour les futures analyses ou expériences menées sur ce corpus (Bänziger *et al.* 2010a). La description du contexte apporte des informations sur les personnes impliquées, telles que leurs genres, personnalités et origines culturelles. Elle comprend également de l'information sur l'environnement qu'il soit physique relié à la tâche (stade atteint par la tâche, degré de son succès), sociale (présence d'autres personnes, normes en vigueurs, etc.), et communicative (mode de communication, phase du discours) (Cowie *et al.* 2010a). Elle inclut également les scripts des énoncés, la technologie des appareils utilisés pour l'enregistrement, la description des annotateurs (nombre, genre, âge) et une description des conditions et outils d'annotation (Bänziger *et al.* 2010a).

2.4.2.3 Validation du corpus

Un corpus des données d'émotion nécessite une validation de la qualité de son contenu émotionnel. Ce processus peut être assuré à divers stades i) au cours de la production, durant laquelle l'énoncé est répété jusqu'à ce qu'il ait une double validation, c'est-à-dire la satisfaction de l'acteur lui-même et celle de l'expert supervisant la simulation s'il s'agit d'un corpus acté ii) post-production à travers l'opération d'annotation des énoncés confiée à une multitude d'annotateurs (Bänziger *et al.* 2010a). En plus du critère exactitude (*accuracy*), deux critères psychométriques standards sont appliqués afin d'évaluer l'opération d'annotation; la validité (*validity*) et la fiabilité (*reliability*) qui peuvent être appliqués selon le type du corpus.

La **validité** consiste à s'assurer que l'enregistrement a été correctement annoté, c'est-à-dire que l'émotion perçue correspond à l'émotion vécue. Ce critère n'est requis que dans certains types d'applications (Cowie *et al.* 2010a).

La **fiabilité** est mesurée à travers le degré d'agrément entre annotateurs dans l'étiquetage d'un même énoncé. Plus le nombre d'agréments entre annotateurs est élevé, plus la fiabilité de l'émotion véhiculée dans l'énoncé est importante. Quand l'objectif du système est de relever l'impression des observateurs évoquée par les signes externes affichés par une personne indépendamment de la véracité de cette perception, il y a relâchement de la contrainte de validité pour le critère de fiabilité (Cowie *et al.* 2010a). Notons qu'un tel accord n'est pas un objectif facile à atteindre. En effet, l'interprétation des mêmes signes est une opération subjective qui peut être influencée par plusieurs facteurs y compris l'humeur, la personnalité, l'intelligence émotionnelle, le genre et la culture de l'annotateur (Cowie *et al.* 2010a).

L'exactitude mesure pour chaque annotateur le pourcentage des annotations correctes. Plus le taux d'exactitude est élevé plus il y a concordance entre le jugement de l'émotion perçue par l'annotateur et l'émotion exprimée (l'intention expressive de l'acteur). Les deux critères fiabilité et exactitude ont été appliquées avec succès pour valider le corpus GEMEP (Bänziger *et al.* 2010a).

Comme souligné dans (Cowie *et al.* 2010a), l'évaluation de la validité est interdépendante des enregistrements et du contexte. Par conséquent, il est difficile de s'assurer que les états émotionnels reconnus soit corrects en dehors d'un contexte très spécifique. À titre d'exemple, des émotions similaires seront probablement exprimées de manières différentes dans des contextes différents (dans un stade, en état de conduite d'un véhicule ou dans une salle d'audience). Un agent conçu pour un contexte particulier ne peut donc être transférable directement vers un autre contexte. Chaque contexte possède sa propre liste de descripteurs d'états émotionnels.

2.4.3 Corpus de données d'émotion dans la revue de littérature

Il existe un ensemble de corpus de parole émotionnelle exploités aussi bien dans le domaine de la psychologie que celui l'informatique affective. Un recensement des corpus les plus

utilisés peut être trouvé dans (Douglas-Cowie *et al.* 2007; El Ayadi *et al.* 2007; Ververidis et Kotropoulos, 2006) ou encore sur le site web de l'association AAAC (Association for the Advancement of Affective Computing, ex-HUMAINE) où la liste des corpus est continuellement mise à jour. Beaucoup de ces corpus sont privés et ne sont pas disponibles à la communauté. En plus du critère de disponibilité, ces corpus peuvent être caractérisés par le type d'émotion véhiculée qui peut être actée, naturelle ou induite, la taille du corpus, la langue dans laquelle il a été collecté et le contenu émotionnel des énoncés (nombre et types de classes d'émotion). Les corpus peuvent se distinguer également par la méthode d'annotation qui peut être de type catégorique (discrète) ou dimensionnelle (continue). Enfin, certains corpus peuvent contenir plusieurs modalités telles la vidéo et le gestuel à côté de la parole. Un des rares corpus ayant d'intéressantes propriétés et est considéré comme la référence, est le corpus FAU AIBO Emotion. C'est un corpus public qui a été mis à la disposition de la communauté scientifique en 2009 lors de la compétition internationale *INTERSPEECH 2009 Emotion Challenge*. Une liste plus exhaustive des corpus ainsi que leurs propriétés sont données à l'ANNEXE I.

2.5 Conclusion

Dans ce chapitre nous avons décrit les émotions vues de la perspective du domaine de la psychologie et de la psychophysiologie. Nous avons soulevé les incertitudes entourant la définition de l'émotion et sa différentiation des autres états affectifs. Nous avons présenté les modèles théoriques d'émotion les plus influents dans le domaine de la RAE dont le choix est déterminant pour les étapes ultérieures. Nous avons décrits les trois principales méthodes de collecte de corpus de parole émotionnelle et les avantages et inconvénients de chacune. Nous avons insisté sur l'importance de considérer le contexte du déploiement du système lors du choix de la stratégie de collecte du corpus, car les types d'émotion et les expressions affichées dépendront du contexte utilisé, qui est généralement non transférable vers un autre type de contexte. Nous avons également rapporté les consignes qui devraient être appliquées lors d'une opération d'annotation d'un corpus d'émotion et comment le valider. L'annotation des émotions est une opération particulièrement difficile quand il s'agit d'étiqueter des

émotions réelles où un énoncé peut véhiculer des émotions mixtes ou ambigües. Dans le chapitre suivant nous allons passer en revue les travaux réalisés dans le développement de systèmes de RAE cette fois-ci de la perspective du domaine d'ingénierie.

CHAPITRE 3

REVUE DE LITTÉRATURE SUR LES SYSTÈMES DE RECONNAISSANCE AUTOMATIQUE DES ÉMOTIONS

3.1 Introduction

Les domaines de la psychologie, la neurobiologie et sciences cognitives ont permis d'avancer notre compréhension de la notion floue des émotions ainsi que leurs modélisations. Des consignes ont été fournis sur les méthodes de constitution d'un corpus des émotions ainsi que sur la procédure de son annotation tel que déjà présenté dans le chapitre précédent. Les étapes suivantes dans la modélisation d'un système de RAE à partir de la parole feront ultérieurement appel au concours de plusieurs autres expertises appartenant à des disciplines aussi variées que celles de la linguistique, la phonétique physiologique, la phonétique acoustique, la phonologie et le domaine de la reconnaissance des formes. La linguistique permet de nous renseigner, par exemple, sur l'existence d'une information émotionnellement saillante utile pour la RAE, au niveau lexical ou au niveau pragmatique (actes de dialogue), ou de nous orienter vers des unités linguistiques, telles que les phonèmes ou mots, en plus de l'énoncé, qui peuvent constituer des unités d'analyse plus pertinentes dans la modélisation des classificateurs. La phonétique physiologique nous permet de mieux comprendre comment le son est produit par le système articulatoire et perçu par le système auditif, afin de choisir le codage le plus approprié lors de l'extraction des traits caractéristiques, qui seront calculés à travers des techniques relevant du domaine de la phonétique acoustique. Dans le domaine de la phonologie, nous nous intéressons à la description des éléments de la prosodie que sont la durée, l'intensité et la fréquence fondamentale des phonèmes. L'étude de l'évolution de ces trois éléments, qui sont perçus respectivement comme étant le rythme, la puissance et la mélodie (*pitch*) de la phrase, permet de déterminer l'effet que pourrait avoir un état émotionnel sur les caractéristiques du son produit. Enfin, la RAE est un problème qui s'apparente à un problème de reconnaissance des formes, d'où la nécessité d'une part, de connaître les techniques de sélection des caractéristiques qui permettent de choisir le sous-ensemble optimal de traits parmi des dizaines ou des centaines de traits candidats et d'autre

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

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

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

Nous commencerons d'abord par présenter dans section 2 les travaux sur les classificateurs de base (unique ou à un seul étage), suivis par les classificateurs multi-étages dans la section 3. Nous aborderons par la suite les stratégies de combinaison dans la section 4 avant de terminer par la présentation de techniques utiles pour l'amélioration des performances de classification des systèmes de RAE.

3.2 Travaux basés sur des classificateurs simples

Nous nous intéresserons dans cette section à présenter les différents traits extraits sur différentes échelles d'unités d'analyse qui ont été modélisés par des classificateurs simples. Par classificateurs simples nous entendons des systèmes basés sur un seul étage ou non composés par des sous classificateurs.

3.2.1 Travaux selon le type d'unité d'analyse

Un critère que nous pouvons examiner pour la classification des systèmes de RAE est le type d'unité d'analyse utilisée dans la reconnaissance des émotions. L'unité d'analyse représente le segment de données de base extrait d'un énoncé et soumis au classificateur pour déterminer sa catégorie d'émotion. Dans les travaux réalisés à ce jour, sept types d'unités ont été expérimentés :

Énoncé : l'unité d'analyse sur laquelle est basée la plupart des travaux. Les vecteurs de traits de la totalité de l'énoncé sont extraits et soumis en une seule entrée au classificateur pour déterminer la catégorie de l'émotion. Parmi les travaux qui se sont basés sur cette unité, citons (Beritelli *et al.* 2006; El Ayadi *et al.* 2007; Grimm et Kroschel, 2005; Inanoglu et Caneel, 2005; Li *et al.* 2007; Lin et Wei, 2005; Pao *et al.* 2005; Petrushin, 2000; Seppänen *et al.* 2003; Sethu *et al.* 2007; Vlasenko *et al.* 2007).

Mot : l'unité d'analyse mot a été testée et comparée avec les performances d'un système basé sur l'unité *énoncé* (Rotaru et Litman, 2005; Schuller *et al.* 2007a). D'après les résultats obtenus dans (Schuller *et al.* 2007a), l'unité *mot* est préférable à l'énoncé à condition qu'un système efficace de segmentation par mot soit disponible. L'étude réalisée dans (Rotaru et Litman, 2005) montre également une amélioration dans la prédiction en utilisant l'unité mot et ce particulièrement en présence de longs tours de parole. Rao et Koolagudi (2012) ont montré que les mots en position finale sont plus discriminants que les mots en début ou en milieu de phrase, aboutissant ainsi à une conclusion similaire à celle observée pour les syllabes.

Phonème : le phonème représente la plus petite unité de son d'une langue. Le choix de cette unité est motivé par l'hypothèse que l'état émotionnel d'un locuteur affecte les phonèmes d'un énoncé avec différentes intensités. L'énoncé est alors segmenté en phonèmes et chaque classe de phonèmes est modélisée séparément. Afin de vérifier cette hypothèse, Lee et ses collègues (Lee *et al.* 2004) ont réalisé deux expériences; dans la première, un classificateur

HMM (*Hidden Markov Model* ou modèles de Markov cachés) émotionnel générique est utilisé. Ce classificateur est entraîné en utilisant les données d'apprentissage de toutes les classes de phonèmes. Dans la deuxième expérience, des HMM par classe de phonèmes sont expérimentés. Le taux de reconnaissance obtenu avec le modèle HMM générique est de 64,77 % alors que les résultats pour les modèles HMM par classe de phonèmes sont respectivement de 72,16 %, 54,86 %, 47,43 %, 44,89 % et 55,11 % pour les classes voyelles, semi-voyelles, nasales, consonnes occlusives et fricatives. Les résultats d'une classification basée sur la combinaison des modèles des cinq classes de phonèmes atteignent 75,57 %. Ces résultats montrent d'une part que les classes des phonèmes ne véhiculent pas, dans les mêmes proportions, la même charge émotionnelle. Les voyelles sont émotionnellement plus saillantes que les autres classes. D'autre part, une classification basée sur une modélisation par classe de phonèmes offre de meilleures performances que la classification à partir d'un modèle générique. Dans (Bitouk *et al.* 2010), les traits à long terme Les types prosodique et spectral (MFCC) ont été modélisés séparément au niveau énoncé et phonème, donnant quatre classificateurs au total. Les phonèmes ont été regroupés en trois catégories : consonnes, voyelles accentuées (*stressed vowels*) et voyelles non accentuées. Les traits des trois classes de phonèmes ont été combinés pour constituer un seul vecteur de traits à l'échelle de l'énoncé. Chacun des systèmes est testé sur deux corpus de données avec les machines à vecteurs de support (*Support Vector Machine*, SVM) comme classificateur. Les résultats obtenus montrent que les performances du système basé sur l'information spectrale extraite à l'échelle des classes de phonèmes dépassent significativement les performances des trois autres systèmes (systèmes basés sur l'information : spectrale à l'échelle de l'énoncé, prosodique à échelle de phonèmes ou de l'énoncé) et ceci pour les deux corpus de données utilisés. La comparaison des performances à l'intérieur des trois groupes de phonèmes montrent que le classificateur basé sur l'information spectrale contenu dans la classe consonne performent significativement mieux comparée aux deux autres classes pour un des deux corpus, alors que la classes des voyelles accentuées performent légèrement mieux pour le second corpus. Dans (Koolagudi et Krothapalli, 2012), les segments de parole sont divisés en trois catégories : consonne, voyelle et région de transition comprise entre la consonne et la voyelle constituant un mot. Les performances du classificateur entrainé avec les données

extraites de la région de transition dépassent largement celles des classificateurs basés sur les segments de types voyelles ou consonnes et obtiennent des résultats comparables au classificateur entrainé avec toutes les données de parole.

Syllabe : pour la même raison ayant conduit à l'expérimentation de l'unité phonème, Schuller et ses collègues ont procédé à une segmentation de l'énoncé basée sur l'unité syllabe (Schuller *et al.* 2007b). Les performances du système basé sur l'unité syllabique sont inférieures à celles de l'énoncé. Dans (Rao et Koolagudi, 2012), les syllabes ont été divisées en trois groupes syllabes de début, du milieu et de fin selon leurs positions dans un mot. La comparaison des performances des systèmes basés sur chacune de ces groupes, montre que les syllabes en position finale contiennent plus d'information discriminante comparée aux autres positions.

Pseudo-syllabe : les pseudo-syllabes représentent le résultat d'une segmentation du contour de la fréquence fondamentale (F0) guidée par les points minima locaux du contour de l'énergie. Une pseudo-syllabe peut correspondre à plusieurs syllabes ou à une partie d'une syllabe. Les informations à long terme de la prosodie et des formants calculées sous forme de coefficients de Legendre ont été extraites pour chaque pseudo-syllabe et modélisées par un GMM dans (Attabi, 2009; et Dumouchel *et al.* 2009).

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

Région voisée / non voisée : dans (Shami et Kamel, 2005), l'énoncé est divisé en une séquence de *N* segments voisés guidée par le contour de la fréquence fondamentale. Un vecteur de trait est extrait pour chaque région voisée. La classification au niveau de l'énoncé est réalisée en calculant la somme des probabilités a postériori calculées pour chaque région

voisée. Les vecteurs de traits, composés des valeurs statistiques de F0, l'énergie, durée et les MFCC sont modelés en utilisant les classificateurs SVM et KNN. Les performances obtenues au niveau énoncé étaient meilleures que celles obtenues au niveau des segments voisés, alors que la combinaison des traits des deux niveaux permettait d'améliorer encore plus les performances.

3.2.2 Travaux selon le type des traits caractéristiques

Nous distinguons principalement trois types d'information paralinguistique utilisée dans le domaine de la RAE : prosodique, spectrale et qualité de la voix.

3.2.2.1 Prosodie

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

La prosodie est un canal parallèle au contenu sémantique du message parlé dans les conversations quotidiennes à travers lequel l'auditeur peut percevoir les intentions et l'état émotionnel de l'orateur, ou encore distinguer une déclaration d'une question ou d'une commande[1] (Huang *et al.* 2001). Les prononciations d'un même mot peuvent avoir des prosodies substantiellement différentes sans affecter l'identité du mot. La prosodie s'intéresse à la relation qui lie la durée, l'amplitude et le pitch au son. Les traits prosodiques sont dits suprasegmentaux, dans le sens où leurs domaines d'interprétation sont au-delà de la limite de l'unité du phone (O'Shaughnessy, 2000).

Pitch et fréquence fondamentale

Le pitch est le phénomène prosodique le plus expressif. Il exprime la hauteur perçue par un humain. Les systèmes de traitement de la parole utilisent la fréquence fondamentale, appelée

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

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

Intensité et amplitude

L'intensité est une sensation auditive basée sur la perception de la force du signal acoustique. L'amplitude du mouvement vibratoire est la contrepartie acoustique de l'intensité. L'amplitude est fonction de la pression sonore; plus celle-ci est grande, plus l'amplitude est grande. La pression sonore représente les variations de pression de part et d'autre d'une pression atmosphérique moyenne (Galarneau *et al.* 2009). L'amplitude est mesurable en watt/cm^2 et est proportionnelle au carré de la pression sonore. L'aire d'audition humaine se mesure sur une échelle logarithmique relative de l'intensité dont l'unité est le décibel (dB). Le décibel est égal à 10 log ((niveau d'intensité en watts/cm^2) / (10^{-16} watts/cm^2)). Le seuil de l'audition varie de 0 à 40 dB selon la fréquence alors que le seuil de la douleur se situe environ à 120 dB (Galarneau *et al.* 2009). L'intensité mesure la quantité de l'énergie dans un signal de la parole et par conséquent reflète l'effort requis pour la production du son (Johnstone *et al.* 2000).

Rythme et débit

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

rapide alors qu'un changement de débit est soit une accélération ou un ralentissement (Galarneau *et al.* 2009).

Les traits prosodiques ont été les premiers à être utilisés dans le domaine de la recherche sur la reconnaissance des émotions (Dellaert *et al.* 1996; McGilloway *et al.* 2000). Plusieurs études ont montré que les traits prosodiques sont de bons descripteurs pour discriminer entre les catégories d'émotion ayant différents niveaux d'*excitation* (*active* versus *passive*). Par ailleurs, ces traits performent moins bien quand il s'agit de différencier entre les émotions ayant un même niveau d'excitation (Cowie *et al.* 2001; Johnstone et Scherer, 2000; Kao *et al.* 2006; Lugger et Yang, 2007; Marchi *et al.* 2012; Nwe *et al.* 2003; Pao *et al.* 2005; Rao et Koolagudi, 2012). Ceci est expliqué par le fait que ces traits semblent être associés aux caractéristiques générales de l'émotion (impliquant la dimension *activation*) plutôt qu'à une catégorie individuelle (Cowie *et al.* 2001). À titre d'exemple, une augmentation de la moyenne et de l'étendue de F0, et une qualité de voix tendue sont associées à une *activation* positive (la joie, la peur, la colère, et dans une moindre mesure la surprise, l'excitation et la perplexité). Par ailleurs, une baisse de la moyenne et une étendue de F0 plus étroite sont associées à une *activation* négative (tristesse, chagrin, et à un degré moindre l'ennui) (Cowie *et al.* 2001).

3.2.2.2 Traits spectraux

Nous avons vu que les traits prosodiques étaient fortement corrélés avec l'axe *activation* mais ne permettaient pas de bien modéliser l'autre dimension représentée par l'axe *valence*. Par conséquent ces traits, à elles seules, ne permettaient pas de discriminer convenablement entre, par exemple, la colère, la joie et l'anxiété ou entre la tristesse, le dégout et le neutre. Ceci a motivé l'utilisation d'autres traits, les traits spectraux en occurrence, qui ont montré un plus grand pouvoir discriminatif entre ces catégories d'émotion. Les expériences menées dans (Marchi *et al.* 2012) représentent un exemple d'étude où de meilleurs résultats de classification d'émotion selon l'axe *valence* sont obtenus avec les traits spectraux et qualité de la voix comparés aux traits prosodiques. Johnstone et Scherer (2000) expliquent la

pertinence des formants, traits spectraux véhiculant principalement de l'information phonétique, par le fait que ces traits demeurent affectés par des changements au conduit vocal qui accompagnent les différents états émotionnels tels que les changements de tension musculaire articulateur et de la salivation. Dans la catégorie des traits spectraux, on retrouve également, à côté des formants, des descripteurs issus des techniques de compression de la parole telles que les MFCC, PLP, les coefficients du codage par prédiction linéaire (Linear Predictive Coding coefficients, LPC), les coefficients cepstraux de prédiction linéaire (*Linear Prediction Cepstral Coefficients*, LPCC) et les coefficients de puissance logarithmique de fréquence (*Log Frequency Power Coefficients*, LFPC).

3.2.2.3 Traits de la qualité de la voix

Les paramètres de qualité de la voix décrivent les propriétés de l'excitation glottale. Outre l'articulation et la prosodie, la phonation est un processus important dans la génération de parole émotionnellement colorée (Lugger *et al.* 2009). Un large éventail de variables phonétiques contribue à l'impression subjective de la qualité de la voix (Laver, 1980). Les descripteurs de la qualité de la voix sont avérés particulièrement utiles pour la distinction entre la joie, la colère et l'anxiété (Lugger et Yang, 2007). La raison de ces bons résultats est que ces trois émotions diffèrent considérablement dans le type de phonation. Pour la production d'un état émotionnel triste, une phonation grinçante (*creaky phonation*) est souvent utilisée. Une voix rauque (*rough voice*) est généralement utilisée pour soutenir un état émotionnel en colère. L'émotion anxiété montre parfois des parties de voix voilée (*breathy voice*) (Lugger et Yang, 2007).

Dans (Cowie *et al.* 2001), deux approches sont citées pour caractériser les traits de la qualité de lavoix. L'approche la plus simple est basée sur les propriétés spectrales (Hammarberg *et al.* 1980) et la seconde sur l'utilisation du filtrage inverse visant à récupérer la forme d'onde de la glotte. Parmi les traits de type qualité de la voix utilisés dans la reconnaissance automatique des émotions, nous retrouvons : taux de non-voisement (proportion de trames non voisées dans un segment), rapport harmoniques-bruit (*harmonics to noise ratio*, HNR)

qui permet de caractériser la contribution du bruit à la parole pendant l'effort vocal, le ratio du temps d'ouverture à la fermeture des cordes vocales et la distribution spectrale d'énergie, et le taux de passage par zéro (*zero-crossing-rate*), vacillement (*jitter*), tremblotement (*shimmer*) (Cowie et al. 2001, Devillers et al. 2010).

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

Les impulsions naturelles glottales ne sont pas réellement périodiques, mais présentent des perturbations appelées tremblotement et vacillement. Le vacillement représente les variations trame par trame dans les périodes de F0. Le tremblotement représente les variations cycle par cycle dans les périodes de l'énergie. Les voix normales ont un vacillement de 0,5 à 1,0 % (p. ex. 1 Hz) et un tremblotement de l'ordre de 0,04 à 0,21 %, ce qui représente un niveau assez bas pour qu'il soit directement perceptible. Bien que le tremblotement et le vacillement soient deux concepts différents, ils sont légèrement corrélés (Huang *et al*. 2001).

Zhang (2008), a combiné les traits prosodiques avec les traits de la qualité de la voix, en plus des formants qu'il a considérés comme étant de types de la qualité de la voix. Les performances obtenues avec un SVM comme classificateur a permis un gain de 10 % par rapport à un système basé uniquement sur les traits prosodiques. Dans (Lugger et Yang, 2007), des traits de type qualité de la voix ont été ajoutés aux traits prosodiques, aux formants et leurs largeurs de bande et la méthode SFFS (*sequential floating forward selection*) a été appliquée pour sélectionner les huit meilleurs traits. Deux des huit traits sélectionnés étaient de type qualité de la voix. Les taux de reconnaissance ont augmenté de 66,7 % à 72,8 % quand le système a été testé sur le corpus *Berlin Emotional Database* en utilisant un classificateur Bayésien.

Le flux glottique (Iliev, 2010) et le signal résidu (Chauhan *et al*. 2010) sont deux types de traits dérivés de la source d'excitation du signal de la parole qui ont été récemment utilisés pour améliorer les performances des systèmes de reconnaissance automatique des émotions. Ces deux types de traits sont extraits du signal de la parole après suppression des

caractéristiques du conduit vocal. Les caractéristiques du conduit vocal sont obtenues en utilisant l'analyse de prédiction linéaire (*Linear Prediction*, LP). L'information du conduit vocal est par la suite supprimée en utilisant le filtrage inverse. Le signal résultant est nommé le résidu LP et contient principalement des informations sur la source d'excitation. Le flux glottique représente le débit d'air expulsé de la trachée et passant à travers les cordes vocales. Les traits de type flux glottique (*Glottal Volume Velocity*, GVV) s'intéresse à l'étude de la forme de l'impulsion décrite à travers les durées de la phase d'ouverture et la phase de fermeture de la glotte, leurs pentes respectives ainsi que le rapport des deux pentes. L'utilisation de l'information sur le signal de source d'excitation pour la RAE a été motivée par des études antérieures qui ont mis en évidence l'existence d'informations spécifiques sur l'émotion dans l'onde glottique. Les résultats obtenus dans ces deux études (Chauhan et al, 2010; Iliev, 2010), confirment que l'information glottique présente une source très efficace pour la reconnaissance automatique de l'émotion et confirment ainsi les résultats d'études antérieures mettant en évidence la présence d'indices sur l'état émotionnel et le style de parole dans l'onde glottique (Cumming, 1995, Laukkanen 1996). Dans (Koolagudi *et al.* 2012a), l'information source (le signal résidu et le flux glottique) a été utilisée individuellement et en combinaison avec l'information prosodique et l'information spectrale à l'échelle de scores de décision. La combinaison de l'information sur la source d'excitation avec les autres types d'information a permis d'obtenir un gain relatif de 6,2 % pour le corpus Emo-DB et de 3,8 % pour le corpus ITKGP-SESC comparé à la combinaison de l'information spectrale et prosodique uniquement.

3.2.2.4 Éclats affectifs

Les éclats affectifs (*affect bursts*) sont définis comme étant de courtes expressions émotionnelles non verbales qui interrompent le discours (Schröder, 2003). Les tests de perception réalisés par Schröder montrent que les éclats affectifs, présentés sans contexte, peuvent transmettre une signification émotionnelle clairement identifiable avec un taux de reconnaissance de 81 %. Les éclats affectifs comprennent à la fois des sons clairs non vocaux (ex., les rires) et des interjections ayant une structure phonologique, tel que « Wow » qui est

utilisée généralement pour exprimer l'admiration, « Hey » pour la menace, « Ja » /jaÖ/ pour l'exaltation, «Hmm» pour l'ennui, « Uff » /ʊf:/ pour le soulagement, « Oh-Oh » pour l'inquiétude, prise rapide de souffle pour l'alarme, grognement /m:/ pour la colère, « Tse » / $ts^{h}ə$/ pour le mépris. Dans (Vidrascu *et al.* 2007), 11 traits caractéristiques de type éclats affectifs et de dysfluidité de la parole ont été étudiés en plus de 118 traits de types prosodiques (F0, énergie, durée), spectrales (formants et leurs largeurs de bande) et de qualité de la voix (vacillement, tremblotement et HNR) extraits d'un corpus d'émotions réelles issus d'un centre d'appel. Les 25 premiers traits plus pertinents des 129 ont été sélectionnés en utilisant un classificateur. Neuf des 25 meilleurs traits sélectionnés étaient de types éclats affectifs (larmes, voix inintelligible) et de dysfluidité (nombre d'hésitations et leurs longueurs). Notons que les traits de types éclats affectifs et de dysfluidité ont été manuellement transcrits.

3.2.2.5 Information linguistique

Les types de traits caractéristiques que nous avons cités jusqu'ici appartiennent tous à la catégorie d'information paralinguistique. Notons que certains auteurs ont introduit dans leurs modélisations l'information linguistique notamment l'information au niveau lexical (saillance émotionnelle des mots) et au niveau du discours (les actes de dialogue tels que le rejet, la répétition, la reformulation ou la demande de répétition). L'information linguistique a été expérimentée séparément dans (Devillers et Vasilescu, 2003) ou en combinaison avec l'information paralinguistique dans (Ang *et al.* 2002; Chen *et al.* 2006; Chuang et Wu, 2004; Forbes-Riley et Litman, 2004; López-Cózar et al. 2008; Lee et Narayanan, 2005; Planet *et al.* 2012; Schuller *et al.* 2005). Dans (Boufaden *et al.* 2008), les auteurs ont montré que l'utilisation d'un vocabulaire composé uniquement des deux mots *Yes* et *No*, donnés comme réponse dans le cadre d'une communication humain-machine d'un centre d'appel, permettait d'améliorer les performances de reconnaissances des émotions négatives versus émotions non-négatives de 13 % quand cette information est combinée avec l'information paralinguistique. L'introduction de l'information linguistique permet donc de réaliser un

certain gain en performance, mais au coût de perdre la propriété d'indépendance du système de RAE de tout langage.

3.2.2.6 Sélection des traits caractéristiques

L'ensemble d'or des traits les plus discriminants, indépendamment du type du corpus (simulé ou spontané) ou des catégories d'émotion véhiculées dans le corpus, si il existe, demeure indéterminé. Ceci a mené plusieurs auteurs à opter pour la méthode de la *force brute*, en procédant à l'extraction du maximum possible de traits acoustiques, que ce soit en quantité (des milliers de traits) ou en qualité (prosodiques, spectraux ou qualité de la voix). Des techniques de sélection de traits sont appliquées par la suite pour élire l'ensemble des traits les plus pertinents pour le cas applicatif étudié. Deux approches d'extraction de traits existent : méthodes basées sur l'extraction et méthodes basées sur la sélection de sous-ensembles de traits. Dans la première approche, les traits sont projetés dans un nouvel espace de taille réduite où les traits sont décorrélés soit en utilisant un apprentissage non supervisé (méthode d'analyse par composantes principales, ou *principal component analysis*, PCA en anglais), ou soit en maximisant la séparation des classes en utilisant un apprentissage supervisé (méthode d'analyse discriminante linéaire, LDA). Les méthodes ACP, LDA ainsi que la combinaison des deux méthodes ont été expérimentées et comparées dans (Hoque *et al.* 2006). Les performances obtenues suite à la combinaison des deux méthodes de réduction d'espace étaient meilleures que prises individuellement.

La deuxième approche basée sur la sélection de sous-ensemble a pour objectif de réduire au plus petit nombre possible les caractéristiques utilisées en utilisant des procédures de recherche et des fonctions d'évaluation. La procédure de recherche permet de générer un sous-ensemble de traits caractéristiques pour son évaluation à partir de l'ensemble des candidats possibles. Les fonctions d'évaluation mesurent la qualité du sous-ensemble de traits candidats généré par les procédures de recherche en vue de trouver l'ensemble de traits optimal. L'ensemble optimal est toujours relatif à la fonction d'évaluation utilisée. Les fonctions d'évaluation peuvent être classées en deux catégories : la catégorie enveloppante

(*wrapper*, telle que la méthode SFS (*Sequential Forward Selection*)) ou la catégorie filtre (*filter*, telle que la méthode RELIEF-F) selon que les sous-ensembles de traits sont évalués dépendamment ou indépendamment de l'algorithme d'apprentissage. Dans ce qui suit nous citerons quelques travaux avec les méthodes de sélection de traits utilisées.

- Sélection de l'avant (*forward selection*) et/ou élimination en arrière (*backward elimination)* (Bhatti, Yongjin et Ling, 2004; Fujie *et al.* 2004a; Huiqin *et al.* 2007; Kwon *et al.* 2003; Lee, 2004; Lin et Wei, 2005; Liu *et al.* 2007b; Pao *et al.* 2005; Sim, Jang et Park, 2007; Xiao *et al.* 2010; Xie *et al.* 2007; You *et al.* 2006; Zhu et Luo, 2007);
- Algorithme génétique (Beritelli *et al.* 2006; Casale, Russo et Serrano, 2007; Noda *et al.* 2006; Oudeyer, 2002; Scherer, 1996; Sim, Jang et Park, 2007);
- Sélection basée sur le facteur de corrélation (*Correlation-based Feature Selection*, CFS) (Vogt et André, 2005; Vlasenko *et al.* 2007); ex. : algorithme Relief-F (Petrushin, 2000; Yu, Aoki et Woodruff, 2004);
- Algorithmes à estimation de distribution (*Estimation of Distribution Algorithm*, EDA) (Alvarez *et al.* 2007);
- Méthode de la variance inexpliquée (*unexplained variance*) (Seppänen, Väyrynen et Toivanen, 2003);

3.2.3 Travaux selon la portée des traits caractéristiques

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

3.2.3.1 Information à court terme

L'information à court terme s'étale généralement sur un intervalle de temps, appelé trame, allant de 10 ms à 30 ms, cadencé à chaque 10 ms. Chaque trame constitue un vecteur de traits caractéristiques. La séquence des vecteurs de l'énoncé véhicule sa structure temporelle. Le vecteur de traits à court terme le plus utilisé est composé des coefficients MFCC, leurs dérivées premières et leurs dérivées secondes.

Parmi les travaux basés sur ce type de vecteur, nous retrouvons ceux de (Hui et al. 2007; Hung et al. 2004; Kim E. H. *et al.* 2007; Neiberg et al. 2006; Nwe et al. 2003; Pirker, 2007; Shafran *et al.* 2003; Vogt et André, 2005; Wahab et al. 2007). D'autres paramètres spectraux ont été également expérimentés, en combinaison avec les coefficients MFCC (Nwe, Foo et De Silva, 2003; Pao *et al.* 2005) tels que les coefficients LPC, LPCC, LFPC et PLP. L'information prosodique a été également utilisée en tant que trait à court terme, parfois seule (Huang et Ma, 2006; Nogueira *et al.* 2001; Schuller *et al.* 2003; Sethu et al. 2007) ou combinée avec les coefficients cepstraux tels les MFCC ou les LPCC (Kwon *et al.* 2003; Li et al. 2007; Lin et Wei, 2005; Nakatsu *et al.* 1999; Nicholson *et al.* 1999; Shafran *et al.* 2003).

3.2.3.2 Information à long terme

L'information à long terme caractérise l'énoncé dans sa globalité. Cette information est représentée sous forme de valeurs statistiques pour une séquence temporelle de valeurs. Ce sont les variables de type prosodique qui sont les plus couramment utilisées, telles que la fréquence fondamentale, l'énergie, le débit de la parole, le nombre et la durée des silences, le rapport de la durée de la région voisée à la région non voisée et celles de type phonétique telles que les formants et leurs bandes passantes, ainsi que des variations de paramètres, dont le tremblotement et le vacillement. La moyenne, l'écart type, le maximum, le minimum, la déviation, la pente, le quartile, le taux de passage par zéro sont des exemples de fonctions statistiques généralement utilisées.

Parmi les travaux basés sur l'utilisation des statistiques de la prosodie, nous citons (Alvarez *et al.* 2007; Bhatti, Yongjin et Ling, 2004; Giripunje et Bawane, 2007; Grimm et Kroschel, 2005; Inanoglu et Caneel, 2005; Lee, Narayanan et Pieraccini, 2002; Petrushin, 2000; Pittermann et Pittermann, 2007; Schuller *et al.* 2007b; Seppänen, Väyrynen et Toivanen, 2003; Yacoub *et al.* 2003; Yu *et al.* 2001; Zhu et Luo 2007). Dans (Pao *et al.* 2005) les paramètres LPC, LPCC, LFPC, PLP, MFCC ont été utilisés comme information spectrale (en plus du vacillement) à long-terme en calculant la valeur moyenne sur l'échelle de l'énoncé. Certains auteurs ont combiné, dans leurs vecteurs de traits de type à long terme, entre les statistiques de la prosodie et les statistiques des coefficients spectraux et qualité de la voix, que nous retrouvons par exemple dans les travaux de (Clavel *et al.* 2006; Kwon *et al.* 2003; Schuller *et al.* 2007a; Schuller *et al.* 2009; Vlasenko *et al.* 2007; Vogt et al. 2006; Zhang, 2008). Rares sont les travaux où seules les statistiques des coefficients cepstraux sont utilisées comme traits caractéristiques à l'instar des travaux de (Beritelli *et al.* 2006).

Dans (Rao et Koolagudi, 2012), les systèmes basés sur l'information prosodique à court terme sont comparés avec les systèmes basés sur l'information prosodique à long terme extraite pour chacune des trois unités : phrase, mot et syllabe. Pour l'information à court terme, chacun des vecteurs de traits de type F0, énergie et durée est modélisé séparément avec un SVM et fusionnés au niveau des scores de décision. Pour l'unité phrase, les performances obtenues par le système basé sur l'information locale (64,38 %) étaient nettement meilleures que celles obtenues par le système basé sur l'information à long terme (43,75 %). Les mêmes tendances de performances ont été observées pour les unités de types mot et syllabe quand les informations globale et locale sont comparées. Pour le système basé sur l'unité mot, les phrases ont été divisées en trois groupes : les mots de début, de milieu et de fin. Les scores obtenus pour chaque groupe de mots ont été fusionnées pour calculer les scores de décision au niveau de l'unité phrase. Le principe de division en trois groupes a été appliqué pour l'unité syllabe.

3.2.4 Travaux selon l'approche de classification

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

3.2.4.1 Approche dynamique

Nous pouvons distinguer deux types d'information temporelle : l'information mesurée sur l'échelle d'un énoncé (intra-énoncé) et l'information sur l'échelle du dialogue (inter-énoncé).

Structure temporelle intra-énoncé

Ce type d'approche a pour objectif la modélisation de la structure temporelle des énoncés. Ce sont les modèles HMM qui sont généralement utilisés dans ce type d'approche (El Ayadi, Kamel et Karray, 2007; Huang et Ma, 2006; Inanoglu et Caneel, 2005; Kwon *et al.* 2003; Lee *et al.* 2004; Li *et al.* 2007; Lin et Wei, 2005; Nogueira *et al.* 2001; Nwe, Foo et De Silva, 2003; Pao *et al.* 2005; Pirker, 2007; Pittermann et Pittermann, 2007; Schuller *et al.* 2003; Sethu, Ambikairajah et Epps, 2007; Vlasenko *et al.* 2007; Wagner, Vogt et André, 2007). Plusieurs types de modèles HMM ont été expérimentés : HMM simples avec architecture gauche-droite (Schuller *et al.* 2003) ou ergodique (Nwe *et al.* 2003), ou encore des modèles de type mixture de HMM expérimenté dans (Fernandez et Picard, 2003) qui ont donné des performances meilleures qu'un HMM simple.

La méthode déformation temporelle dynamique multidimensionnelle (*Multi-Dimensional Dynamic Time Warping*) a été également utilisée comme mesure de similarité entre les modèles dynamiques du contenu émotionnel calculé à l'échelle de l'énoncé. Ces modèles sont représentés à travers une série temporelle de vecteurs de type EP (*Emotion Profile*) constituant une matrice appelée *Ematogram* (Kim et Provost, 2013).

Structure temporelle inter-énoncé

L'information temporelle utile pour la classification des émotions ne se limite pas à la structure temporelle interne d'un énoncé (structure intra-énoncé) mais se manifeste également à travers le schéma de succession des états émotionnels véhiculé par l'historique des énoncés (structure inter-énoncés). Les modèles HMM ont été utilisés dans (Meng et Bianchi-Berthouze, 2011) pour modéliser cette structure inter-énoncés. Dans (Wollmer *et al.* 2009), ce sont les réseaux de neurones récurrents à longue mémoire à court terme (*Long Short-Term Memory*, LSTM) qui ont été utilisés pour la modélisation de l'information contextuelle entre les énoncés successifs de la parole émotionnelle. Dans (Wollmer *et al.* 2010), une modélisation basée sur une architecture combinant les réseaux bidirectionnels avec les réseaux LSTM (appelée LSTM *bidirectionnel*, BLSTM), a permis d'obtenir de meilleures performances que LSTM. Les réseaux bidirectionnels BLSTM sont constitués de deux couches LSTM, un pour le traitement en avant et l'autre pour le traitement en arrière.

3.2.4.2 Approche statique

Dans l'approche statique des classificateurs de type statique, tels que les GMM, SVM, KNN et les réseaux de neurones sont utilisés dans la modélisation. Avec ce type d'approche, il est possible également de récupérer l'information sur la structure temporelle de l'énoncé en optant pour l'information de portée à long terme comme traits caractéristiques. SVM donnent de meilleures performances avec l'approche statique basée sur l'information à long-terme. Le classificateur GMM modélisant l'information à court terme offre quant à lui de meilleures performances comparé aux autres classificateurs. Le modèle GMAVR (*gaussian mixture vector autoregressive models*) a été proposé par (El Ayadi *et al.* 2007), afin de modéliser la structure temporelle de la parole en intégrant le processus vectoriel autorégressif avec les GMM. Les performances obtenues avec GMAVR étaient supérieures à celles des HMM, KNN et réseaux de neurones multicouches. Dans (Zhu et Luo, 2007), un réseau de neurones modulaire (*Modular Neural Network*, MNN) constitué de sous-réseaux, a été expérimenté. Chaque réseau de neurones modulaire est spécialisé dans une classe d'émotion particulière.

Les résultats obtenus montrent que les performances du MNN dépassent celles d'un réseau de neurones standard. Parmi les travaux basés sur l'approche statique, nous retrouvons (Alvarez *et al.* 2007; Beritelli *et al.* 2006; Clavel *et al.* 2006; Kwon *et al.* 2003; Lee, Narayanan et Pieraccini, 2002; Lin et Wei, 2005; Neiberg, Elenieus et Laskowski, 2006; Pao *et al.* 2005; Petrushin, 2000; Schuller *et al.* 2007b; Seppänen, Väyrynen et Toivanen, 2003; Ververidis et Kotropoulos, 2005; Yacoub *et al.* 2003).

3.2.4.3 Approche logique floue

La troisième approche expérimentée est basée sur un système d'inférence flou. Ce choix est motivé par les incertitudes qui caractérisent les émotions, et particulièrement l'absence de frontières claires entre les différentes catégories d'émotions ainsi qu'au problème de chevauchement des classes dans la réalisation acoustique des émotions. Plusieurs travaux ont examiné cette approche (Giripunje *et al.* 2007; Grimm *et al.* 2005; Lee *et al.* 2003).

3.2.4.4 Approche basée sur la similarité

Une approche qui peut constituer une solution particulièrement intéressante pour le domaine de la RAE est celle basée sur la similarité. À notre connaissance, l'usage des descripteurs basés sur la similarité demeure encore une voie non exploitée dans le domaine de la RAE et les travaux de (Mower *et al.* 2011) demeurent parmi les très rares études qui s'inscrivent dans cette approche. Une méthode appelée *Emotion Profile (EP) based representation* a été proposée par Mower et ses collègues (2011). Dans cette méthode, les émotions sont exprimées en termes de présence ou d'absence d'un ensemble de composants d'émotions telles que la colère, la joie, la neutralité et la tristesse. Les EPs sont construits en utilisant un SVM avec une fonction à base radiale (RBF). Un SVM est entrainé pour chaque classe d'émotion versus les autres classes d'émotion. Chaque EP contient *n*-composants, un pour chaque sortie SVM spécifique à une émotion. Les profils sont créés en pondérant chacune des n sorties (± 1) par la distance séparant un point particulier de la frontière de l'hyperplan.

L'émotion finale est sélectionnée en classifiant le profil généré dans un mode dépendant du locuteur en utilisant l'approche naïve bayésienne (*Naive Bayes*).

Les scores de vraisemblance générés par les modèles d'ancrage sont un autre exemple de traits basés sur la similarité (les modèles d'ancrage, qui seront étudiés en détails dans les chapitres suivants, est la terminologie utilisée dans le domaine du traitement de la parole). Vu que les méthodes que nous allons proposer dans cette thèse sont principalement basées sur l'approche de similarité et sur les méthodes d'ancrage, nous allons présenter dans cette section une revue de littérature sur les travaux basés sur ces modèles réalisés essentiellement dans d'autres domaines de recherche. Les modèles d'ancrage ont été introduits pour l'indexation des locuteurs dans les grandes bases de données audio (Sturim *et al.* 2001), puis prolongé pour l'identification du locuteur (Mami *et al.* 2002), la vérification du locuteur (Yang *et al.* 2006) et plus récemment pour la classification des traits du locuteur tels que les traits de personnalité (Attabi et Dumouchel, 2012). Dans (Mami *et al.* 2003), l'application de l'analyse discriminante linéaire (LDA) aux scores de vraisemblance a permis aux modèles d'ancrage de dépasser les performances des GMM. En outre, une combinaison de modèles d'ancrage basés sur des approches probabilistes et déterministes a été proposée dans (Collet *et al.* 2005b). Les performances obtenues à travers cette combinaison ont dépassé celles obtenues avec un système à base de GMM-UBM. L'approche probabiliste vise à modéliser la variabilité intra-locuteur. Au lieu de représenter l'emplacement des énoncés d'un locuteur par un seul point dans l'espace des modèles d'ancrage, ils sont modélisés à l'aide d'une distribution normale. Parfois un classificateur SVM est placé en aval pour modéliser les vecteurs de scores de vraisemblance. Une telle architecture hybride (combinant un modèle génératif et discriminatif, GMM- ou HMM-SVM) a déjà été appliquée avec succès pour la classification des chiffres manuscrits (Abou-Moustafa *et al.* 2004) et pour la vérification de signature hors ligne (Batista *et al.* 2010). Dans les deux études un HMM est utilisé comme système frontal pour calculer les scores de probabilité alimentant un SVM (ou ensemble de SVM). Une amélioration relative de 1,23 % a été obtenue par rapport à un HMM pour le problème des chiffres manuscrits tandis que le taux d'erreurs individuel pouvait être réduit de 10 % pour le problème de la vérification de signature. Pour la tâche de reconnaissance du

locuteur, l'architecture GMM-SVM a été également expérimentée dans (Zhao *et al.* 2007) et (Lei *et al.* 2006). Dans (Zhao *et al.* 2007), le système GMM-SVM atteint des performances comparables au système GMM-UBM. Dans (Lei *et al.* 2006), où les distributions gaussiennes dans l'UBM ont été utilisées comme un espace de référence, les résultats du système GMM-SVM étaient meilleurs que les modèles d'ancrage basés sur la distance euclidienne à la fois pour les problèmes de vérification et d'identification du locuteur.

3.3 Combinaison de classificateurs

Nous avons déjà vu que la reconnaissance des émotions à partir de la parole est caractérisée par une importante confusion, non seulement pour les systèmes de RAE mais aussi pour les humains. Une façon que l'humain parfois utilise pour lever cette ambigüité est de rechercher d'autres indices au niveau du visage ou des gestes du locuteur. Si nous désirons faire un parallèle avec le domaine de la reconnaissance des formes et des machines d'apprentissage, ceci correspond à une combinaison de classificateurs. Par conséquent, la combinaison de classificateurs devient un moyen à explorer pour améliorer la robustesse des systèmes de RAE. Cependant, dans le domaine de la RAE à partir de la parole appliqué par exemple dans un contexte d'un centre d'appels, les autres sources additionnelles d'information telles que l'image du visage ou les gestes du locuteur ne sont pas disponibles et la seule source d'information pour lever l'ambiguïté se limite à l'utilisation de la voix. L'apport attendu par la combinaison de classificateurs proviendrait donc essentiellement de la diversité de l'évidence engendrée par une multitude de systèmes basés sur différents types de descripteurs acoustiques, extraits éventuellement au niveau d'unités d'analyses différentes, mesurés sur diverses échelles et/ou modélisés avec différents modèles de classification. La stratégie de combinaison de ces sources d'indices aura également un grand impact sur les performances de la fusion. La stratégie de combinaison comprend la méthode de combinaison (algorithme) et le stade de la fusion (niveau des traits vs niveau des scores). Notons qu'il n'est pas important d'avoir des classificateurs très performants individuellement pour obtenir une combinaison performante, mais plutôt d'avoir des classificateurs complémentaires dans leurs prises de décision. Dans (Lugger *et al.* 2009), trois méthodes de combinaisons ont été citées

et comparées : hiérarchique, série et parallèle. Dans cette thèse nous allons organiser les méthodes de combinaison dans la revue de la littérature des systèmes de RAE en quatre groupes. En plus des trois méthodes citées précédemment nous ajouterons une quatrième méthode de combinaison qu'on appellera combinaison en cascade. Pour plus d'informations sur les méthodes de combinaison des classificateurs en général nous référons le lecteur à (Kuncheva, 2004).

3.3.1 Combinaison en cascade

Une combinaison en cascade est basée sur une succession de classificateurs où les sorties de chaque classificateur sont utilisées comme données d'entrée pour le classificateur suivant. L'objectif de cette architecture est la recherche de descripteurs de plus haut niveau ayant une plus grande capacité de discrimination entre les classes d'émotion. Ces traits de haut niveau sont obtenus en sortie d'un classificateur placé en amont ayant traité les descripteurs de plus bas niveau, et qui seront utilisés à leur tour par un autre classificateur subséquent. Généralement ce type d'architecture sont composées de deux à trois niveaux (classificateurs). Les deux architectures GMM- ou HMM-SVM et DNN-HMM sont deux exemples d'une telle architecture.

Dans (Dumouchel *et al.* 2009; Hu *et al.* 2007; Lefter *et al.* 2010), les supervecteurs sont utilisés comme traits de haut niveau afin de discriminer les émotions. Les supervecteurs sont des vecteurs obtenus après concaténation des moyennes de chaque gaussienne d'un GMM pour former un vecteur de haute dimension. Un SVM basé sur les supervecteurs offrent de meilleures performances que celles d'un GMM standard et ceci pour les deux modes d'évaluation indépendant ou dépendant du genre dans (Hu *et al.* 2007 et Lefter *et al.* 2010). L'intégration des fonctions exhaustives (*sufficient statistics*) du second ordre des termes de la covariance (décrivant la forme de la distribution) aux côtés des fonctions exhaustives du premier ordre (la moyenne) dans les supervecteurs a permis dans (Nwe *et al.* 2013) d'améliorer les performances de classification.

Dans (Chandrakala et Sekhar, 2009), un GMM est également utilisé comme modèle génératif en amont d'un classificateur de type discriminatif, un SVM en occurrence. Deux approches différentes ont été utilisées pour la classification des séries temporelles des vecteurs MFCC. Dans la première, chaque série temporelle des données d'entrainement est modélisée par un GMM. Si *M* est le nombre de séries de données d'entrainement, chaque énoncé est représenté par un vecteur composé de *M* valeurs de probabilité de vraisemblance associées aux *M* modèles GMM. Ces vecteurs sont soumis comme données d'entrée au SVM. Dans la deuxième approche, chaque série temporelle des données est divisée en un nombre fixe, *L*, de segments. Chaque segment est modélisé par un GMM. Les paramètres de chacune des gaussiennes des GMM (moyenne, covariance, pondération) d'un segment donné sont concaténés pour former un vecteur caractérisant un segment. La série temporelle est finalement modélisée par un vecteur de taille fixe composé de la concaténation des *L* vecteurs représentant les *L* segments de la série temporelle. Les vecteurs obtenus sont également modélisés par un SVM. Les résultats obtenus montrent que le système basé sur la deuxième approche où chaque segment est modélisé par une seule gaussienne avec une matrice de covariance pleine donne de meilleures performances. Par ailleurs, ces deux systèmes combinant GMM et SVM offrent des performances largement supérieures aux systèmes basés uniquement sur un seul modèle SVM ou GMM.

Dans (Ortego-Resa *et al.* 2009), les scores d'un système GMM-SVM (SVM entrainé avec les supervecteurs) basé sur l'information prosodique à court terme et les scores d'un second système SVM basé sur l'information prosodique à long terme sont fusionnés et entrainés avec un troisième SVM. Cette méthode de combinaison, appelée fusion de modèles d'ancrage, a permis d'améliorer significativement les performances de reconnaissance lorsque testée sur deux des trois corpus par rapport à une fusion au moyen de la règle de la *somme*.

Dans (Meng et al. 2011), un système composé de trois étages a été proposé pour la classification des unités émotionnelles de type mot selon les niveaux des dimensions affectifs. Chaque étage admet comme entrées, les sorties (scores) des classificateurs de

l'étage précédent. Une combinaison de 13 classificateurs KNN (qui diffèrent dans le nombre de voisins, *k*) sont utilisés comme classificateurs du premier étage. En deuxième étage, un HMM discret modélisant l'information temporelle entre les différentes unités est apparié à chacun des classificateurs KNN. Les prédictions de l'ensemble des 13 classificateurs HMM sont combinées à travers un autre HMM bâti en troisième étage. Les résultats ont montré que les systèmes à trois étages modélisant l'information temporelle entre les unités d'expression affective, améliorent significativement les performances par rapport à un système à un seul étage ne tenant pas compte de cette information temporelle.

Le modèle neuro markovien profond (ou DNN-HMM, *Deep Neural Network - Hidden Markov Model*) est un nouveau modèle d'architecture hybride qui a été récemment expérimenté dans le domaine de la reconnaissance des émotions (Le et Provost, 2013; Li *et al.* 2013) après avoir été testé avec succès dans le domaine de la reconnaissance de la parole. Les réseaux de neurones profonds (DNN), sont des réseaux MLP classiques avec plusieurs couches, où l'apprentissage est généralement initialisé par un algorithme de préapprentissage. Cet ensemble de couches est capable de capturer la relation non linéaire sous-jacente entre les données. La nouveauté de cette nouvelle famille de réseaux de neurones réside dans la façon dont les couches cachées sont entrainées. L'apprentissage d'un réseau de neurones profond est réalisé en deux étapes et a été proposé dans (Bengio *et al.* 2007; Hinton *et al.* 2006). Les premières couches cachées sont entrainées de manière non supervisée en utilisant en général des machines de *Boltzman* restreintes (RBM, on parle alors de réseaux de croyance profonde), successivement verrouillés et empilés de l'entrée jusqu'à la dernière couche. L'estimation des poids d'une couche cachée est qualifiée de préapprentissage. La dernière couche (couche de décision) est ensuite ajoutée au réseau. Pour apprendre les paramètres de cette couche, tous les poids des couches du modèle sont déverrouillés et une rétropropagation classique est effectuée sur l'ensemble du réseau. Cette opération permet d'apprendre la fonction de décision discriminante et d'affiner les paramètres du réseau. RBM est un type de modèle graphique non orienté construit à partir d'une couche d'unités cachées stochastiques binaires et une couche d'unités visibles stochastiques avec une distribution gaussienne pour traiter les valeurs réelles des données d'entrée de la parole. Afin de rendre

un HMM plus discriminant, les mélanges de gaussiennes (GMM) des HMM sont remplacés par un réseau de neurones profond. Ainsi, le pouvoir génératif et modélisant des HMM est combiné avec la capacité discriminante d'un réseau de neurones profond (DNN). Dans (Le et Provost, 2013), 39 systèmes basés sur l'architecture DNN-HMM et qui diffèrent dans le nombre de trames à l'intérieurs des fenêtres contextuelles ainsi que le nombre d'états dans les HMM, ont été expérimentés sur le corpus FAU AIBO. Le système basé sur un seul état avec une fenêtre de 37 trames a obtenu 45.08% en termes de la moyenne non pondérée des rappels (UAR), dépassant les résultats de l'état de l'art. Dans la même étude, le meilleur résultat obtenu au moyen de la combinaison de plusieurs classificateurs DNN-HMM était de 45.60 %. Dans (Li *et al.* 2013), plusieurs variantes de systèmes basés sur une architecture hybride avec HMM ont été comparées. Le système DNN-HMM basé sur un pré-entrainement supervisé a obtenu de meilleures résultats comparé à un système pré-entrainé d'une manière non supervisée et a également dépassé les performances des systèmes de type *shallow*-NN-HMM, MLP-HMM et GMM-HMM.

3.3.2 Combinaison hiérarchique

Une architecture hiérarchique est composée de plusieurs mini-classificateurs distribués sur plusieurs étages utilisée comme alternative à un classificateur unique multi-classe. Les mini-classificateurs servent à discerner des sous-groupes de classes dans des étages situés en amont avant de procéder à une classification plus raffinée dans les étages subséquentes. La motivation derrière cette architecture repose sur l'idée qu'en commençant par la classification des classes plus faciles, nous réduisons le nombre de classes à discriminer et par conséquent on réduit la propagation d'erreurs et on augmente la probabilité de classifier les classes restantes qui sont autrement difficilement discriminables. L'autre avantage consiste en la possibilité d'utiliser à chaque étage les traits les plus discriminants adaptés aux sous-ensembles des classes sous-jacentes (Breazeal et Aryananda, 2002; Lugger et Yang, 2007). L'ordre de séparation des classes peut être motivé par deux choix (i) connaissance a priori sur les modèles théoriques des émotions i.e., classes d'émotion à haute *activation* vs

classes d'émotion à faible *activation* (ii) ou à travers un test empirique en se fiant à la matrice de confusion obtenue d'un classificateur préliminaire (Lee *et al.* 2011).

Dans (Fu *et al.* 2008), une architecture à deux étages comprenant un classificateur SVM en amont utilisé pour la classification des classes en deux sous-groupes d'émotion est suivie par une seconde classification plus raffinée utilisant une combinaison de quatre HMM. L'architecture à deux étages a permis d'améliorer le taux de reconnaissance par rapport à une architecture à un seul étage (constituée de HMM) de 57,8 % à 76,1 %. Dans (Lugger et Yang, 2007), le classificateur de base est comparé à un classificateur basé sur une architecture à deux étages motivée par le modèle bidimensionnel. Dans le premier étage, un classificateur est entrainé pour classifier les émotions selon l'axe *activation* (élevé vs bas) du modèle d'émotion dimensionnel en utilisant les traits prosodiques. Dans le second étage, un classificateur est créé pour classifier les émotions de chacun des sous-groupes issus de l'étape précédente, en utilisant la combinaison de tous les types de traits. Le taux obtenu avec cette architecture est de 74,5 % dépassant le taux obtenu avec l'architecture à un seul étage qui lui est égal à 66,7 %. Les mêmes auteurs ont expérimenté une architecture à trois étages, motivée cette fois-ci par le modèle d'émotion tridimensionnel comprenant les axes : *activation*, *emprise* et *évaluation* (Lugger et Yang, 2008). Dans le premier étage les émotions sont classées selon le niveau d'*activation*. Dans le second étage, la classification est réalisée selon le niveau *emprise* (*potency*) à l'intérieur de chaque classe d'*activation*. En troisième étage, les émotions sont discriminées selon la dimension *évaluation*. Un gain absolu de 5.5 % a été apporté par l'architecture à trois étages par rapport à l'architecture à deux étages (passant de 83,5 % à 88,8 %). Dans (Xiao *et al.* 2010), une architecture multi-étage guidée par le modèle dimensionnel des émotions a permis d'améliorer les performances de 65,12 % à 73,58 % pour le corpus Berlin en utilisant un système dépendant du genre. Dans (Breazeal et Aryananda, 2002), une architecture à trois étages a également donné de meilleures performances qu'un classificateur à un seul étage.

3.3.3 Combinaison parallèle

La combinaison de plusieurs classificateurs, connue sous le nom de combinaison parallèle, est efficace quand le groupe de classificateurs est très diversifié et est négativement dépendants, c.-à-d., les erreurs sont commises sur des objets fortement différents (Kuncheva *et al.* 2000). Contrairement au *bagging* ou au *boosting*, où la diversité est basée sur les différentes instances d'apprentissage, la diversité peut également être obtenue en choisissant différents sous-ensembles de caractéristiques pour chaque membre du groupe des classificateurs. Afin de simplifier la présentation des travaux, nous proposons une organisation selon la motivation recherchée derrière cette stratégie de combinaison. Nous proposons d'utiliser les quatre critères déjà utilisés pour la présentation des travaux basés sur un classificateur simple.

3.3.3.1 Diversification dans les types de traits

Dans (Fu *et al.* 2008), quatre classificateurs HMM basés sur différents ensembles de traits (incluant information prosodique et spectrale) ont été combinés en utilisant une nouvelle variante du vote classé (*ranked voting*) appelée vote d'ordre pondéré (*weighted order voting*). Les HMM sont basés sur des modèles discrets gauche-droite. Un taux de reconnaissance de 57,8 % a été obtenu après fusion alors que celui du meilleur système HMM pris individuellement est de 40,5 %. Les expériences ont été évaluées en utilisant la base de données *Beihang University Mandarin Emotion Speech*, un corpus d'émotion simulée.

3.3.3.2 Diversification dans la portée temporelle de l'information acoustique

Dans (Rao et Koolagudi, 2012), la combinaison de l'information prosodique à court terme avec l'information prosodique à long terme extraites au niveau de la phrase a permis de réaliser un gain relatif de 2,52 %. Dans la même étude, la combinaison du même type d'information extraite cette fois-ci sur les échelles d'unité mot et syllabe a également amélioré les résultats de classification.

3.3.3.3 Diversification des unités d'analyse

La combinaison des traits extraits à différents niveaux (trame, syllabe et mot) permet d'améliorer les performances comparées au meilleur système n'utilisant qu'un ou deux types d'unités seulement quand appliquée à la langue chinoise (Mandarin), (Kao et Lee, 2006). Dans (Clavel *et al.* 2006), l'information extraite de la région voisée est combinée avec l'information de la région non-voisée au niveau des scores de décision. Les résultats obtenus montrent que l'apport des régions non-voisées n'est significatif que si les segments classés sont totalement non-voisés ce qui met en évidence le pouvoir discriminatif des régions voisées par rapport aux régions non-voisées.

3.3.3.4 Diversification des unités d'analyses et des types de descripteurs

Dans (Bitouk *et al.* 2010), l'information spectrale extraite au niveau du phonème et l'information prosodique extraite au niveau de l'énoncé ont été combinées en un seul vecteur de traits (fusion précoce). Les résultats obtenus montrent un gain relatif de 3 % pour le premier corpus et une baisse relative de l'ordre de 1 % pour le second corpus comparés au système basé sur l'information spectrale à l'échelle du phonème.

3.3.3.5 Diversification des modèles de classification

Dans (Pao *et al.* 2007) des vecteurs de traits composés des descripteurs MFCC, LPCC et LPC ont été utilisés pour construire différents modèles de classificateurs; KNN, KNN *Weighted* (WKNN), *Weighted Discrete* KNN (W-DKNN), *Weighted Average Patterns of Categorical*, KNN (WCAP), et SVM. La combinaison des classificateurs en utilisant les règles du vote majoritaire ou du *maximum* a permis d'améliorer le taux de classification du meilleur classificateur (W-DKNN en occurrence) de 0,9 % à 6,5 %. Chacun des classificateurs peut mieux reconnaitre certaines classes d'émotions que d'autres classificateurs.

3.3.3.6 Diversification des types et portées de traits, d'unités d'analyse et de modèles de classification

Dans (Hu *et al.* 2007), l'information à long-terme de la prosodie et de la qualité de la voix est combinée avec l'information spectrale (MFCC) à court terme. Le système basé sur l'information à long-terme est basé sur le classificateur SVM. L'information à court terme est modelé par un GMM pour extraire un supervecteur formé de la valeur moyenne de chaque gaussienne. Les supervecteurs sont utilisés comme entrée pour le classificateur SVM. Les performances du système basé sur l'information à court terme (82,5 %) étaient supérieures à celles du système à long terme (79,2 %). La combinaison des deux systèmes permet de réduire le taux d'erreur du meilleur système de 25,6%, 23 % et 22,9 % pour les femmes, hommes et indépendant du genre respectivement. Dans (Lefter *et al.* 2010), quatre classificateurs ont été combinés en utilisant la régression logistique linéaire. Il s'agit d'un SVM basé sur l'information prosodique à long terme, un deuxième classificateur UBM-GMM basé sur les traits acoustiques à court terme (*Relative Spectral Perceptual Linear Predictive*, RASTA PLP), un troisième classificateur basé sur les supervecteurs modélisés avec un SVM et enfin un quatrième classificateur connu sous le nom «*dot-scoring*» (qui est une approximation linéaire d'un UBM-GMM). Le taux d'erreur a été réduit à 4,2 % alors que le taux d'erreur du meilleur classificateur était de 15,5 %. Dans (Vlasenko *et al.* 2007), l'information spectrale (MFCC) à court terme est modélisée avec un GMM. Les probabilités de vraisemblance des énoncées obtenues avec le modèles GMM sont combinées avec les traits à long terme de type prosodique, spectral et qualité de la voix pour former un seul vecteur de traits utilisé comme vecteur d'entrée à un SVM. Les résultats obtenus montrent que la combinaison des scores du modèle GMM aux autres traits permettait d'apporter des gains relatifs de l'ordre de 8,1 % pour le corpus EMODB et 0,6 % pour le corpus SUSAS.

Dans (Dumouchel *et al.* 2009), les scores de trois systèmes basés sur différentes propriétés ont été testés sur le corpus FAU AIBO Emotion. Le premier système utilise l'information spectrale à court terme modélisé avec un GMM. Dans le deuxième système, l'information prosodique à l'échelle de l'unité *pseudosyllabe* a été modélisée à travers un modèle GMM-

UBM. Un SVM basé sur les supervecteurs a été utilisé comme troisième système. Les trois systèmes ont été combinés en utilisant la régression logistique. Les résultats obtenus après fusion a permis d'améliorer légèrement les résultats du meilleur système individuel (le premier système). Dans (Kim, Georgiou *et al.* 2007), l'information spectrale à court-terme modélisée avec un GMM a été combinée avec l'information prosodique à long-terme modélisée avec l'algorithme kNN. La fusion des scores des deux systèmes a aussi permis une réduction du taux d'erreur du système de détection des émotions en temps réel.

3.3.4 Combinaison série

La combinaison série consiste à exploiter des classificateurs en file. Chaque classificateur reconnaitra un sous ensemble d'instances et soumettra les instances restantes au classificateur suivant éventuellement plus compétent pour reconnaitre une partie des instances restantes.
Dans (Lugger et Yang, 2009), trois stratégies de combinaisons ont été évaluées: hiérarchique, série et parallèle (chaque classificateur fournit une prédiction indépendamment des autres). Les résultats obtenus en utilisant les traits prosodiques, spectraux et qualité de la voix extraits du corpus Berlin Emotional Database montrent la dominance de la combinaison parallèle suivie par la méthode série.

3.4 Techniques d'amélioration des performances des systèmes de RAE

Dans cette section nous allons présenter quelques techniques utiles pour l'amélioration des performances des systèmes de RAE. On peut classer ces méthodes principalement en deux catégories : techniques exploitant l'information du mode opératoire des systèmes et méthodes remédiant au problème de la rareté des données.

3.4.1 Techniques basées sur l'exploitation de l'information sur le mode opératoire

3.4.1.1 Mode dépendant- versus indépendant du locuteur

Plusieurs études ont montré que l'adaptation d'un système de reconnaissance d'émotion à un locuteur particulier permet d'améliorer les performances en comparaison à un système opérant en mode indépendant du locuteur. Les résultats obtenus dans (Rybka *et al.* 2013) montrent que l'importance du taux d'amélioration dépend du classificateur utilisé (élevé pour SVM et bas pour KNN). Il est montré également que l'amélioration a été particulièrement observable quand le taux de reconnaissance de base d'un locuteur donné était faible. L'intégration de l'information a priori sur un locuteur peut se faire par l'intégration des données d'apprentissage du locuteur dans la construction du modèle d'apprentissage (Rybka *et al.* 2013) ou à travers une normalisation des traits dépendante du locuteur (Cao *et al.* 2012; Le et Provost, 2013; Schuller *et al.* 2010). Dans un contexte d'apprentissage multi-corpus, trois méthodes de normalisation ont été comparées dans (Schuller *et al.* 2010): par locuteur, par corpus, combinaison des deux normalisations et sans normalisation. Les meilleures performances ont été obtenues par les systèmes basés sur une normalisation par locuteur.

3.4.1.2 Mode dépendant versus indépendant du genre

La création de modèles dépendants du genre, afin de prendre en considération les différences des traits acoustiques entre hommes et femmes, permet d'améliorer encore plus les performances de reconnaissance (Lee et Narayanan, 2005; Ververidis et Kotropolos, 2004; Lin et Wei, 2005). Dans certains travaux, les auteurs ont supposé que l'information sur le genre est déjà disponible et utilise le classificateur correspondant pour la prédiction (Hu *et al.* 2007). Dans d'autres études, aucune information préalable n'est supposée connue et un détecteur automatique du genre, utilisé en amont du système dans une architecture à deux étages ou plus, est intégré pour la détection du genre (Shahin, 2013; Vogt et André, 2006; Xiao *et al.* 2010). En plus des données spécifiques au genre utilisées dans l'apprentissage du modèle de classification, un système dépendant du genre peut profiter de la phase

d'extraction des traits pour sélectionner les traits les plus pertinents en fonction du genre (Lee et Narayanan, 2005; Ververidis et Kotropolos, 2004; Vogt et André, 2006; Xiao *et al.* 2010). Dans (Vogt et André, 2006), l'utilisation d'un détecteur de genre automatique ayant un taux de reconnaissance égal à 90,26 % pour le corpus Berlin et 91,85 % pour SmartKom, a permis une amélioration relative de 2 % et 4 % pour chacun des deux corpus respectivement par rapport un classificateur indépendant du genre. Dans (Shahin, 2013), les performances d'un classificateur utilisant un système de détection automatique du genre, avec un taux de reconnaissance égal 96,77 %, est comparé avec un deuxième système de reconnaissance d'émotion générique et un troisième système muni de l'information correcte sur le genre. Le système muni d'un système de détection automatique du genre a obtenu un taux de reconnaissance moyen de 86,79 % et des taux de 78,06 % et 82,31 % pour le deuxième et troisième système respectivement quand les systèmes sont testés avec le corpus LDC (*Linguistic Data Consortium*) *Emotional Prosody Speech and Transcripts*. Des performances similaires ont été obtenues quand les systèmes ont été évalués avec un second corpus collecté par les auteurs. Dans (Xiao *et al.* 2010), une amélioration de 4,78 % est obtenue sur le corpus Berlin avec un système dépendant du genre par rapport à un système mixte. Ce dernier cause particulièrement une mauvaise prédiction de l'émotion *peur* par rapport aux autres émotions (Xiao *et al.* 2010).

3.4.2 Techniques basées sur le traitement du problème de rareté des données d'apprentissage

Un des problèmes majeurs dans le domaine de la RAE à partir de la parole est la difficulté de collecter de larges quantités de données émotionnelles réelles (Schuller *et al.* 2011). Afin de contourner ce problème, des solutions basées soit sur l'utilisation d'autres corpus déjà disponibles et annotés ou soit sur l'utilisation de données non étiquetées ont été proposées.

3.4.2.1 Combinaison de plusieurs corpus

Une des méthodes proposées pour améliorer la robustesse des systèmes entrainés avec une quantité faible de données est l'utilisation des données d'apprentissage collectées dans un autre contexte ou même dans une langue différente (Lefter *et al.* 2010; Shami et Verhelst, 2007; Vidrascu et Devillers, 2008). Dans (Shami et Verhelst, 2007), des systèmes entrainés sur des corpus et testés sur un corpus différent n'offraient que peu de pouvoir de généralisation. Quand tous les corpus sont fusionnés, les performances des systèmes sont légèrement dépassées par les systèmes entrainés et testés sur le même corpus. Dans (Lefter *et al.* 2010), les performances de systèmes entrainés et testés avec le même corpus (approche intra-corpus), utilisés comme système de référence sont comparés avec (i) des systèmes entrainés avec deux corpus et testés sur un troisième corpus (approche corpus croisés) et (ii) des systèmes entrainés avec tous les corpus, y compris les données d'entrainement du corpus utilisé pour le test (approche corpus mixte). Dans le cas où l'ensemble des corpus ne contiennent que des émotions simulées, les performances obtenues avec les systèmes à corpus croisés sont en général plus faibles que celles des systèmes de références (avec quelques exceptions) alors que les performances avec l'approche à corpus mixte sont meilleures dans la plupart des conditions. Dans le cas où un corpus d'émotion réelle est utilisé comme quatrième corpus dans les expériences, le taux d'erreur obtenu avec l'approche corpus croisé est le double du taux obtenu avec l'approche intra-corpus quand les classificateurs sont entrainés avec les données d'émotion actée et testés avec les données d'émotion réelle. Avec l'approche mixte, les performances sont meilleures pour un corpus sur trois.

3.4.2.2 Co-apprentissage

L'idée est de minimiser l'effort humain d'annotation en utilisant le co-apprentissage (co-training) (Liu *et al.* 2007; Maeireizo *et al.* 2004; Zhang *et al.* 2013). Le co-apprentissage est un algorithme d'apprentissage semi-supervisé et itératif permettant d'annoter les données d'une manière automatique à partir d'une quantité limitée de données manuellement

étiquetées en se basant sur la sortie de deux classificateurs. Les deux classificateurs sont entrainés en utilisant deux ensembles de traits qui sont conditionnellement indépendants étant donnée la classe (i.e., traits prosodiques à long terme vs traits spectraux à courts terme). À chaque itération de l'algorithme, le sous-ensemble des données correctement classifiées avec le plus haut niveau de confiance par les deux classificateurs est ajouté aux données d'apprentissage pour entrainer de nouveau les deux classificateurs. Dans (Liu *et al.* 2007), un sous ensemble de données égal à 5 % de la taille des données globales d'apprentissage est utilisé comme données étiquetées pour l'entrainement du modèle de départ des deux classificateurs. Le co-apprentissage a permis d'apporter une amélioration de 9 % pour le modèle des femmes et de 7,4 % pour le modèle des hommes, comparé aux modèles entrainés d'une manière supervisée.

3.4.2.3 Étiquetage actif

L'étiquetage actif (active labelling) est une méthode de sélection introduite pour réduire la quantité de données étiquetées nécessaire pour l'apprentissage, en sélectionnant les données les plus informatives qui permettraient d'améliorer les performances. Les données non étiquetées ainsi sélectionnées sont par la suite manuellement annotées. Cette méthode s'avère particulièrement utile quand la distribution des classes est fortement déséquilibrée. Dans la méthode d'étiquetage actif proposée dans (Zhang *et al.* 2013), la valeur informative véhiculée par une donnée est estimée à travers la valeur du taux d'agrément entre annotateurs d'un ensemble de données de départ généralement en petite quantité. À travers ces valeurs, un modèle de régression est entrainé pour prédire les valeurs de données non étiquetées à sélectionner. Le critère de sélection peut être basé soit sur la valeur de confiance de la donnée ou soit sur la classe minoritaire. Les expériences menées dans (Zhang *et al.* 2013) sur le corpus FAU AIBO Emotion pour résoudre le problème à deux classes, ont montré que la seconde option permettait d'obtenir de meilleurs résultats. Avec un ensemble de départ de 500 instances annotées (3 % de la taille totale du corpus), la sélection active suivie d'une annotation humaine de 2,4k d'instances supplémentaires de la classe minoritaire a permis d'améliorer les performances de 2 % comparée au système obtenu avec toutes les données

d'apprentissage (d'une taille de 9459), c'est-à-dire en réduisant le taille des données d'entrainement de 70 %.

3.5 Conclusion

Dans ce chapitre nous avons présenté une revue de la littérature sur les systèmes de RAE. Nous avons commencé par présenter les différents systèmes basés sur des classificateurs simples selon la nature et la portée de l'information acoustique extraite, le type d'unité d'analyse émotionnelle et l'approche de classification. Nous avons vu que les descripteurs acoustiques de bas niveau ne sont généralement que peu discriminants pour servir à caractériser n'importe quel type ou nombre de classes d'émotion particulièrement lorsque ces descripteurs sont modélisés par des classificateurs de base. C'est pourquoi l'utilisation de stratégies de combinaison de classificateurs représente une solution intéressante pour améliorer la robustesse des systèmes de RAE. Nous avons vu dans plusieurs études que la combinaison hiérarchique permettait d'améliorer les performances en comparaison avec un système à étage unique. Nous avons également vu que la combinaison en cascade permettait de générer des descripteurs de plus haut niveau plus discriminants pour la classification. La majorité des travaux se sont intéressés à la combinaison parallèle cherchant à combiner des classificateurs de base indépendants en diversifiant un ou plusieurs critères sur lesquels est basée la conception d'un système de RAE. Indépendamment de la meilleure méthode de combinaison, il est en ressort clairement qu'un système multiétage est largement plus discriminant qu'un système à étage unique. Nous avons terminé le chapitre par la présentation de quelques techniques très intéressantes tenant en compte les spécificités du domaine de la RAE pour rendre les systèmes de reconnaissance encore plus robustes.

CHAPITRE 4

MÉTHODOLOGIE ET APPROCHE BASÉE SUR LA SIMILARITÉ POUR LA CLASSIFICATION DES ÉMOTIONS

4.1 Introduction

Afin de répondre à la complexité du problème de la reconnaissance automatique des émotions, nous proposons dans cette thèse une approche basée sur la similarité. Nous distinguons deux niveaux auxquels le concept de similarité peut être appliqué i) niveau des traits caractéristiques et ii) niveau des méthodes de classification. Dans une approche où les traits sont basés sur la similarité, les objets sont caractérisés relativement à d'autres objets pris comme ensemble de référence. Les traits basés sur la similarité, non profondément investigués dans le domaine de la RAE, nous paraissent tout indiqués en absence d'un ensemble optimal de traits acoustiques discriminants indépendamment du type du corpus d'émotion. La robustesse des traits basés sur la similarité peut être renforcée si ces derniers sont extraits à partir d'un espace de descripteurs de haut niveau.

Dans ce chapitre nous allons introduire et motiver l'approche basée sur la similarité pour le problème de la RAE (section 2). Nous décrirons dans la section 3, le corpus de données sélectionné pour entrainer et évaluer les modèles proposés et les considérations qui nous ont menées vers le choix de ce corpus. Nous décrirons dans la section 4 le protocole d'expérimentation et les critères d'évaluation des systèmes proposés. Nous consacrerons la section 5 pour l'investigation d'un ensemble de descripteurs de haut niveau qui peuvent constituer des candidats potentiels pour mesurer la similarité. Le type de descripteurs de haut niveau le plus pertinent aux spécificités de notre problématique sera sélectionné et utilisé dans les systèmes proposés dans les chapitres subséquents.

4.2 Approche basée sur la similarité

4.2.1 Motivation

La méthodologie proposée pour réduire la confusion entre classes et ainsi améliorer les performances des systèmes de RAE repose sur une approche basée sur la similarité. Cette approche est motivée en premier lieu par les connaissances a priori sur les modèles théoriques des émotions du domaine de la psychologie où principalement deux modèles traditionnels sont en compétition pour la représentation des émotions et ayant chacun ses propres avantages et inconvénients à savoir le modèle discret et le modèle dimensionnel. L'idée est de reconnaitre les émotions catégoriques (modèle discret) en utilisant un espace continu (modèle dimensionnel), combinant ainsi les deux modèles au moyen d'une approche basée sur la similarité. La méthodologie basée sur la similarité représente une approche naturelle pour aborder le problème de la reconnaissance de l'émotion car le concept de proximité entre classes est bien présent et illustré dans la cartographie des émotions catégoriques dans l'espace bi- ou tridimensionnel des émotions. Ainsi, dans cet espace, chaque classe d'émotion peut être considérée comme proche (similaire) ou lointaine (dissimilaire) par rapport à d'autres catégories selon la dimension considérée. La notion de proximité et sa nature possède donc une existence conceptuelle propre en psychologie indépendamment de la configuration expérimentale des systèmes de RAE (tel que les types des traits acoustique extraits). Le deuxième concept que nous pouvons relever du modèle dimensionnel est l'existence de dimensions théoriques telles que les dimensions *valence*, *excitation* et *contrôle* qui représentent les critères selon lesquels la (dis)similarité entre émotions est mesurée. L'ensemble des dimensions (relations) présentes peut varier selon l'ensemble des états émotionnels présents dans le corpus qui dépendent de la source et du contexte de sa collecte. Sur le plan expérimental, ces dimensions théoriques peuvent se traduire en variables représentant des descripteurs de haut niveau calculés à partir de descripteurs de bas niveau tels que les traits acoustiques dans le domaine de reconnaissance de forme. Malgré le grand potentiel des méthodes basées sur la similarité pour la tâche de la RAE, celles-ci ne sont que peu explorées. La plupart des méthodes basées sur la similarité

dans la RAE sont surtout implémentés au niveau des classificateurs (k-plus proches voisins en l'occurrence), et peu au niveau des traits.

4.2.2 Traits basés sur la similarité

Dans une approche basée sur la similarité, la représentation d'un objet est basée sur la description de la relation de proximité qui le lie avec d'autres objets plutôt que sur une description basée sur les traits caractéristiques de l'objet. Par conséquent, un objet est décrit par des similarités par paires évaluées à travers les distances le séparant d'un ensemble d'objets appelé ensemble de représentants. Un vecteur de distance, $d(., p_i)$, représentant la dissimilarité avec un prototype p_i sera interprété comme un trait caractéristique. L'objectif par la suite est d'entrainer le classificateur à travers ces données relationnelles, qui peuvent capturer des caractéristiques différentes du problème, en comparaison avec l'approche typique de classification basée sur les vecteurs de traits (Pekalska *et al.* 2006).

Bâtir les classificateurs dans un espace de similarité est justifié d'une part par le fait que la similarité est importante entre objets similaires, c.-à-d. appartenant à la même classe, et faible pour les objets de classes différentes, ce qui offre une possibilité de discrimination (Pekalska *et al.* 2006). Ainsi, le fait que deux observations, $\mathbf{X}_i$ et $\mathbf{X}_j$, présentent à la fois le même degré de similarité envers un ensemble d'observations et le même degré de dissimilarité envers d'autres observations, va renforcer l'hypothèse que les deux observations $\mathbf{X}_i$ et $\mathbf{X}_j$ appartiennent à la même classe d'observations (Bicego *et al.* 2004). D'autre part, les dimensions d'un espace de dissimilarité véhiculent de l'information ayant le même niveau d'importance contrairement à une représentation générale basée sur les traits caractéristiques possédant en général différents attributs (i.e., différentes grandeurs physiques) et plages, comme par exemple le poids ou la longueur (Pekalska *et al.* 2006). Dans une expérience faisant intervenir des données synthétiques (Pekalska *et al.* 2002), un problème complexe dans un espace 2D discriminable à travers un classificateur quadratique, se transforme en un problème linéairement séparable dans un espace de similarité (Bicego *et al.* 2004). La similarité est également intéressante lorsqu'il n'y a pas de moyen simple de définir les

caractéristiques, lorsque les traits comprennent à la fois des mesures catégoriques et continues ou lorsqu'ils ne sont dotés que d'un faible pouvoir de discrimination (Pekalska *et al.* 2001).

4.2.3 Méthodes de classification

Le concept de similarité peut être aussi appliqué au niveau de la méthode de classification des données après avoir été appliqué pour leurs représentations. Une méthode couramment utilisée pour classifier les instances par mesure de similarité est la règle des *k*-plus proches voisins (KNN). Indépendamment de sa simplicité, la règle KNN donne souvent des résultats compétitifs comparativement à plusieurs autres classificateurs plus complexes. Cependant, la méthode de recherche dans KNN souffre de trois inconvénients : i) capacité de stockage et complexité de calcul relativement élevées exigées durant la phase de test, ii) sa sensibilité aux échantillons bruités qui conduit à une perte potentielle de précision iii) sa sensibilité au problème de distribution biaisée des classes (même si c'est un problème qui touche également beaucoup d'autres classificateurs puissants). Une multitude de recherches ont été consacrées pour surmonter ces limitations parmi celles-ci l'utilisation de techniques d'optimisation de prototypes (KD-trees or ball-trees (Friedman *et al.* 1977)), de métriques plus puissantes telles que la distance de Mahalanobis (Goldberger *et al.* 2004) ou de fonctions de pondération de contribution des voisins (Pao *et al.* 2007). L'algorithme KNN a été la première méthode à avoir été expérimentée pour la classification des données représentées par les dissimilarités dans (Jain et Zongker, 1997). La règle KNN donne souvent des résultats compétitifs à d'autres classificateurs plus complexes quand les données de référence sont représentatives et une bonne mesure de dissimilarité est utilisée (Pekalska *et al.* 2006). Quand ces conditions ne sont pas vérifiées les performances se détériorent et l'utilisation de classificateurs plus puissants est nécessaire. Dans cette thèse nous proposons d'utiliser une autre méthode de classification qui est basée sur la mesure de similarité à l'instar de KNN mais qui permet de pallier les problèmes de représentativité des instances et la complexité du calcul et de stockage en phase de test, appelée *classificateur du plus proche centroïde* (Fukunaga, 1990).

Le classificateur du plus proche centroïde

Le classificateur du plus proche centroïde ou moyenne (*Nearest Mean Classifier*, NMC) est un algorithme de classification simple et rapide appliqué avec succès à de nombreux problèmes de classification, y compris sur des données de grande dimension, montrant à la fois une bonne et solide performance (Shin et Kim, 2009). Cet algorithme utilise des données d'apprentissage pour calculer la moyenne μ de chaque classe, $\mu_k = \frac{1}{n_k}\sum_{i \in I_k} x_i$, qui sera utilisée comme prototype représentant la classe lors de la phase de test. Une donnée de test *z* est assignée à la classe qui minimise la distance, *d()*, séparant son prototype et la donnée de test (équation (4.1)).

$$emotion\_cible = \text{argmin}_k d(\mu_k, z) \tag{4.1}$$

Le prototype représentant la classe sera calculé dans l'espace des vecteurs de similarité. Cette méthode se prête bien au modèle théorique des émotions catégoriques représentées dans un espace bi- ou tridimensionnel. Chaque émotion dans cet espace est représentée par un centroïde. La classification d'une donnée de test dépendra de sa proximité à un des centroïde de cet espace. Cette proximité peut être évaluée en mesurant la distance (euclidienne, voir (4.2)) entre vecteurs. La mesure de l'angle (la similarité cosinus, voir (4.3)) peut constituer une meilleure alternative si le bruit présent dans les données affecte plus le module que la direction des vecteurs de traits. Une autre solution est d'utiliser une mesure qualitative telle que les rangs des scores à la place d'une mesure quantitative.

Métrique euclidienne

$$\mathrm{d}(\mathbf{SV}_1, \mathbf{SV}_2) = \sqrt{\left|\mathbf{SV}_1 - \mathbf{SV}_2\right|^2} \tag{4.2}$$

Métrique cosinus

$$\mathrm{d}(\mathbf{SV}_1, \mathbf{SV}_2) = 1 - \frac{\langle \mathbf{SV}_1, \mathbf{SV}_2 \rangle}{\|\mathbf{SV}_1\| \|\mathbf{SV}_2\|} \tag{4.3}$$

où $\langle \mathbf{SV}_1, \mathbf{SV}_2 \rangle$ représente le produit scalaire des vecteurs $\mathbf{SV}_1$ et $\mathbf{SV}_2$.

Notons que bâtir un système autour d'une méthode (algorithme) de classification basée sur la mesure de proximité ne signifie pas l'exclusion des autres approches (ex. : génératives ou discriminatives) dans la modélisation globale de l'architecture du système. En effet, l'algorithme basé sur la similarité peut être implanté par un module placé au dernier étage d'un système d'une architecture multi-étages.

4.3 Corpus de parole émotionnelle FAU AIBO Emotion

Étant donné que les méthodes proposées sont basées sur une approche statistique guidée par les données, il est important de choisir un corpus de parole émotionnelle qui réunit les conditions permettant d'entrainer et d'évaluer des systèmes de RAE opérant dans des conditions de vie réelle. Le corpus FAU AIBO Emotion (Steidl, 2009), est l'un des rares corpus de parole émotionnelle réunissant ces critères. En effet, FAU AIBO Emotion est un corpus multi-classes, d'émotion spontanée, avec des catégories de classes d'émotion non complètes (*full-blown*). Les énoncés sont de courtes durées véhiculant des émotions subtiles et caractérisées par une large variation dans l'expression de la même classe d'émotion. Enfin la distribution des classes est sévèrement biaisée en ce sens que certaines classes d'émotion sont surreprésentées par rapport à d'autres. Un tel corpus nous permettra une évaluation plus juste de systèmes conçus pour des scénarios plus réalistes.

Le corpus FAU AIBO Emotion a été présenté et rendu public en 2009 à l'occasion de la compétition *INTERSPEECH 2009 Emotion Challenge* afin de fournir à la communauté une base de données de taille moyenne contenant de la parole émotionnelle plus spontanée et moins prototypique pour tenir compte des scénarios plus réalistes. C'est un corpus d'enregistrements en langue allemande de 51 enfants âgés de 10 à 13 ans interagissant avec AIBO, un chien robot de compagnie. Les expériences ont été menées selon le paradigme WOZ où AIBO était entièrement contrôlé à distance par l'expérimentateur à l'insu des enfants. Les enfants étaient instruits de parler à AIBO comme s'ils parlaient à un vrai chien. Le corpus, constitué de 8,9 heures de parole et 48 401 mots, a été recueilli dans deux écoles

différentes *OHM* et *MONT*. Le signal de la parole a été enregistré à un taux d'échantillonnage de 48 kHz et une quantification de 16 bits et sous-échantillonné par la suite à 16 kHz.

Les enregistrements ont été segmentés automatiquement en 13 642 tours de parole en utilisant un seuil de pause d'une seconde. En moyenne, un tour de parole est de longueur de 3,5 mots. Notons que l'état émotionnel de l'enfant peut changer à l'intérieur d'un tour de parole. Ni le tour de parole ni le mot ne représente l'unité optimale d'analyse émotionnelle, mais une unité intermédiaire syntaxiquement et sémantiquement significative. Les tours de parole sont donc manuellement segmentés en *segments* (*chunk* en anglais) sur la base de critères syntaxiques-prosodique (des exemples de tours de paroles segmentés en segments sont données en ANNEXE II). La taille moyenne d'un segment est de 2,66 en nombre de mots et de 1,7 en secondes. Le nombre total de segments est de 18 216 dont 9 959 proviennent de l'école OHM et 8 257 de l'école MONT. Outre le contenu verbal du corpus, les sons non-verbaux tels que les bruits respiratoires, le rire, la toux, le bruit et les hésitations vocales ou nasales ont été également manuellement transcrits.

Les données sont étiquetées par cinq annotateurs en dix classes d'émotions : en colère (*angry*), susceptible/irrité (*touchy*/*irritated*), joyeux (*joyful*), surpris (*surprised*), ennuyé (*bored*), impuissant (*helpless*), mamanais (*motherese*), réprimandant (*reprimanding*), emphatique (*emphatic*), autre (*other*) en plus de la classe neutre (*neutral*). Une description plus détaillée de ces catégories d'émotions est donnée dans le Tableau-A II-1 de l'ANNEXE II. Le *mot* a été choisi comme unité d'annotation afin de permettre de capturer les changements rapides d'émotion à l'intérieur d'un même tour de parole. Seuls les mots ayant obtenu un vote majoritaire par les annotateurs ont été sélectionnés pour les expérimentations. Les différents taux d'accord entre les cinq annotateurs pour chacune des classes sont donnés au Tableau-A II-2 de l'ANNEXE II. Tel que l'on pourrait l'observer (voir Tableau-A II-2 de l'ANNEXE II), les fréquences de certaines classes d'émotion dans le corpus sont très rares et ne peuvent suffire pour un apprentissage machine robuste des modèles d'émotion. C'est pourquoi les classes moins fréquentes sont regroupées en famille de classes d'émotions plus large pour former au total cinq classes d'émotion : neutre (**N**); emphatique (**E**); colère

(*Anger*, **A**) regroupant les états en colère, susceptible et réprimandant; la classe positive (**P**) regroupant mamanais et joyeux et enfin la classe reste (**R**) qui regroupe toutes les autres classes d'émotion restantes. Le Tableau 4.1 présente la matrice de confusion de l'opération d'étiquetage entre annotateurs pour les nouvelles classes d'émotion que nous avons compilée à partir des valeurs de la matrice de confusion des classes élémentaires du Tableau-A II-3 de l'ANNEXE II.

Tableau 4.1 Matrice de confusion d'étiquetage (en %) des cinq annotateurs du corpus FAU AIBO Emotion après regroupement des petites catégories d'émotions (compilé à partir des valeurs du Tableau-A II-3)

	A	**E**	**N**	**P**
A	64,3	13,9	20,7	1,1
E	13,9	54,1	30,2	1,7
N	4,1	14,0	78,6	3,2
P	2,2	5,1	30,7	62,0

Les pourcentages de chacune des cinq classes pour chacune des partitions, entrainement et test, sont représentés dans la Figure 4.1. Cette figure met clairement en évidence la distribution fortement déséquilibrée des cinq classes d'émotion en faveur de la classe **N**. Plus d'information sur le robot AIBO, le scénario d'induction des émotions, la segmentation et la transcription enregistrements ainsi que leurs annotations est donnée dans l'ANNEXE II.

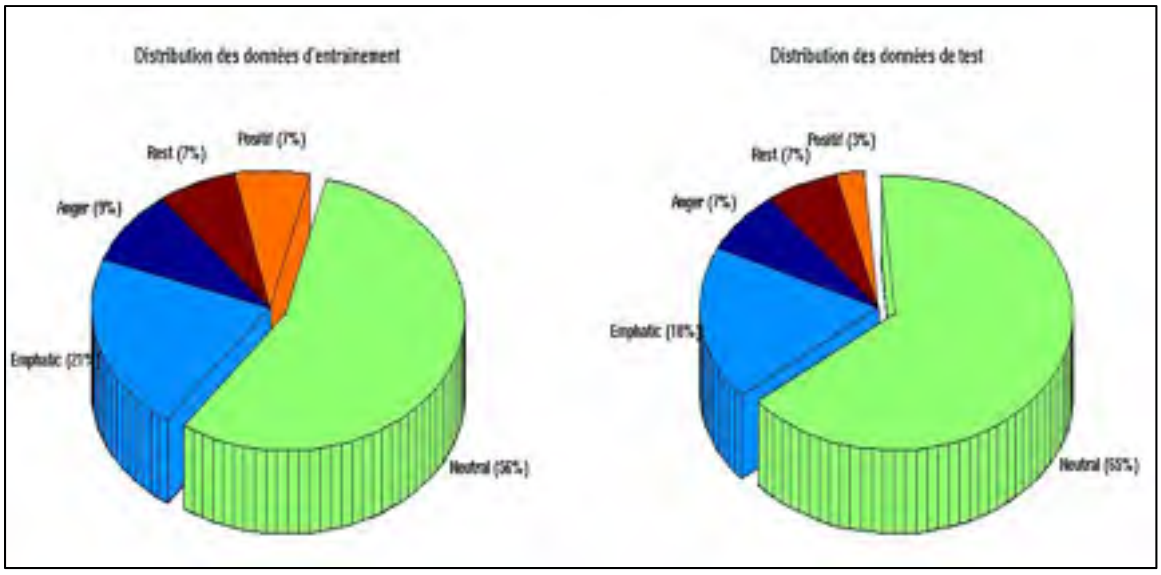

Figure 4.1 Répartition des classes du corpus FAU AIBO Emotion pour les partitions d'entrainement et de test

4.4 Protocole d'expérimentation

Les paramètres des modèles sont ajustés sur la base des données d'entraînement en utilisant le protocole de la validation croisée à neuf plis. Afin d'assurer la stricte indépendance du système de RAE vis-à-vis du locuteur, chacune des neuf partitions contient un ensemble disjoint de locuteurs. Étant donné que les classes du corpus FAU AIBO Emotion sont fortement déséquilibrées, les résultats seront optimisés selon le critère de la moyenne non pondérée du rappel (*Unweighted Average Recall*, UAR) en premier lieu suivi du critère de la moyenne pondérée du rappel (*Weighted Average Recall*, WAR), appelé encore exactitude (*accuracy*). L'évaluation des performances se fera sur les données de test. Les deux critères sont calculés comme suit.

Soit $MC = (mc_{ij}),\ i,j = 1,\dots,C$, la matrice de confusion d'un classificateur d'un problème à C classes d'émotion où mc_{ij} est l'élément de la i-ième ligne et de la j-ième colonne de la matrice et représente le nombre de données de la classe i prédites par le classificateur comme appartement à la classe j.

$$UAR = \frac{1}{C}\sum_{i=1}^{c}\left(\frac{mc_{ii}}{\sum_{j=1}^{C} mc_{ij}}\right) \times 100\% \quad (4.4)$$

$$WAR = \frac{\sum_{i=1}^{C} mc_{ii}}{\sum_{i}\sum_{j}^{C} mc_{ij}} \times 100\% \quad (4.5)$$

Notons qu'un classificateur qui prédit toutes les données de test comme étant de la même classe que la classe majoritaire, à savoir la classe **N** (neutre), obtiendra 65 % en terme de WAR, mais seulement 20 % en UAR. Ce sont également les mêmes mesures qui ont été adoptées pour la compétition internationale d'émotion INTERSPEECH 2009. Par conséquent, nos résultats peuvent être comparés avec l'état de l'art.

Après avoir décrit notre méthodologie, nous allons chercher dans le reste de ce chapitre le type d'espace de traits de haut qui permettra de mieux répondre aux contraintes des corpus de parole émotionnelle.

4.5 Choix des descripteurs de haut niveau

Plusieurs types d'espace de descripteurs peuvent être candidats pour mesurer la similarité entre classes. Le choix le plus trivial est d'utiliser les traits acoustiques extraits directement du signal de parole. Ces traits acoustiques de type long terme sont calculés par exemple au moyen de paramètres statistiques sur les descripteurs prosodiques, spectraux et qualité de la voix. Toutefois, tel que nous l'avons déjà vu dans le chapitre précédent, ces paramètres acoustiques sont des mesures assez brutes et les classes d'émotion représentées dans un tel espace sont caractérisées par un important chevauchement et sont par conséquent difficilement discernables. C'est pourquoi nous proposons d'utiliser des descripteurs d'un niveau d'abstraction supérieur obtenus après traitement des traits acoustiques de bas niveau. Les techniques de traitement comprendront non seulement des opérations de normalisation mais aussi d'extraction de paramètres de plus haut niveau obtenus par exemple après une modélisation de ces traits acoustiques. Un tel espace plus raffiné permettra de mettre en

évidence et faire ressortir des relations de similarité entre classes qui se rapprocheront des relations renfermées dans le modèle théorique dimensionnel telle que la relation *plus* **agréable** ou **désagréable** que, *plus* **active** ou **passive** que, ou *plus* **dominant** que.

Dans un premier temps nous allons inspecter la pertinence des *supervecteurs* comme type de descripteurs de haut niveau obtenus après modélisation de traits de bas niveau (principalement les traits cepstraux MFCC) utilisés dans le domaine du traitement de la parole. Nous examinerons également deux types de vecteurs de traits dérivés des supervecteurs obtenus après projection dans de ces traits dans des espaces plus réduits, les *eigenvoices* et *fishervoices*.

4.5.1 Supervecteurs et dérivées

Les supervecteurs sont des traits qui représentent les fonctions exhaustives des observations (*sufficient statistics* en anglais) de chacune des composantes (c'est la moyenne qui est utilisée en général) d'un modèle de mélange Gaussien, (Gaussian Mixture Model, GMM). Si m dénote le nombre total de composantes du mélange Gaussien et D la dimension du vecteur acoustique, nous pouvons alors concaténer les vecteurs moyennes, $\boldsymbol{\mu}$, de chaque gaussienne pour former un vecteur $\mathbf{sv}$, de haute dimension (égale à $m \times D$), appelé supervecteur.

$$\mathbf{sv} = [\boldsymbol{\mu}_1^T \quad \boldsymbol{\mu}_2^T \quad \ldots \quad \boldsymbol{\mu}_m^T]^T \tag{4.6}$$

Ainsi, un segment de parole émotionnel est représenté par un seul point dans un espace de haute dimension. Les descripteurs MFCC à l'échelle de trames sont extraits comme traits de bas niveau à partir des énoncés de tailles variables pour former des supervecteurs de taille fixe.

4.5.1.1 Modélisation par mélange de gaussiennes

La modélisation par mélange de gaussiennes permet d'approximer une fonction de densité de probabilité de complexité quelconque en choisissant un nombre suffisant de composantes gaussiennes. Le choix des modèles GMM dans le domaine du traitement du signal de parole en général et celui de la RAE en particulier est motivé par la notion intuitive que chaque densité de composante d'un mélange de gaussiennes permet de modéliser une ou un certain nombre de classes acoustiques telles les voyelles ou les fricatives. Ces classes acoustiques reflètent un aspect général de la configuration du système de la production de la parole (poumons, cordes vocales et conduit vocal) sous l'effet de l'émotion ressentie. Étant donné que les données d'apprentissage et de test ne sont pas « phonétiquement » annotées, les classes acoustiques sont considérées comme cachées dans le sens où la classe des données observées est inconnue. Par conséquent, la densité des vecteurs de traits générée de ces classes acoustiques cachées prête bien à un mélange de gaussiennes.

Une densité de probabilité d'un GMM est une somme pondérée de *M* composantes de densités normales et s'écrit sous la forme mathématique suivante :

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i b_i(\mathbf{x}) \tag{4.7}$$

où **x** est un vecteur de données de dimension *d*, λ est le modèle GMM, w_i représente la proportion de chaque gaussienne du mélange avec les contraintes $\sum_{i=1}^{M} w_i = 1$ et $w_i \geq 0$; et $b_i(\mathbf{x}), i = 1, \cdots, M$, sont les densités normales multidimensionnelles données par :

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\mathbf{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \sum_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right) \tag{4.8}$$

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

modèle GMM λ est défini par $\lambda = \{w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}$ où $w_m, \boldsymbol{\mu}_m$ et $\boldsymbol{\Sigma}_m$ représentent respectivement la pondération, le vecteur de la moyenne et la matrice de covariance de chacune des *M* composantes gaussiennes constituant le mélange de gaussiennes λ.

L'objectif de la phase d'apprentissage d'un modèle GMM est d'estimer les paramètres λ qui modélisent le mieux la distribution des données d'apprentissage. Il existe plusieurs techniques pour l'estimation des paramètres d'un GMM et la méthode la plus populaire et bien établie est la méthode d'estimation du maximum de vraisemblance (ML, *Maximum Likelihood estimation*). La petite taille des segments de paroles des corpus émotionnels ne permet pas une estimation fiable des fonctions exhaustives de chacune des composantes gaussiennes modélisant le segment de parole. Par conséquent, un modèle du monde appelé UBM (*Universal Background Model*) est créé comme modèle initial à partir des données d'apprentissage de toutes les classes d'émotion. Le modèle UBM est entrainé en utilisant la méthode du maximum de vraisemblance via l'algorithme *espérance-maximisation* (EM, ou *Expectation-Maximization* en anglais). Le modèle UBM est par la suite adapté pour chaque segment en utilisant une quantité limitée de données extraites de ces segments. Dans cette étude nous allons expérimenter deux méthodes d'adaptation ; MAP (*Maximum A Posteriori*) et MLLR (*Maximum Likelihood Linear Regression*) (Leggetter et Woodland, 1995). Nous examinerons succinctement dans ce qui suit chacune de ces méthodes d'estimation des paramètres. Plus de détails peuvent être trouvés dans l'ANNEXE III.

4.5.1.2 Méthode d'estimation du maximum de vraisemblance

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

$$p(\mathbf{X} \mid \lambda) = \prod_{n=1}^{N} p(\mathbf{x}_n \mid \lambda) \tag{4.9}$$

Il n'existe pas de méthode analytique connue pour résoudre le problème de maximisation de cette fonction non linéaire du paramètre λ. Cependant, nous pouvons choisir $\lambda = \{w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}$ telle que la vraisemblance $p(\mathbf{X}|\lambda)$ est un maximum local en utilisant une méthode itérative telle que la méthode EM. La méthode EM est décrite dans l'ANNEXE III.

4.5.1.3 Adaptation MAP

L'adaptation MAP (Reynolds *et al.* 2000 ; Huang *et al.* 2001) permet d'ajuster les paramètres du modèle préentrainé (UBM) de manière à ce que de nouvelles données modifient les paramètres du modèle, guidé par la connaissance a priori. En utilisant les données observées **X**, l'estimation MAP peut être formulée comme suit :

$$\hat{\lambda} = \arg\max_{\lambda}\left[p(\lambda|\mathbf{X})\right] = \arg\max_{\lambda}\left[p(\mathbf{X}|\lambda)p(\lambda)\right] \tag{4.10}$$

En absence de l'information a priori, c'est-à-dire aucune connaissance sur le modèle λ et si $p(\lambda)$ possède une distribution uniforme alors l'estimation MAP devient identique à l'estimation ML. La MAP est une adaptation dépendante des données, par conséquent les paramètres du mélange de gaussiennes de l'UBM sont adaptés avec différentes grandeurs. Si une gaussienne est bien représentée par les nouvelles données à utiliser pour l'adaptation, alors ces données auront un poids plus important dans l'estimation des nouveaux paramètres. Dans le cas contraire, c'est-à-dire quand une gaussienne est mal représentée par les nouvelles données, les nouveaux paramètres estimés seront plus influencés par les anciennes valeurs, qui représentent les paramètres du modèle UBM qui eux sont mieux entraînés. Les formules d'adaptation des paramètres sont données en ANNEXE III.

4.5.1.4 Adaptation MLLR

MLLR est une méthode basée sur des fonctions de transformation de régression linéaire pour faire le transfert des paramètres des gaussiennes des anciens modèles vers les nouveaux

modèles adaptés. La matrice de transformation est calculée de manière à maximiser la vraisemblance des données d'adaptation par rapport au modèle via l'algorithme EM. Un k^{ième} vecteur moyenne $\boldsymbol{\mu}_k$ est transformé selon l'équation

$$\hat{\boldsymbol{\mu}}_k = \mathbf{A}_c \boldsymbol{\mu}_k + \mathbf{b}_c \tag{4.11}$$

où $\mathbf{A}_c$ représente la matrice de régression et $\mathbf{b}_c$ un vecteur de biais additif. Contrairement à l'adaptation MAP où les paramètres (les moyennes) de chaque gaussienne sont estimés individuellement, dans MLLR ce sont les paramètres des gaussiennes d'une grande classe de régression qui subissent la même transformation. Cette grande classe peut aller jusqu'à contenir toutes les composantes gaussiennes du GMM. C'est pourquoi le processus d'adaptation MAP exige une plus grande quantité de données d'adaptation que MLLR. L'adaptation MLLR est recommandée lorsque la quantité de données d'adaptation est très petite (une adaptation rapide et grossière est obtenue en résultat). Par contre la méthode MAP est plus performante que MLLR dans le cas où les données d'adaptation sont en grande quantité (car le processus d'adaptation MLLR atteint un point de saturation et les performances ne peuvent être améliorées après un certain point même en présence de quantités supplémentaires de données d'adaptation (Goronzy et Kompe, 1999; Huang *et al.* 2001; Sam, 2011).

4.5.1.5 Combinaison de MLLR et MAP

Considérant les avantages et inconvénients des deux méthodes MAP et MLLR, il serait bénéfique de combiner les deux méthodes d'adaptation en incorporant le principe d'adaptation MAP dans MLLR. Cette combinaison permet d'avoir à la fois une adaptation rapide avec les fonctions de transformation MLLR et d'avoir une modification directe des paramètres du modèle qui converge avec l'estimation ML dans le cas d'une augmentation des données d'apprentissage. L'estimation du vecteur de la moyenne selon la combinaison des deux méthodes est réalisée selon l'équation suivante (Blouet *et al.* 2004):

$$\hat{\boldsymbol{\mu}}_k = \alpha\,(\mathbf{A}_c\boldsymbol{\mu}_k + \mathbf{b}_c) + (1-\alpha)\boldsymbol{\mu}_k \tag{4.12}$$

où α représente le facteur de contrôle de l'adaptation. Étant donnée la grande dimension des supervecteurs, nous proposons également de procéder à une réduction de la taille de ces vecteurs de traits et à une décorrélation entre ces traits en utilisant soit (i) un apprentissage non supervisé à travers une analyse en composantes principales (ACP) ou (ii) un apprentissage supervisé afin de maximiser les distances interclasse dans le nouvel espace de projection à travers la méthode d'analyse par discrimination linéaire (LDA).

Voix propre

L'application d'ACP sur les supervecteurs donne en sortie des vecteurs propres appelés voix propres, *eigenvoices* en anglais, (par analogie à eigenfaces utilisés pour la reconnaissance faciale, (Kirby et Sirovich, 1990)). ACP, est une technique non supervisée qui a pour objectif de trouver un sous-espace de faible dimension où le maximum de la variance est préservé. La plupart de la variance dans les données est conservée dans les premiers vecteurs propres, c'est pourquoi seuls les *k* premiers sont conservés. L'approche eigenvoice a été appliquée avec succès dans le domaine de la reconnaissance de la parole à petit vocabulaire (Kuhn e al. 2000), la reconnaissance du locuteur (Thyes *et al.* 2000) et du regroupement en locuteurs (Castaldo *et al.* 2008).

Fishervoice

Un vecteur de trait de type fishervoice représente un vecteur de faible dimension obtenu après projection des supervecteurs dans un espace construit à partir d'une analyse discriminante linéaire (LDA, *linear discriminant analysis*). LDA est une technique supervisée qui a pour objectif de trouver un sous espace de faible dimension tout en maximisant la séparabilité entre classes. Le terme fishervoice a été également utilisé par analogie au terme *fisherface* apparu dans le domaine de la reconnaissance faciale (Belhumeur *et al.* 1997). Dans (Chu *et al.* 2009), les performances obtenues avec fishervoice dans le

regroupement en locuteurs dépassaient significativement celles qui ont été obtenues avec eigenvoices.

Deux types de vecteurs fishervoice seront expérimentés. Dans le premier, les supervecteurs sont directement projetés dans l'espace LDA. Le second est obtenu en procédant au préalable à une réduction de la dimension des supervecteurs au moyen de ACP suivie par la suite d'une projection dans l'espace fishervoice. Le recours à une réduction de dimension de *n-C* avant LDA est justifié par le problème de singularité de la matrice de dispersion interclasse, $\boldsymbol{S}_W$, qui possède un rang égal à *n-C*.

4.5.1.6 Expérimentations

Dans ce qui suit nous présentons les résultats exploratoires sur les performances des supervecteurs, voix propres et fishervoices. Différentes configurations ont été expérimentées pour extraire le vecteur optimal. Ces configurations dépendent de la méthode d'adaptation (MAP vs combinaison MLLR et MAP), le nombre de composantes gaussiennes dans les GMM (2, 4, 8, 16, 32, 64, 128, 256, 512, 1024) et par conséquent la dimension des supervecteurs (égale à #Gauss. X #mfcc) ainsi que les techniques de normalisation utilisées. Trois techniques de normalisations ont été expérimentées ; la normalisation de la moyenne par soustraction du vecteur moyenne UBM, z-score (normalisation de la moyenne et de la variance) et enfin une pondération des vecteurs eigenvoice par les valeurs propres des composantes principales utilisées afin de mettre plus d'accent sur l'importance des directions ACP avec plus de variabilité (Shum *et al.* 2011).

Méthode basée sur la similarité

Les performances des supervecteurs utilisés comme traits de haut niveau ont été évaluées sur le corpus FAU AIBO Emotion. Pour mesurer la similarité, nous avons représenté chaque classe d'émotion par un supervecteur entrainé avec toutes les données de cette classe. Le supervecteur d'un énoncé de test est obtenu par l'adaptation du vecteur moyenne UBM. Trois

métriques ont été expérimentées pour mesurer les distances séparant un supervecteur de test avec l'ensemble des vecteurs de référence : euclidienne, cosinus et une métrique dérivée de la distance Kullback-Leibler introduite dans (Campbell *et al.* 2006) et définie comme suit :

$$\mathrm{d}(\mathbf{SV}_1, \mathbf{SV}_2) = \sum_{i=1}^{m} w_i \left(\boldsymbol{\mu}_i^1 - \boldsymbol{\mu}_i^2\right)^T \boldsymbol{\Sigma}_i^{-1} \left(\boldsymbol{\mu}_i^1 - \boldsymbol{\mu}_i^2\right) \tag{4.13}$$

où w_i et $\boldsymbol{\Sigma}_i^{-1}$ représentent le poids et la matrice de covariance de la i-ième mixture de l'UBM et $\boldsymbol{\mu}_i^1$ représente le vecteur moyenne de la gaussienne i du GMM de l'énoncé représenté par le supervecteur $\mathbf{SV}_1$.

Le Tableau 4.2 présentent les performances de classification et les configurations associées obtenus pour les deux méthodes d'adaptation lorsque évaluées sur les données de test du corpus FAU AIBO Emotion.

Tableau 4.2 Performances des supervecteurs en fonction de la méthode d'adaptation utilisée pour entrainer l'extracteur. Ces résultats sont obtenus sur les données de test du corpus d'émotion FAU AIBO. Le nombre (*D*) représente la dimension optimisée du supervecteur

	UAR	**WAR**	**# Gauss. (*D*)**	**Métrique**
MAP	36,70 %	25,95 %	2 (78)	Cosinus
MLLR_MAP	39,91 %	36,32 %	8 (312)	Euclidienne

Trois observations peuvent être tirées de ces résultats. Nous constatons que i) la méthode MLLR combinée avec MAP donne de meilleurs résultats en comparaison avec une adaptation basée sur la MAP uniquement; ii) les meilleures performances des supervecteurs sont atteintes avec 2 gaussiennes seulement pour la MAP comparées à 8 gaussiennes pour la méthode MLLR_MAP; iii) le nombre de gaussiennes utilisées pour extraire les supervecteurs est considérablement faible comparé au domaine de reconnaissance du locuteur par exemple où de meilleures performances peuvent être atteintes avec un nombre de gaussiennes beaucoup plus élevé (2048 gaussiennes est une valeur de référence). En fait chacune de ces trois constatations présente à elle seule un argument suffisant soutenant que la taille des

segments de parole émotionnelle est assez courte pour alimenter en quantité suffisante de données l'opération d'adaptation des modèles et assurer par conséquent une extraction robuste des supervecteurs.

Les supervecteurs adaptés à travers la combinaison MLLR et MAP ont été utilisés pour calculer les traits de types eigenvoice et fishervoice. Les fishervoices calculés après application d'une projection ACP ont donné de meilleures performances en comparaison aux fishervoices calculés directement des supervecteurs. Les performances de reconnaissance obtenues avec les vecteurs eigenvoices (UAR=40,15 % et WAR=35,87 %) sont équivalentes à celles obtenues avec les supervecteurs alors que celles obtenues avec fishervoices (UAR=40,50 % et WAR=43,09 %) sont légèrement supérieures en terme de UAR à celles des supervecteurs alors qu'il a eu une nette amélioration par rapport au critère WAR. Dans la prochaine section, nous allons continuer à évaluer les performances des supervecteurs cette fois-ci avec un modèle probabiliste génératif.

4.5.1.7 Analyse discriminante linéaire probabiliste (PLDA)

Les limites de performances obtenues avec les supervecteurs et ses dérivées ne sont pas liées à l'approche de classification qui est basée sur la similarité mais à la capacité d'extraire de robustes supervecteurs. Afin de confirmer cette hypothèse, nous avons modélisé les supervecteurs par une méthode de classification récente et puissante appelée l'analyse discriminante linéaire probabiliste (*Probabilistic Linear Discriminant Analysis*, PLDA). Le modèle PLDA représente l'état de l'art dans le domaine de la reconnaissance du locuteur (Kenny, 2010) et a été initialement appliqué pour la reconnaissance faciale (Prince et Elder, 2007), comme il a été également expérimenté en reconnaissance des langages (Brummer *et al.* 2009). PLDA est un modèle génératif qui permet de modéliser séparément la variabilité utile des variabilités nuisibles dans les vecteurs afin d'éliminer la variabilité (nuisible) et de se concentrer sur la variabilité interclasse utile pour la discrimination des différentes classes d'émotion.

Soit $\mathbf{x}_{ij}$ le vecteur de traits caractérisant le *j*-ième segment de parole des données d'apprentissage de la *i*-ième classe d'émotion. La modélisation des données est réalisée selon le modèle suivant:

$$\mathbf{x}_{ij} = \boldsymbol{\mu} + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij} + \varepsilon_{ij} \tag{4.14}$$

La partie $\boldsymbol{\mu} + \mathbf{F}\mathbf{h}_i$ de la formule modélise la variation qui existe entre les classes d'émotion. Elle dépend uniquement de la classe de l'émotion *i* indépendamment des différences qui existent dans la manière d'exprimer cette émotion. La deuxième partie, $\mathbf{G}\mathbf{w}_{ij} + \varepsilon_{ij}$ modélise le bruit à l'intérieure de la même classe d'émotion. Par bruit, on entend toutes variabilités qui existent à l'intérieure de la classe quelle qu'elles soient leurs origines (diversité entre individus dans l'expression de l'émotion, variabilité causée par le langage ou par les appareils d'acquisition et de transmission du signal). Le terme $\boldsymbol{\mu}$ représente le vecteur de la moyenne du modèle UBM. La matrice $\mathbf{F}$ contient une base pour le sous-espace interclasses d'émotion et le terme $\mathbf{h}_i$ représente la position dans ce sous-espace. La matrice $\mathbf{G}$ contient une base pour le sous-espace intraclasse. Toute variabilité restante à l'intérieure d'une classe d'émotion qui n'est pas expliquée par $\mathbf{G}\mathbf{w}_{ij}$ est modélisée par le bruit résiduel ε_{ij}. Dans le jargon d'analyse de facteur, les matrices $\mathbf{F}$ et $\mathbf{G}$ contiennent les facteurs et les variables latentes, $\mathbf{h}_i$ et $\mathbf{w}_{ij}$ représentent les facteurs de chargement. Le facteur $\mathbf{h}_i$ représente ce que nous pourrions appeler l'identité de l'émotion *i*, où toutes les instances de la classe d'émotion sont générées par le même $\mathbf{h}_i$. Les variables $\mathbf{h}_i$ et $\mathbf{w}_{ij}$ sont des variables gaussiennes centrées réduites, c.-à-d., $\mathbf{h}_i \propto N(0,\mathbf{I})$ et $\mathbf{w}_{ij} \propto N(0,\mathbf{I})$ et la probabilité conditionnelle est donnée comme suit:

$$\Pr(\mathbf{x}_{ij} \,|\, \mathbf{h}_i, \mathbf{w}_{ij}, \theta) = N(\mu + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij}, \Sigma) \tag{4.15}$$

La phase d'entrainement de ce modèle consiste à apprendre les paramètres $\theta = \{\boldsymbol{\mu}, \mathbf{F}, \mathbf{G}, \Sigma\}$ à partir des données d'apprentissage $\mathbf{x}_{ij}$. Étant donné que les variables latentes $\mathbf{h}_i$ et $\mathbf{w}_{ij}$ ne sont pas connues, il n'est pas possible de calculer directement les valeurs de ces paramètres. L'algorithme EM est alors utilisé pour estimer d'une manière itérative les deux ensembles de paramètres. Dans la phase estimation, la distribution postérieure complète sur les variables latentes $\mathbf{h}_i$ et $\mathbf{w}_{ij}$ est calculée pour des valeurs fixes des paramètres. Dans la phase maximisation, l'estimation des valeurs des paramètres est optimisée. Le détail des formules est identique à celui utilisé dans (Prince et Elder, 2007).

Résultats avec PLDA

Pour les expériences, des supervecteurs avec différents nombres de composantes gaussiennes ont été expérimentées. Nous avons également normalisés les variances des supervecteurs avant la modélisation PLDA. Les taux de reconnaissance obtenus sont 39,71 % pour l'UAR et 46,75 % pour le WAR. Une décorrélation des traits avec ACP avant modélisation PLDA n'a pas permis d'améliorer les performances. Nous observons que les résultats UAR obtenus avec PLDA sont légèrement inférieurs aux résultats obtenus avec la méthode basée sur la similarité (39,91 %, confirmant ainsi d'une part notre hypothèse sur le problème de la robustesse des supervecteurs extraits et d'autre part la pertinence de l'approche basée sur la similarité pour la RAE.

4.5.2 Scores de vraisemblance comme traits de haut niveau

4.5.2.1 Motivation

Les supervecteurs ou ses dérivées peuvent constituer de puissants traits de haut niveau quand les données d'adaptation sont en quantité suffisante, une condition en général difficilement satisfaite quand il s'agit de corpus d'émotion spontanée. Avec des segments de parole émotionnelle de courtes durées (en phase de test), il est difficile d'estimer d'une manière

fiable les paramètres d'un modèle GMM même en utilisant des techniques d'adaptation. Par ailleurs, les données des classes émotionnelles sont en quantité suffisante pour l'apprentissage des modèles durant la phase d'entrainement. Par conséquent, il est possible de substituer les paramètres des modèles GMM utilisés en tant que points dans l'espace des supervecteurs durant les phases d'entrainement et de test, par une modélisation basée sur une distribution statistique, i.e., un modèle GMM, construit durant la phase d'apprentissage pour chaque classe d'émotion. Les scores de vraisemblance (ou logarithme de vraisemblance) obtenus par l'évaluation de ces modèles pour un segment de parole constitueront les nouveaux vecteurs de traits de haut niveau. De cette manière le problème engendré par la faiblesse de la taille des segments avec les supervecteurs sera résolu par l'évaluation d'un modèle déjà estimé. L'espace de vraisemblance engendré par exemple par un modèle génératif tel qu'un GMM ou HMM possèdera une capacité de discrimination potentielle. Cette hypothèse est appuyée par les résultats de classification des émotions, basée sur les modèles GMM et *Bayes* (le maximum de vraisemblance, ML) comme règle de décision.

Table 4-1 Comparaison de performances entre différents systèmes évalués avec les données de test du corpus FAU AIBO Emotion.

	UAR	**WAR**
Supervecteurs	39,91 %	36,32 %
Eigenvoices	40,15 %	35,87 %
Fishervoices	40,50 %	43,09 %
PLDA	39,71 %	46,75 %
GMM-*Bayes*	**41,65 %**	42,49 %

Les résultats du Table 4-1 montrent que les performances obtenues avec la règle du maximum de vraisemblance sur les scores des GMM (*Bayes' rule*) dépassent en terme UAR

tous les systèmes déjà décrits (systèmes basés sur les supervecteurs et ses dérivés ainsi que le système PLDA), ce qui consolide notre hypothèse de la pertinence de traits générés dans l'espace de probabilité de vraisemblance. Dans ce qui suit nous décrirons comment les scores de décision sont formulés en vecteurs et utilisés comme entrées pour un classificateur d'un second étage.

4.5.2.2 Scores de vraisemblance et mesure de similarité

Soit $\mathbf{X}$ un énoncé de parole émotionnelle représenté par une séquence de T trames, $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_T\}$. Le nouveau vecteur $\mathbf{L}(\mathbf{X})$ obtenu par mappage de $\mathbf{X}$ dans l'espace de vraisemblance, est calculé comme suit :

$$\mathbf{L}(\mathbf{X})=\begin{bmatrix}\frac{1}{T}\log P(\mathbf{X}|\lambda_1)\\ \vdots \\ \frac{1}{T}\log P(\mathbf{X}|\lambda_C)\end{bmatrix}, \tag{4.16}$$

où $\log P(\mathbf{X}|\lambda_i)$ est le logarithme de la probabilité de vraisemblance du vecteur de traits $\mathbf{X}$ étant donné le modèle GMM λ_i de la i-ième classe des C catégories d'émotion. La valeur $1/T$ représente le facteur de normalisation temporelle introduite pour prendre en considération la taille variable des énoncés de parole. En supposant l'indépendance des trames, $\log P(\mathbf{X}|\lambda_i)$ est calculé selon l'équation :

$$\log P(\mathbf{X}|\,\lambda_i)=\sum_{n=1}^{T}\log P(\mathbf{x}_n\,|\,\lambda_i) \tag{4.17}$$

Dans ce qui suit nous appellerons $\mathbf{L}(\mathbf{X})$ *vecteur de caractérisation de l'émotion* (VCE). Il est intéressant de noter que le vecteur de vraisemblance VCE est plus qu'un vecteur caractéristique de haut niveau, il constitue en fait une mesure de similarité. En effet, une valeur de vraisemblance $P(\mathbf{X}|\lambda_i)$ (ou $\log P(\mathbf{X}|\lambda_i)$) mesure le degré de similarité entre un énoncé $\mathbf{X}$ et une classe d'émotion représentée par un modèle GMM λ_i. Plus un énoncé est

similaire à une classe, plus la valeur de $P(\mathbf{X}|\lambda_i)$ sera élevée et plus il est dissimilaire à cette classe moins la valeur de vraisemblance est importante. À travers un vecteur VCE, il est donc possible de reconnaitre le degré de similarité ou de dissimilarité de $\mathbf{X}$ à l'ensemble des catégories d'émotion. Le même principe s'applique également pour deux vecteurs VCE quelconques $\mathbf{L}(\mathbf{X}_1)$ et $\mathbf{L}(\mathbf{X}_2)$ correspondants aux deux énoncés $\mathbf{X_1}$ et $\mathbf{X_2}$ respectivement. Si les deux vecteurs possèdent des valeurs élevées pour les mêmes modèles et de faibles valeurs pour les mêmes autres modèles il serait alors probable que les deux énoncés appartiennent à la même classe d'émotion.

En utilisant la règle de décision *Bayes* dans les résultats précédents, la décision d'affecter un énoncé à une classe a été basée uniquement sur la valeur de vraisemblance maximale de tous les modèles. Ainsi, seule une partie de l'information sur la similarité qui est utilisée dans le processus de prise de décision alors que l'information sur le rapport qui existe entre les différents scores ainsi que leurs rangs est restée inexploitée. Par conséquent, en utilisant le vecteur VCE comme vecteur de traits d'entrée à un autre classificateur, c'est tous les scores ainsi que les relations qui les relient qui seront pris en considération dans la modélisation. C'est pourquoi on s'attendra à ce que les performances des systèmes GMM-*Bayes* soient améliorées par les systèmes utilisant les VCE comme traits caractéristiques où les modèles GMM auront pour rôle de mesurer la similarité entre les énoncés et les classes d'émotion.

4.5.2.3 Vecteur de traits VCE et l'analyse des émotions

Il est intéressant de noter qu'au-delà du rôle joué en tant que traits discriminants utile pour la classification, les vecteurs VCE peuvent aussi servir dans le domaine de la psychologie comme un moyen de représentation compacte et efficace et servir d'outil d'aide à l'analyse du contenu émotionnel des énoncés. D'une part, toute émotion secondaire (non primaire) pourra par exemple être représentée au moyen d'un ensemble de modèles d'émotions dites primaires. Il est possible de déterminer l'identité des émotions primaires qui composent une émotion secondaire par l'analyse des éléments du vecteur VCE correspondant. Les émotions primaires composant l'émotion secondaire correspondront aux éléments ayant les valeurs les

plus élevées (dépassant vraisemblablement un certain seuil) du vecteur VCE associé aux données de l'émotion secondaire. D'autre part, les vecteurs VCE sont très pertinents pour représenter, analyser et/ou détecter les émotions mixtes (plus d'une émotion véhiculée dans un énoncé) ou encore les émotions ambiguës. La détection de telles émotions permettra par exemple de réserver un traitement particulier pour leurs prises en charge. Notons que dans le jargon de la communauté de la recherche de la parole, l'espace généré par les modèles GMM/HMM est connu sous le nom d'espace d'ancrage (*anchor space*) et l'ensemble des modèles de référence est appelé les modèles d'ancrage (*anchor models*).

4.6 Conclusion

Dans ce chapitre nous avons présenté la méthodologie proposée dans cette thèse afin d'améliorer les performances des systèmes de RAE. Nous avons vu qu'une approche basée sur la similarité présente une alternative intéressante pour représenter et classifier les énoncés émotionnels. Les vecteurs caractéristiques d'émotion obtenus par projection des traits acoustiques dans un espace d'ancrage (formé par un référentiel de modèles GMM des classes d'émotion) représentent à la fois des traits de hauts niveaux discriminants et des descripteurs basés sur la similarité. Ils sont également particulièrement préconisés pour domaine de la RAE où la taille limitée des énoncés ne permet pas d'extraire des paramètres statistiques fiables. D'ailleurs, il a été montré dans le domaine de vérification du locuteur, que les performances des modèles d'ancrage dépassent celles des GMM dans les conditions où la taille des données est insuffisante (Mami et Charlet, 2002).

Dans les chapitres suivants nous présenterons les systèmes proposés dans cette thèse construits sur la base des traits VCE. Les systèmes proposés seront également bâtis sur des méthodes de classification basées sur la similarité. Dans le chapitre suivant un nouveau système intitulé *plus proche patron de similarité pondéré* est proposé. Le système utilise une mesure qualitative pour estimer la proximité entre énoncés et/ou classes émotionnelles. Dans le CHAPITRE 6 et CHAPITRE 7, les modèles d'ancrage basés sur des mesures quantitatives, principalement euclidienne et cosinus seront étudiés.

CHAPITRE 5

MÉTHODE DU PLUS PROCHE PATRON DE SIMILARITÉ PONDÉRÉ

5.1 Introduction

Dans ce chapitre, nous introduisons une nouvelle méthode de classification intitulée le *plus proche patron de similarité pondéré* (*Weighted Ordered Classes-Nearest Neighbors*, WOC-NN) (Attabi et Dumouchel, 2011; 2012). WOC-NN est une méthode basée sur le concept de similarité qui lierait un énoncé avec une classe d'émotion. Cette similarité est reflétée à travers le degré de concordance qui existerait entre les membres de leurs voisinages (classes voisines). Plutôt que de représenter un énoncé ou une catégorie d'émotion par des vecteurs acoustiques bruts, un vecteur caractéristique de haut niveau composé de classes d'émotion ordonnées selon leurs ordres de proximité est créé comme *représentation qualitative* de l'émotion. Rappelons que la notion de proximité et d'éloignement entre classes d'émotion est bien présente dans le modèle dimensionnel des émotions en psychologie. Dans nos travaux, plutôt que de forcer la projection des modèles de classes d'émotion dans un espace dimensionnel selon les modèles théoriques d'émotion existants, nous allons bâtir nos modèles sur une approche qui est guidée par les données (*data-driven approach*). Par conséquent, le voisinage de chaque classe sera exclusivement appris à partir des données d'apprentissage et aucune règle basée sur les connaissances a priori ne sera prise comme hypothèse de départ.

Après une description générale de la méthode du *plus proche patron de similarité pondéré* dans la section 2, nous allons décrire plus en détails comment sont construits les patrons de similarités dans la section 3. La section 4 décrit comment les patrons de similarité sont pondérés et utilisés pour la classification. Un second modèle de patron de proximité de type non linéaire est proposé dans la section 5, dans le but de modéliser les interactions qui existeraient entre les différents éléments (classes) composant les patrons de similarité. Enfin, dans la section 6, nous allons rapporter les résultats expérimentaux de classification des

émotions avec la méthode WOC-NN et comparer les résultats avec les performances des systèmes de référence.

5.2 Vue d'ensemble du système WOC-NN

Le système WOC-NN est composé de deux niveaux : modélisation et classification. Le niveau modélisation constitue le système frontal responsable de l'extraction des patrons de proximité (ou de similarité) pondérés. Le niveau classification concerne la prédiction de la classe de sortie d'un énoncé sur la base du niveau de concordance entre le patron de proximité de la donnée de test et celui de la classe cible. Durant la phase d'apprentissage, chaque classe d'émotion E_i est caractérisée par un patron de proximité qui représente l'ensemble des classes d'émotion ordonnées en fonction de leurs degrés de similarité avec la classe E_i comme le montre la Figure 5.1. Le rang d'une classe à l'intérieur d'un patron de proximité reflète le degré de similarité avec la classe E_i par rapport à l'ensemble des classes. Les rangs des classes dans un patron de proximité n'ont pas tous le même pouvoir de discrimination, par conséquent chaque rang de classe dans le patron est pondéré par un coefficient proportionnel à son pouvoir de discrimination. Enfin, une opération de sélection de traits est appliquée sur les classes composant le patron de proximité en vue de ne garder que les rangs renforçant la discrimination des patrons de proximité.

Durant la phase de test, le patron de proximité associé à la donnée de test est généré puis comparé au patron de proximité de chacune des classes. La donnée de test est affectée à la classe dont le patron de proximité retourne la distance minimale, c'est à dire, la classe ayant le plus grand nombre de classes voisines en commun avec la donnée de test. La métrique de distance utilisée dans la comparaison est la somme pondérée des différences entre les rangs des classes des deux patrons, une variante de la distance de *Hamming*.

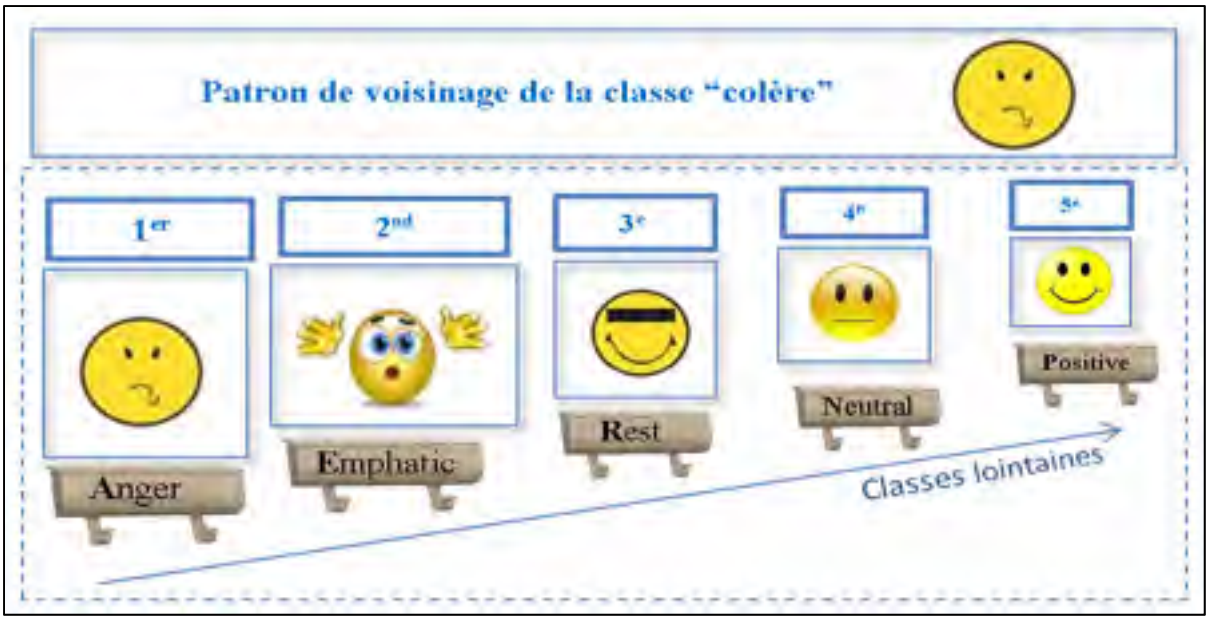

Figure 5.1 Exemple d'un patron de voisinage calculé pour la classe *colère* à partir des données d'apprentissage du corpus FAU AIBO Emotion

Dans la règle de décision *Bayes*, une donnée de test est associée à la classe qui minimise le risque de *Bayes*. Lorsque tous les types d'erreurs ont les mêmes coûts, la formule est simplifiée et la classe prédite correspond à la classe qui maximise la probabilité a posteriori ou encore le score de vraisemblance si les classes ont les même probabilités a priori. L'information sur le score du maximum de vraisemblance peut s'avérer insuffisante pour garantir que la classe prédite est la vraie classe particulièrement pour les données ambiguës. Avec la méthode proposée, plutôt que de baser la classification sur un rang de score unique, celui du maximum en occurrence, les rangs des scores de probabilité des autres classes sont inclus dans le processus de décision mais avec différents niveaux de prépondérance. Par conséquent, la règle de décision basée sur la valeur du maximum de vraisemblance peut être considérée comme un cas particulier de la règle de décision du système WOC-NN où le niveau d'importance de chaque rang du modèle est nul, sauf pour le rang maximisant les scores de probabilité.

5.3 Patron de proximité

Le patron de proximité ou de voisinage d'une classe d'émotion ou d'un énoncé est considéré comme étant un vecteur caractéristique de haut niveau dans l'architecture WOC-NN. Un élément du vecteur de patron de proximité d'un énoncé représente l'indice d'une classe d'émotion et le rang de cet élément dans le vecteur reflète le degré de similarité de cette classe d'émotion avec cet énoncé. Le degré de similarité d'une classe par rapport à une autre est estimé dans l'espace des probabilités de vraisemblance. Tout modèle de classification capable de générer des scores de probabilité fiables peut être utilisé à ce stade. Pour le système proposé, le modèle GMM est utilisé comme fonction d'estimation de densité de probabilité pour générer les probabilités de vraisemblance des traits acoustiques.

Soit $\{\lambda_1, \lambda_2, \ldots, \lambda_C\}$ l'ensemble des modèles GMM associés aux C classes d'émotion. Soit $\mathbf{X}$ un énoncé de parole émotionnelle représenté par une séquence de T trames, et $\mathbf{L}(\mathbf{X})$ le vecteur de caractérisation de l'émotion (VCE) de $\mathbf{X}$.

$$\mathbf{L}(\mathbf{X}) = \begin{bmatrix} \frac{1}{T}\log P(\mathbf{X}|\lambda_1) \\ \vdots \\ \frac{1}{T}\log P(\mathbf{X}|\lambda_C) \end{bmatrix}, \tag{5.1}$$

Les scores de probabilité $l_i = \frac{1}{T}P(X|\lambda_i)$ des éléments du vecteur $\mathbf{L}(\mathbf{X})$ sont triés dans l'ordre décroissant, $l_{r_1} \geq l_{r_2} \geq \cdots \geq l_{r_C}$. Les indices de classes, r_i pris dans cet ordre, sont concaténés pour former le vecteur $\mathbf{r}$, que nous appellerons patron de proximité (ou de similarité) de $\mathbf{X}$:

$$\mathbf{r} = [r_1\, r_2 \cdots r_C]^T \tag{5.2}$$

Un patron de proximité $\mathbf{r}$ peut être soit un représentant *d'une instance* de données si ce patron est estimé à partir d'un énoncé ou soit un représentant de *classe* s'il est estimé avec les

données d'entrainement d'une classe d'émotion et qui jouera le rôle d'un patron représentatif d'une catégorie d'émotion.

5.4 Métrique de mesure de similarité

La dissimilarité (ou similarité) entre une classe d'émotion et un énoncé de test peut être mesurée en utilisant la distance de *Hamming* entre les deux vecteurs de patron de proximité qui leurs sont associés. Le nombre d'éléments (rangs de classes) qui diffèrent entre les deux vecteurs de patron de proximité (par une comparaison par paire entre éléments des deux vecteurs) représente la distance de *Hamming*. Soit $\mathbf{r}$ et $\mathbf{r}'$ deux vecteurs qui dénotent le patron de proximité d'un énoncé de test **X** et celui de la classe d'émotion E_i respectivement. La distance de *Hamming* entre $\mathbf{r}$ et $\mathbf{r}'$ est formulée comme suit :

$$D_H(\mathbf{r},\mathbf{r}') = \#\{j : r_j \neq r'_j\} \tag{5.3}$$

où #*E* représente le cardinal de l'ensemble *E*. Lors de la phase de test, l'énoncé de test **X** est affecté à la classe d'émotion qui minimise la distance de *Hamming* entre son patron de proximité et celui de **X**. Pour simplifier la notation et la rendre plus compacte, nous introduisons un vecteur **v** de dimension *C* que nous appellerons *patron de distance* (ou *vecteur de distance*), avec des éléments $\mathbf{v}_j = 1$ si $r_j \neq r'_j$ et $\mathbf{v}_j = 0$ autrement.

La Figure 5.2 illustre le processus de génération du patron de proximité $\mathbf{r}$ associé à un énoncé **X** ainsi que le vecteur de distance **v** séparant $\mathbf{r}$ de $\mathbf{r}^k$, le patron de proximité de la classe d'émotion *k*. Le cas où tous les éléments composant un patron de distance **v** sont nuls, correspond au cas idéal où **X** et E_i partagent le même patron de proximité (**X** appartient à la classe E_i). Dans la pratique le vecteur **v** contiendra une ou plusieurs valeurs à 1 même si $\mathbf{r}$ et $\mathbf{r}'$ appartiennent à la même classe.

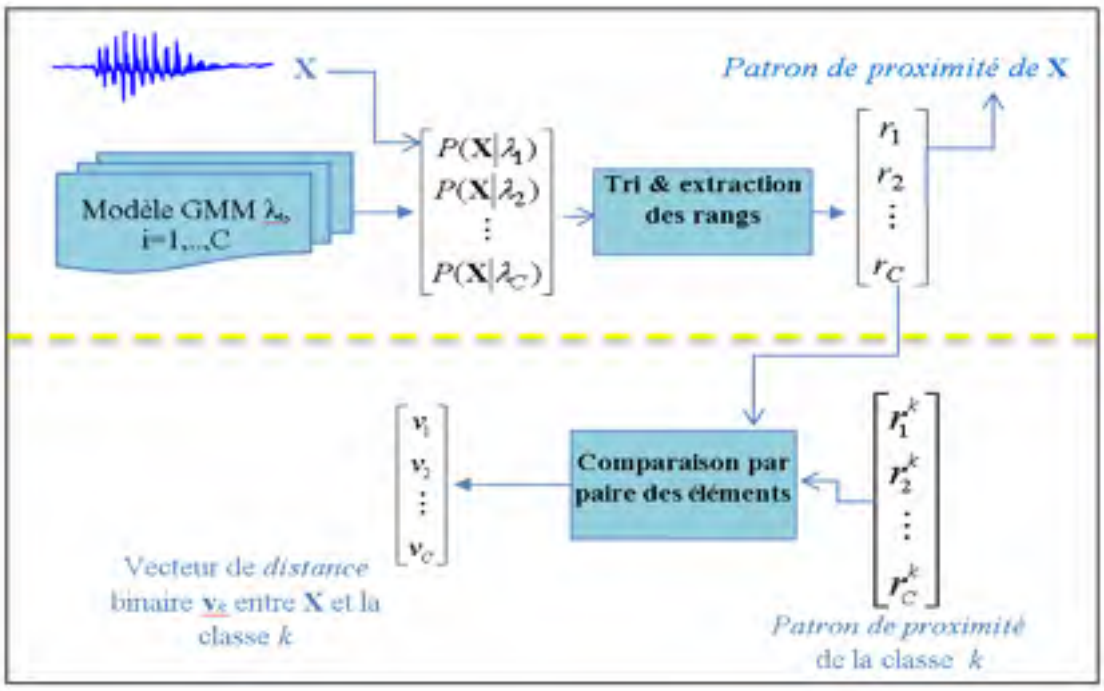

Figure 5.2 La partie supérieure de cette figure illustre le processus de génération de patron proximité d'un énoncé **X**. La partie inférieure montre comment le vecteur de distance entre ce modèle et le patron de proximité de la classe d'émotion *k* est calculé

Un exemple de calcul du vecteur de distance entre une donnée de test **X** et la classe *colère* est donné dans la Figure 5.3. La nouvelle formule de la métrique, est récrite comme suit :

$$\mathcal{D}_H(\mathbf{r},\mathbf{r}') = \mathbf{v}^T\,\mathbf{v} \tag{5.4}$$

5.4.1 Pondération des rangs de classe

L'utilisation directe de la distance de *Hamming* pour mesurer la dissimilarité entre les patrons de proximité lors de la classification peut souffrir de deux inconvénients dans le contexte des systèmes de RAE. La première concerne la quantité importante de données qui seront classées comme ambiguës (car plus d'un patron de proximité de classe retournera une distance de *Hamming* minimale). Cela est dû d'une part à la nature de la métrique qui renvoie

des valeurs discrètes et d'autre part, au nombre limité de classes d'émotions à classifier. Dans le cas d'un problème à *C*-classes, les valeurs possibles retournées par la métrique sont $\{0,1,\ldots,C\}$ c.-à-d. $(C+1)$ différentes valeurs possibles. Quand la valeur de C est basse la chance d'avoir plus d'une classe retournant la distance minimale pour une donnée de test est plus importante. Lorsque la distance de *Hamming* est appliquée pour évaluer les données de test du corpus FAU AIBO Emotion, 30,5 % des données sont classifiées comme ambiguës. Le deuxième inconvénient découle de l'hypothèse implicite que chaque rang de classe d'un patron de proximité a le même pouvoir de discrimination que les autres rangs que suppose l'utilisation directe de la distance de *Hamming*. Cette hypothèse est particulièrement fausse pour le premier rang de classe (le modèle qui maximise les scores de probabilité), comme nous le verrons plus loin. Ces considérations motivent l'utilisation d'une variante de distance de *Hamming* où nous introduisons un vecteur de coefficients de pondération pour chaque patron de proximité $\mathbf{w}_i = [w_{i0}\, w_{i1}\, w_{i2} \cdots w_{iC}]^T$, où w_{ij} représente le coefficient de pondération du rang *j* du patron de proximité associé à la classe d'émotion *i*.

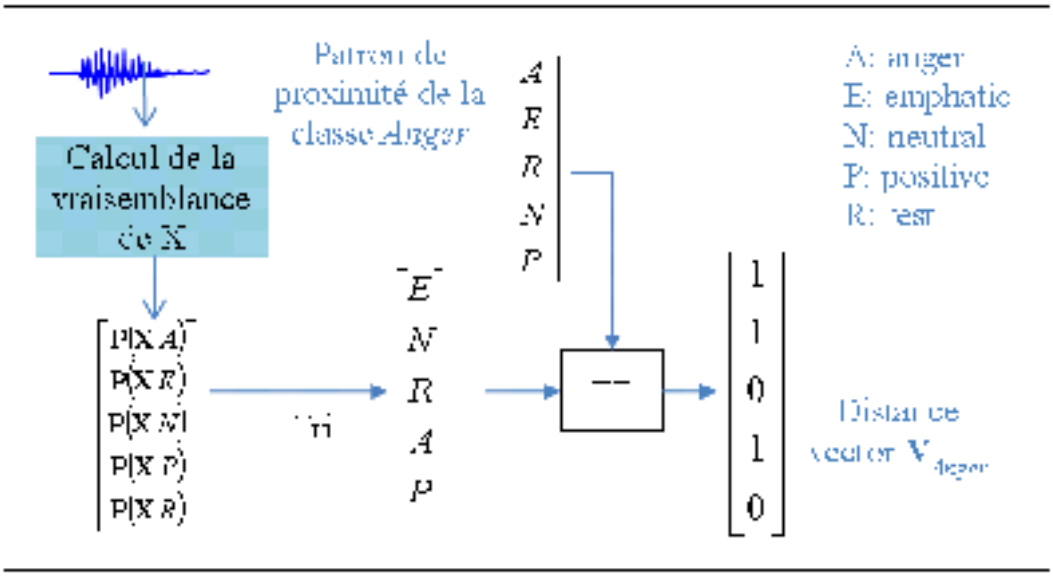

Figure 5.3 Exemple de patron de distance calculé pour un énoncé **X** et le patron de proximité de la classe *colère*

Deux approches différentes peuvent être utilisées pour l'estimation des coefficients de pondération des rangs de classe dans un patron de proximité : la première est basée sur l'estimation d'un vecteur de pondération global partagé par les patrons de proximité de toutes les classes d'émotion. Dans la seconde approche, un vecteur de pondération spécifique est estimé séparément pour chaque patron de proximité de classe afin de prendre en considération la spécificité du modèle de voisinage de chaque catégorie d'émotion. Les résultats de classification obtenus montrent que les performances d'un système basé sur un vecteur de pondération spécifique à chaque classe sont légèrement supérieures à celles d'un système basé sur un vecteur de pondération partagé par toutes les classes. Dans ce qui suit nous allons montrer comment estimer le poids de chaque élément du vecteur de distance afin d'optimiser les performances de classification. À cet effet, un nouveau modèle de classification à deux-classes basé sur la régression logistique est créé. Les variables prédictives ou indépendantes de ce modèle vont correspondre aux rangs de classe des patrons de proximité et les coefficients de ces variables correspondront aux poids associés à ces rangs. Le modèle de régression logistique aura pour objectif de discriminer entre les patrons de distance corrects versus les patrons incorrects.

5.4.2 Modèle de régression logistique

Le vecteur de poids $\mathbf{w}_i$ associé au patron de proximité de la classe i est estimé par régression logistique, un modèle discriminant de la forme suivante:

$$p\left(y = 1 \middle| \mathbf{w}_i, \mathbf{v}\right) = \frac{\exp\left(\mathbf{w}_i^T \mathbf{v}\right)}{1 + \exp\left(\mathbf{w}_i^T \mathbf{v}\right)} \tag{5.5}$$

où y représente l'étiquette de la classe cible du vecteur de distance $\mathbf{v}$ (y = 1, correspond à un patron de distance correct et 0 à un vecteur incorrect). La valeur du coefficient w_{ij} représente le degré de robustesse de l'association de la variable (rang de la classe) v_{ij} au modèle correct. L'équation (5.5) peut être récrite comme une fonction *log odds*:

$$\text{logit}\left(y=1|\mathbf{w}_i,\mathbf{v}\right)=\log\frac{P\left(y=1|\mathbf{w}_i,\mathbf{v}\right)}{1-P\left(y=1|\mathbf{w}_i,\mathbf{v}\right)}=\mathbf{w}_i^T\mathbf{v}\,, \tag{5.6}$$

qui représente le logarithme du rapport de la probabilité de reconnaître un patron correct sur la probabilité qu'il ne sera pas reconnu. Étant donné que les valeurs de **v** sont des variables binaires (0 ou 1), la valeur maximale de *logit* en faveur de reconnaître un patron correct pour une classe d'émotion donnée est bornée par une valeur maximale égale à la somme des valeurs positives des éléments du vecteur de pondération $\mathbf{w}_i$. Il est intéressant de noter que cette valeur maximale peut être considérée comme un indicateur sur le pouvoir de discrimination du patron de proximité pour la classification des classes dans le système WOC-NN. La valeur de corrélation, égale à 0,86, mesurée entre les valeurs maximales du *logit* de chaque classe et les taux de reconnaissance des classes correspondantes du système WOC-NN vient confirmer cette forte association.

La mesure de la distance de *Hamming* pondérée après introduction du vecteur de pondération est formulée comme suit :

$$\mathcal{D}_{wH}\left(\mathbf{r},\mathbf{r}'\right) \;=\; \mathbf{w}_i^T\;\mathbf{v}' \tag{5.7}$$

où le vecteur $\mathbf{v}'$ de dimension (C +1) représente le vecteur **v** (de dimension C) augmenté de la valeur 1 pour représenter le coefficient de la constante (*intercept* en anglais). La règle de décision pour classifier un énoncé de test **X** s'écrit comme suit :

$$emotion\_cible=\underset{i=1,\ldots,C}{\arg\min}(\mathbf{w}_i^T\;\mathbf{v}_i') \tag{5.8}$$

où $\mathbf{v}_i'$ représente le vecteur de distance entre le patron de proximité de l'énoncé **X** et celui de la classe d'émotion i, et $\mathbf{w}_i$ le vecteur de pondération associé à la même classe d'émotion i.

5.4.3 Génération des données d'entrainement

L'estimation du vecteur de pondération w_i nécessite l'utilisation des données d'apprentissage pour entrainer le modèle de régression logistique. Dans cette section nous décrirons la méthode utilisée pour générer les données d'apprentissage spécifiques à une classe d'émotion et qui seront assignées à l'entrainement des coefficients de pondération associés à cette classe. Les données d'apprentissage pour une classe d'émotion *k* utilisées pour l'estimation du vecteur de pondération $\mathbf{w}_k$ sont générées selon le schéma bloc de la Figure 5.4. Pour chaque donnée d'apprentissage **X** appartenant à la classe *k*, un vecteur de patron de proximité **r** est calculé. *C* vecteurs de patrons de distances $\mathbf{v}_k^i, i = 1,\ldots,C$ sont par la suite générés pour le vecteur **r** à travers une opération de comparaison par paires avec le patron de proximité de chacune des *C* classes d'émotion. La comparaison par paires entre le vecteur **r** et le vecteur de patron de proximité de la classe *k* va former un vecteur de patron de distance de type correct ($y = 1$), alors que la comparaison de **r** avec les *C-1* patrons de proximité des autres classes d'émotion va constituer *C-1* vecteurs de patron de distance de type incorrect ($y = 0$). Ainsi pour une classe d'émotion de n_k données d'apprentissage, $(n_k \times C)$ vecteurs de distance sont générés pour l'estimation du vecteur de coefficients de pondération associés à la classe *k*. Ce processus est répété pour l'apprentissage du vecteur de pondération de chacune des classes.

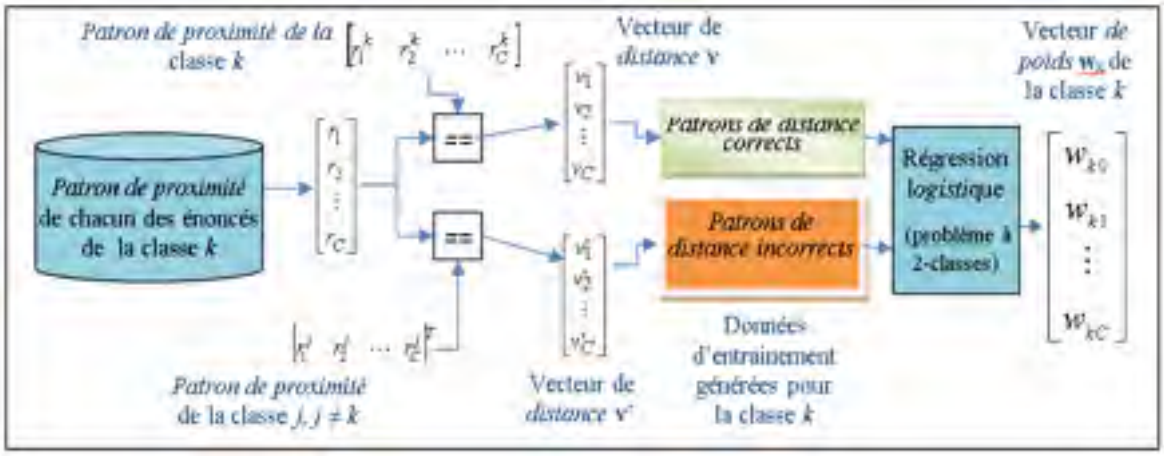

Figure 5.4 Schéma bloc de la méthode de génération des données d'entrainement utilisées l'apprentissage des coefficients de pondération de la classe *k* en utilisant la régression logistique

La génération des données pour l'estimation du vecteur de pondération indépendant de toute classe revient à regrouper les données d'entrainement générées pour chacune des classes en un seul ensemble de données et estimer par la suite le vecteur de pondération partagé par toutes les classes.

5.4.4 Réduction de la dimensionnalité

Dans le système WOC-NN, chaque patron de proximité est considéré comme un vecteur de traits caractéristiques. En tant que tel, certains rangs de classe (traits caractéristiques) dans un patron de proximité peuvent être bruités ou non suffisamment discriminants dans la caractérisation du voisinage d'une classe et devraient être écartés du patron (ou inhibés). Il existe plusieurs approches et méthodes dédiées à la réduction de la dimensionnalité telles que présentées dans le CHAPITRE 3. Il est intéressant de noter qu'avec l'architecture du système WOC-NN aucun autre algorithme n'est requis pour réaliser cette opération. En effet, la méthode utilisée pour la génération des données d'entrainement et en particulier le choix du type des classes de sortie (modèles corrects versus modèles incorrects) nous permettent de faire usage des résultats de la régression logistique, appliquée pour l'estimation des coefficients de pondération, pour sélectionner les rangs de classe les plus pertinents. En

régression logistique, un coefficient de pondération à valeur positive indique que la variable prédictive indépendante qui lui est associée possède un impact positif sur la probabilité *logit* de la prédiction de la classe $y=1$, donc dans la reconnaissance des modèles corrects. A l'inverse, un rang de classe avec un coefficient négatif a pour effet de diminuer la probabilité de reconnaître les modèles corrects et peut donc être considéré comme trait caractéristique nuisible et doit être retiré du modèle.

Nous introduisons également une deuxième condition de sélection de rangs de classe permettant d'assurer que seules les variables qui contribuent de manière significative au modèle sont élues. A cet effet, un test statistique de *Chi-deux* appelé test de *Wald* est appliqué pour chaque variable prédictive. Avec ce second critère, l'ensemble des rangs de classe sélectionnés ne comprendra que les rangs ayant une association significative avec le patron de proximité correct (pour un seuil de signification donné, égal à 0,05 dans notre étude). Le nouveau vecteur de poids est calculé comme suit :

$$\hat{\mathbf{w}}_i = \begin{bmatrix} \hat{w}_{i1} \\ \hat{w}_{i2} \\ \vdots \\ \hat{w}_{iC} \end{bmatrix}, \quad \text{où } \hat{w}_{ij} = \begin{cases} w_{ij} & \text{si } (w_{ij} \geq 0) \text{ et } (wald_{ij} \geq seuil) \\ 0 & \text{sinon} \end{cases} \tag{5.9}$$

5.4.5 Normalisation de la pondération

Lorsque le vecteur de pondération est estimé séparément pour chaque patron de proximité, la somme des coefficients de pondération associée à un patron de proximité d'une classe peut être supérieure ou inférieure à la somme associée à un patron de proximité d'une autre classe d'émotion. Une classe bien définie théoriquement et bien représentée expérimentalement (à travers des données d'apprentissage représentatives) lui sera généralement associée des coefficients de poids élevés reflétant ainsi la capacité discriminative du patron de proximité. L'écart de la somme des coefficients entre patrons de classes peut s'accentuer davantage sous l'effet de l'étape de la sélection des traits. Les patrons de proximité associés aux différentes

classes d'émotion peuvent avoir un nombre différents de rangs de classes sélectionnés ce qui va accroitre la valeur du *logit* des patrons de proximité ayant un plus grand nombre de rangs sélectionnés. L'évaluation de la distance en utilisant la formule (5.8) pour la prédiction des données, sans normalisation préalable des coefficients de pondération, va favoriser dans une certaine mesure la reconnaissance des classes ayant les plus petites sommes des coefficients de pondération et pénalisera les classes bien-entrainées. Afin de prévenir que les performances de reconnaissance soient biaisées en faveur de certaines classes, les valeurs du vecteur de pondération de chaque patron de proximité sont normalisées de sorte que leur somme soit égale à 1. Le nouveau vecteur de pondération $\overline{\mathbf{w}}_i$ est calculé comme suit :

$$\overline{\mathbf{w}}_i = \begin{bmatrix} \overline{w}_{i0} \\ \vdots \\ \overline{w}_{iC} \end{bmatrix}, \quad \text{où } \overline{w}_{ij} = \hat{w}_{ij} / \sum_{k=0}^{C} \hat{w}_{ik} \tag{5.10}$$

5.5 Interaction entre les classes dans un patron de proximité

Dans les sections précédentes nous avons considéré l'absence ou la présence d'une classe d'émotion à un rang donné dans un patron de proximité comme étant un facteur de prédiction de l'émotion cible indépendamment de la présence des autres classes d'émotion à leurs rangs respectifs. Dans cette section, nous allons examiner l'interaction qui existe entre les rangs de classe dans un patron de proximité et son effet sur les performances de classification. Ceci revient à étudier l'absence simultanée de deux classes d'émotion, ou éventuellement plus, à leurs rangs respectifs comme caractéristique de dissimilarité dans un patron de proximité. La Figure 5.5 est un exemple de patron de proximité de type double interaction calculé pour la classe *colère*, estimé à partir des données d'apprentissage du corpus FAU AIBO Emotion. L'investigation de l'interaction est motivée par l'idée que les classes d'émotion peuvent être liées les unes aux autres dans leur espace de proximité. Par exemple, quand deux patrons de proximité appartenant à deux énoncés différents sont comparés, on peut s'attendre à ce qu'il ait probablement des couples d'émotion du patron de proximité qui vont être simultanément

présents ou absents en fonction de l'appartenance ou non des deux énoncés à la même classe d'émotion.

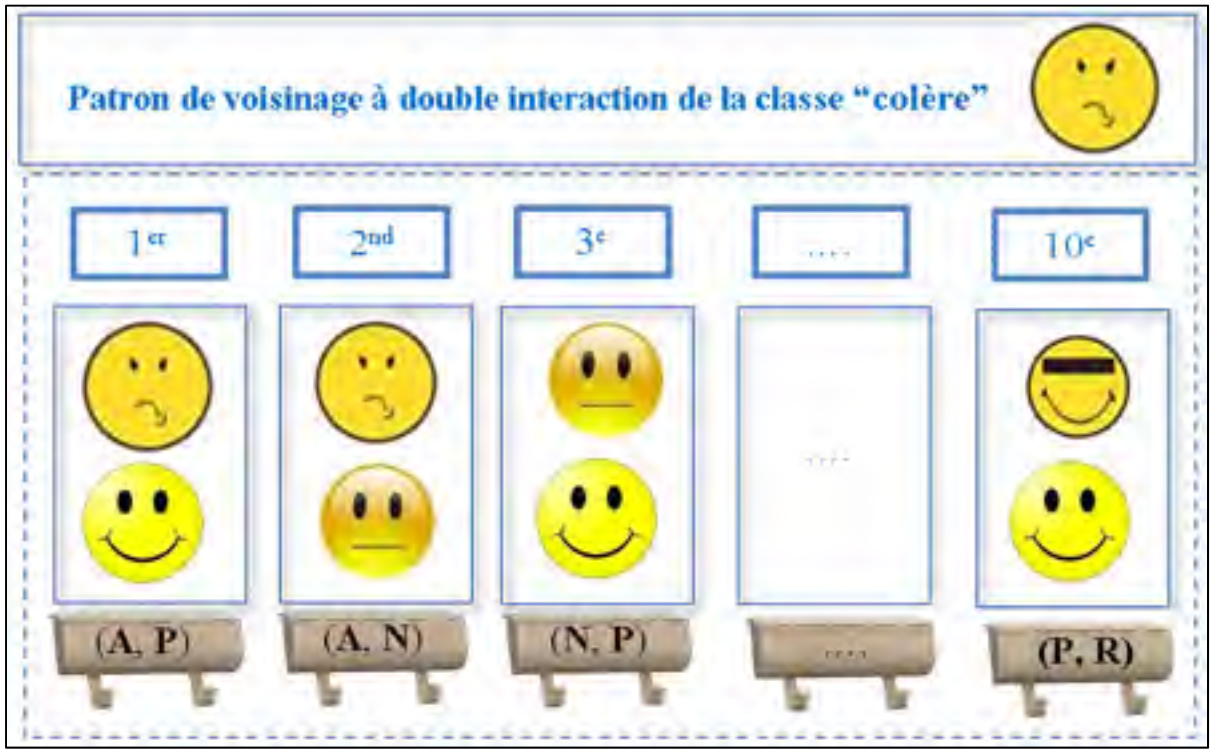

Figure 5.5 Exemple de patron de proximité à interaction double entre les rangs de classe calculé pour la classe *colère*

L'interaction entre les classes est modélisée à travers un modèle non-linéaire. Nous nous intéresserons dans cette étude à deux types de modèles non-linéaires : modèle basé sur l'interaction entre deux rangs de classes, que nous appellerons modèle à double interaction, et un deuxième modèle faisant intervenir trois rangs de classe et qui sera nommé modèle à triple interaction.

La distance dans la nouvelle règle de décision du modèle non-linéaire à double interaction peut-être formulée comme suit :

$$\mathcal{D}_{wH}(\mathbf{r},\mathbf{r}') = \sum_{i\neq j} w_{ij}\, v_i\, v_j \qquad (5.11)$$

où $i,j = 1,..,C$, w_{ij} représente le poids associé à l'importance du rang des classes i et j combinées, v_i et v_j représentent respectivement le i-ième et j-ième éléments du vecteur de distance $\mathbf{v}$ séparant les patrons $\mathbf{r}$ et $\mathbf{r}'$. Afin de rendre la formule de la règle de décision plus compacte sous une forme vectorielle et garder les mêmes formules des algorithmes d'apprentissage et de test utilisées pour le modèle linéaire, nous allons introduire un nouveau vecteur de distance qui intégrera la non-linéarité du modèle. Soit $\hat{\mathbf{v}}$ le nouveau vecteur de distance modélisant la double interaction. Chaque élément du vecteur $\hat{\mathbf{v}}$ représente le produit de deux variables prédictives du vecteur de distance $\mathbf{v}$ du modèle linéaire, calculé pour toutes les combinaisons de paires de classes distinctes, formulé comme suit:

$$\hat{\mathbf{v}} = \begin{bmatrix} \hat{v}_{12} \\ \vdots \\ \hat{v}_{ij} \\ \vdots \\ \hat{v}_{(C-1)C} \end{bmatrix}, \quad \text{où } \hat{v}_{ij} = (v_i \times v_j), \quad i = 1,..,C-1 \text{ et } j = i+1,..,C \tag{5.12}$$

Le nombre d'éléments du vecteur $\hat{\mathbf{v}}$ est égal au nombre de combinaisons possibles de deux classes parmi l'ensembles de C classes, évalué à $\binom{C}{2} = \frac{C!}{2!(C-2)!}$. La nouvelle dimension du vecteur $\hat{\mathbf{v}}$ passe donc de la taille C à une dimension plus large égale à $l = ((C-1)\times C)/2$. Pour une valeur de C égal à 5 (nombre de classes du corpus FAU AIBO Emotion), l est égale à dix. Avec la nouvelle forme du vecteur $\hat{\mathbf{v}}$, la même formule (5.8) est utilisée comme règle de décision.

Nous nous sommes également intéressés à l'investigation du modèle basé sur l'interaction triple entre classes d'émotion (interaction entre trois variables prédictives) sur les performances du système. Un élément $\hat{v}_{ijk}$ d'un tel vecteur de distance est formulé comme suit :

$$\hat{\mathbf{v}} = \begin{bmatrix} \hat{v}_{123} \\ \vdots \\ \hat{v}_{ijk} \\ \vdots \\ \hat{v}_{(C-2)(C-1)C} \end{bmatrix}, \quad \text{où } \hat{v}_{ijk} = (v_i \times v_j \times v_k);\ i = 1,..,C-2;\ j = i+1,..,C-1;\ k = j+1,..,C \qquad (5.13)$$

La dimension du vecteur de distance à triple interaction est de $\binom{C}{3} = \frac{C!}{3!(C-3)!}$. Il est également égal à dix pour un nombre de classes *C* égal à cinq.

5.6 Résultats expérimentaux

WOC-NN sera expérimenté sur le corpus FAU AIBO Emotion. Le calcul des patrons de proximité de chacune des classes, les vecteurs de poids qui leurs sont associés ainsi que l'opération de la sélection des traits seront basés sur les données d'apprentissage. Différentes variantes de systèmes WOC-NN seront expérimentées afin d'étudier l'effet de la pondération des rangs, de la sélection des traits et l'utilisation d'un vecteur de pondération commun versus spécifique pour chaque patron de proximité. Nous comparerons également les performances du système WOC-NN-1 basé sur un simple rang de classe avec les systèmes basés sur l'interaction double et triple des rangs de classe que nous désignerons par WOC-NN-2int et WOC-NN-3int respectivement. Les performances des différentes variantes du système WOC-NN ont été optimisées séparément selon le protocole de la validation croisée à neuf plis. Le nombre de composantes optimisé du GMM du système frontal est de 64 gaussiennes. Enfin, un système GMM-*Bayes* est développé en tant que système de référence.

5.6.1 Patrons de proximité des classes d'émotion du corpus FAU AIBO Emotion

Le Tableau 5.1 présente les patrons de proximité de chacune des cinq classes du corpus FAU AIBO Emotion. Nous remarquons que les données les plus similaires à une classe d'émotion sont les données appartenant à sa propre classe qui est considérée par conséquent comme étant sa première classe voisine. Cette constatation est valide pour toutes les classes

d'émotion (ceci pourrait ne pas être toujours vrai si le patron de voisinage est construit à partir des données de test par exemple). Si nous prenons la classe **E** (*emphatique*) comme exemple, la classe qui lui est la plus similaire, après sa propre classe, est la classe **A** (*colère*) suivie de la classe **N** (*neutre*) à la position médiane de son voisinage. Par contre la classe la plus lointaine, donc considérée comme étant la plus dissimilaire est la classe **P** (*positive*). Il est également intéressant de constater que la classe **P** est la classe la plus dissimilaire dans les patrons de proximité de toutes les classes d'émotion (excepté dans son propre patron). Ceci suggère que la classe **P** forme un bloc d'émotion distinct alors que les quatre classes restantes forment un autre bloc d'émotions opposé qui partagent certaines caractéristiques communes.

Tableau 5.1 Patron de proximité du modèle linéaire de chacune des classes du corpus FAU AIBO Emotion ainsi que le poids associé à chaque rang appris à partir des données d'entrainement

Rangs	**1er**	**2nd**	**3e**	**4e**	**5e**
A	**A** 1.91	**E** -0.38	**R** 0.19	**N** 0.58	**P** 0.48
E	**E** 1.76	**A** -0.55	**N** 0.16	**R** 0.04	**P** 0.55
N	**N** 1.51	**R** 0.36	**E** -0.68	**A** 0.49	**P** 0.05
P	**P** 1.57	**R** 0.97	**N** 0.75	**E** -0.18	**A** 0.81
R	**R** 0.66	**N** -0.46	**E** -0.82	**A** 0.73	**P** 0.09

Rang sélectionné. | Rang avec un poids négatif. | Critère statistique *Wald* non satisfait.

La pondération associée à chaque rang ainsi que le résultat de l'opération de sélection de traits sont également affichés dans le Tableau 5.1. Le nombre en dessous du nom de classe correspond au coefficient de pondération associé à ce rang. Plusieurs observations intéressantes peuvent être relevées à partir de ces valeurs. D'abord, nous constatons que la pondération à l'intérieure d'un même patron de proximité diffère d'un rang à un autre. Le premier rang de classe, par sa valeur élevée pour la plupart des patrons, représente de loin la

variable la plus importante dans le processus de décision pour la plupart des classes d'émotion excepté pour la classe **R** (*reste*). Notons que dans le cas où tous les rangs de classes possèderaient des valeurs négligeables par rapport au premier rang et ce pour tous les patrons de proximité, la règle de décision du système WOC-NN correspondrait à celle du système GMM-*Bayes* et les deux systèmes obtiendraient par conséquent des performances similaires (le système GMM-*Bayes* devient un cas particulier du système WOC-NN).

Par ailleurs, afin d'expliquer pourquoi la classe **R** ne possède pas le poids du rang le plus pondérant dans son propre patron de proximité, nous avons analysé la matrice de confusion du système frontal basé sur le modèle GMM et la règle de décision *Bayes*. La matrice de confusion montre que la classe **R** est la classe la plus difficile à reconnaitre avec un taux de rappel de 23 %. Par conséquent, le modèle de la classe ne génère pas des scores fiables ce qui est reflété à travers la valeur du poids de la classe à l'intérieur du patron de proximité. On notera également que l'importance de la contribution d'une classe dans un patron de proximité ne dépend pas de son rang à l'intérieur de ce patron, elle est par conséquent indépendante du niveau similarité ou de dissimilaire entre deux classes d'émotion. Par ailleurs, le signe négatif de certains coefficients nous renseigne sur l'existence de rangs de classes qui peuvent être considérés comme traits néfastes pour la classification.

Les résultats de l'opération de la sélection des traits sont représentés dans le Tableau 5.1 à travers le motif des cellules. Les cellules blanches représentent les rangs de classe pertinents retenus après sélection. Les cellules grises et les cellules hachurées en diagonales représentent les rangs de classe qui ne répondent pas aux critères de positivité du coefficient de pondération et au test de statistique de *Wald* respectivement. Dans le Tableau 5.2 sont présentés les patrons de proximité des classes d'émotion du modèle à double interaction (système WOC-NN-2int). Pour la classe **P** par exemple, la paire de classes (**E**, **R**) possède le poids le plus important, et par conséquent l'absence simultanée de ces deux classes à leurs rangs respectifs dans un patron de distance fournira l'indice le plus fort que la donnée de test correspondant à ce modèle de distance ne fait pas partie de la classe **P**.

Tableau 5.2 Pondération des rangs du modèle non-linéaire des patrons de proximité à double interaction de chacune des classes du corpus FAU AIBO Emotion apprise à partir des données d'entrainement

	1er	2nd	3e	4e	5e	6e	7e	8e	9e	10e
A	A, P 0.98	A, N 0.93	N, P 0.54	A, E 0.53	A, R 0.47	E, R	E, P	N, R	E, N	P, R
E	E, R 1.07	A, P 1.07	A, E 0.86	E, N 0.40	E, P	A, N	N, R	N, P	A, R	P, R
N	A, E 1.06	N, P 0.92	N, R 0.51	A, P 0.35	E, N 0.30	A, N 0.23	E, R	E, P	P, R	A, R
P	E, R 1.65	A, P 1.50	A, N 1.10	N, P 0.43	P, R	N, R	E, N	A, E	E, P	A, R
R	A, R 2.20	A, E 0.67	N, R	P, R	E, N	A, N	A, P	E, R	N, P	E, P

Rang sélectionné | Rang avec un poids négatif | Critère statistique *Wald* non satisfait

5.6.2 Résultats de la classification

Les résultats sont optimisés pour maximiser la moyenne du rappel non pondérée (UAR) comme première mesure suivie de la moyenne du rappel pondérée (WAR) étant donné que les classes d'émotion FAU AIBO sont fortement déséquilibrées. Les résultats de la classification des données de test pour les différentes configurations du vecteur du poids du modèle linéaire du système WOC-NN sont présentés dans le Tableau 5.3. Nous observons que le système sans pondération de rangs (le système OC-NN) donne les plus mauvais résultats. L'application de coefficients de pondération partagés par tous les patrons de proximité des classes d'émotion mais sans sélection de traits (système SW-OC-NN) donne des performances similaires à un système GMM basé sur la règle *Bayes*. L'application d'une pondération dépendante du patron de proximité de chacune des classes (système WOC-NN-WFS) améliore légèrement les performances. Finalement, l'application de la procédure de

sélection de rangs à l'intérieur du patron de proximité (système WOC-NN) améliore les résultats et donne un gain relatif de 3,46 % par rapport au système GMM-*Bayes*.

Tableau 5.3 Résultats de classification du système WOC-NN obtenus sur les données de test en fonction des différentes configurations du vecteur de pondération du modèle linéaire

Systèmes	**UAR**	WAR
GMM-*Bayes*	41,05 %	41 %
OC-NN	37,43 %	24,97 %
SW-OC-NN	41,05 %	41,37 %
WOC-NN-WFS	41,33 %	41,20 %
WOC-NN	**42,47** %	41,10 %

5.6.3 Résultats du modèle non linéaire

Le Tableau 5.4 compare les résultats du modèle linéaire et non linéaire des systèmes WOC-NN. D'abord, nous constatons que la normalisation de poids améliore la performance du système WOC-NN indépendamment du modèle utilisé (avec ou sans interaction de rangs de classe). Les résultats montrent également que le système WOC-NN-2int dépasse légèrement les performances de classification du système WOC-NN-1int en terme UAR. Ceci peut signifier que tenir compte de l'absence (ou présence) simultanée d'un rang constitué d'une paire de classes dans le modèle de distance peut-être plus informatif que l'absence d'un rang de classe individuel pris indépendamment des autres. En outre, WOC-NN-3int donne les plus mauvais résultats. Ceci s'explique par le fait qu'il n'y a pas assez de classes d'émotion pour former un rang de trois classes qui peuvent caractériser un patron de proximité d'une manière fiable, comme nous pouvons le constater du Tableau 5.1, où la classe **R** par exemple ne contient dans son patron de proximité que deux classes d'émotion discriminantes.

Tableau 5.4 Effets de la normalisation des poids et de l'interaction entre les rangs de classes sur les performances des systèmes WOC-NN testés sur le corpus d'émotion FAU AIBO

Systèmes	**Normalisation**	**UAR**	**WAR**
WOC-NN-1int	-	42,47 %	41,10 %
	+	42,71 %	41,03 %
WOC-NN-2int	-	42,41 %	37,12 %
	+	**43,14** %	35,33 %
WOC-NN-3int	-	37,50 %	29,67 %
	+	37,57 %	34,09 %

Le Tableau 5.5 compare les résultats du système proposé avec les meilleurs systèmes de la compétition internationale *INTERSPEECH 2009 Emotion Challenge* et le système de référence (GMM-*Bayes*). Tout d'abord, nous constatons que WOC-NN surpasse le système GMM-*Bayes* avec 5,1 % en gain relatif. Rappelons que WOC-NN est basé sur le système GMM à l'exception de l'utilisation des autres rangs des scores dans la prise de décision. Ces résultats indiquent que les rangs des scores autres que celui qui maximise la probabilité a posteriori contiennent de l'information pertinente et complémentaire dans la prise de décision de classification d'une donnée de test. Enfin, les résultats obtenus avec le système proposé WOC-NN dépassent ceux du meilleur système unique (Lee *et al.* 2009) et systèmes combinés (Kockmann *et al.* 2009) de la compétition *INTERSPEECH 2009 Emotion Challenge*, de 3,45 % et 4,46 % respectivement.

Tableau 5.5 Comparaison des résultats du système proposé avec les meilleurs systèmes de la compétition *INTERSPEECH 2009 Emotion Challenge*, testés sur le corpus FAU AIBO Emotion

Systèmes	**UAR**	WAR
GMM-*Bayes*	41,05 %	41 %
(Lee *et al.* 2009) (Régression logistique bayésienne)	41,3 %	43,9 %
(Kockmann *et al.* 2009) (fusion de 2 systèmes de types *analyse factorielle jointe*)	41,7 %	-
WOC-NN	**43,14** %	35,33 %

5.7 Conclusion

Dans ce chapitre nous avons présenté une nouvelle méthode de classification des émotions basée sur un descripteur de haut niveau appelé le *patron de proximité pondéré* construit à partir des vecteurs VCE. Dans la règle de décision de WOC-NN, tous les rangs de classe participent au processus de prise de décision avec différents poids contrairement à la règle de décision où seul le rang de classe maximisant la probabilité des scores est prépondérant. Deux différents modèles de WOC-NN ont été construits; linéaire et non linéaire. Dans le modèle non-linéaire l'interaction entre les rangs de classe dans un même patron de proximité sont modélisés. Les résultats expérimentaux réalisés sur le corpus FAU AIBO Emotion montrent que le système WOC-NN obtient de meilleures performances de classification, et ce indépendamment du modèle utilisé, par rapport au système de référence GMM-*Bayes* ou comparé aux meilleurs systèmes de la compétition *INTERSPEECH 2009 Emotion Challenge*.

Le patron de similarité créé dans cette étude est basé sur un modèle de voisinage à une seule dimension. Il est possible de concevoir des patrons de similarité à deux dimensions ou plus qui simuleront les axes du modèle dimensionnel des émotions. Notons que dans ce cas, la méthode que nous avons proposée peut supporter ces nouveaux patrons multidimensionnels

sans grand changement. Il suffit de concaténer les patrons des différentes dimensions en un seul patron unidimensionnel de grande taille et de maintenir les autres étapes inchangées. Par ailleurs, ce principe de concaténation de patrons de similarité, peut être également être utilisé pour combiner plusieurs patrons de voisinage issus de différents classificateurs. Ainsi WOC-NN peut servir de méthode de combinaison de classificateurs ayant chacun son propre système frontal.

CHAPITRE 6

MODÈLES D'ANCRAGE POUR LA RECONNAISSANCE MUTICLASSES D'ÉMOTION

6.1 Introduction

Dans ce chapitre, nous allons étudier les modèles d'ancrage appliqués pour la reconnaissance des émotions à partir de la parole dans un contexte d'un problème multi-classe. Les modèles d'ancrage sont basés sur l'utilisation de vecteurs de traits basés sur la similarité que nous avons appelé les vecteurs de caractérisation d'émotion (VCE). Le système frontal, modélisé avec des modèles GMM, agit comme un extracteur de traits VCE. La classification d'un énoncé de test dépendra de la proximité de son vecteur VCE avec le VCE représentant chacune des classes d'émotion. Contrairement à la méthode WOC-NN décrite dans le chapitre précédent où une mesure qualitative (rangs des classes) est utilisée pour mesurer la proximité, les modèles d'ancrage étudiés dans ce chapitre sont basés sur des mesures de similarité quantitatives; euclidienne et cosinus. Les trois parties qui composent les modèles d'ancrage seront décrites dans la section 2. Les données dans l'espace d'ancrage sont bruitées, affectées particulièrement au niveau de leurs modules, tel que nous le montrerons dans la section 3. L'utilisation de métriques basées sur les magnitudes est par conséquent particulièrement problématique. Nous verrons dans la section 4, que l'introduction de la normalisation permettant de réduire la variance intraclasse a pour effet de réduire l'impact du bruit et d'améliorer les performances de classification. Nous étudierons dans la section 5, la représentativité des classes en termes du nombre de vecteurs VCE optimal pris comme référence pour représenter une classe lors de la mesure de la proximité avec la donnée de test. Dans la dernière section, nous allons montrer la puissance des modèles de référence pour classifier des émotions spontanées en comparant leurs performances à des classificateurs plus complexes.

6.2 Modèles d'ancrage

Dans un système de modèles d'ancrage, une classe d'émotion est caractérisée par son degré de similarité avec d'autres classes d'émotion. L'ensemble des modèles de classes d'émotion pris comme référence est appelé *modèles d'ancrage* et forme un *espace d'ancrage*. Trois étapes caractérisent la conception d'un système de modèle d'ancrage : la construction de l'espace d'ancrage, la projection des caractéristiques acoustiques dans l'espace d'ancrage et enfin la classification des énoncés émotionnels.

6.2.1 Construction de l'espace d'ancrage

Dans un problème de reconnaissance de formes où le nombre de classes est illimité tel qu'un problème de vérification de locuteur, c'est un sous-ensemble de locuteurs ou de locuteurs virtuels (obtenus par regroupement de locuteurs) le plus représentatif de l'ensemble de tous les locuteurs qui est utilisé pour générer les modèles de référence. Lorsque le problème en main implique un nombre limité de classes, comme c'est le cas pour la reconnaissance des émotions, nous avons la possibilité de modéliser l'ensemble des classes d'émotion. Ainsi, dans ce type de problème à catégories multiples, toutes les classes ont l'avantage d'être bien représentées dans l'espace d'ancrage. Nous pouvons souligner deux différences principales entre les modèles d'ancrage appliqués à la reconnaissance du locuteur et ceux appliqués à la reconnaissance des émotions. Tout d'abord, pour le cas de la reconnaissance du locuteur, l'espace d'ancrage possède une dimension élevée composée d'une centaine de modèles de locuteurs. Pour la RAE, la dimension de l'espace d'ancrage est beaucoup moins élevée en raison du nombre limité de classes d'émotion disponibles. Deuxièmement, en reconnaissance du locuteur, la personne à caractériser dans l'espace d'ancrage lors de la phase d'entrainement ou de test, n'appartient généralement pas à l'ensemble des modèles d'ancrage. Par contre en RAE, l'émotion de l'énoncé de test appartient déjà à l'ensemble des modèles d'ancrage étant donné que tous les modèles des classes d'émotion sont utilisés comme modèles d'ancrage.

Si chacune des C classes d'émotion est modélisée par un GMM λ_i, l'espace de référence pourrait être défini par l'ensemble $\Gamma=\{\lambda_1,\lambda_2,\cdots,\lambda_C\}$ ($\Gamma=\{\lambda_A,\lambda_E,\lambda_N,\lambda_P,\lambda_R\}$ si le corpus FAU

AIBO Emotion est pris comme exemple). Notons que nous avons opté pour l'utilisation des GMM plutôt que les modèles de Markov cachés (HMM) à la lumière des résultats précédents obtenus sur le corpus FAU AIBO Emotion. Pour l'ensemble des systèmes HMM étudiés dans (Schuller *et al.* 2009), les performances obtenues avec un HMM à un seul état (c.-à-d. un GMM) sont légèrement meilleures en comparaison avec un HMM à trois états et légèrement plus faibles comparé à un HMM à cinq états. Dans (Dumouchel *et al.* 2009), où un plus grand nombre de gaussiennes sont utilisées (plutôt que deux utilisés dans (Schuller *et al.* 2009)), le modèle GMM surpasse le HMM.

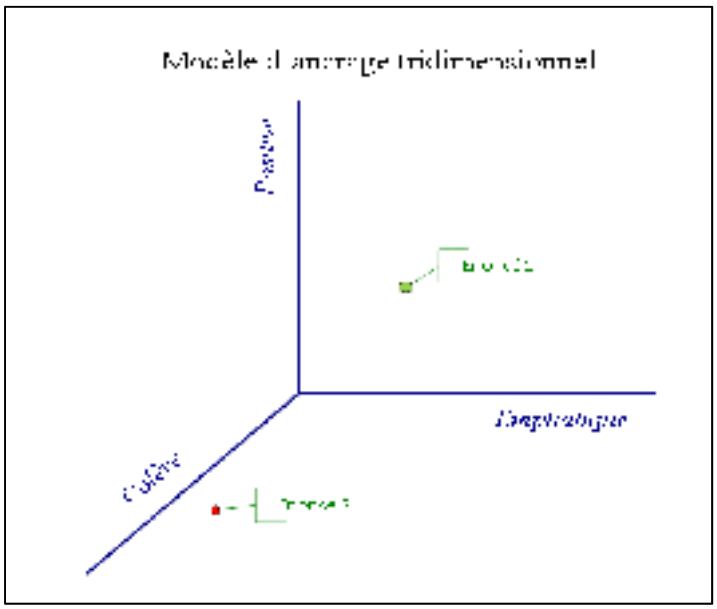

Figure 6.1 Exemple d'espace d'ancrage tridimensionnel engendré par les modèles des classes d'émotion *colère*, *emphatique* et *positive*

6.2.2 Mappage dans l'espace d'ancrage

Soit **X** un énoncé de parole émotionnelle représenté par une séquence de trames. **X** est projeté dans l'espace d'ancrage à travers le calcul de son vecteur VCE, L(**X**) :

$$\mathbf{L}(\mathbf{X}) = \begin{bmatrix} \frac{1}{T}\log P(\mathbf{X}|\lambda_1) \\ \vdots \\ \frac{1}{T}\log P(\mathbf{X}|\lambda_C) \end{bmatrix}, \tag{6.1}$$

où $\log P(\mathbf{X}|\lambda_i)$ représente le logarithme de la probabilité de vraisemblance de **X** étant donné le modèle GMM de la classe *i* appartenant à l'ensemble des modèles des classes $\{\mathbf{A}, \mathbf{E}, \mathbf{N}, \mathbf{P}, \mathbf{R}\}$.

Deux types de vecteurs VCE sont calculés en utilisant la formule (6.1) en fonction des valeurs de **X**; (i) vecteur VCE représentant une classe d'émotion calculé lors de la phase d'apprentissage et (ii) un vecteur représentant un énoncé de test en général calculé durant la phase de prédiction. Un vecteur représentant d'une classe d'émotion *i* est estimé en utilisant tous les énoncés d'entrainement de la classe *i* selon l'équation (6.2):

$$\mathbf{L}^i = \frac{1}{n_i}\sum_{q}^{n_i} \mathrm{L}(\mathbf{X}_q^i) \tag{6.2}$$

où $\mathbf{X}_q^i$ représente le *q*-ième énoncé de la classe *i* et n_i le nombre d'énoncés d'entrainement de la classe *i*. La Figure 6.1 illustre un exemple de deux vecteurs VCE représentants deux énoncés projetés dans un espace d'ancrage à trois dimensions engendré par les modèles des classes d'émotion *colère*, *emphatique* et *positive*.

6.2.3 Classification des énoncés émotionnels

Pour classifier un énoncé de test, la distance entre le vecteur VCE de la donnée de test et le VCE représentant chacune des classes est calculée en utilisant une mesure de similarité. Deux métriques sont expérimentées dans nos travaux : les mesures euclidienne et cosinus définies comme suit :

- **Métrique euclidienne**

$$\mathrm{d}(\mathbf{L}_1, \mathbf{L}_2) = \sqrt{|\mathbf{L}_1 - \mathbf{L}_2|^2} \tag{6.3}$$

- **Métrique cosinus**

$$\mathrm{d}(\mathbf{L}_1, \mathbf{L}_2) = 1 - \frac{\langle \mathbf{L}_1, \mathbf{L}_2 \rangle}{\|\mathbf{L}_1\| \|\mathbf{L}_2\|} \tag{6.4}$$

où $\langle \mathbf{L}_1, \mathbf{L}_2 \rangle$ est le produit scalaire des vecteurs $\mathbf{L}_1$ et $\mathbf{L}_2$. Dans cette section, chacune des C classes d'émotion est représentée durant la phase de test par un vecteur VCE unique. La règle de décision est formulée comme suit:

$$\text{emotion} = \underset{i=1,\ldots,C}{\arg\min}(\mathrm{d}(\mathbf{L}_T, \mathbf{L}_i)) \tag{6.5}$$

où d représente la métrique utilisée pour calculer la distance entre $\mathbf{L}_T$, le VCE de la donnée de test, et $\mathbf{L}_i$ le VCE représentant la classe d'émotion i.

6.3 Configuration expérimentale

Les performances des modèles d'ancrage sont évalués à la fois pour la mesure euclidienne et la mesure cosinus. Les paramètres des modèles, tels que le nombre de composantes gaussiennes des GMMs sont ajustés sur la base des données d'entraînement en utilisant le protocole de la validation croisée à neuf plis. Les résultats sont optimisés selon UAR comme premier critère suivi de WAR.

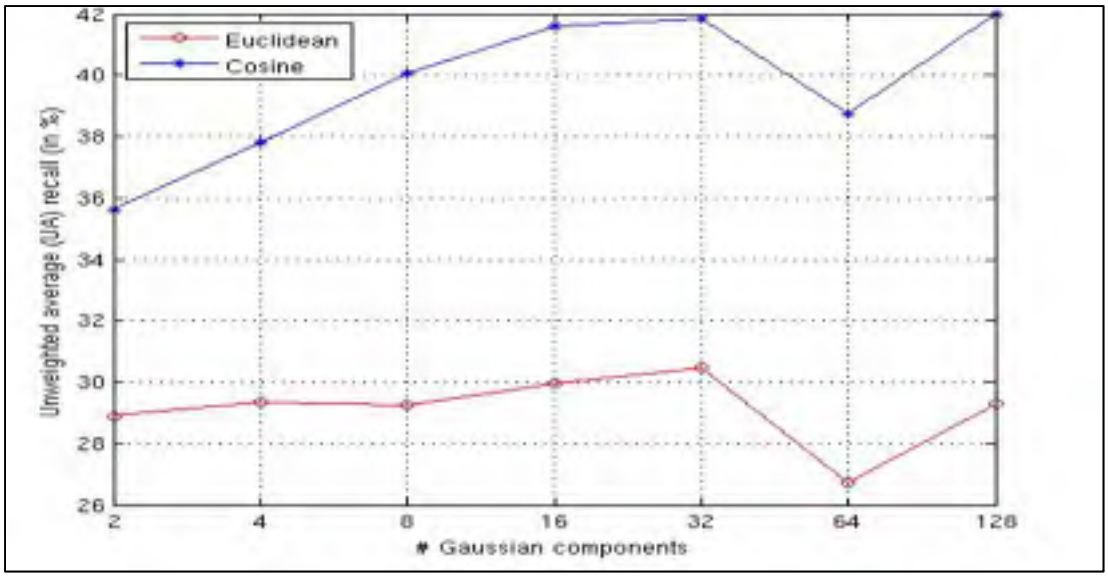

Figure 6.2 Résultats UAR obtenus en utilisant la validation croisée à 9 plis sur les données d'entrainement du corpus FAU AIBO Emotion en fonction du nombre de gaussiennes du GMM. Deux systèmes de modèles d'ancrage sont comparés : un basé sur la distance euclidienne et le second sur la similarité cosinus

La Figure 6.2 montre les résultats des modèles d'ancrage obtenus pour les deux métriques évalués en utilisant la validation croisée à neuf plis sur les données d'apprentissage. Nous observons que le modèle d'ancrage basé sur la distance euclidienne réalise de très faibles performances par rapport au système basé sur la similarité cosinus. Les résultats suggèrent que les énoncés d'émotion projetés dans l'espace d'ancrage sont plus discernables à travers leurs directions plutôt que par leurs magnitudes. Cela implique que le bruit affecte plus les traits dans leurs modules que dans leurs angles.

Pour illustrer l'ampleur du bruit sur les magnitudes des vecteurs **L**, nous avons représenté dans la Figure 6.3 la moyenne et la variance des valeurs cartésiennes de chacune des variables des vecteurs VCE. Les données d'entrainement sont utilisées pour calculer les statistiques de chacune des classes d'émotion dans les graphiques (a) à (c). Pour un vecteur VCE représentatif d'une classe d'émotion donnée, *k*, on peut s'attendre à ce que la composante (c.-à-d. la variable égale à la moyenne des logarithmes des probabilités des données d'une classe d'émotion) associée au modèle de classe *k* ait la valeur maximale dans ce vecteur si les modèles sont bien entrainés. Pour les autres composantes du vecteur, plus un

modèle d'une autre classe d'émotion est proche au modèle de la classe *k*, plus le logarithme de la probabilité de vraisemblance est plus élevé et vice- versa. L'importance du score reflète le degré de similarité d'une classe donnée relativement à d'autres classes.

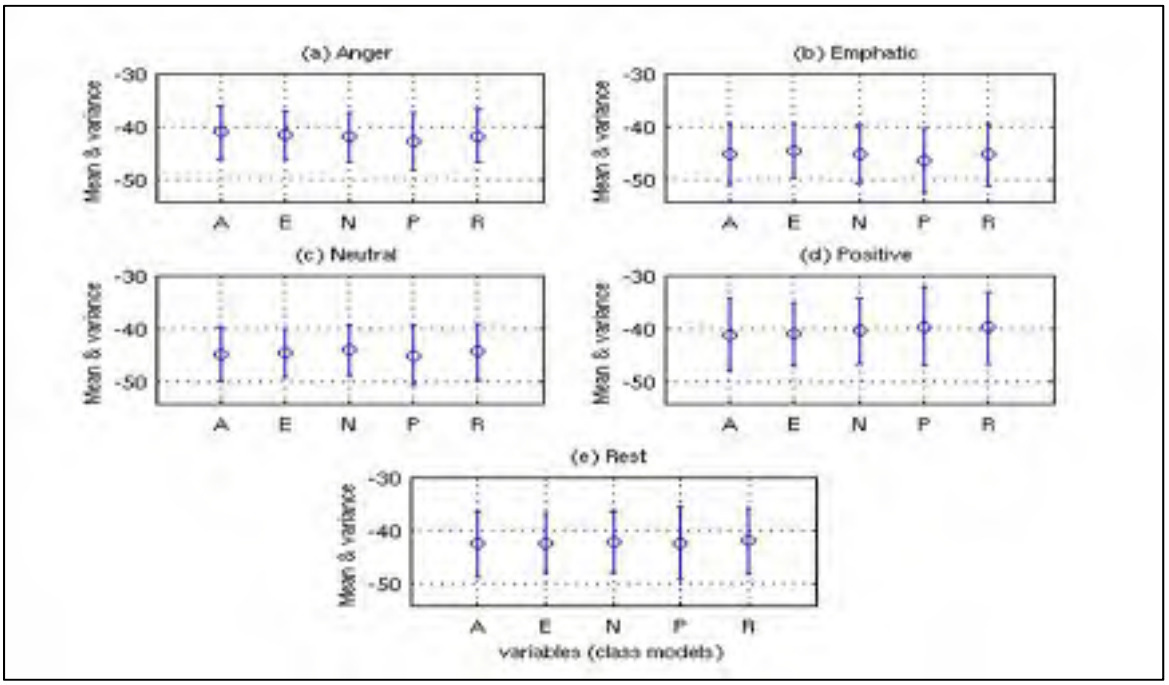

Figure 6.3 Dans chaque graphique sont tracées les valeurs de la moyenne et de la variance des scores de vraisemblance des données des classes d'émotion en fonction de chacune des dimensions de l'espace d'ancrage (modèle d'émotion composant les vecteurs VCE). Dans cette figure, les valeurs tracées représentent les valeurs statistiques des coordonnées cartésiennes des vecteurs VCE (voir l'équation (6.1))

Par ailleurs, nous avons tracé dans la Figure 6.4 la moyenne et la variance de la valeur angulaire des variables représentants les coordonnées cartésiennes des vecteurs VCE cartographiés dans la Figure 6.3. La *i*-ième variable du vecteur angulaire représente l'angle entre le vecteur VCE et le *i*-ième axe de l'espace euclidien. Formellement, le vecteur angulaire d'un vecteur **L** est calculé comme suit. Soit $\mathbf{L} = (l_1, l_2, \ldots, l_C)^T$ un vecteur VCE d'un énoncé et $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_C)$ la base standard où $\mathbf{e}_1 = (1,0,\ldots,0)^T, \ldots,$ et $\mathbf{e}_C = (0,0,\ldots,1)^T$. La valeur angulaire entre **L** et le *i*-ième axe standard est égale à:

$$\text{angle}\left(\mathbf{L},\mathbf{e}_i\right)=\arccos\left(\frac{<\mathbf{L},\mathbf{e}_i>}{\|\mathbf{L}\|\times\|\mathbf{e}_i\|}\right); \tag{6.6}$$

après simplification, nous obtenons:

$$\text{angle}\left(\mathbf{L},\mathbf{e}_i\right)=\arccos\left(\frac{l_i}{\|\mathbf{L}\|}\right) \tag{6.7}$$

Nous avons également:

$$\cos\left(\text{angle}(\ \mathbf{L},\mathbf{e}_i)\right)=\frac{l_i}{\|\mathbf{L}\|} \tag{6.8}$$

Figure 6.3 révèle que les variables (qui représentent la sortie de la fonction de densité des modèles GMM des classes d'émotion) ne sont pas discriminantes par leurs valeurs cartésiennes notamment en raison de l'étendue de leurs variances. Par ailleurs, le problème de l'étendue de la variance est beaucoup moins important pour les valeurs angulaires des variables telles que représentées dans la Figure 6.4.

Il est intéressant de noter que (6.8) ressemble à la formule de la normalisation de la longueur qui a été récemment utilisée dans (Dehak, Kenny *et al.* 2011; Garcia-Romero et Espy-Wilso, 2011) en tant qu'étape de prétraitement pour améliorer les performances de reconnaissance du locuteur. En fait, la normalisation de la longueur revient à calculer le cosinus de l'angle de la variable qui est plus discriminant tel qu'illustré par les Figure 6.3 et Figure 6.4. Dans la section suivante, nous chercherons à déterminer et à expliquer la source du bruit affectant particulièrement la magnitude des vecteurs VCE.

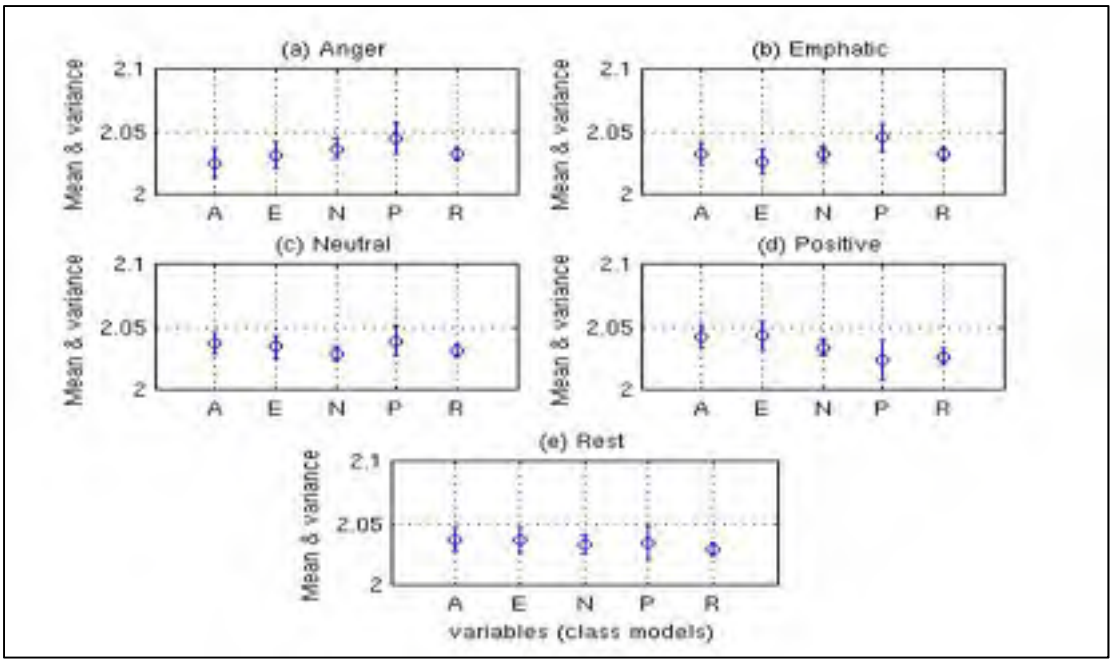

Figure 6.4 Dans chaque graphique sont tracées la moyenne et la variance des scores de vraisemblance des données d'une classe d'émotion à l'égard de chaque modèle d'émotion (composantes des vecteurs VCE). Dans cette figure, les valeurs tracées représentent les valeurs statistiques des valeurs angulaires des vecteurs VCE par rapport à la base standard (voir l'équation (6.7))

6.4 Problème des données bruitées avec la métrique euclidienne

Tel que nous l'avons observé, les distances euclidiennes sont sujettes à une importante distorsion visible à travers l'étendue de la variance caractérisant les classes et par la chute de performance en comparaison avec les systèmes GMM-*Bayes* ou avec les modèles d'ancrage basés sur la similarité cosinus. Notons, que la faiblesse des performances obtenues avec la distance euclidienne en comparaison avec le cosinus est également rapportée dans plusieurs études en reconnaissance du locuteur (Collet *et al.* 2005a; Collet *et al.* 2005b; Mami *et al.* 2002; Yang *et al.* 2006). Nous montrerons dans cette section, que le bruit est en fait introduit par l'opérateur du logarithme utilisé dans le calcul de la probabilité de vraisemblance. L'application du logarithme est une étape importante dans les calculs car non seulement il permet de simplifier l'estimation des probabilités du point de vue arithmétique, mais aussi il

permet de prévenir le problème de soupassement de capacité (*arithmetic underflow*) des machines lors du calcul du produit des probabilités des trames de parole. Le logarithme étant une fonction continue et monotone, il est possible de substituer le maximum des probabilités dans la règle de décision *Bayes* par le maximum des logarithmes de probabilités étant donné que $\underset{i=1,\ldots,C}{\operatorname{argmin}}(P_i) = \underset{i=1,\ldots,C}{\operatorname{argmin}}\left(\log(P_i)\right)$. Cependant, quand l'opérateur du logarithme est appliqué dans le calcul de la distance euclidienne, l'hypothèse stipulant que :

$$\left\|\mathbf{x}-\mathbf{L}_i\right\| < \left\|\mathbf{y}-\mathbf{L}_j\right\| \quad \Rightarrow \quad \left\|\log(\mathbf{x})-\log(\mathbf{L}_i)\right\| < \left\|\log(\mathbf{y})-\log(\mathbf{L}_j)\right\| \tag{6.9}$$

est implicitement endossée. Une telle hypothèse n'est en réalité pas toujours vraie. La Figure 6.5 illustre un exemple où la distance euclidienne séparant un point **x** d'un point **a** est plus petite que la distance séparant **x** de **b** sur une échelle linéaire alors que sur l'échelle logarithmique, le point **x** devient plus proche du point **b** que du point **a**.

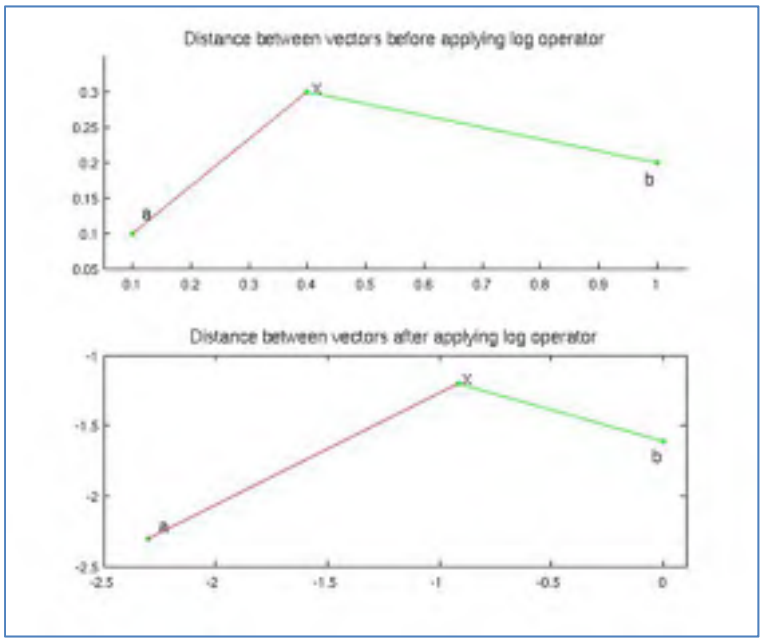

Figure 6.5 Exemple où la distance euclidienne séparant un point **x** d'un point **a** est plus petite que la distance séparant **x** de **b** sur une échelle linéaire alors que sur l'échelle logarithmique, le point **x** devient plus proche du point **b** que du point **a**

Afin d'étudier comment la fonction du logarithme altère les distances euclidiennes et déterminer l'allure de cette altération, nous avons tracé dans la Figure 6.6 un graphique 3D montrant l'évolution des valeurs de distances entre deux variables unidimensionnelles *x* et *y* pour toutes les valeurs de probabilité possibles *x* et *y* comprises entre 0 et 1. L'allure de cette évolution est illustrée pour les valeurs *x* et *y* avant et après l'application du logarithme.

Dans le graphique supérieure de la Figure 6.6, nous observons que la contribution des variables *x* et *y* dans le calcul de la distance est proportionnelle à la différence entre les deux variables indépendamment des valeurs *x* et *y*. Dans le domaine logarithmique, nous observons que l'allure de la contribution des variables dans le calcul de la distance obéit à un autre schéma. D'une part, pour une même valeur de différence entre *x* et *y*, sa contribution dans le calcul de la distance n'aura pas le même impact. En effet, quand les valeurs de probabilités de *x* et *y* sont proches de la valeur 1, la contribution de leurs différence est moins importante que si les valeurs de *x* et *y* étaient plus proches de zéro (côté aplati près du demi plan contenant le point (1,1) et surélevé du côté du demi-plan contenant le point (0,0)). D'autre part, comme deuxième différence, nous constatons que la contribution dans le calcul de la distance n'est plus linéairement proportionnelle à la différence de valeurs entre *x* et *y*. Les petites différences entre *x* et *y* possèdent un poids plus important que les grandes différences entre *x* et *y* (effet d'exagération des petites différences et atténuation de l'apport des grandes différences.

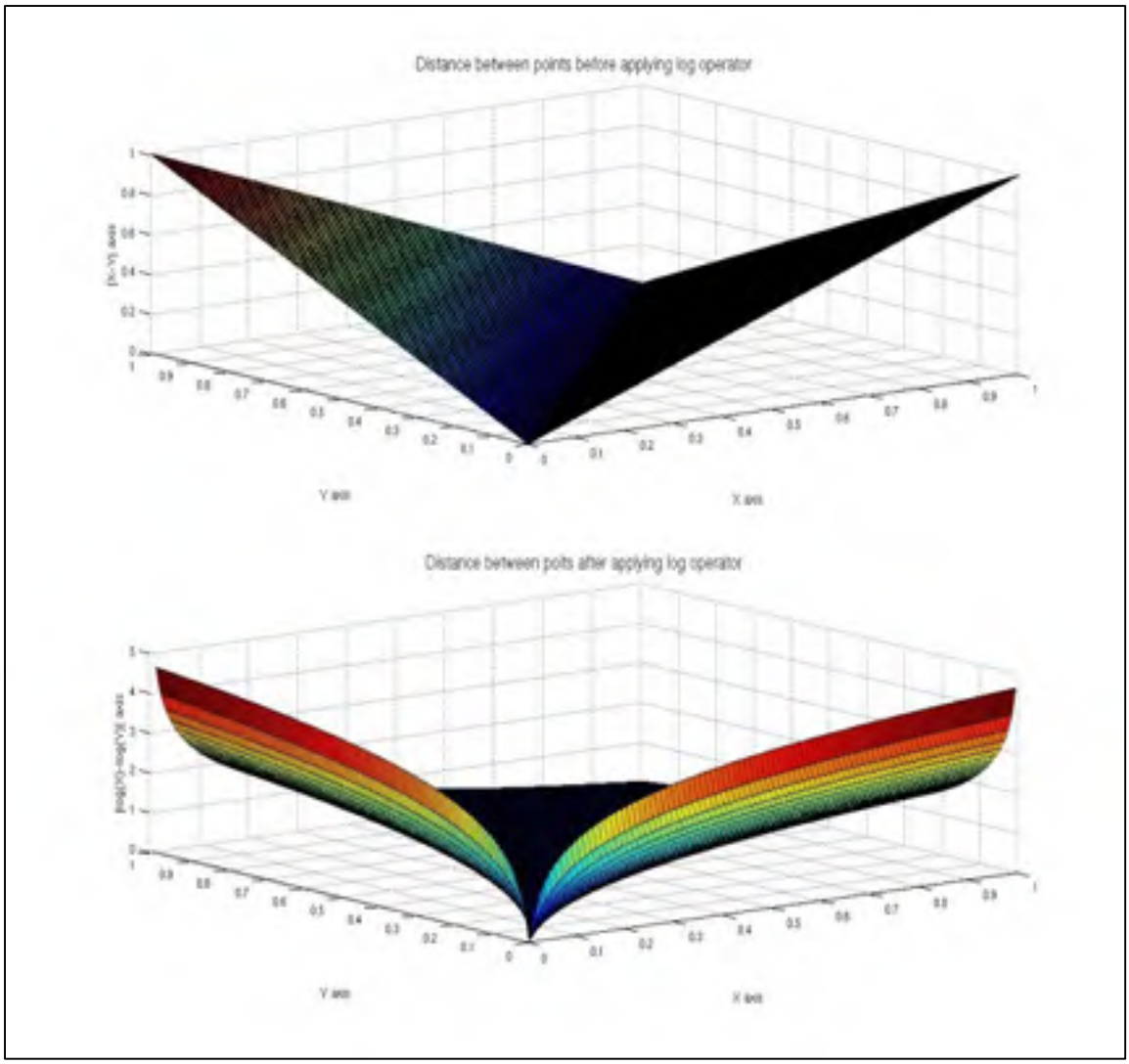

Figure 6.6 Graphique 3D montrant l'évolution des valeurs de distances entre deux variables unidimensionnelles x et y pour toutes les valeurs de probabilité possibles x et y comprises entre 0 et 1. L'allure de cette évolution est illustrée pour les valeurs x et $\boldsymbol{y}$ avant (graphique du haut) et après (graphique du bas) l'application du logarithme

Ce sont ces deux caractéristiques de l'opérateur logarithmique appliqué à la différence de valeurs qui font que l'hypothèse (6.9) n'est pas satisfaite, ce qui explique les mauvaises performances des modèles d'ancrage utilisés avec la similarité euclidienne. Dans la section suivante, nous allons traiter le problème de la variance à travers la normalisation de la covariance intraclasse (*Within-Class Covariance Normalization*, WCCN) afin d'augmenter la séparabilité entre les différentes classes.

6.5 Normalisation des scores

6.5.1 Normalisation de la covariance intraclasse

WCCN est une technique introduite dans (Hatch et Stolcke, 2006) pour entrainer un noyau linéaire généralisé d'un système à base de SVM afin de minimiser les erreurs de type faux positifs et faux négatifs attendus. Le noyau linéaire généralisé $k(\mathbf{L}_1, \mathbf{L}_2)$ est formulé comme suit :

$$k(\mathbf{L}_1, \mathbf{L}_2) = \mathbf{L}_1^t \, \mathbf{R} \, \mathbf{L}_2 \tag{6.10}$$

où $\mathbf{L}_1$ et $\mathbf{L}_2$ sont deux instances de données, et $\mathbf{R}$ est une matrice positive semi-définie. La résolution analytique passe par la substitution de $\mathbf{R}$ par $\mathbf{W}^{-1}$ ($\mathbf{R} = \mathbf{W}^{-1}$), où $\mathbf{W}$ est la matrice de covariance intraclasse des données définie comme suit:

$$\mathbf{W} = \sum_{i=1}^{C} p(i) \cdot \mathbf{S}_i \tag{6.11}$$

où $p(i)$ et $\mathbf{S}_i$ représentent respectivement la probabilité a priori et la matrice de covariance de la classe *i*. La décomposition de *Cholesky* nous permet d'écrire $\mathbf{W}^{-1}$ sous la forme $\mathbf{A}\mathbf{A}^T = \mathbf{W}^{-1}$, $\mathbf{A}$ est une matrice triangulaire inférieure. Les nouvelles métriques euclidienne et cosinus sont formulées comme suit après introduction de la normalisation WCCN :

- **Métrique euclidienne :**

$$d(\mathbf{L}_1, \mathbf{L}_2) = \left[\mathbf{A}^T(\mathbf{L}_1 - \mathbf{L}_2)\right]^T \left[\mathbf{A}^T(\mathbf{L}_1 - L_2)\right] \tag{6.12}$$

- **Métrique cosinus :**

$$d(\mathbf{L}_1, \mathbf{L}_2) = 1 - \frac{(\mathbf{A}^T\mathbf{L}_1)^T (\mathbf{A}^T\mathbf{L}_2)}{\left\|\mathbf{A}^T\mathbf{L}_1\right\| \left\|\mathbf{A}^T\mathbf{L}_2\right\|} \tag{6.13}$$

La normalisation WCCN a été appliquée avec succès dans l'espace des traits *i*-vecteurs (Dehak *et al.* 2011). Un *i*-vecteur est une représentation de faible dimension d'un supervecteur. WCCN a été également appliquée dans (Zhao *et al.* 2007) afin d'améliorer les performances des SVM ayant comme caractéristiques d'entrée les probabilités de vraisemblance. Enfin, une méthode de normalisation proche de WCCN, appelée normalisation vectorielle Z, a été introduite dans (Charlet *et al.* 2007) et qui a été appliquée pour le problème de la vérification du locuteur en utilisant les modèles d'ancrage. La norme VZ est une extension de la norme Z au cas multivarié et qui vise à normaliser les scores contre la variabilité intra-locuteur.

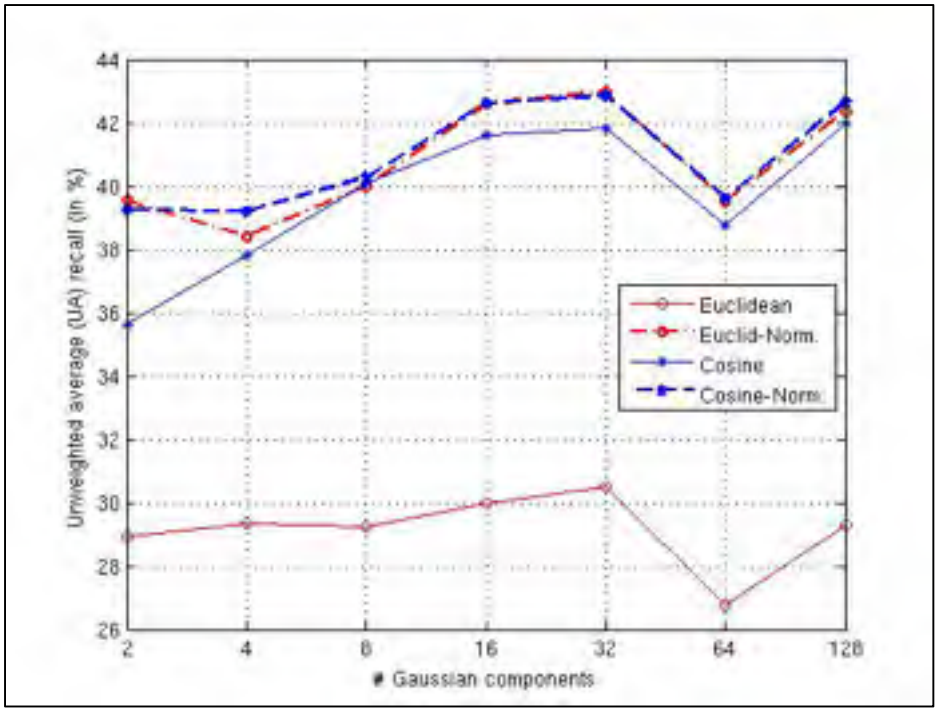

Figure 6.7 Effet de la normalisation WCCN sur les performances UAR des modèles d'ancrage en fonction du nombre de gaussiennes des GMMs. Les résultats sont obtenus en utilisant la validation croisée à neuf plis sur les données d'apprentissage du corpus FAU AIBO Emotion

6.5.2 Résultats et discussion

La Figure 6.7 montre les résultats de classification pour les modèles d'ancrage, basés sur les métriques euclidienne et cosinus avant et après la normalisation WCCN, évalués sur les données d'apprentissage à l'aide de la validation croisée à neuf plis. Tout d'abord, nous observons que WCCN améliore les performances des deux métriques. Cette amélioration est beaucoup plus importante pour la distance euclidienne. Le gain relatif est de l'ordre de 3,3 % et 40 % pour le système à base de cosinus et euclidienne respectivement. Il est également intéressant de noter qu'après application de la normalisation, les systèmes à base des mesures euclidienne et cosinus montrent des performances similaires. Les meilleures performances sont obtenues pour les deux mesures avec un modèle GMM de 32 gaussiennes utilisé comme système frontal des modèles d'ancrage. En conséquence, le même nombre de composantes de gaussiennes sera utilisé pour les expériences de test.

Pour visualiser l'effet de la normalisation WCCN sur les données (scores des logarithmes des probabilités de vraisemblance), nous avons tracé dans la Figure 6.8 la distribution des données de la classe *colère* (**A**) pour chaque dimension (modèles d'émotion) avant et après normalisation. Dans le haut de la Figure 6.8, nous observons que la distribution des scores des données de la classe **A** générés par son propre modèle GMM est similaire aux distributions des scores de **A** mais générés par les modèles des autres classes. Cette similitude dans le comportement des modèles rend difficile l'extraction d'information discriminante par comparaison des scores d'un énoncé face aux différents modèles de classe, l'idée de base sur laquelle est basée l'utilité les modèles d'ancrage.

Observons maintenant la distribution des scores après normalisation telle qu'illustrée dans le graphique du bas de la même figure. Nous constatons que WCCN a eu pour effet de maximiser la capacité discriminative des modèles dans l'espace d'ancrage, comme le montre cette répartition plus uniforme des modèles sur les valeurs possibles de l'espace d'ancrage.

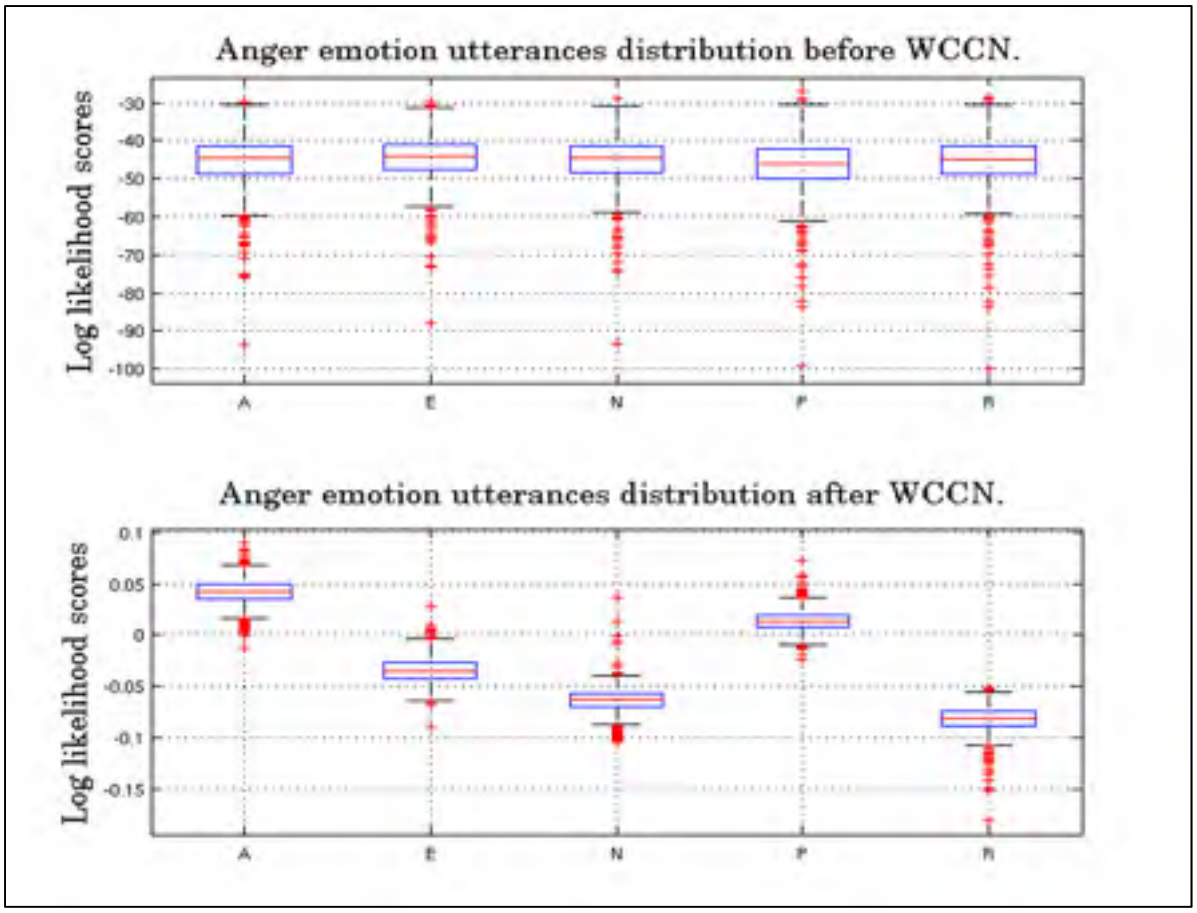

Figure 6.8 Diagramme à surfaces de la distribution des scores des énoncés de la classe *colère* (A) pour les cinq modèles d'émotion avant (en haut) et après (en bas) la normalisation WCCN. Sur chaque boîte, la marque centrale représente la médiane, les bords les 25e et 75e percentiles, les moustaches (*whiskers*) indiquent les données les plus extrêmes, et les valeurs aberrantes sont représentées individuellement

Le Tableau 6.1 donne les résultats de classification des données de test avec les modèles d'ancrage. Les modèles utilisés dans la phase de test ainsi que la matrice **A** de WCCN sont entrainés avec toutes les données d'apprentissage. On constate que les observations émises sur les résultats obtenus avec les données d'entraînement s'étendent aussi sur les données de test. WCCN améliore significativement les performances pour les deux mesures : euclidienne et cosinus. On note également que les systèmes d'ancrage basés sur la similarité euclidienne et cosinus atteignent des performances comparables après normalisation. Il est intéressant d'observer que les résultats obtenus pour les données de test sont meilleures que celles obtenus pour les données d'apprentissage. Ceci s'explique par le fait que les modèles GMM

utilisés durant la phase de test sont plus robustes que ceux utilisés avec le protocole de la validation croisée. Pour les modèles de test, nous avons utilisé neuf partitions au lieu des huit utilisées dans la validation croisée à neuf plis, soit l'ajout des données de trois locuteurs supplémentaires. Pour vérifier cette assertion, nous avons évalué les données de test en utilisant les mêmes modèles utilisés pour l'évaluation de données d'apprentissage. Les résultats obtenus avec le modèle d'ancrage basé sur la distance euclidienne par exemple, chutent de 44,19 % à 42,18 %, ce qui donne des performances plus basses que celles obtenues par les données d'apprentissage. Ce résultat confirme notre assertion et souligne l'importance d'avoir plus de données et plus de locuteurs pour un développement plus robuste des systèmes de RAE.

Tableau 6.1 Résultats des différents modèles d'ancrage évalués avec les données de test du corpus FAU AIBO Emotion

Systèmes modèles d'ancrage	**UAR**	WAR
Cosinus	42,25 %	33,57 %
Euclidienne	26,59 %	23,00 %
Cosinus + WCCN	43,91 %	46,01 %
Euclidienne + WCCN	**44,19** %	47,44 %

6.6 Vecteurs représentatifs des classes

Une différence majeure entre les modèles d'ancrage présentés dans cette étude et la méthode du *k* plus proches voisins est le type de points de données d'entraînement auxquelles sont comparées les données de test durant la phase de classification. Dans la méthode KNN, la classification est approximée localement et elle est basée sur les *k* plus proches exemples (classification basée sur des instances). Un inconvénient majeur de cet algorithme réside dans sa sensibilité envers les valeurs aberrantes contenues dans les données d'entrainement. Le modèle d'ancrage offre l'avantage de comparer les données de test à un vecteur plus fiable, fixe et global utilisé en tant que représentant de la classe. Ce représentant est calculé lors de

l'étape d'apprentissage en utilisant toutes les données d'apprentissage. Dans la section suivante nous étudierons l'effet d'augmenter le nombre de représentants sur les performances de classification.

6.6.1 Représentant unique versus représentants multiples

Dans la conception des modèles d'ancrage présentés dans les sections précédentes, chaque classe d'émotion a été représentée par un vecteur VCE unique. Une autre alternative consiste à représenter chaque classe d'émotion par un ensemble de vecteurs représentatifs. Un modèle avec des représentants de classe multiples pourrait être particulièrement utile lorsque les données ont une distribution multimodale. Dans cette section, nous étudions l'impact de dupliquer le nombre de vecteurs représentatifs sur les performances des modèles d'ancrage. Dans le reste de ce chapitre, nous nommerons *multiplis* (*multifold*) le système basé sur plus d'un vecteur représentatif par opposition au système à plis unique (*unifold*) basé sur un représentant unique.

Considérons $\{\mathbf{L}_1^i, \mathbf{L}_2^i, \cdots, \mathbf{L}_r^i\}$ comme étant l'ensemble des vecteurs représentatifs d'une classe d'émotion $\mathbf{E}_i$. La règle de décision appliquée au système *multiplis* s'écrira alors comme suit :

$$\text{emotion} = \arg\min_{i=1,\ldots,C} \left(\sum_{j=1}^{r} \mathrm{d}\left(\mathbf{L}_T, \mathbf{L}_i^j\right) \right) \tag{6.14}$$

où $\mathbf{L}_i^j$ représente le *j*-ième vecteur représentatif de la classe d'émotion $\mathbf{E}_i$. Différentes méthodes peuvent être utilisées pour sélectionner les représentants des classes d'un système *multiplis*. Trois méthodes seront étudiées et leurs performances comparées à celles d'un système à plis unique :

1. **Sélection aléatoire** : De chaque classe d'émotion, $r\,(r \geq 2)$ énoncés (vecteurs VCE) sont choisis aléatoirement à partir de ses données d'entraînement. Les performances sont par la suite évaluées pour cet ensemble de représentants. Ce processus peut se

répéter plusieurs fois pour prendre finalement la moyenne des évaluations de l'ensemble des itérations.

2. **Regroupement** : Les données d'entrainement de chaque catégorie d'émotion sont regroupées en *r* grappes (*clustering*) sur la base de la distance séparant leurs valeurs VCE. Le vecteur de la moyenne de chaque grappe est pris comme un représentant de cette classe.
3. **Regroupement pondéré** : Nous étudions également une version pondérée de la méthode de regroupement. L'objectif est de réduire l'effet (le poids) des grappes composées de données aberrantes. La contribution de chaque pôle dans le calcul de la distance est proportionnelle au nombre d'instances dans la grappe. La nouvelle règle de décision adaptée à la pondération des grappes est formulée alors comme suit:

$$\text{emotion} = \arg\min_{i=1,\ldots,C}\left(\sum_{j=1}^{r}\frac{n_i^j}{n_i}\times \mathrm{d}\left(\mathbf{L}_T,\mathbf{L}_i^j\right)\right) \tag{6.15}$$

où n_i représente la taille des données d'entrainements de la classe *i* et n_i^j la taille de la grappe *j* de la classe *i*.

6.6.2 Résultats expérimentaux

Afin de déterminer la valeur du nombre de représentants par classe des systèmes *multiplis*, nous allons effectuer d'abord des expériences sur les données d'entrainement en augmentant le nombre de vecteurs représentatifs utilisés par classe. Pour le système *multiplis* basé sur la sélection aléatoire, 50 itérations sont exécutées. À chaque itération, un nouveau sous-ensemble de vecteurs représentatifs de classe est choisi au hasard et les performances UAR sont évaluées. Les moyennes des cinquante exécutions sont calculées et tracées dans la Figure 6.9. Nous observons une augmentation de performance à mesure que le nombre de représentants par classe augmente jusqu'à une valeur moyenne de 150 vecteurs pour lesquels les meilleurs résultats sont atteints pour les modèles d'ancrage avant normalisation WCCN.

Pour les versions des systèmes normalisés, les performances ne cessent de s'améliorer lentement jusqu'à atteindre 450 vecteurs.

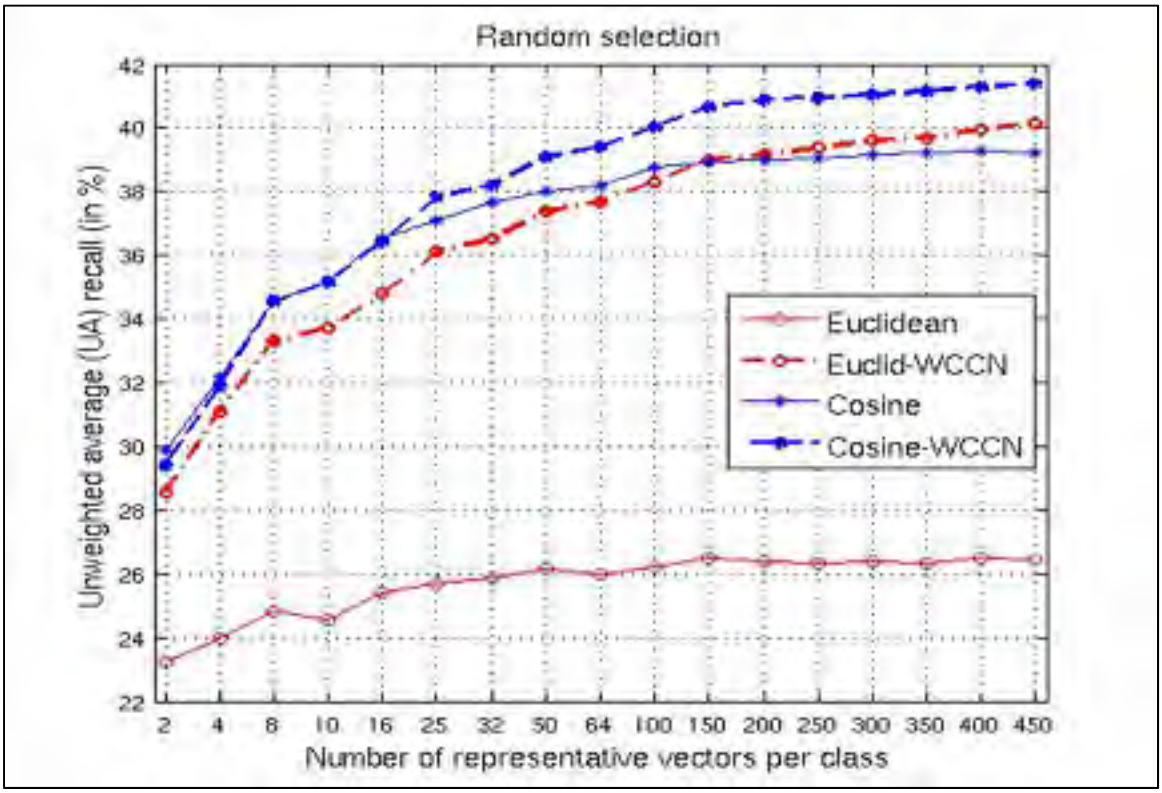

Figure 6.9 Résultats UAR moyenne de 50 exécutions des modèles d'ancrage en fonction du nombre de vecteurs VCE sélectionnés comme vecteurs représentatifs de classe. À chaque itération, un nouveau sous-ensemble des données d'entraînement est aléatoirement sélectionné comme classe vecteurs représentatifs. Les performances sont évaluées sur les données d'entrainement en utilisant la validation croisée à neuf plis

Les performances des systèmes *multiplis* basés sur une sélection par regroupement sont affichées dans la Figure 6.10. Les meilleures performances sont atteintes avec seulement deux grappes et les performances peuvent baisser de manière drastique pour un plus grand nombre de grappes. En outre, lorsque les grappes sont pondérées proportionnellement à leurs tailles, les performances sont plus stables au changement du nombre de grappes, comme le montre la Figure 6.11. Les performances deviennent alors moins sensibles aux grappes composées de valeurs aberrantes, bien que les meilleurs résultats ne se soient pas améliorés suite à l'opération de pondération.

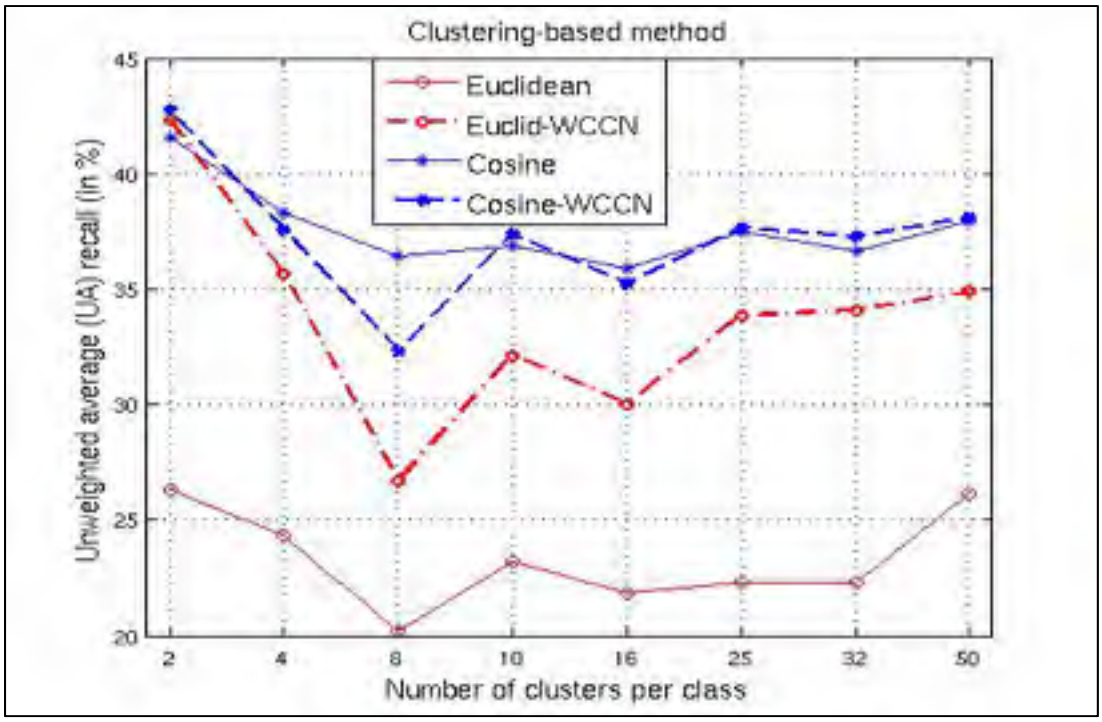

Figure 6.10 Résultats UAR des modèles d'ancrage en fonction du nombre de grappes par classe. Les centres de grappes sont comme utilisés vecteurs représentatifs de classe. Les performances sont évaluées sur les données d'apprentissage en utilisant la validation croisée à neuf plis

Les résultats obtenus sur les données de test sont rapportés dans le Tableau 6.2. Comme on peut le constater, l'augmentation du nombre de vecteurs représentatifs n'améliore pas les performances, et ce indépendamment de la méthode de sélection des vecteurs. Ce résultat suggère que les données peuvent être traitées comme ayant une distribution *unimodale*.

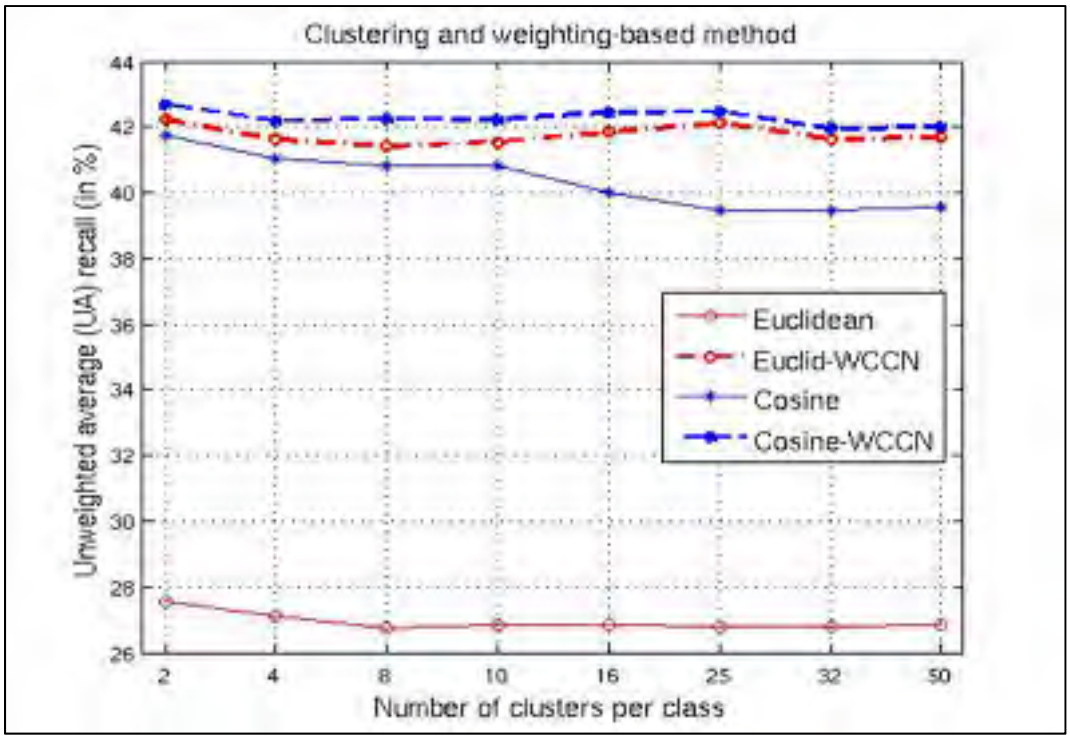

Figure 6.11 Résultats UAR des modèles d'ancrage en fonction du nombre de pôles par classe utilisées comme vecteurs représentatifs. Les grappes sont pondérées par une valeur proportionnelle à la taille de la classe. Les performances sont évaluées sur les données d'entrainement en utilisant la validation croisée à neuf plis

Tableau 6.2 Comparaison des modèles d'ancrage de type *multifold* et *unifold* évalués sur les données de test du corpus FAU AIBO Emotion

Système	**UAR**	**WAR**
Aléatoire	42,55 %	45,40 %
Grappes	43,94 %	48,84 %
Grappes-pondérées	43,41 %	49,73 %
Unifold	**44,19 %**	47,44 %

6.7 Comparaison avec des systèmes dorsaux plus complexes

Les valeurs de probabilités de vraisemblance calculées en utilisant les modèles GMM pourraient être utilisées directement comme des scores de classement finaux à l'aide de la règle de décision de *Bayes* (premier type d'architecture de système). Dans le second type d'architecture, les valeurs de probabilités pourraient être considérées comme des traits de haut niveau alimentant (des entrées) un classificateur additionnel placé en aval, quoique simple (démuni de phase d'apprentissage). L'algorithme KNN et les modèles d'ancrage basés sur des métriques sont des exemples d'une telle architecture. Dans le troisième type d'architecture, les scores de probabilité sont utilisés comme entrées pour un système dorsal plus complexe ayant un algorithme d'apprentissage plus sophistiqué comme par exemple les SVM ou les réseaux neuronaux multicouches (MLP). Dans cette section, nous déterminerons laquelle des trois architectures (voir Figure 6.12) est plus pertinente pour le problème de la RAE à partir de la parole spontanée.

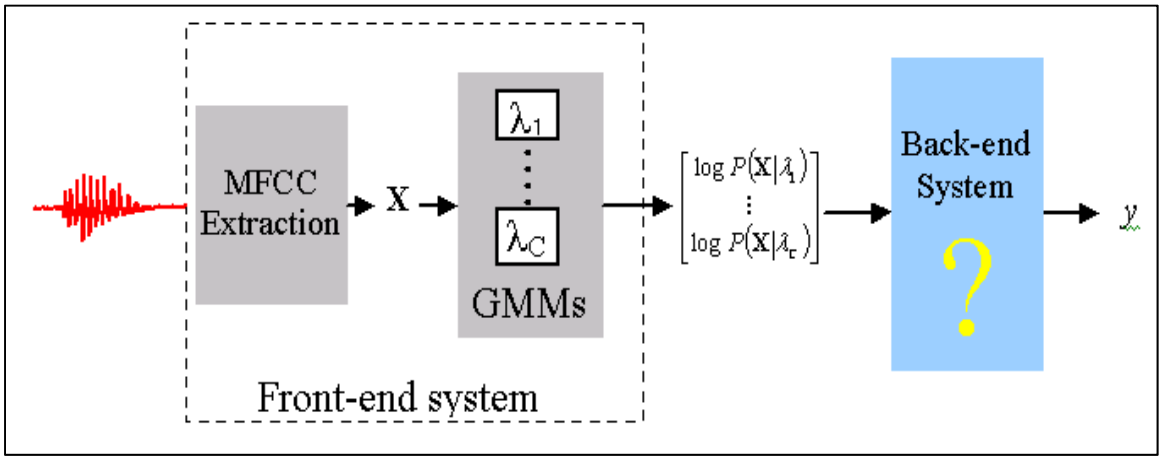

Figure 6.12 Types d'architectures de systèmes basés sur les scores du logarithme de probabilité de vraisemblance

6.7.1 Traitement du problème de distribution biaisée des classes de données

Quand un modèle discriminant comme SVM est utilisé comme classificateur nous avons particulièrement besoin de traiter le problème de la distribution déséquilibrée des classes de données. Sans les techniques d'échantillonnage de données, les performances de classification seront biaisées en faveur de la classe majoritaire au détriment des autres classes. Plusieurs méthodes ont été proposées dans la littérature pour atténuer l'impact d'une distribution asymétrique des classes :

1. le sous-échantillonnage en réduisant la taille de la classe majoritaire à celle de la classe minoritaire (Weiss et Provost, 2001);
2. le suréchantillonnage en générant de nouveaux échantillons de la classe minoritaire en utilisant des algorithmes tels que SMOTE (Synthetic Minority Oversampling Technique) (Chawla *et al.* 2002);
3. l'échantillonnage d'ensemble (Yan *et al.* 2003) qui consiste à sous-échantillonner la classe majoritaire en un ensemble de sous-ensembles de données. Chaque sous-ensemble est utilisé pour entrainer un classificateur séparé.

Dans (Rosenberg, 2012), ces trois méthodes d'échantillonnage ont été évaluées et comparées en utilisant SVM comme classificateur. SVM a été entrainé en utilisant l'ensemble des traits caractéristiques de référence de la compétition *INTERSPEECH 2009 Emotion Challenge* extrait du corpus FAU AIBO Emotion. En termes de performance UAR, ces techniques sont classées comme suit: SMOTE, le sous-échantillonnage suivi d'échantillonnage d'ensemble.

Dans la même étude, l'auteur a montré que la pondération de l'importance (*importance weighting*) présente une meilleure alternative que les méthodes d'échantillonnage pour optimiser la moyenne non pondérée du rappel (UAR) appliquée aux données ayant une distribution asymétrique. Cette technique consiste à appliquer une pondération d'importance pour chacune des données d'apprentissage dans la fonction objective optimisée durant la phase d'apprentissage. La valeur de la pondération est inversement proportionnelle à la taille

de la classe. La fonction objective du SVM entrainé avec *hinge-loss* est exprimée comme suit :

$$\min \ \mathbf{V}^T\mathbf{V} + c\sum_j \xi_j \ , \tag{6.16}$$

où c représente la pente de la fonction *hinge*, $\mathbf{v}$ un vecteur normal à la frontière de décision et ξ_j est une variable d'écart (*slack variable*). Après introduction de la pondération d'importance, la fonction objective est réécrite comme suit:

$$\min \ \mathbf{V}^T\mathbf{V} + c\sum_j \gamma_j\xi_j \ , \tag{6.17}$$

où γ_j représente la valeur de pondération associée au point de données j qui est égale à $\frac{1}{n_j}$, l'inverse de la taille de la classe à laquelle le point j appartient.

6.7.2 Résultats expérimentaux

Nous allons dans cette section évaluer les performances de classification des trois architectures avec et sans techniques d'échantillonnage. A cet effet, un modèle GMM est évalué en utilisant la règle de décision de *Bayes*, avec des valeurs de probabilité a priori égales, en guise de premier type architecture. Un énoncé de test est classifié selon la classe d'émotion qui maximise la valeur du logarithme de probabilité de vraisemblance de la donnée de test :

$$\text{emotion} = \arg\max_{i=1\ldots C}\left(\log P\left(\mathbf{X}\middle|\ \lambda_i\right)\right) \tag{6.18}$$

Tableau 6.3 Comparaison des trois différents types d'architectures évaluées sur les données de de test du corpus FAU AIBO Emotion

Systèmes	Rappel (%)					UAR
	A	E	N	P	R	(%)
Sans système dorsal						
GMM-Bayes	46,97	49,27	41,83	46,95	23,23	41,65
Systèmes dorsaux simples						
KNN (k=5)	28,3	37	74,3	11,3	2,2	30,62
KNN-S (k=41)	49,3	45	38,5	49,3	18	40,02
KNN-D1 (k=21)	30,8	63	56,7	23,9	2,2	35,32
KNN-D2 (k=75)	53,5	45,1	48,5	52,1	10,6	**41,96**
KNN-W (k=211)	51,1	48,1	47,2	53,1	15,1	**42,92**
Modèles d'ancrage	55,97	47,02	49,79	55,35	12,82	**44,19**
Systèmes dorsaux Complexes						
SVM *(linéaire)*	0	4,8	98,9	0	0	20,74
SVM-S *(linéaire)*	40,8	52,9	53	32,9	16,7	39,26
SVM-D1 *(linéaire)*	3,3	67,8	68,4	3,7	0	28,64
SVM-D2 *(linéaire)*	56,6	42,7	47,1	59,2	14,3	**43,98**
SVM-W *(linéaire)*	83,8	48,1	0	0	0	26,38
SVM *(polynom. d=3)*	8,2	15,6	95,8	0	0	23,92
SVM-S *(d=3)*	39,1	53,1	55,5	32,9	16,2	39,36
SVM-D1 (*d=3*)	17,3	65,8	67,6	11,7	0	32,48
SVM-D2 (*d=3*)	55,3	47,8	50,2	52,1	13,4	**43,76**
SVM-W *(d=3)*	100	0	0	0	0	20
SVM *(RBF kernel)*	0	0	100	0	0	20
SVM-S *(RBF)*	60,1	31	28,4	23	55,6	39,62
SVM-D1 *(RBF)*	0	76,1	56,8	0	0	26,58
SVM-D2 *(RBF)*	70,9	3,4	1,4	0	75,1	30,16
SVM-W *(RBF)*	100	0	0	0	0	20
Logistique	21,1	31,8	87,4	9,4	0	29,94
Logistique-S	40,1	52,7	53,1	34,7	17,5	39,62
Logistique-D1	28,8	66,6	62,1	28,2	0	37,14
Logistique-D2	57,6	45,4	44,5	59,6	13,8	**44,18**
Logistique-W	58,6	47,3	41,8	59,2	15,1	**44,40**
MLP	18,2	22,5	91,1	10,3	0	28,42
MLP-S	38,6	40	65	35,7	11	38,06
MLP-D1	20,8	78,1	48,5	25,8	0	34,64
MLP-D2	55	60,7	32,5	62,9	0,7	**42,36**
MLP-W	0	80,2	0	0	5,2	26,44
Forêts d'arbres	26,8	37,1	75,5	9,4	2,4	30,24
Forêts d'arbres-S	38,3	44,3	51	28,2	13,6	35,08
Forêts d'arbres-D1	33,7	57,2	47,7	23,5	5,8	33,58
Forêts d'arbres-D2	48,6	39,2	33,9	32,4	14,1	33,64
Forêts d'arbres-W	36,7	50,1	48,7	20,7	6,5	32,54

Dans *Classificateur-XX*, *XX* désigne la méthode d'échantillonnage; S : SMOTE, W: pondération d'importance, S1 et S2 les deux variantes du sous-échantillonnage

Notez que nous avons déjà testé le système GMM adapté à partir d'un modèle du monde UBM en utilisant les deux techniques d'adaptation MAP et MLLR-MAP et les résultats obtenus avec un modèle entrainé directement avec la méthode ML sont légèrement meilleures. Comme deuxième type d'architecture, nous étudierons la méthode KNN en plus des modèles d'ancrage décrits dans la section 6. Enfin, pour le troisième type d'architectures utilisant un système frontal plus sophistiqué, nous expérimenterons quatre classificateurs : SVM, MLP, régression logistique, et les forêts d'arbres décisionnels (ou forêts aléatoires de l'anglais *random forest*). Afin de pallier au problème des données asymétriques, les trois meilleures méthodes présentées dans (Rosenberg, 2012) sont étudiées, à savoir, la pondération de l'importance, SMOTE et le sous-échantillonnage. Deux variantes de la méthode du sous-échantillonnage sont testées. Dans la première, la classe majoritaire (*neutre* dans le cas du corpus FAU AIBO Emotion) est réduite à la taille de la deuxième plus fréquente classe (*Emphatique*). Dans la seconde, toutes les classes majoritaires sont réduites à la taille de la classe minoritaire (*positive*).

Les résultats des différents systèmes évalués sur les données de test sont rapportés dans le Tableau 6.3. Plusieurs observations peuvent être relevées. Premièrement, nous constatons que lorsque les données d'apprentissage sont utilisées sans recours aux techniques d'échantillonnage ou de pondération, les systèmes basés sur des systèmes dorsaux complexes donnent les plus mauvais résultats. Cela est particulièrement vrai pour SVM dont les performances sont comparables à un système basé sur la chance, i.e., assignant toutes les données à la classe majoritaire (**N**). Deuxièmement, nous observons que l'introduction de la pondération de l'importance et les techniques d'échantillonnage remédie en général au problème de la distribution asymétrique des données. Bien que la meilleure technique dépend du type de classificateur utilisé, la technique du sous-échantillonnage à la taille de la classe la moins fréquente conduit généralement à de meilleurs résultats. Le sous-échantillonnage à la taille de la classe la moins fréquente appliqué aux systèmes MLP et SVM avec un noyau polynômial de degré un ou trois surpasse non seulement les autres techniques mais aussi représente la seule technique qui permet de dépasser les résultats du système GMM-*Bayes*. Pour les classificateurs SVM avec fonction à base radiale (*Radial Basis Function*, RBF) et

les forêts d'arbres décisionnels, les meilleurs résultats sont obtenus avec la technique SMOTE. Cependant les performances UAR demeurent inférieures à celle du GMM-Bayes. Quant à la pondération de l'importance, cette technique se comporte différemment selon le classificateur utilisé. Pour le SVM et en particulier avec un noyau RBF, l'effet de la distribution asymétrique observé avant application de la pondération est inversé après application, c.-à-d., toutes les données de l'une des classes minoritaires sont correctement classées au détriment des autres classes. D'autre part, la pondération de l'importance vient en première position par rapport aux autres techniques et parvient à atténuer le problème des données asymétrique lorsqu'elle est testée avec KNN et la régression logistique. Nous observons également que le sous-échantillonnage à la taille de la classe la moins fréquente obtient de bons résultats aussi, mais légèrement inférieurs à ceux de la pondération de l'importance. Les deux techniques lorsqu'elles sont utilisées avec les classificateurs KNN et régression logistique surpassent GMM-*Bayes*. Si nous comparons à présent entre les deux variantes du sous-échantillonnage, nous constatons que le sous-échantillonnage à la taille de la classe la moins fréquente donne toujours de meilleurs résultats par rapport au sous-échantillonnage à la taille de la deuxième plus fréquente classe.

Ces résultats nous enseignent que la meilleure méthode pour pallier au problème des données asymétriques dépend d'une part du type de classificateur choisi pour modéliser les traits. D'autre part, la comparaison de nos résultats avec ceux obtenus dans (Rosenberg, 2012) avec le classificateur SVM (seul classificateur expérimenté dans cette étude) montre que la meilleure méthode d'échantillonnage dépend également du type de traits caractéristiques modélisés. En effet, dans (Rosenberg, 2012), où les traits de types suprasegmentaux sont modélisés avec SVM (linéaire), c'est la méthode de pondération de l'importance qui a permis d'optimiser les performances UAR contrairement à notre étude où c'est la méthode du sous-échantillonnage qui a retourné les meilleurs résultats lorsque les scores de probabilités sont pris comme traits caractéristiques. Par conséquent, la meilleure méthode pour pallier au problème des classes non balancées dépendra des données ainsi que le classificateur choisis.

Enfin, il est intéressant de noter que les modèles d'ancrage basés sur de simples métriques comme la distance euclidienne ou cosinus est le seul système capable d'améliorer les performances obtenues avec GMM-*Bayes* (gain relative de 6,2 %) sans utilisation d'aucune technique d'échantillonnage ou de pondération de l'importance. En outre, la métrique euclidienne surpasse tous les autres systèmes plus complexes, même lorsque ces différentes techniques sont appliquées à l'exception de la régression logistique lorsque testés avec la pondération de l'importance qui donne des résultats légèrement supérieurs (44,4 %). Ces résultats permettent de conclure que les modèles d'ancrage utilisés avec de simples métriques sont moins sensibles à la répartition non balancée des classes grâce à une utilisation d'un nombre de vecteurs représentatifs équilibré, fiables et de portée globale pour chacune des classes d'émotion. Une comparaison des modèles d'ancrage avec le meilleur système de type simple ou combiné de *INTERSPEECH 2009 Emotion Challenge* montre un gain relatif de 7,1 % et 6,1 % respectivement.

6.8 Conclusion

Dans ce chapitre, nous avons proposé les modèles d'ancrage basés sur les similarités euclidienne et cosinus pour la classification d'émotions spontanées. Nous avons montré que les distances séparant les vecteurs dans un espace d'ancrage sont affectés par du bruit introduit par l'utilisation de l'opérateur logarithmique au cours du calcul de la probabilité de vraisemblance. Dans ces conditions et en absence d'une normalisation préalable, la similarité cosinus est recommandée étant moins sensible à ce type de bruit en comparaison avec la distance euclidienne. La normalisation WCCN a permis d'améliorer davantage les bonnes performances déjà obtenues avant normalisation avec la similarité cosinus. Nous avons constaté également qu'après normalisation, les mesures euclidienne et cosinus obtiennent des performances équivalentes. Une comparaison de performances avec d'autres systèmes plus complexes tel que SVM et MLP montrent une supériorité des modèles d'ancrage dans un contexte où la distribution des classes de données est fortement déséquilibrée. Leurs insensibilité au problème de la distribution biaisée des données (échantillonnage non requis), à la courte durée des énoncés ainsi que l'absence de phase d'entrainement dans la partie

dorsale de leur système (donc moins de paramètres à ajuster) font des modèles d'ancrage de puissants systèmes de classification des émotions. Dans le chapitre suivant nous allons étudier les modèles d'ancrage dans un contexte d'un problème à deux classes et déterminer dans quelle mesure les bonnes performances obtenues dans un contexte multi-classe sont transférables au cas d'une classification binaire.

CHAPITRE 7

MODÈLES D'ANCRAGE : PROPRIÉTÉS ET APPLICATION À UNE CLASSIFICATION BINAIRE

7.1 Introduction

Dans le chapitre précédent, nous avons proposé les modèles d'ancrage pour la classification des émotions dans un contexte de classification multi-classe. Dans ce chapitre nous allons étendre l'étude de leurs performances pour les problèmes à deux classes et déterminer si les améliorations obtenues dans un contexte multi-classe sont généralisables pour une classification binaire. Nous nous intéresserons également à étudier les propriétés des modèles d'ancrage et à déterminer les paramètres qui influencent sur leurs performances de classification ainsi que le lien qui pourrait exister entre les frontières de décision du système frontal (GMM en occurrence) basé sur la règle de décision *Bayes* et le système dorsal basé sur les mesures euclidienne et cosinus. Nos travaux sont aussi motivés par les résultats préliminaires présentés dans la Figure 7.1. Cette figure présente les résultats de classification des systèmes GMM-*Bayes* et des modèles d'ancrage basés sur la similarité cosinus (MA-COS) pour les dix expériences à deux classes d'émotion du corpus FAU AIBO Emotion. Des performances similaires aux modèles d'ancrage basé sur le cosinus sont obtenues quand la distance euclidienne est utilisée après normalisation des scores. Nous observons que quatre des dix expériences affichent des performances de classification très similaires (en termes UAR) entre les systèmes GMM-*Bayes* et MA-COS et qui suggèrent : (i) l'existence d'un lien entre les deux classificateurs et (ii) que l'ajout de la partie dorsale au système GMM n'améliore pas souvent les performances de classification. Afin d'expliquer géométriquement ces résultats, nous allons déterminer au préalable les équations des frontières de décision des modèles d'ancrage basés sur les deux métriques et déterminer ensuite dans quelles conditions les systèmes GMM-*Bayes* et les modèles d'ancrages peuvent produire des résultats similaires (coïncidence de leurs régions de décision).

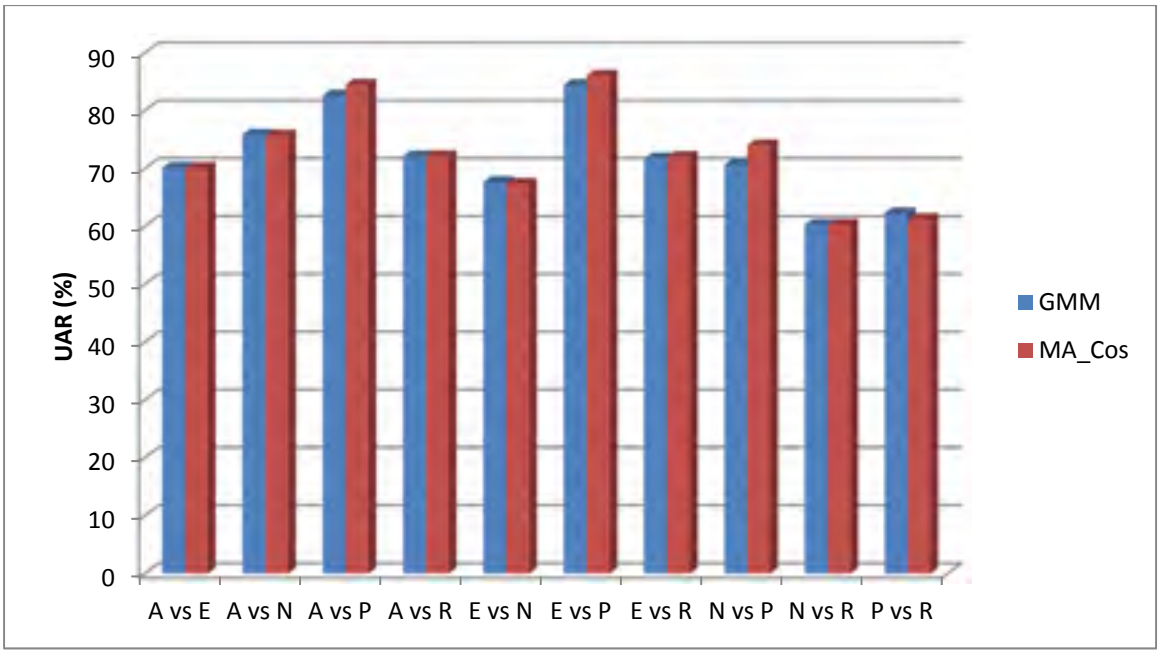

Figure 7.1 Résultats de 10 expériences de classification binaire réalisées sur les données de test du corpus FAU AIBO Emotion. La figure présente les résultats UAR des systèmes GMM-Bayes et les modèles d'ancrage basés sur la similarité cosinus

Nous commencerons donc par déterminer dans la section 2 les équations des frontières de décision des modèles d'ancrage basés sur les métriques euclidienne et cosinus dans un espace d'ancrage bidimensionnels (généré par deux modèles GMM). Nous étudierons dans section 3 les propriétés des modèles d'ancrage en fonction des valeurs des vecteurs représentatifs des classes d'émotion. Dans la section 4, nous allons comparer entre les régions de décision des systèmes GMM-Bayes et les modèles d'ancrage basés sur les mesures euclidienne et cosinus avant d'évaluer dans la section 5 les performances de ces systèmes dans un espace d'ancrage bidimensionnel. Dans la section 6, nous allons généraliser les équations des frontières de décision des modèles d'ancrage générées dans un espace bidimensionnel pour des dimensions supérieures et les performances des modèles d'ancrage seront évaluées dans un espace d'ancrage à cinq dimensions (section 7). Enfin, nous allons comparer la complexité algorithmique en temps de calcul entre les systèmes GMM-Bayes avec les modèles d'ancrage dans la section 8.

7.2 Analyse géométrique des modèles d'ancrage dans espace bidimensionnel

Dans cette section, nous présentons une analyse géométrique des modèles d'ancrage basés sur les mesures de distance euclidienne et cosinus. Nous commençons par définir la forme des régions de décision dans le contexte de la classification binaire réalisée dans un espace d'ancrage à deux dimensions avant de généraliser pour un espace de plus haute dimension. Soit un espace d'ancrage à deux dimensions engendré par l'ensemble $\Gamma=\{\lambda_i,\lambda_j\}$, c'est à dire, par les fonctions de densité de probabilité associées aux modèles GMM des classes *i* et *j* (les modèles des classes *i* et *j* représenteront l'axe des *x* et *y* respectivement). Soit $\mathbf{L}_i=(l_1^i,l_2^i)$ et $\mathbf{L}_j=(l_1^j,l_2^j)$ les vecteurs VCE représentatifs des classes *i* et *j*, respectivement. Sans perte de généralité, nous supposerons que $l_2^j > l_2^i$ (à savoir, la somme des logarithmes des probabilités de vraisemblance de la classe *j* par rapport au modèle λ_j est supérieure à la somme des logarithmes des probabilités de vraisemblance de la classe *i* par rapport au modèle λ_j). Soit $\mathbf{y}=(y_1,y_2)$ le vecteur VCE d'une donnée de test où y_1 et y_2 représentent les scores du logarithme de vraisemblance contre les modèles des classes *i* et *j*, respectivement. Étant donné les valeurs négatives des scores de logarithme de vraisemblance, tous les points VCE sont situés dans un quart de l'hyperespace (espace d'ancrage). Notons également que, selon la règle de décision de *Bayes*, la frontière de décision optimale est délimitée par une droite de pente positive (égale à l'unité comme expliqué dans la section 3) et les données de la classe *i* devraient être situées au-dessous de la droite de décision et au-dessus pour les données de la classe *j*. Un point de test *y* est classifié comme étant de la classe *i* si la distance séparant *y* et le représentant de la classe *i* est inférieure à la distance séparant *y* et le représentant de la classe *j*, i.e., $\mathrm{d}(\mathbf{y},\mathbf{L}_i) < \mathrm{d}(\mathbf{y},\mathbf{L}_j)$.

7.2.1 Métrique euclidienne

Dans cette section, nous allons dériver les régions de décision des modèles d'ancrage basés sur la mesure de la distance euclidienne. Un vecteur de test VCE *y* est classifié comme appartenant à la classe *i* si et seulement si:

$$\|\mathbf{y}-\mathbf{L}_i\| < \|\mathbf{y}-\mathbf{L}_j\|, \tag{7.1}$$

i.e.,

$$\sqrt{(y_1-l_1^i)^2+(y_2-l_2^i)^2} < \sqrt{(y_1-l_1^j)^2+(y_2-l_2^j)^2}$$

Après simplification et réarrangement des termes, on obtient:

$$y_2 < \frac{l_1^i-l_1^j}{l_2^j-l_2^i}y_1 + \frac{1}{2}\left[\frac{(l_1^j)^2+(l_2^j)^2-(l_1^i)^2-(l_2^i)^2}{(l_2^j-l_2^i)}\right] \tag{7.2}$$

qui peut être écrite de façon équivalente comme:

$$y_2 < \frac{l_1^i-l_1^j}{l_2^j-l_2^i}y_1 + \frac{\|\mathbf{L}_j\|^2-\|\mathbf{L}_i\|^2}{2\left(l_2^j-l_2^i\right)}, \tag{7.3}$$

en définissant $a_{euc}=\frac{l_1^i-l_1^j}{l_2^j-l_2^i}$ et $b_{euc}=\frac{\|\mathbf{L}_j\|^2-\|\mathbf{L}_i\|^2}{2\left(l_2^j-l_2^i\right)}$ nous obtenons :

$$y_2 < a_{euc}y_1 + b_{euc}. \tag{7.4}$$

La partie droite de cette inégalité est linéaire en y_1. En conséquence, le modèle d'ancrage basé sur la distance euclidienne est un discriminant linéaire qui a a_{euc} comme coefficient directeur (pente) et b_{euc} comme ordonnée à l'origine.

Notons également que la droite de décision traverse le milieu du segment [$l_i l_j$] ayant le point $\left(\frac{l_1^i+l_1^j}{2},\frac{l_2^i+l_2^j}{2}\right)$ comme coordonnées. Cela signifie que le tracé de la frontière de décision dans un modèle d'ancrage basé sur la distance euclidienne est basé sur le moment du premier ordre des paramètres statistiques des scores (c.-à-d., le point représentatif de la classe qui

correspondant à la valeur moyenne de la classe) et considère que les classes ont la même valeur de variance (moment du deuxième ordre).

L'inégalité (7.2) est dérivée sous l'hypothèse $(l_2^j - l_2^i) \neq 0$, c.-à-d., les vecteurs représentatifs des deux classes ont des valeurs différentes sur le deuxième axe de l'espace d'ancrage (dimension engendrée par le deuxième modèle de classe). Dans le cas contraire, le second modèle dégénère à un modèle non-discriminant dans l'espace d'ancrage pour la métrique euclidienne et la règle de décision est restreinte à la valeur du premier modèle. La droite de décision est perpendiculaire au premier axe et la traverse au milieu du segment $[l_1^i l_1^j]$ tel que dérivée dans (7.5) à partir de la formule (7.1) :

$$\begin{cases} y_1 > \dfrac{l_1^i + l_1^j}{2}, & \text{si } (l_1^i > l_1^j) \\ y_1 < \dfrac{l_1^i + l_1^j}{2}, & \text{si } (l_1^i < l_1^j) \end{cases} \tag{7.5}$$

Pour calculer (7.5), nous avons supposé que $l_1^i \neq l_1^j$. Dans le cas contraire nous sommes confrontés avec deux classes ayant le même vecteur représentant. Dans un tel cas, un modèle d'ancrage basé sur une simple métrique n'est plus pertinent pour aborder le problème.

7.2.2 Similarité cosinus

Dans cette section, nous allons dériver l'équation de la frontière de décision des modèles d'ancrage basés sur la mesure cosinus. Une donnée de test **y** appartient à la classe *i*, si et seulement si:

$$1 - \frac{\langle \mathbf{y}, \mathbf{L}_i \rangle}{\|\mathbf{y}\| \|\mathbf{L}_i\|} < 1 - \frac{\langle \mathbf{y}, \mathbf{L}_j \rangle}{\|\mathbf{y}\| \|\mathbf{L}_j\|}, \tag{7.6}$$

i.e.,

$$\frac{y_1 l_1^i + y_2 l_2^i}{\left\|\mathbf{L}_i\right\|} > \frac{y_1 l_1^j + y_2 l_2^j}{\left\|\mathbf{L}_j\right\|}$$

en définissant $\alpha = \frac{\left\|\mathbf{L}_i\right\|}{\left\|\mathbf{L}_j\right\|}$, avec $\alpha l_1^j - l_1^i \neq 0$, nous obtenons :

$$\begin{cases} y_2 < -\dfrac{\alpha l_1^j - l_1^i}{\alpha l_2^j - l_2^i} y_1, & \text{si } l_2^i < \alpha l_2^j \\ y_2 > -\dfrac{\alpha l_1^j - l_1^i}{\alpha l_2^j - l_2^i} y_1, & \text{sinon} \end{cases} \tag{7.7}$$

Ainsi,

$$y_2 = a_{\cos}\, y_1 \tag{7.8}$$

où $a_{\cos} = -\frac{\alpha l_1^j - l_1^i}{\alpha l_2^j - l_2^i}$, représente l'équation de la frontière de décision et correspond à une droite passant par l'origine avec une pente égale à $a_{\cos}$.

Le dénominateur s'annule dans le cas où $l_2^i = \frac{\left\|\mathbf{L}_i\right\|}{\left\|\mathbf{L}_j\right\|} l_2^j$. Ce cas se produit quand $\frac{l_1^i}{l_2^i} = \frac{l_1^j}{l_2^j}$, c.-à-d., les vecteurs représentatifs des deux classes ont la même pente (pointent dans la même direction). Dans un tel cas, l'utilisation de la similarité cosinus n'est pas pertinente pour les modèles d'ancrage. À noter que le rôle de α dans la formule $a_{\cos}$ est de normaliser la grandeur du vecteur $\mathbf{L}_j$ à celle du vecteur $\mathbf{L}_i$.

7.2.3 Relation entre les vecteurs représentatifs de classe et la métrique de similarité

Les vecteurs représentatifs des classes représentent des éléments clés dans un système de modèle d'ancrage. Outre les propriétés de chacun de ces points pris séparément (tel que leurs

coordonnées, magnitudes et angles), les vecteurs représentatifs peuvent également être caractérisés lorsqu'ils sont reliés entre eux, à travers la longueur et la pente des segments reliant ces points. Dans cette section, nous déterminerons les propriétés les plus importantes de ces vecteurs qui peuvent influencer sur le fonctionnement du système des modèles d'ancrage et qui peuvent éventuellement nous fournir des indices précurseurs sur les performances de classification.

Examinons maintenant l'équation (7.4) qui établit la région de décision du système métrique basée sur la distance euclidienne. La valeur du coefficient a_{euc} peut aussi être récrite sous forme $a_{euc} = -1 \Big/ \frac{l_2^j - l_2^i}{l_1^j - l_1^i}$. Il est intéressant de noter que le dénominateur de a_{euc} n'est autre que la valeur de la pente du segment [$l_i l_j$] reliant les points représentatifs des classes *i* et *j* formulée comme suit :

$$a_{\mathbf{L}_i\mathbf{L}_j} = \frac{l_2^j - l_2^i}{l_1^j - l_1^i} \tag{7.9}$$

Ainsi, la valeur de la pente de la frontière de décision, a_{euc}, est en effet inversement proportionnelle à la valeur de la pente $a_{\mathbf{L}_i\mathbf{L}_j}$ (c'est-à-dire $a_{euc} = \frac{-1}{a_{\mathbf{L}_i\mathbf{L}_j}}$), et (7.4) peut maintenant s'écrire en termes de $a_{\mathbf{L}_i\mathbf{L}_j}$ de la façon suivante :

$$y_2 < \frac{-1}{a_{\mathbf{L}_i\mathbf{L}_j}} y_1 + b_{euc} \tag{7.10}$$

Par conséquent, la valeur de la pente de la frontière de décision des modèles d'ancrage à base de la distance euclidienne dépend exclusivement de la valeur de la pente du segment [$l_i l_j$].

Prenons maintenant $\hat{\mathbf{L}}_j = (\hat{l}_1^j, \hat{l}_2^j)$ comme étant le vecteur $\mathbf{L}_j$ normalisé par α; $\hat{\mathbf{L}}_j = (\alpha l_1^j, \alpha l_2^j) = \alpha \mathbf{L}_j$. Dans (7.8), l'équation modélisant la région de décision du système basé sur la mesure cosinus, la valeur de la pente $a_{\cos}$ peut être écrite comme $a_{\cos} = -1 \Big/ \frac{\hat{l}_1^j - l_1^i}{\hat{l}_2^j - l_2^i}$, où $\hat{\mathbf{L}}_j$ et $\mathbf{L}_i$ sont d'une même grandeur et ne diffèrent que par leurs directions ($\hat{\mathbf{L}}_j$ et $\mathbf{L}_i$ se trouvent sur le même quadrant de rayon $\|\mathbf{L}_i\|$). Encore une fois, nous constatons que $a_{\cos}$ est tout simplement inversement proportionnelle à la valeur de la pente du segment reliant les points représentatifs des classes *i* et *j* (après normalisation de la taille pour le cas de la similarité cosinus) dénommé $a_{\mathbf{L}_i\hat{\mathbf{L}}_j}$ (c.-à-d., $a_{\cos} = \frac{-1}{a_{\mathbf{L}_i\hat{\mathbf{L}}_j}}$). L'équation (7.8) peut ainsi être exprimée sous la forme équivalente suivante :

$$y_2 = \frac{-1}{a_{L_i\hat{L}_j}} y_1 \tag{7.11}$$

À partir des formules précédentes, on en déduit que la précision de la frontière de décision estimée par les modèles d'ancrage basés sur les deux métriques dépend de la valeur de la pente du segment reliant les vecteurs représentatifs des classes (avec normalisation au préalable de la taille des vecteurs s'il s'agit de la similarité cosinus).

7.2.4 Propriétés des vecteurs représentatifs de classe

Dans ce qui suit, nous discuterons des performances des modèles d'ancrage en fonction de l'orientation de la pente de $\mathbf{L}_i\mathbf{L}_j$ et de la métrique sélectionnée :

- **Pente négative** ($a_{\mathbf{L}_i\mathbf{L}_j} < 0$)

La pente de $\mathbf{L}_i\mathbf{L}_j$ est négative lorsque la différence des scores (logarithmes de vraisemblances) entre les données des deux classes contre le premier modèle est de signe

opposé à la différence de leurs scores par rapport au deuxième modèle, à savoir, $(l_1^i - l_1^j)(l_2^i - l_2^j) < 0$. Ceci correspond au cas idéal où le logarithme de vraisemblance des données de la première classe par rapport à son propre modèle est supérieure à celui des données de la seconde classe par rapport au premier modèle et vice-versa, c'est-à-dire $l_1^i > l_1^j$ et $l_2^j > l_2^i$. Dans ce cas, les classes sont convenablement représentées par leurs vecteurs représentatifs et fournissent une information a priori importante pour mieux estimer la frontière de décision des classes. Lorsque la pente de $\mathbf{L}_i\mathbf{L}_j$ est négative, les frontières de décision des deux métriques euclidienne et cosinus possèdent des valeurs de pente positives ($a_{euc} > 0$ et $a_{\cos} > 0$) comme c'est le cas pour un classificateur basé sur la règle de décision de *Bayes*. La forme de la pente du segment reliant les classes **A** (colère) et **P** (positive) dans le graphique (a) de la Figure 7.2 est un exemple de pente négative. Contrairement au segment $\mathbf{L}_i\mathbf{L}_j$ qui peut également avoir une pente positive, on notera que $\mathbf{L}_i\hat{\mathbf{L}}_j$ a toujours une pente de valeur négative, c'est-à-dire $(l_1^i - \hat{l}_1^j)(l_2^i - \hat{l}_2^j) < 0$, comme on peut aisément le démontrer étant donné que $\hat{\mathbf{L}}_j$ et $\mathbf{L}_i$ ont le même signe et la même valeur de magnitude (tous deux sont situés sur le troisième quadrant du cercle de rayon $\|\mathbf{L}_i\|$). Par conséquent, la frontière de décision des modèles d'ancrage basé sur la similarité cosinus a toujours une pente de valeur positive. Ce cas est illustré dans la Figure 7.2 (f).

- **Pente positive** ($a_{\mathbf{L}_i\mathbf{L}_j} > 0$)

C'est l'inverse du cas précédent et ne s'applique qu'au modèle d'ancrage à base d'une distance euclidienne. Il survient généralement lorsque les logarithmes de vraisemblance des données de la première classe pour le premier et second modèles sont plus petites que les logarithmes de vraisemblance des données de la deuxième classe par rapport aux modèles de la première et deuxième classe respectivement, à savoir ($l_1^i < l_1^j$ et $l_2^i < l_2^j$) ou ($l_1^i > l_1^j$ et $l_2^i > l_2^j$). La forme du segment reliant les vecteurs représentatifs des classes **E** et **P** de la Figure 7.2 (d) est un exemple de pente positive. Il faut souligner que ce cas de figure représente un cas problématique influant négativement sur les performances des modèles d'ancrage. En effet, la

pente de la frontière de décision a dans ce cas-ci une valeur négative ($a_{euc} < 0$, voir (7.10)), contrairement à la forme de la pente de la frontière de décision optimale qui est positive conformément à la règle de décision *Bayes*. Ceci explique géométriquement les mauvais résultats obtenus avec les modèles d'ancrage basés sur la distance euclidienne en comparaison avec la mesure cosinus si les données VCE ne sont pas traitées au préalable.

L'allure de la pente peut également coïncider avec certains cas extrêmes qui sont problématiques pour certaines métriques :

- **Pente verticale** ($a_{\mathbf{L}_i\mathbf{L}_j} \to \infty$) **ou horizontale** ($a_{\mathbf{L}_i\mathbf{L}_j} \approx 0$)

Si la métrique euclidienne est utilisée, une pente verticale ou horizontale indiquerait qu'aucune information discriminante ne peut être fournie par les vecteurs représentatifs des classes selon l'axe des *x* (dimension engendrée par le modèle de la classe *i*) ou selon l'axe des *y* (dimension engendrée par le modèle de la classe *j*), respectivement. Dans ce cas, des modèles supplémentaires sont nécessaires pour élargir la dimension de l'espace d'ancrage afin d'améliorer les performances de reconnaissance. Le graphique représentant la classification de **E** vs **N** de la Figure 7.2 (c) est un exemple de pente verticale où la classe **N** (axe des *x*) ne joue aucun rôle dans la discrimination des vecteurs représentatifs des deux classes. D'autre part, le graphique de Figure 7.2 (b) représentant la classification de **A** vs **R** fournit un exemple de pente horizontale, où le modèle de classe **R** (axe des *y*) ne contribue pas à la discrimination des vecteurs représentatifs des classes. Notez que la pente $a_{\mathbf{L}_i\hat{\mathbf{L}}_j}$ ne peut être ni verticale ni horizontale car $\mathbf{L}_i$ et $\hat{\mathbf{L}}_j$ se trouvent sur le même quadrant; donc l'utilisation de la similarité cosinus est plus pertinente dans ce cas, sauf si les deux points sont très proches.

- **Pente à valeur unitaire** ($a_{\mathbf{L}_i\mathbf{L}_j} \approx 1$):

Une pente proche de l'unité indique une très forte corrélation linéaire entre les deux classes. Ce type de pente représente un cas problématique pour les modèles d'ancrage à base de similarité cosinus, étant donné que les vecteurs représentatifs des deux classes ont alors la même direction et le point $\mathbf{L}_i$ est confondu avec $\hat{\mathbf{L}}_j$. Le graphique de la Figure 7.2 (e))

illustrant la frontière de décision du classificateur de la classe **N** vs **R** est un exemple de cas où la valeur de la pente est proche de un.

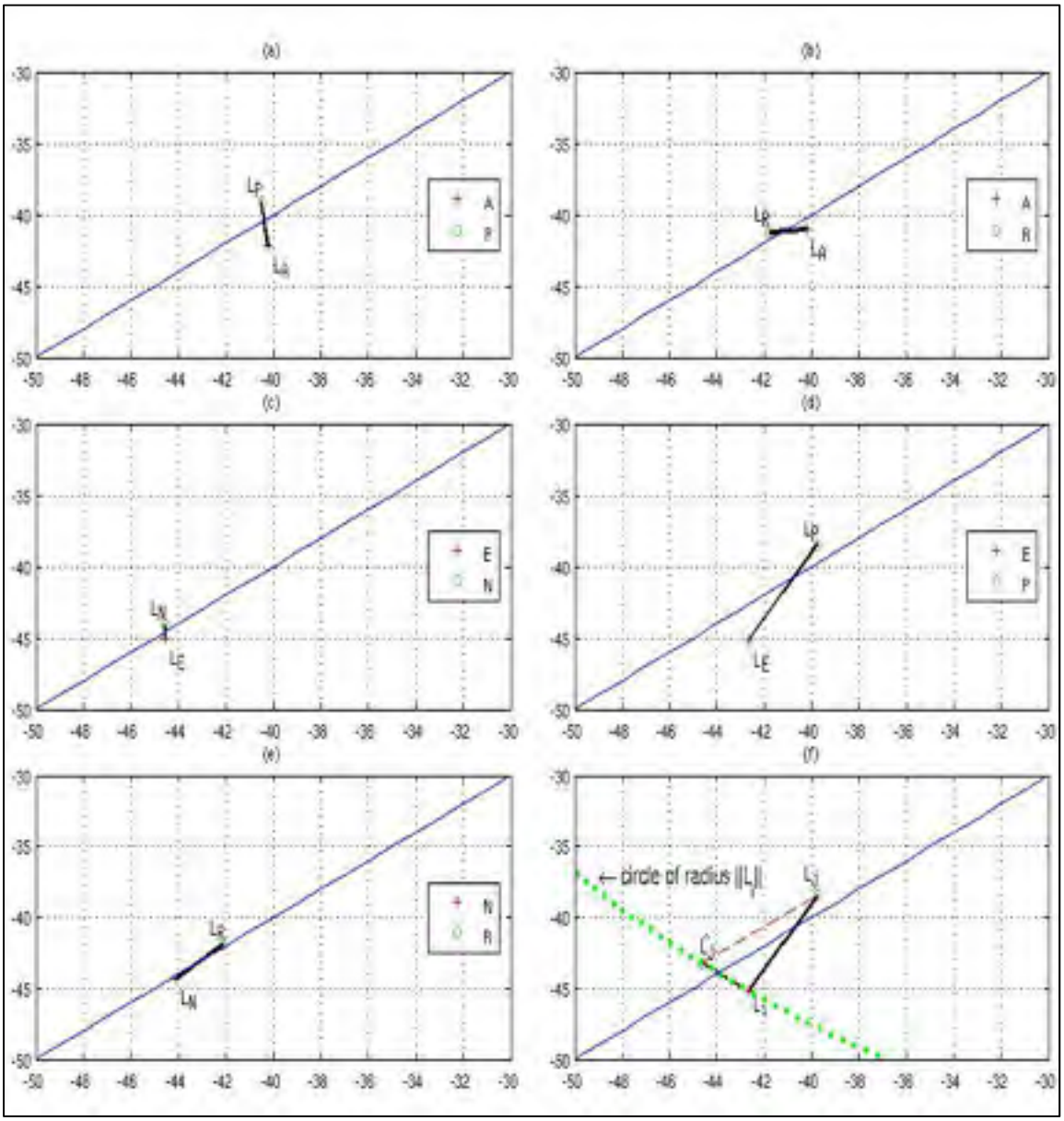

Figure 7.2 Exemples de pentes des segments reliant les vecteurs représentatifs de deux classes dans l'espace d'ancrage. Le vecteur représentant de chaque émotion est calculée en utilisant les données d'entrainement du corpus FAU AIBO Emotion

Le Tableau 7.1 résume la pertinence des mesures euclidienne et cosinus, lorsqu'elles sont utilisées comme métriques de similarité pour les modèles d'ancrage, en fonction des valeurs de la pente du segment reliant les vecteurs représentatifs des classes.

Tableau 7.1 Pertinence des métriques euclidienne et cosinus pour les modèles d'ancrage en fonction de la valeur de la pente du segment reliant les vecteurs représentatifs des classes

$a_{\mathbf{L}_i\mathbf{L}_j}$	$a_{\mathbf{L}_i\hat{\mathbf{L}}_j}$	a_{euc}	$a_{\cos}$	**Pertinence Euclidienne**	**Pertinence Cosinus**
$a_{\mathbf{L}_i\mathbf{L}_j} < 0$	$a_{\mathbf{L}_i\hat{\mathbf{L}}_j} < 0$	$a_{euc} > 0$	$a_{\cos} > 0$	√	√
$a_{\mathbf{L}_i\mathbf{L}_j} > 0$	$a_{\mathbf{L}_i\hat{\mathbf{L}}_j} < 0$	$a_{euc} < 0$	$a_{\cos} > 0$		√
$a_{\mathbf{L}_i\mathbf{L}_j} \to \infty$	$a_{\mathbf{L}_i\hat{\mathbf{L}}_j} = C$	$a_{euc} \approx 0$	$a_{\cos} > 0$		√
$a_{\mathbf{L}_i\mathbf{L}_j} \approx 0$	$a_{\mathbf{L}_i\hat{\mathbf{L}}_j} \neq 0$	$a_{euc} \to \infty$	$a_{\cos} > 0$		√
$a_{\mathbf{L}_i\mathbf{L}_j} \approx 1$	Indéfinie	$a_{euc} \approx -1$	Indéfinie		

Nous avons montré dans la discussion précédente que la valeur de la pente reliant les points représentatifs des classes fournit une information préliminaire sur la capacité de discrimination des classes des modèles d'ancrage. Cette information peut être utilisée pour expliquer ou prédire avant les expérimentations, si un modèle d'ancrage a la capacité d'améliorer les performances de classification et quelle mesure de similarité est plus pertinente pour les données disponibles.

7.3 Comparaison entre des règles décision Bayes et modèles d'ancrage

Dans certaines expériences, tel que nous l'avons exposé dans l'introduction, les performances d'un système GMM basé sur la règle de décision *Bayes* donnaient des performances similaires à celles des modèles d'ancrage. Dans cette section, nous allons comparer entre les frontières de décision de ces différents systèmes et établir dans quelles

conditions cette similitude peut se produire. Selon le théorème de *Bayes*, la probabilité a posteriori d'un modèle λ pour une observation $\mathbf{X}_T$ est donnée par :

$$P(\lambda|\mathbf{X}_T) = \frac{P(\mathbf{X}_T|\lambda) * P(\lambda)}{P(\mathbf{X}_T)} \tag{7.12}$$

où $P(\mathbf{X}_T|\lambda)$ représente la probabilité de vraisemblance, $P(\lambda)$ la probabilité a priori de λ et $P(\mathbf{X}_T)$ la probabilité marginale. Un énoncé de test $\mathbf{X}_T$ est classifié dans la classe *i* plutôt que *j* si et seulement si la probabilité postérieure de λ_i est supérieure à celle de λ_j, c.-à-d. $P(\lambda_i|\mathbf{X}_T) > P(\lambda_j|\mathbf{X}_T)$. En prenant le logarithme de part et d'autre et en remplaçant la probabilité postérieure par la partie droite de (7.12) on obtient:

$$y_2 < y_1 + \log\frac{P(\lambda_i)}{P(\lambda_j)} \tag{7.13}$$

où $y_1 = \log P(X_T|\lambda_1)$ et $y_2 = \log P(X_T|\lambda_2)$. La frontière de décision est une droite de pente égale à 1 ($a_{Bayes} = 1$). Le logarithme du rapport des deux probabilités à priori des deux classes détermine la valeur de la constante ($b_{Bayes} = \log\frac{P(\lambda_i)}{P(\lambda_j)}$). Si $P(\lambda_i)$ est supérieure $P(\lambda_j)$ la frontière de décision est déplacée vers le haut, priorisant la classe *i*. Dans le cas contraire, la frontière est déplacée vers le bas et c'est la classe *j* qui est avantagée dans la classification. Dans le cas où les deux classes possèdent la même valeur de probabilité a priori, la frontière de décision correspond à la bissectrice du plan. Notons que dans la plupart des problèmes réels les valeurs des probabilités à priori $P(\lambda_i)$ et $P(\lambda_j)$ sont inconnues.

Comme nous pouvons le voir à partir des équations (7.4), (7.8) et (7.13), les modèles d'ancrage basés sur les mesures euclidienne et cosinus ainsi que le système basé sur la règle de décision *Bayes*, font tous usage d'un discriminant linéaire pour séparer les régions de décision. Rappelons qu'il s'agit ici de discriminant linéaire dans l'espace des scores de

vraisemblance et non des traits acoustiques où les frontières de décision sont plus complexes (basées sur la modélisation GMM).

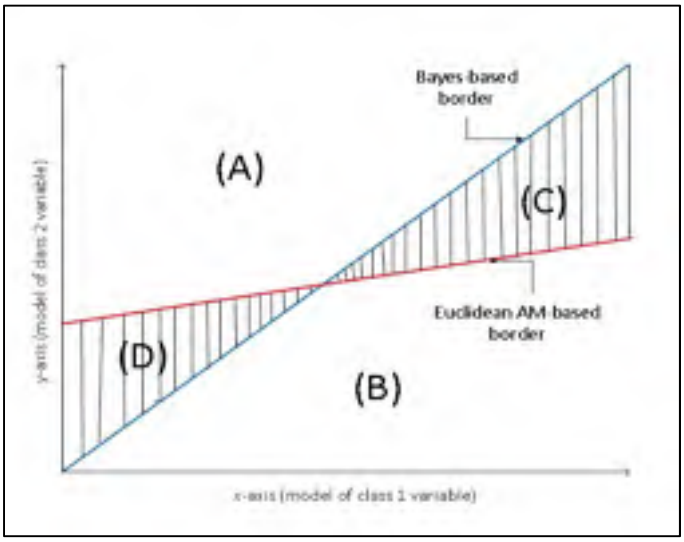

Figure 7.3 Exemples de régions de décision des classificateurs GMM-*Bayes* (bleu) et le MA-EUC (rouge). Les deux classificateurs prédisent les mêmes classes pour les régions (A) et (B) et des classes différentes pour (C) et (D)

La Figure 7.3 illustre les régions de décision d'un classificateur basé sur la règle de décision *Bayes* et celui d'un modèle d'ancrage basé sur une distance euclidienne (MA-EUC). Les frontières de décisions des deux classificateurs divisent l'espace d'ancrage en quatre régions. La zone (*A*) représente la région où les données sont classifiées comme appartenant à la classe 2 par les deux classificateurs. Les données de la région (*B*) sont également assignées conjointement par les deux classificateurs à la classe 1. Les données situées dans la région (*C*) sont prédites comme appartenant à la classe 1 par le classificateur basé sur la règle *Bayes* et à la classe 2 par le modèle d'ancrage, et vice versa pour les données appartenant à la région (*D*). Plus larges sont les régions (*C*) et (*D*) (donc plus de données sont contenues à l'intérieur), plus les deux classificateurs sont différents. Plus petites sont ces deux régions (donc moins de données contenues à l'intérieur), plus proches sont les résultats des deux classificateurs. Une meilleure performance de reconnaissance pour les modèles d'ancrage indique une plus grande précision dans l'estimation de la frontière de décision à travers les

coefficients a_{euc} et b_{euc} comparée à la règle de de décision Bayes grâce à l'utilisation de l'information a priori fournie par les vecteurs représentatifs des classes.

Les trois systèmes, GMM-*Bayes* et les modèles d'ancrage basé sur les similarités cosinus et euclidienne, ajustent tous leurs frontières de décision en tenant compte des connaissances a priori sur la distribution des classes. La façon dont cet ajustement est mis en œuvre varie entre les trois méthodes. Comme mentionné précédemment, dans la règle de décision de *Bayes*, la frontière est ajustée par un déplacement linéaire selon l'axe des *x* tandis que la pente est maintenue fixe. Dans les modèles d'ancrage à base de cosinus, la frontière de décision est ajustée à travers la valeur de la pente de la droite $a_{\cos}$. La frontière passe toujours par l'origine et traverse également le segment reliant les deux points représentatifs des classes L_i et L_j. Lorsque la valeur de la pente $a_{\cos}$ est très proche d'un, les modèles d'ancrage à base de cosinus se comportent de façon similaire au système basé sur *Bayes* utilisant des probabilités a priori égales (performances similaires).

Pour les modèles d'ancrage à base de la distance euclidienne, la frontière de décision est paramétrée par deux coefficients : la pente et la constante. La règle de décision de *Bayes* peut être considérée comme un cas particulier des modèles d'ancrage à base de la distance euclidienne lorsque la valeur de la pente a_{euc} de (7.4) est égal à un. Ce cas correspond à la condition où $l_1^i - l_1^j = l_2^j - l_2^i$, qui signifie que la somme du logarithme des probabilités de vraisemblance des données de la classe *i* contre les deux modèles est égale à celle des données de la classe *j*. Enfin, la frontière de décision de la règle de *Bayes* basée sur des probabilités a priori égales et les modèles d'ancrage basés sur la distance euclidienne se confondent quand $a_{euc} = 1$ et $b_{euc} = 0$. Cela se produit lorsque $l_1^i = l_2^j$ et $l_1^j = l_2^i$, comme on peut facilement le dériver, ce qui signifie que les deux points représentatifs sont symétriques par rapport à la droite de pente égale à un (frontière de décision de *Bayes*).

7.4 Expérimentation des modelés à ancrage à espace bidimensionnel

7.4.1 Configuration expérimentale

Nous allons évaluer les performances des modèles d'ancrage à base des métriques euclidienne et cosinus et les comparer au système basé sur la règle *Bayes*. Pour évaluer les performances des modèles d'ancrage dans un contexte d'un problème à deux classes, une expérience est réalisée pour chaque paire de classes sélectionnées à partir de l'ensemble des cinq classes du corpus FAU AIBO Emotion (ce qui donne un total de dix expériences). Les résultats sont optimisés en maximisant en premier le critère UAR suivi de WAR étant donné que la distribution des classes du corpus est déséquilibrée. Rappelons, qu'un classificateur des classes **N** vs **P** par exemple, qui prédira toutes les données de test comme étant de la classe dominante **N**, atteindra 96,2 % en termes de WAR, mais seulement 50 % en termes UAR. Le nombre de composantes gaussiennes des modèles GMM est optimisé séparément pour chaque classificateur binaire sur la base des données d'entraînement utilisant le protocole de validation croisée à neuf plis indépendants du locuteur. Les nombre de composantes gaussiennes permettant d'optimiser les performances de classification pour chacun des modèles GMM sont : 32 pour **N** vs **R**, **P** vs **R** et **E** vs **N**; 64 pour **A** vs **N**, **A** vs **P**, **A** vs **R** et **E** vs **R**; 128 pour **A** vs **E** et **N** vs **P** et 256 pour **E** vs **P**. Pour comparer les performances des modèles d'ancrage avec les modèles GMM-*Bayes*, qui servira d'indicateur sur la proximité relative des frontières de décision des différents classificateurs, les résultats de la reconnaissance sont présentés en termes de gain relatif du critère UAR ($UAR_{MA}/UAR_{GMM-Bayes} * 100$).

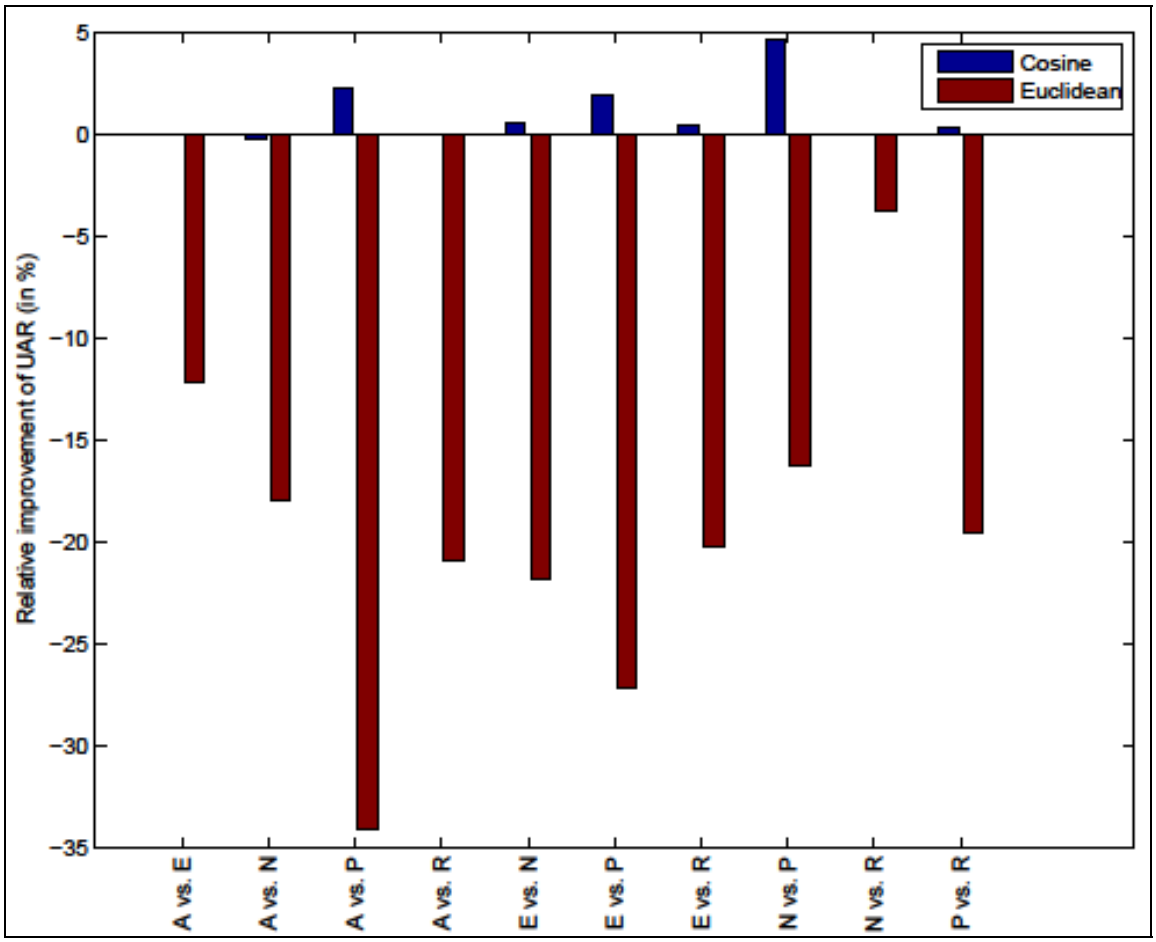

Figure 7.4 Gain/perte relatif en UAR pour les modèles d'ancrage à base des mesures cosinus et euclidienne par rapport au système GMM-Bayes. Les systèmes sont évalués sur les données de test du corpus FAU AIBO Emotion

7.4.2 Résultats avant la normalisation WCCN

La Figure 7.4 montre le gain relatif UAR obtenus par les modèles d'ancrage par rapport aux systèmes GMM-Bayes pour chacune des dix expériences. Les performances sont évaluées sur les données de test. Nous constatons que lorsque la distance euclidienne est utilisée comme mesure pour les modèles d'ancrage, les performances baissent considérablement par rapport à la règle de *Bayes* (gain relatif négatif). Pour la métrique cosinus, nous observons une amélioration de performance pour des expériences **A** vs **P** (2,3 %), **E** vs **P** (1,9 %) et **N** vs **P** (4,7 %), alors que des résultats équivalents sont obtenus pour les autres expériences,

comme pour **A** vs **E**, **A** vs **R** et **N** vs **R**. Globalement, une amélioration moyenne de 1 % est obtenue pour les dix expériences.

Pour expliquer ces résultats, nous avons tracé à la Figure 7.5 les régions de décision des trois classificateurs pour chacune des dix expériences. Tout d'abord, nous observons que les pentes des segments reliant les points représentatifs des classes sont positives pour les dix expériences ($a_{\mathbf{L}_i\mathbf{L}_j} > 0$). Par conséquent, la pente de frontière de décision des modèles d'ancrage basés sur la distance euclidienne est négative ($a_{euc} < 0$), à l'opposé de la pente d'une frontière de décision optimale ce qui explique les mauvais résultats des modèles d'ancrage à base d'une distance euclidienne en absence d'une étape de prétraitement des données. D'autre part, on observe que pour les expériences **A** vs **E**, **A** vs **N**, **E** vs **R**, **N** vs **R** et **P** vs **R**, les pentes des segments reliant les points représentatifs ont des valeurs proches de un, indiquant que les vecteurs représentatifs des classes pointent vers la même direction. Cela marque une très forte corrélation linéaire entre les modèles de classe. Pour cette forme de pente, les frontières de décision des classificateurs GMM-*Bayes* et les modèles d'ancrage basés sur la similarité cosinus se recouvrent ce qui explique la similitude des résultats obtenus pour ces cas. Dans le cas des expériences **A** vs **R** et **E** vs **N**, la pente du segment $a_{\mathbf{L}_i\mathbf{L}_j}$ est respectivement horizontale et verticale, ce qui indique qu'une seule dimension de l'espace d'ancrage peut fournir des informations discriminantes lorsque la similarité euclidienne est utilisée. Outre la forme de la pente, les deux points représentatifs sont très proches l'un de l'autre (les deux vecteurs pointent donc plus ou moins vers la même direction), inhibant ainsi la capacité du système basé sur la similarité cosinus d'améliorer les performances. Pour les expériences restantes, à savoir **A** vs **P**, **E** vs **P** et **N** vs **P**, on constate que la frontière de décision du système basé sur la mesure cosinus est notablement distincte de celle du système GMM-*Bayes*, ce qui explique l'amélioration relative apportée aux résultats obtenus avec la règle de décision *Bayes*.

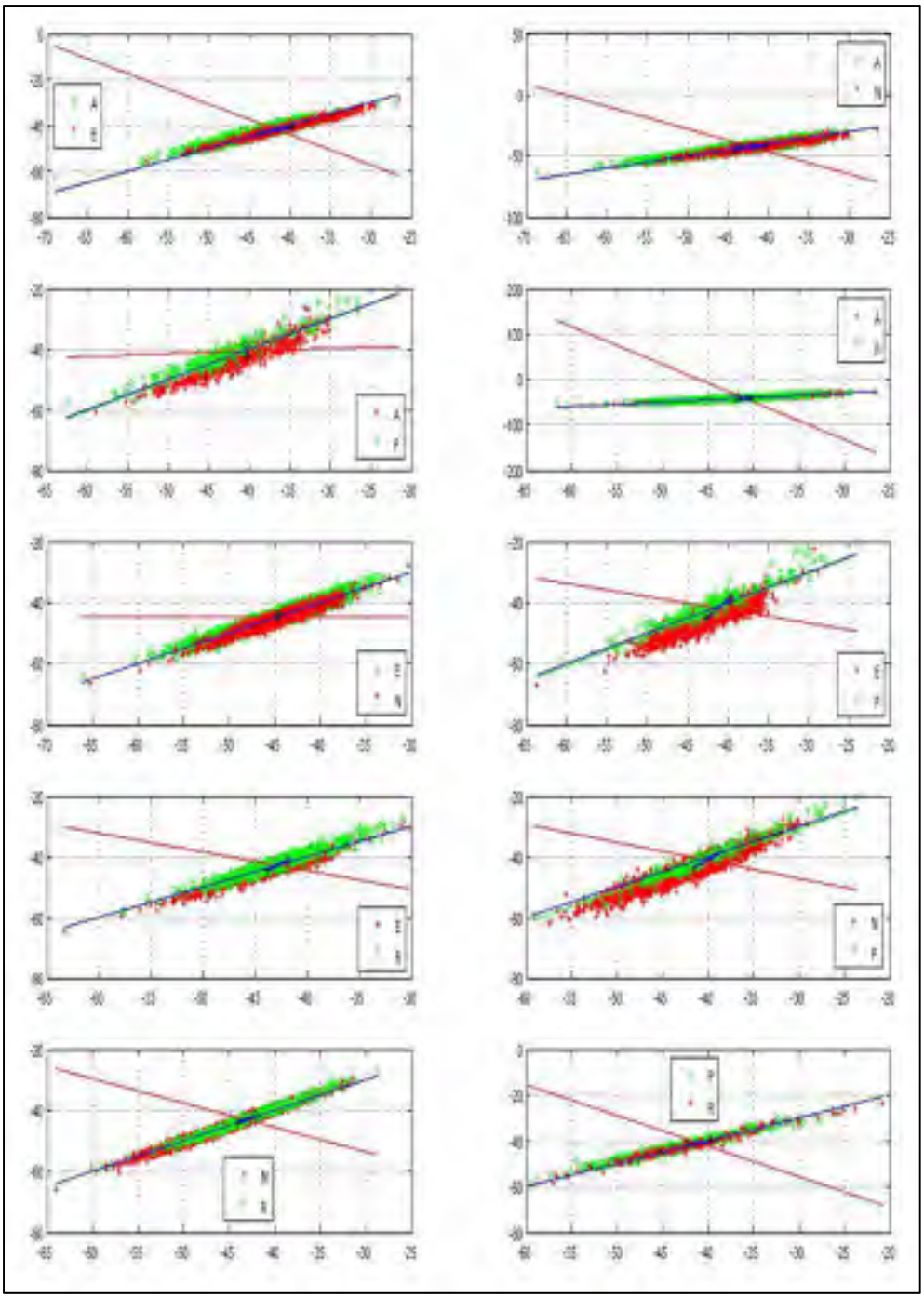

Figure 7.5 Les régions de décision sont établies pour chacun des dix problèmes à deux classes avant normalisation. La frontière de décision est représentée en bleu pour la règle Bayes, en rouge pour la distance euclidienne et en vert pour le cosinus. Les points représentatifs des classes sont reliés par une ligne continue

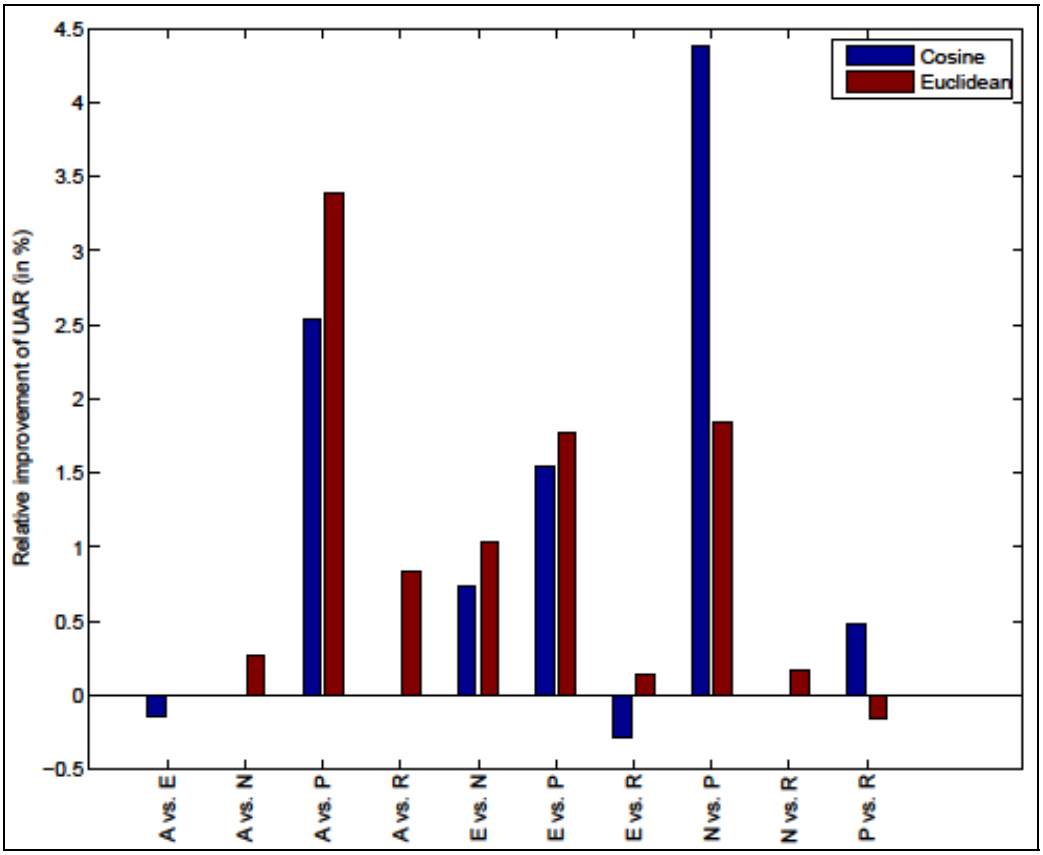

Figure 7.6 Gain relatif en performance UAR pour les modèles d'ancrage à base des mesures cosinus et euclidienne par rapport au système GMM-*Bayes* après la normalisation WCCN. Les systèmes sont évalués sur les données de test du corpus FAU AIBO Emotion

7.4.3 Effet géométrique de la normalisation WCCN

La Figure 7.6 montre l'amélioration relative apportée par les modèles d'ancrage basés sur les métriques cosinus et euclidienne par rapport aux systèmes GMM-Bayes après application de la normalisation WCCN. D'abord, on constate que l'amélioration relative moyenne obtenue par les systèmes MA-EUC dans les dix expériences passe de -19,4 % à 0,9 %, atteignant pratiquement les performances des modèles d'ancrage à base de cosinus des expériences précédentes (1 %). En outre, les performances globales du système basé sur le cosinus demeurent pratiquement inchangées à la suite de WCCN.

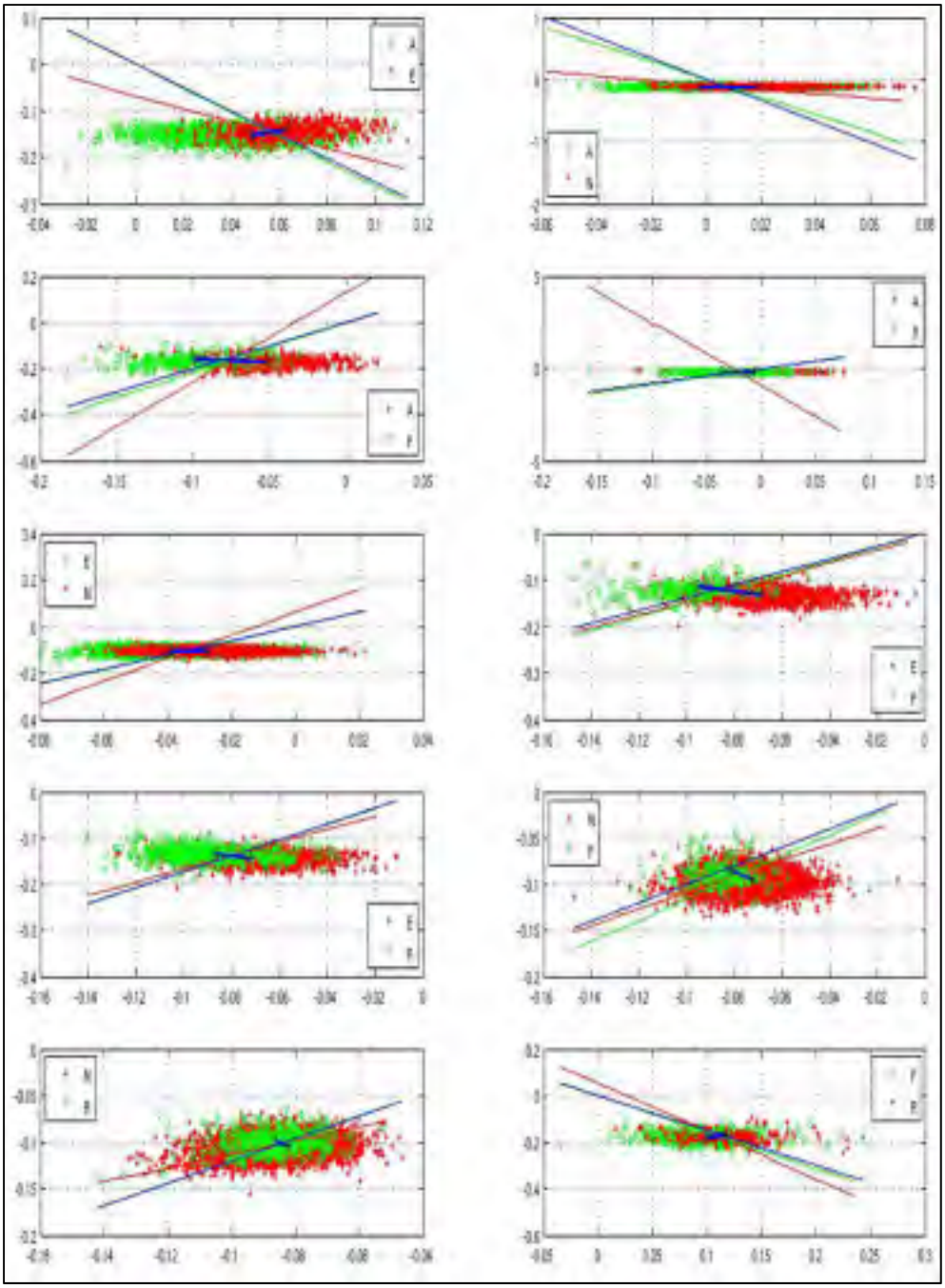

Figure 7.7 Les régions de décision sont établies pour chacun des dix problèmes à deux classes après la normalisation WCCN. La frontière de décision est dessinée en bleu pour le système GMM-*Bayes*, en vert pour MA-COS et en rouge pour MA-EUC. Les points représentatifs des classes sont reliés par une ligne continue

La Figure 7.7 montre l'effet géométrique de la normalisation WCCN sur les régions de décision des dix expériences. Tout d'abord, nous observons que WCCN a introduit un effet de décorrélation sur les classes, reflétée à travers les formes des segments reliant les points représentatifs de classe. Nous observons également que la variance le long de chaque axe est réduite. En conséquence, la frontière de décision associée à la métrique euclidienne est ajustée et sa pente possède le même signe que celui du classificateur basé sur la règle *Bayes*, ce qui explique l'amélioration obtenue par le système basé sur la métrique euclidienne après normalisation. Enfin, un chevauchement important caractérise les régions de classe, ce qui met en évidence les limites de discrimination d'un espace d'ancrage bidimensionnel. Ce problème peut être résolu par l'extension de l'espace d'ancrage avec d'autres modèles, comme on le verra dans la section suivante.

7.5 Espace d'ancrage multidimensionnel

Nous avons vu dans la section précédente, qu'un espace d'ancrage généré par deux modèles de classe peut s'avérer non suffisamment discriminant pour améliorer les performances du système frontal GMM-*Bayes* pour certains problèmes (exemples de pente de segment $[l_i l_j]$ verticale, horizontale ou égale à un). Pour rendre l'espace plus discriminant particulièrement dans un contexte de classification binaire, nous proposons de construire un espace d'ancrage d'une plus grande dimension en incluant des modèles de classes qui ne sont pas directement impliquées dans le problème de classification. Ces nouveaux modèles permettront d'engendrer de nouvelles dimensions qui vont capter des dissimilarités non couvertes par les dimensions déjà existantes.

Soit *r*, la nouvelle dimension de l'espace d'ancrage engendré par les *r* modèles GMM, et soient $\mathbf{L}_i = (l_1^i,\ldots,l_r^i)$ et $\mathbf{L}_j = (l_1^j,\ldots l_r^j)$ les vecteurs représentatifs des classes *i* et *j*, respectivement. Soit $\mathbf{y} = (y_1,\ldots,y_r)$, le nouveau vecteur VCE des données de test. Nous allons maintenant généraliser les formules des frontières de décision dérivées pour les mesures euclidiennes et cosinus vues dans l'espace à deux dimensions pour un espace de dimension élevée, *r*.

7.5.1 La distance euclidienne

Un vecteur de test **y** est classé comme appartenant à la classe *i* si et seulement si :

$$\|\mathbf{y}-\mathbf{L}_i\| < \|\mathbf{y}-\mathbf{L}_j\|,$$

à savoir,

$$\sum_{k=1}^{r}(y_k - l_k^i)^2 < \sum_{k=1}^{r}(y_k - l_k^j)^2 \ . \tag{7.14}$$

Après simplification et réarrangement des termes, on obtient la formule suivante :

$$\sum_{k=1}^{r}(l_k^i - l_k^j)y_k + \sum_{k=1}^{r}(l_k^j)^2 - \sum_{k=1}^{r}(l_k^i)^2 > 0\,, \tag{7.15}$$

qui peut être exprimée de façon plus compacte en :

$$(\mathbf{L}_i - \mathbf{L}_j)\mathbf{y}^T + \|\mathbf{L}_j\|^2 - \|\mathbf{L}_i\|^2 > 0\,. \tag{7.16}$$

L'ensemble des points

$$\left\{\mathbf{y}\,\middle|\,\mathbf{a}_{euc}\,\mathbf{y}^T = b_{euc}\right\} \tag{7.17}$$

représente l'équation de la frontière de décision où $\mathbf{a}_{euc} = \mathbf{L}_i - \mathbf{L}_j$ et $b_{euc} = \|\mathbf{L}_i\|^2 - \|\mathbf{L}_j\|^2$. Il s'agit d'une équation d'hyperplan dans $\Re^r$ avec un vecteur normal $\mathbf{a}_{euc}$ et une constante b_{euc} qui détermine le décalage de l'hyperplan de l'origine. L'hyperplan détermine deux demi-espaces. Le hyperespace déterminé par (7.17) qui s'étend dans la direction de a_{euc} représente la région de décision de la classe *i*; l'autre hyperespace, qui s'étend dans la direction de $-a_{euc}$ représente la région de décision de la classe *j*.

7.5.2 La mesure cosinus

Basée sur la similarité angulaire dans l'espace d'ancrage de dimension *r*, une donnée de test **y** est classifiée dans la classe *i*, si et seulement si :

$$\frac{\langle \mathbf{y}, \mathbf{L}_i \rangle}{\|\mathbf{y}\| \|\mathbf{L}_i\|} > \frac{\langle \mathbf{y}, \mathbf{L}_j \rangle}{\|\mathbf{y}\| \|\mathbf{L}_j\|},$$

à savoir,

$$\frac{\sum_{k=1}^{r} y_k l_k^i}{\|\mathbf{y}\| \|\mathrm{L_i}\|} > \frac{\sum_{k=1}^{r} y_k l_k^j}{\|\mathbf{y}\| \|\mathrm{L_j}\|}. \tag{7.18}$$

Après simplification et réarrangement des termes, on obtient :

$$\sum_{k=1}^{r} \left(\frac{l_k^i}{\|\mathrm{L_i}\|} - \frac{l_k^j}{\|\mathrm{L_j}\|} \right) y_k > 0, \tag{7.19}$$

ou en utilisant la notation vectorielle :

$$\left(\frac{\mathbf{L}_i}{\|\mathrm{L_i}\|} - \frac{\mathbf{L}_j}{\|\mathrm{L_j}\|} \right) \mathbf{y}^T > 0. \tag{7.20}$$

L'ensemble des points

$$\left\{ \mathbf{y} \,\middle|\, a_{\cos} \mathbf{y}^T = 0 \right\} \tag{7.21}$$

où $a_{\cos} = \frac{\mathbf{L}_i}{\|\mathrm{L_i}\|} - \frac{\mathbf{L}_j}{\|\mathrm{L_j}\|}$, représente la région de décision des deux classes qui est également un hyperplan dans $\Re^r$, avec un vecteur normal $a_{\cos}$ passant par l'origine.

7.5.3 Résultats expérimentaux dans un espace d'ancrage à cinq dimensions

Dans cette section, nous allons présenter les résultats de la classification binaire utilisant un espace d'ancrage à cinq dimensions en utilisant trois modèles de classes supplémentaires dans chacune des dix expériences. La matrice de covariance intraclasse est estimée en utilisant les données d'entraînement des cinq classes d'émotion et elle est partagée par l'ensemble des expériences. La Figure 7.8 montre les gains obtenus par les modèles d'ancrage avant et après WCCN. Pour le système MA-EUC, nous constatons que la normalisation WCCN a permis une augmentation substantielle des performances, similairement à ce qui a été observée avec l'espace d'ancrage à deux dimensions. On note également que WCCN a amélioré légèrement les performances du système d'ancrage à base de cosinus.

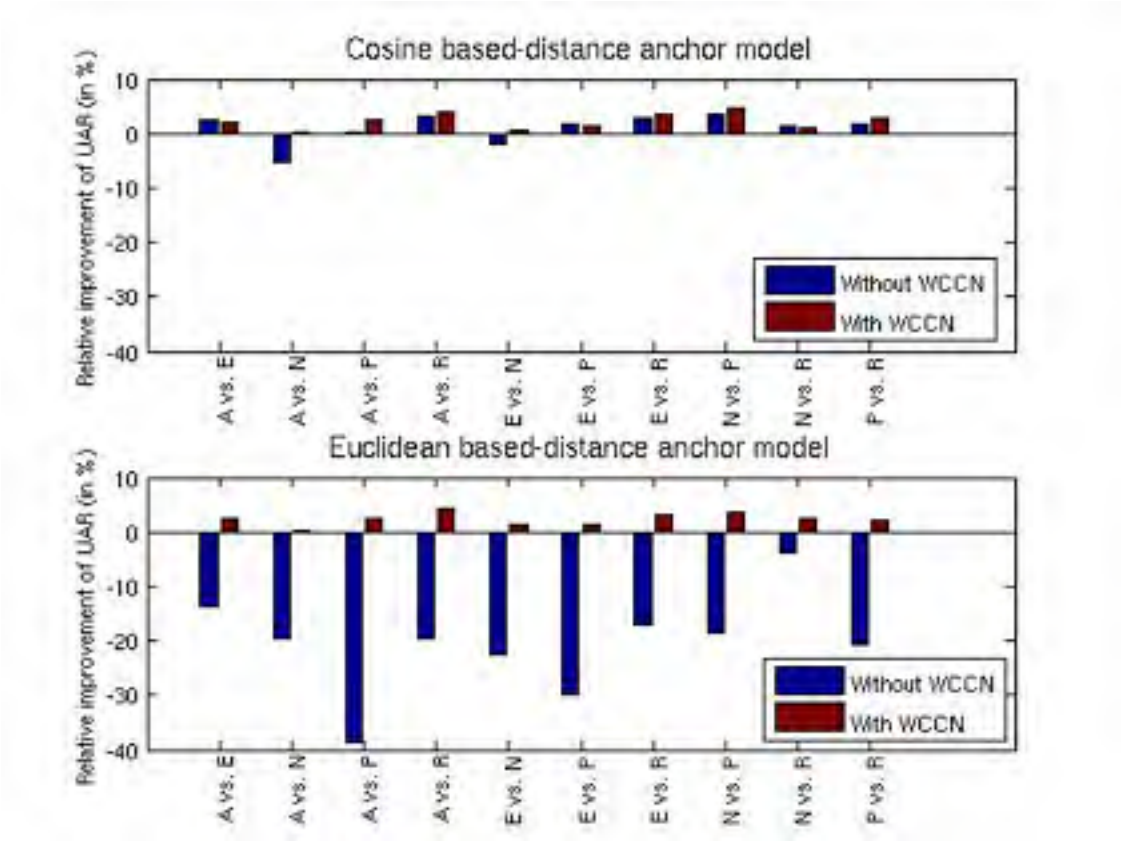

Figure 7.8 Comparaison des gains relatifs des systèmes MA-COS et MA-EUC par rapport à GMM-*Bayes* dans un espace à cinq dimensions avant et après la normalisation WCCN. Les systèmes ont été évalués sur les données de test du corpus FAU AIBO Emotion

La Figure 7.9 montre l'amélioration relative des modèles d'ancrage par rapport aux modèles GMM-*Bayes* dans un espace à cinq dimensions comparée à ce qui a été obtenu dans l'espace à deux dimensions. En premier, nous constatons un gain pour toutes les expériences et ce pour les deux mesures contrairement aux résultats obtenus dans l'espace à deux dimensions où nous avions noté une perte de performance pour certaines expériences. Deuxièmement, nous observons que toutes les améliorations obtenues par les modèles d'ancrage dans un espace à deux dimensions sont également conservées dans l'espace à cinq dimensions. On constate également que toutes les expériences où les systèmes d'ancrage n'ont pas réussi à améliorer les résultats dans l'espace d'ancrage à deux dimensions, montrent une amélioration de l'ordre de ~ 2 % ou plus dans l'espace à cinq dimensions, à l'exception de **A** vs **N**, où les améliorations sont marginales.

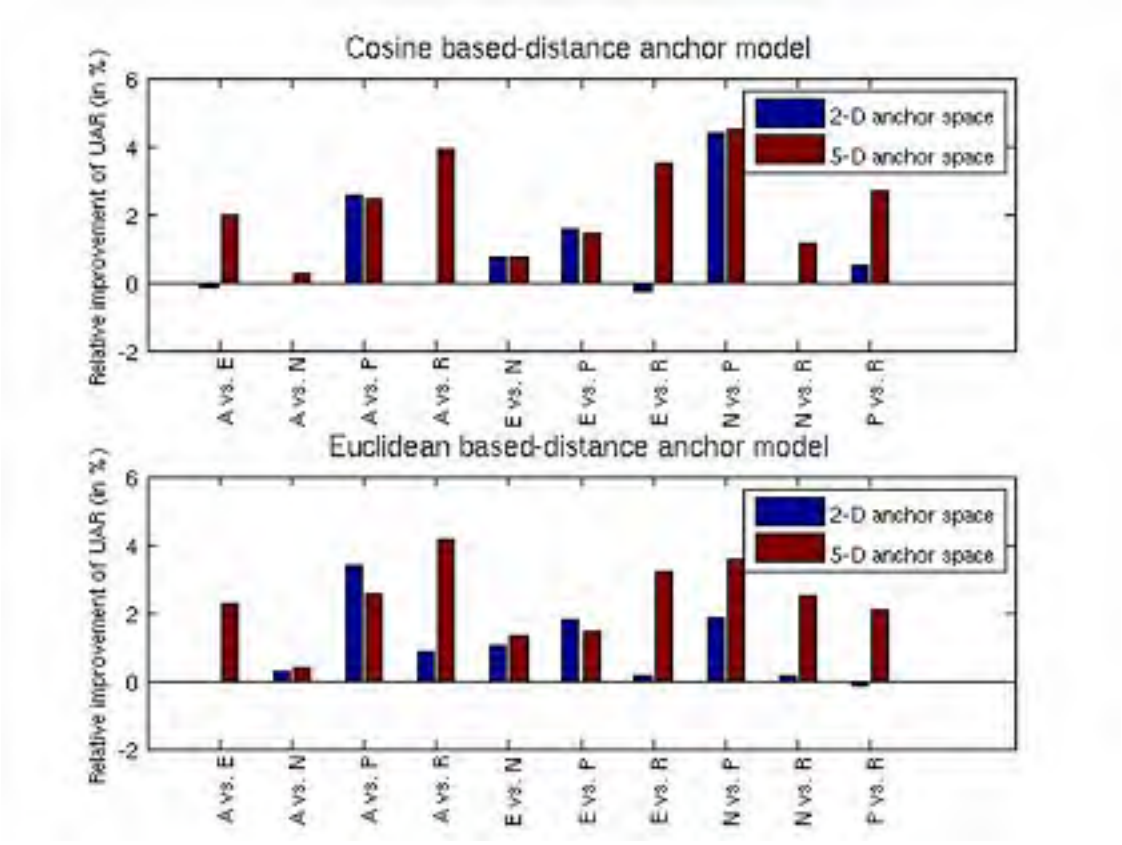

Figure 7.9 Comparaison du gain relatif en UAR par rapport aux systèmes GMM-*Bayes* pour les modèles d'ancrage construits dans un espace à deux versus cinq dimensions. Les performances des modèles d'ancrage sont évaluées après la normalisation WCCN en utilisant les données de test du corpus FAU AIBO Emotion

Une amélioration relative moyenne de 2,3 % est atteinte pour les deux systèmes à base euclidienne et cosinus sur les dix expériences dans l'espace à cinq dimensions comparée à 0,9 % dans l'espace bidimensionnel. Une amélioration relative maximale de 4,5 % est atteinte pour l'expérience **N** vs **P** par le système basé sur la mesure cosinus et un gain relatif de 4,2 % pour l'expérience **A** vs **R** par le système basé sur la distance euclidienne. Dans l'ensemble, ces résultats montrent que les trois nouveaux modèles utilisés pour étendre la dimensionnalité de l'espace d'ancrage ont augmenté le pouvoir discriminatif des modèles d'ancrage dans le cadre de la classification binaire. Cela témoigne du fait que l'espace d'ancrage peut aider à estimer avec une plus grande précision les paramètres des frontières de décision des classificateurs, quand la dimension de l'espace d'ancrage est suffisamment large pour capter les dissimilarités.

Les résultats détaillés en termes d'UAR et WAR obtenus par les systèmes MA-CAOS, MA-EUC et GMM-*Bayes* pour chacune des dix expériences sont donnés au Tableau 7.2.

Tableau 7.2 Résultats UAR et WAR pour les dix expériences à deux classes évaluées sur les données de test du corpus FAU AIBO Emotion. Les performances des modèles d'ancrage sont évaluées dans un espace d'ancrage à cinq dimensions

	Bayes		**Euclidean**		**Cosine**	
	UAR(%)	WAR(%)	UAR(%)	WAR(%)	UAR(%)	WAR(%)
A vs **E**	70,1	72,6	**71,7**	73,7	71,5	73,0
A vs **N**	75,7	75,6	**76,0**	77,8	75,9	77,3
A vs **P**	82,6	85,3	**84,7**	85,2	84,6	85,1
A vs **R**	72,1	72,2	**75,1**	75,0	74,9	74,9
E vs **N**	67,4	68,5	**68,3**	69,1	67,9	68,9
E vs **P**	84,5	90,1	**85,7**	86,2	**85,7**	86,0
E vs **R**	71,7	71,2	74,0	74,5	**74,2**	74,5
N vs **P**	70,7	86,0	73,2	80,5	**73,9**	78,9
N vs **R**	60,3	62,9	**61,8**	69,5	61,0	67,6
P vs **R**	63,0	68,7	64,3	67,1	**64,7**	66,8

7.6 Comparaison de la complexité algorithmique et optimisation

Dans cette dernière section, nous allons comparer la complexité algorithmique des systèmes GMM-*Bayes* et des modèles d'ancrage en termes de temps de calcul pour les étapes d'entraînement et de test. Dans la phase d'entraînement, le système GMM-*Bayes* nécessite l'entraînement d'un modèle GMM pour chacune des classes d'émotion ainsi que le calcul des logarithmes des probabilités de vraisemblance pour les données de validation afin d'ajuster le nombre de gaussiennes qui optimise les performances. Pour les modèles d'ancrage, aucun apprentissage supplémentaire n'est requis excepté l'estimation du vecteur représentatif de chacune des classes d'émotion en calculant la moyenne des logarithmes des probabilités de vraisemblance des données de validation.

Dans la phase de test, les scores des logarithmes des probabilités de vraisemblance des données de test sont calculés pour les deux systèmes. Pour le système GMM-*Bayes* la décision est prise par simple comparaison entre les scores de vraisemblance des modèles de classe. Pour les modèles d'ancrage, le calcul de distance entre le vecteur représentatif de chaque classe et celui de la donnée de test est requis comme étape supplémentaire. Le calcul des métriques de distance pour les modèles d'ancrage ne présente en fait pas une charge de calcul significative étant donné que les mesures sont d'une complexité polynomiale.

En outre, le coût de l'évaluation des métriques euclidienne et cosinus peut être optimisée davantage en utilisant les formules (7.16) et (7.20) au lieu de (7.14) et (7.18), respectivement. Les valeurs a_{euc}, b_{euc} et $a_{\cos}$ peuvent être calculées une seule fois durant la phase d'apprentissage et le coût de l'évaluation des données de test est ramené au produit scalaire de deux vecteurs. La complexité en termes d'opérations nécessaires pour l'évaluation des distances passe de $12r$ à $2r$ (6 fois moins d'opérations) pour la mesure cosinus et de $6r$ à $2r$ pour la mesure euclidienne. Ainsi, l'évaluation de l'une des deux mesures n'aura que $2r$ comme coût de calcul supplémentaire par rapport à la règle de décision de *Bayes*. Cette optimisation peut être particulièrement utile lorsqu'il s'agit d'évaluer les performances d'un système de reconnaissance en mode batch ou lorsque le système est déployé sur de petits

appareils accrochés sur des vêtements biométriques intelligents (contraintes de temps de réponse et consommation de l'énergie de la batterie) utilisés pour le suivi en temps réel du changement d'état émotionnel.

7.7 Conclusion

Dans ce chapitre nous avons étudié sur le plan théorique et expérimental les modèles d'ancrage basés sur la mesure euclidienne et cosinus dans un contexte de classification binaire. À partir des formules des frontières de décision des modèles d'ancrage, nous avons montré que la pente du segment reliant les vecteurs représentatifs des classes, possède un impact important sur les performances des modèles d'ancrage et peut nous instruire sur le meilleur choix de la mesure de similarité à adopter. Si les modèles d'ancrage basés sur les deux mesures similarité et le modèle GMM-*Bayes* appartiennent à la même catégorie de classificateurs, à savoir des discriminants linéaires, ces derniers diffèrent dans le type de paramètres à estimer pour ajuster la frontière de décision. Nous avons également étudié dans quelles conditions les modèles d'ancrage peuvent coïncider avec les modèles GMM-*Bayes*.

Sur le plan expérimental, nous avons montré que dans un espace d'ancrage à deux dimensions, les deux modèles d'ancrage peuvent s'avérer non suffisamment discriminants pour tirer profit de l'approche de similarité. Nous avons proposé d'ajouter de nouveaux modèles basés sur des données externes pour étendre la dimensionnalité de l'espace d'ancrage et augmenter le pouvoir discriminatif des modèles d'ancrage. Notons que les données utilisées pour entrainer les modèles additionnels ne devraient pas forcément provenir d'autres classes d'un même corpus. Tout corpus externe qu'il soit émotionnel ou non peut servir pour l'entrainement des modèles additionnels. En effet, dans (Attabi *et al.* 2012a), nous avons combiné des données de classes appartenant à trois différents types de corpus pour étendre l'espace d'ancrage de trois problèmes de classification binaire : les traits de personnalité, l'amabilité du locuteur et la détection de pathologie à partir de la parole. Les résultats expérimentaux que nous avons réalisés montrent que les modèles d'ancrage, indépendamment des deux métriques utilisées, obtiennent de meilleures performances qu'un

système GMM-*Bayes*. Fait intéressant, ces améliorations de performances sont obtenues simplement par ajustement de la frontière de décision, sans aucune étape d'apprentissage supplémentaire ni une augmentation importante de complexité en temps de calcul durant la phase de test.

CHAPITRE 8

LES MODÈLES D'ANCRAGE POUR LA COMBINAISON DE CLASSIFICATEURS

8.1 Introduction

L'extraction de traits caractéristiques discriminants et robustes aux bruits et à la variabilité intraclasse est une étape importante dans la construction de systèmes de RAE robustes et demeure un domaine de recherche très active dans la communauté. La diversification des différents types de traits et leurs fusion est un autre axe exploré par plusieurs auteurs pour améliorer les performances de classification. Par ailleurs le choix de la meilleure stratégie de combinaison de classificateurs adaptée au problème de la reconnaissance des émotions continue de captiver l'attention des chercheurs. Dans ce chapitre, nous allons présenter nos travaux qui essayent d'apporter une contribution à chacun de ces trois points. Sur le plan des traits caractéristiques, nous proposons deux nouveaux types de descripteurs spectraux expérimentés pour la première fois en reconnaissance des émotions. Ces deux types de traits, appelés (i) les traits MFCC et PLP de type *multitaper* (Attabi *et al.* 2013) et (ii) les coefficients cepstraux de modulation d'amplitude (AMCC, *amplitude modulation cepstral coefficients*) basés sur l'opérateur non linéaire lissé de l'énergie (Alam *et al.* 2013), seront présentés dans la section 2. Outre les bonnes performances obtenues par ces nouveaux traits expérimentés individuellement, nous montrerons dans la section 3 que les systèmes bâtis sur ces traits véhiculent des informations complémentaires discriminantes utiles pour la fusion. Enfin, nous mettrons en évidence dans la section 4, l'autre propriété intéressante des modèles d'ancrage quand ceux-ci sont utilisés comme méthode de combinaison de classificateurs.

8.2 Nouveaux traits spectraux pour la reconnaissance des émotions

Les paramètres MFCC (Davis *et al.* 1980) et PLP (Perceptual Linear Prediction, (Hermansky, 1990)) sont deux paramètres spectraux intensivement utilisés et reconnus également pour donner de bonnes performances dans le domaine du traitement de la parole en général et celui de la reconnaissance des émotions en particulier. Les techniques de

paramétrage MFCC et PLP ont pour but de simuler la manière dont un son est perçu par un humain. Habituellement, le spectre est estimé en utilisant un périodogramme fenêtré via la transformée de Fourier discrète (DFT). Malgré leur faible biais, une conséquence de fenêtrage des données est l'augmentation de la variance du spectre estimé (Riedel *et al.* 1995; Thomson, 1982). Une technique élégante pour réduire la variance spectrale est de substituer l'estimation du périodogramme fenêtré par une estimation spectrale basée sur une multitude de fenêtres connue sous le nom de l'estimation multitaper du spectre (Riedel *et al.* 1995; Thomson, 1982). Dans le processus d'estimation spectrale multitaper, un ensemble de fenêtres orthogonales est appliqué au signal de parole de courte durée et les estimations spectrales résultantes sont mises en moyenne, ce qui réduit la variance spectrale. La méthode de multitaper a été utilisée dans les applications géophysiques (Prieto *et al.* 2007), de l'amélioration de la parole (*speech enhancement*, (Hu *et al.* 2004)) et dans la reconnaissance du locuteur et de parole (Alam *et al.* 2013; Kinnunen *et al.* 2012). Les performances obtenues avec cette méthode d'estimation spectrale dépassaient celles du périodogramme fenêtré.

La deuxième méthode que nous proposons afin de réduire l'effet du bruit sur la dégradation des performances des MFCC est l'utilisation de l'opérateur d'énergie non linéaire (*nonlinear energy operator*, NEO) pour l'extraction de nouvelles représentations cepstrales. L'avantage de NEO est qu'il utilise seulement quelques échantillons du signal d'entrée pour estimer l'énergie requise pour générer un signal AM-FM et le séparer en deux composantes : amplitude et fréquence. Avec l'approche basée sur NEO, l'hypothèse du signal stationnaire n'est plus requise comme c'est le cas pour la prédiction linéaire (LP) ou la transformée de Fourier (Mitra *et al.* 2012). La modulation d'amplitude et de fréquence (AM, FM) d'un signal de parole joue un rôle important dans la perception et la reconnaissance de la parole (Georgogiannis *et al.* 2012). Récemment, le modèle AM-FM a été appliqué dans le domaine de la reconnaissance de l'émotion (Georgogiannis *et al.* 2012; Wu *et al.* 2011). Dans (Wu *et al.* 2011), la modulation de traits spectraux à long terme basée sur la transformée de Hilbert a été proposée alors que c'est les traits MFCC à base de l'opérateur d'énergie *Teager* (TEO) qui ont été expérimentés dans (Georgogiannis *et al.* 2012). L'approche de démodulation basée sur l'opérateur d'énergie non linéaire a de nombreuses propriétés intéressantes comme

la simplicité, l'efficacité et l'adaptabilité aux variations de signal instantané (Alam *et al.* 2013; Maragos *et al.* 1993). Dans (Alam *et al.* 2013) nous avons proposé d'utiliser les coefficients cepstraux de modulation d'amplitude (AMCC) basés sur l'opérateur non linéaire lissé de l'énergie (SNEO) (Mukhopadhyay *et al.* 1998; Potamianos, 1995) comme descripteurs pour la classification des émotions.

8.2.1 Estimation du Spectrum multitaper

Dans les applications de traitement de la parole, le spectre de puissance est souvent estimé à l'aide d'un estimateur du spectre fenêtré directe. Pour la *m*-ième trame et la *k*-ième série (bin) de fréquence, une estimation du périodogramme fenêtré (*simple taper* en anglais) peut être formulée comme suit :

$$\hat{S}_d(m,k) = \left| \sum_{j=0}^{N-1} w(j)\, s(m,j)\, e^{-\frac{i2\pi jk}{N}} \right|^2, \tag{8.1}$$

où $k \in \{0,1,\cdots,K-1\}$ désigne l'indice de la série de fréquence, N est la longueur de trame, $s(m, j)$ est le signal de parole dans le domaine temporel et $w(j)$ désigne la fonction de fenêtrage dans le domaine temporel, également connu sous le nom de fenêtre de pondération. Une fenêtre d'observation, telle qu'une fenêtre de *Hamming*, est généralement symétrique et amoindrie aux bordures de la trame.

Le fenêtrage réduit le biais, à savoir, la valeur prévue de la différence entre le spectre estimé et le spectre réel, mais ne réduit pas la variance de l'estimation spectrale (Kay, 1988). Pour réduire la variance de l'estimateur MFCC et PLP, nous utiliserons l'estimation spectrale de type multitaper au lieu de l'estimation de périodogramme fenêtré (Hansson-Sandsten *et al.* 2009; Riedel *et al.* 1995; Thomson, 1982). L'estimateur de spectre multitaper, qui utilise M fonctions de fenêtrage orthogonales au lieu d'une seule fenêtre, peut être exprimé comme suit :

$$\hat{S}_{MT}(m,k)=\sum_{p=1}^{M}\lambda(p)\left|\sum_{j=0}^{N-1}w_p(j)s(m,j)e^{-\frac{i2\pi jk}{N}}\right|^2, \tag{8.2}$$

où N représente la longueur de trame et w_p le *p-ième* périodogramme fenêtré de données (p = 1, 2, ...,. M) utilisé pour l'estimation spectrale $\hat{S}_{MT}(\cdot)$. Enfin, $\lambda(p)$ est le coefficient de pondération associé au *p-ième* périodogramme fenêtré. Les fenêtres $w_p(j)$ sont généralement choisies de sorte qu'elles soient orthonormées. Le spectre multitaper est donc estimé via une moyenne pondérée de M sous-spectres individuels. L'idée derrière le multifenêtrage est de réduire la variance des estimations spectrales en calculant la moyenne de M estimations spectrales directes, chacune avec une fenêtre de données distincte. Si tous les M périodogrammes fenêtrés sont orthogonaux deux à deux et bien conçus pour prévenir les fuites spectrales, les estimations multitaper résultantes dépasseront le périodogramme fenêtré en termes de la variance réduite, en particulier, lorsque le spectre d'intérêt a une plage dynamique élevée ou des variations rapides (McCoy *et al.* 1998). Par conséquent, la variance des traits MFCC et PLP calculés à travers une estimation spectrale multitaper sera également faible.

Divers périodogrammes fenêtrés ont été proposés dans la littérature pour l'estimation du spectre. Dans ce chapitre, nous allons présenter la famille de périodogrammes fenêtrés *sinus*, qui sont faciles à calculer et sont orthogonaux deux à deux, et peuvent être formulés comme suit (Riedel *et al.* 1995) :

$$w_p(j)=\sqrt{\frac{2}{N+1}}\sin\left(\frac{\pi p(j+1)}{N+1}\right),\ j=0,1,....,N-1 \tag{8.3}$$

La constante multiplicative assure l'orthogonalité des périodogrammes fenêtrés. Les périodogrammes fenêtrés *sinus* sont appliqués avec une pondération optimale pour l'analyse de cepstre, appelé estimateur sinusoïdal pondéré de cepstre (*Sinusoidal Weighted Cepstrum Estimator*, SWCE) dans (Hansson-Sandsten *et al.* 2009). Plus de détails sur d'autres familles de périodogrammes fenêtrés peuvent être trouvés dans (Attabi *et al.* 2013 ; Alam *et al.* 2013).

La Figure 8.1 illustre une famille de M (M = 6) périodogrammes fenêtrés dans le domaine temporel et fréquentiel.

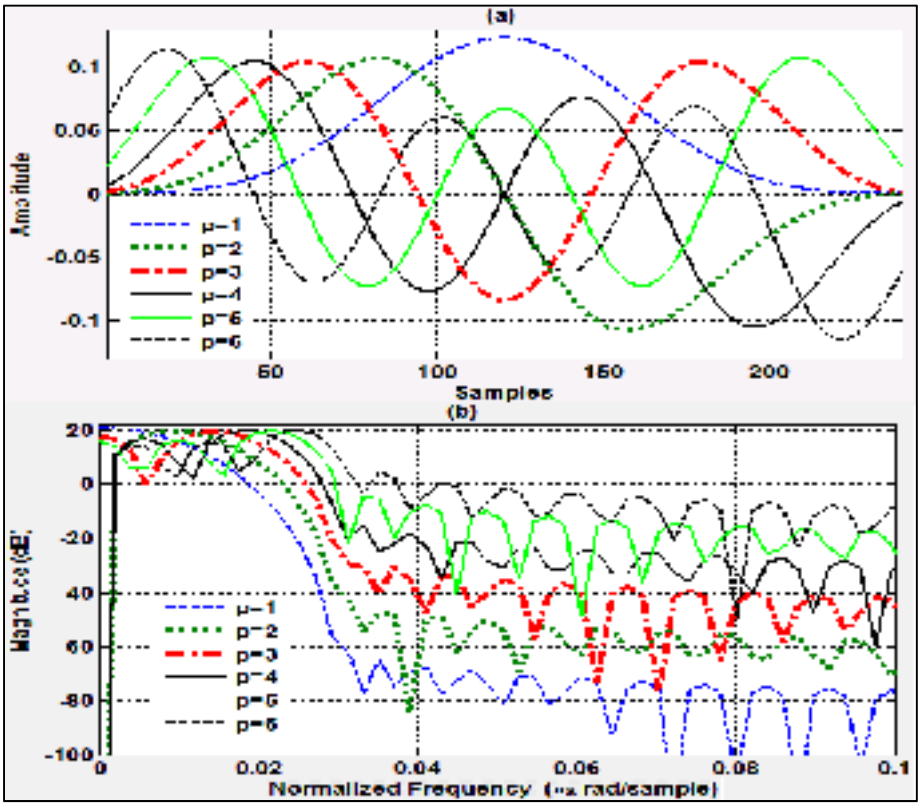

Figure 8.1 Multitapers pour N = 256, M = 6 des domaines (a) du temps et (b) de fréquence
Tirée de Attabi et al. (2013)

8.2.2 Extraction des MFCC et PLP multitaper

Le processus d'extraction des traits MFCC et PLP de type multitaper est présenté à travers le bloc-diagramme de la Figure 8.2. Comme prétraitement, le décalage du courant continu (*DC offset* en anglais) est supprimé et le spectre du signal est pré-accentué. Le signal de parole est par la suite décomposé en une série de trames de 20 à 30 *ms* qui se chevauchent avec un décalage de trame de 10 *ms*. Chaque trame est ensuite multipliée par une fenêtre unique (lorsque M = 1), tel qu'une fenêtre de *Hamming* ou par plusieurs fenêtres afin de réduire l'effet de la discontinuité introduite par le processus de tramage. Le spectre de puissance est

estimé par le calcul du carré de la magnitude de la transformée de Fourier discrète (DFT) de la trame. Le spectre du signal de parole est ensuite filtré par un groupe de filtres triangulaires passe-bande qui simulent les caractéristiques de l'oreille humaine appelés échelle de Mel.

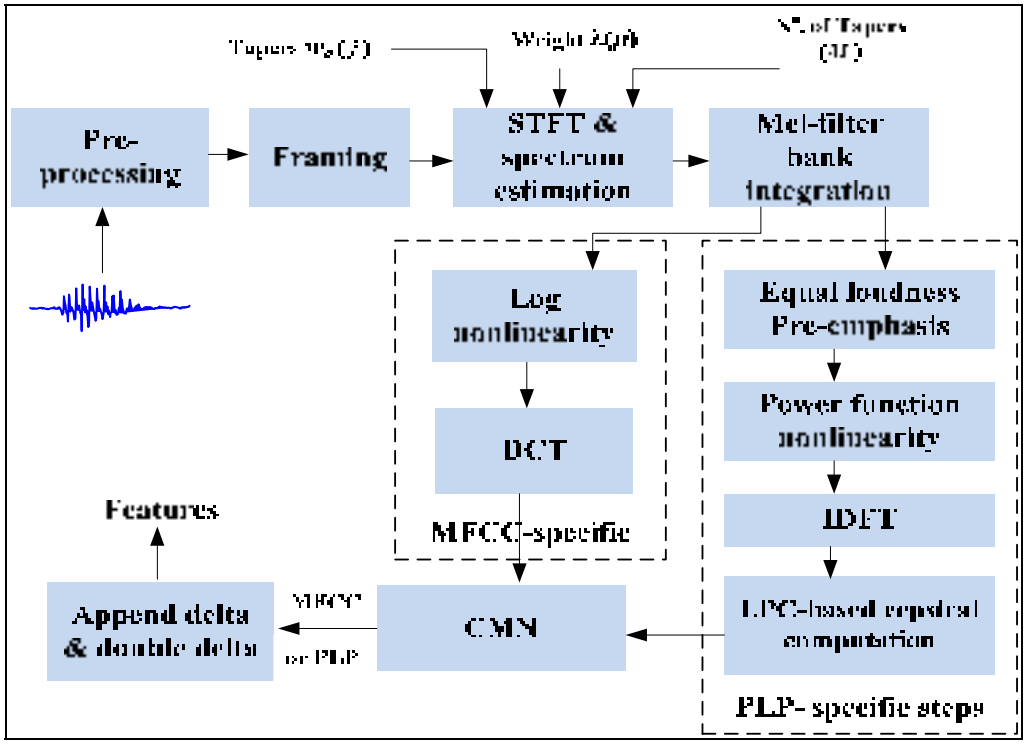

Figure 8.2 Schéma illustrant l'extraction des traits MFCC et PLP basée sur l'estimation spectrale de type multitaper

Après ces étapes d'extraction de traits communes aux MFCC et PLP, le processus de calcul PLP se poursuit par l'application de la loi de la puissance proposée par (Hermansky, 1990). La transformée de Fourier discrète inverse (IDFT) est ensuite appliquée et les coefficients de la prédiction linéaire sont générés en utilisant la récursivité cepstrale (Gold *et al.* 1999). Quant aux MFCC, le processus d'extraction se poursuit par le calcul du logarithme de l'énergie de chaque filtre. Le cepstre à l'échelle de fréquence *Mel* est obtenu par le calcul de la transformée en cosinus discrète du logarithme de la sortie des filtres (reconversion du log-Mel-spectre vers le domaine temporel). Une fois les traits MFCC et PLP statiques sont extraits, les premières et secondes dérivées des traits sont calculées et ajoutées au vecteur de traits.

8.2.3 Extraction des traits AMCC

Les traits de type AMCC sont extraits selon le bloc diagramme de la Figure 8.3. Après suppression du silence, le signal de parole est segmenté en trames et un fenêtrage de type *Hamming* est appliqué. Un banc de filtres *Gammatones* est par la suite appliqué pour décomposer le signal en *C* canaux (*C* = 40) couvrant la plage de fréquences de 100-7200 Hz (pour une fréquence d'échantillonnage de 16 000 Hz). Le spectre de puissance AM pour chaque canal est alors estimé en utilisant l'opérateur d'énergie non linéaire lissée (SNEO).

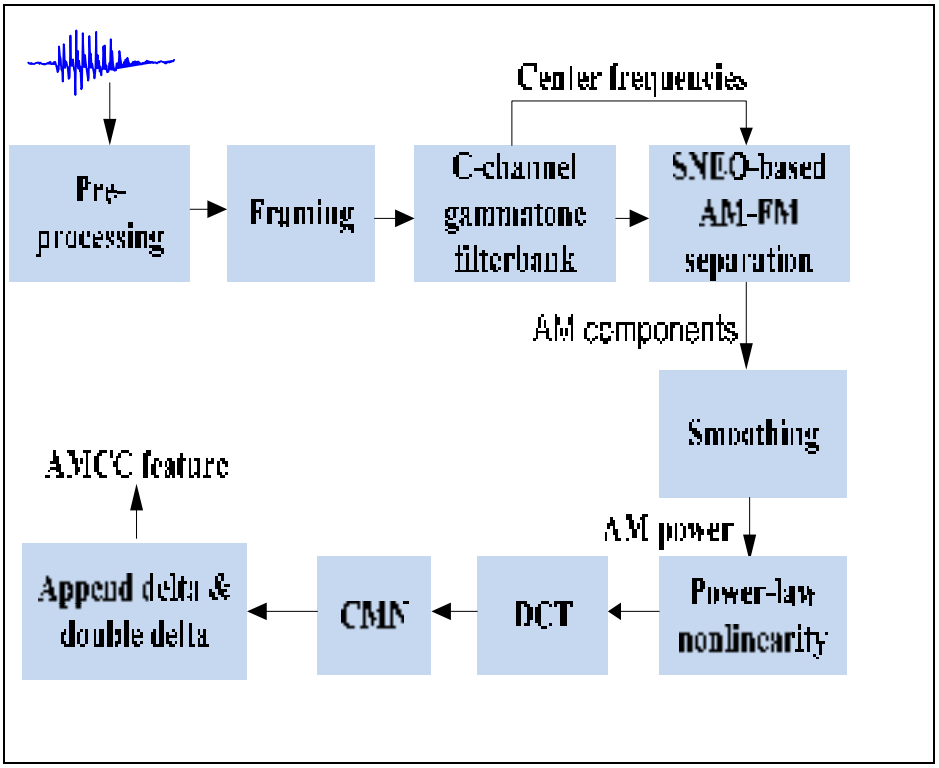

Figure 8.3 Bloc diagramme du processus d'extraction des traits AMCC

Soit $x(c,n)$ la trame vocale du *c*-ième canal où c, $c = 1, \dots, C$, est l'indice du canal du banc de filtres *Gammatones* à *C* canaux, n, $n = 1, \dots, N$, est l'indice temporel discret et *N* est la taille de trame en nombre d'échantillons. Le NEO standard de $x(c,n)$ peut être exprimé comme suit :

$$\Psi\big(x(c,n)\big) = x(c,n)\,x(c,n) - x(c,n-1)\,x(c,n+1) \tag{8.4}$$

La version lissée (SNEO) de NEO est obtenue en utilisant une fenêtre de lissage *w(n)* selon (8.5) où $\otimes$ représente le symbole de l'opérateur de convolution. Un filtre de lissage binomial à sept points avec une réponse impulsionnelle (1, 6, 15, 20, 15, 6, 1) est appliqué (Maragos *et al.* 1993). La valeur absolue de la valeur NEO permet de s'assurer que la valeur de l'énergie est toujours positive.

$$\Psi^s\big(x(c,n)\big) = \big|\Psi\big(x(c,n)\big)\big| \otimes\ w(n) \tag{8.5}$$

Pour le *c*-ième canal, la composante AM est estimée en utilisant l'algorithme de séparation d'énergie discret (DESA) comme suit :

$$|\hat{a}(c,n)| = \left(\frac{\Psi^s\big(x(c,n)\big)}{1 - \left(1 - \dfrac{\Psi^s\big(y(c,n)\big) + \Psi^s\big(y(c,n+1)\big)}{4\,\Psi^s\big(x(c,n)\big)}\right)^2} \right)^{1/2} \tag{8.6}$$

où $y(c,n) = x(c,n)\ - x(c,n-1)$. Afin de réduire la plage dynamique, la composante AM estimée est lissée en utilisant un filtre de la moyenne avec une fenêtre coulissante de taille égale à cinq. La puissance AM de la *m*-ième trame du *c*-ième canal est calculée comme suit :

$$P(m,c) = \sum_{n=1}^{N} (|\hat{a}(c,n)|^2) \tag{8.7}$$

Un vecteur de traits de dimension 13 est obtenu après application de la fonction *puissance* avec un coefficient égal à 0,07 et de la transformée de cosinus discrète (DCT). Les traits sont ensuite normalisés par soustraction de la moyenne des cepstraux et finalement les première et seconde dérivées sont ajoutées aux traits cepstraux statiques.

8.2.4 Évaluation des performances individuelles des traits proposés

Dans cette section, nous évaluerons individuellement les performances des nouveaux traits en utilisant les modèles GMMs et les données du corpus FAU AIBO Emotion. Les performances ont été optimisées selon le critère UAR en utilisant la validation croisée à neuf plis sur les données d'apprentissage. Afin d'évaluer les performances des nouveaux traits, nous avons pris les traits MFCC et PLP basés sur une fenêtre unique de type *Hamming* comme traits de référence. Les traits PLP ont été extraits selon le processus de base de HTK (Young *et al.* 2006), dont l'analyse spectrale auditif est basé sur un banc de filtres *Mel* à la place d'un banc de filtres Bark en forme de trapèze.

8.2.4.1 Résultats des traits multitaper et des traits AMCC

Les traits MFCC et PLP de type multitaper sont calculés en utilisant un nombre de périodogrammes fenêtrés $M=6$, le même nombre qui a permis d'optimiser les performances pour les problèmes de la reconnaissance vocale et de la vérification du locuteur comme indiqué dans (Alam *et al.* 2013; Kinnunen *et al.* 2012). Tous les vecteurs MFCC et PLP, à fenêtre unique ou multiple sont composés des 12 premiers coefficients statiques plus l'énergie et leurs premières et secondes dérivées. Les résultats obtenus pour chaque type de traits sont données au Tableau 8.1. Nous observons que l'estimation du spectre basée sur le multifenêtrage permet d'améliorer les scores UAR comparé à une fenêtre unique que ce soit pour les traits MFCC ou PLP. Des gains relatifs de 3,84 % et 4,02 % sont obtenus respectivement pour les traits MFCC et PLP. Par ailleurs, les traits AMCC basés sur l'opérateur d'énergie non-linéaire lissé permet d'améliorer encore plus les performances UAR comparé aux autres traits en obtenant un gain relatif de 6,7 % par rapport aux traits MFCC ce qui met en évidence la puissance de cette méthode d'estimation de l'énergie.

Tableau 8.1 Résultats des traits MFCC et PLP de type multitaper (MFCC_MT et PLP_MT) ainsi que les traits AMCC comparés aux traits MFCC et PLP conventionnels

	UAR	**WAR**
MFCC	41,70 %	42,50 %
MFCC_MT	43,30 %	40,62 %
PLP	42,34 %	41,47 %
PLP_MT	43,50 %	41,80 %
AMCC	44,50 %	42,58 %

8.3 Complémentarité des traits

Dans cette section nous allons étudier la complémentarité des traits proposés dans la discrimination des classes d'émotion à travers une fusion tardive de ces traits. Dans une fusion tardive (*late fusion* en anglais) la combinaison est réalisée en niveau des scores (sorties) des classificateurs par opposition à une fusion hâtive (*early fusion*) où la combinaison est faite au niveau des vecteurs des traits d'entrée.

8.3.1 Analyse des matrices de confusion

Dans la Figure 8.4, nous avons tracé les taux de rappel obtenus pour chacune des classes d'émotion en fonction du type de traits utilisé. Ces taux ont été calculés à partir des matrices de confusion des expériences précédentes. Nous constatons qu'il existe une complémentarité dans le type de classe d'émotion reconnue et ce particulièrement entre le classificateur basé sur les traits AMCC d'un côté et les classificateurs basés sur les autres traits de l'autre côté. Ainsi AMCC donnent les meilleurs résultats pour les classes **A**, **N** et **P** et les plus basses performances pour les classes **E** et **R**. Les meilleurs résultats pour les classes **E** et **R** sont obtenus par contre par PLP et MFCC_MT respectivement. Nous constatons également que la

différence de reconnaissance de classe en absolue entre le meilleur et le pire classificateur peut aller de 6,5 % pour la classe **N** jusqu'à 22,6 % pour la classe **E**.

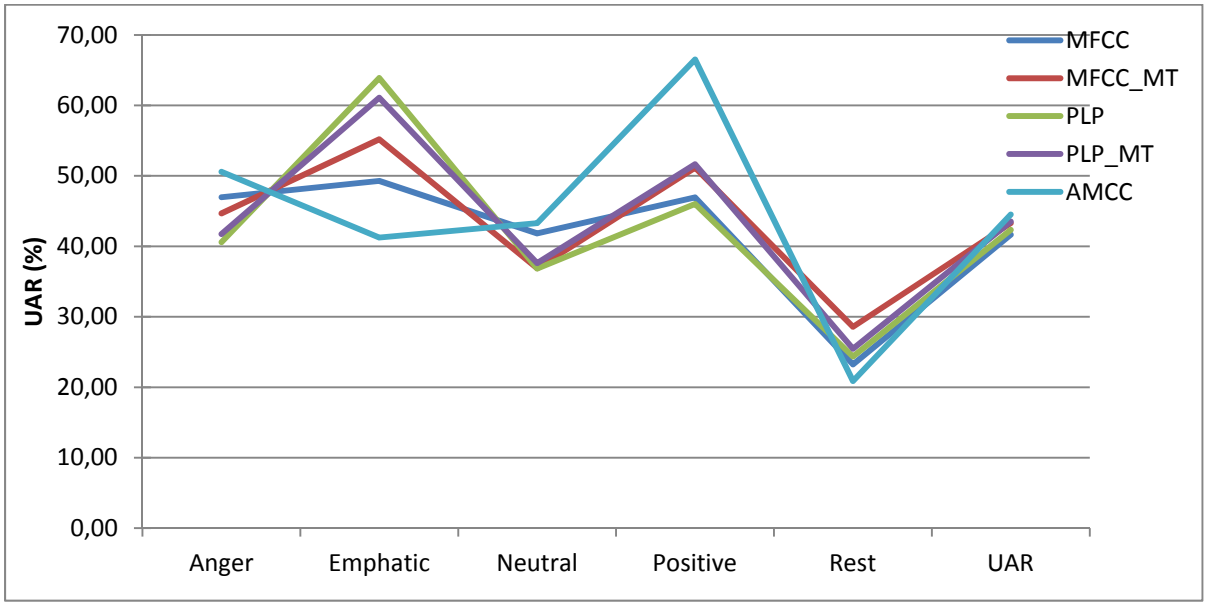

Figure 8.4 Résultats de classification par classe d'émotion en fonction du type de traits utilisé

8.3.2 Combinaison des traits

Cette complémentarité de compétence dans la reconnaissance des classes peut se traduire par une augmentation de la capacité d'inférence en cas de fusion. Nous avons donc implanté la fusion dans cette section en combinant les classificateurs au niveau des scores avec la méthode naïve bayésienne.

Afin de contrer le problème de la distribution non balancée des classes de données, Nous avons échantillonné les données. Les performances de classification obtenues sur les données de test du corpus FAU AIBO Emotion après fusion sont de 46,56 % et 38,55 % pour les critères UAR et WAR respectivement. Ces résultats montrent un gain relatif de 4,63 % obtenu en terme UAR par rapport au meilleur système individuel, AMCC en occurrence. La

complémentarité entre les différents traits dans la reconnaissance des classes d'émotions s'est traduite donc par une amélioration du taux classification à travers la fusion des classificateurs.

8.4 Fusion avec les modèles d'ancrage

Dans les deux chapitres précédents nous avons vu que les modèles d'ancrage permettaient d'améliorer les performances des modèles GMMs utilisés comme partie frontale du système de reconnaissance des émotions. Dans cette section, nous montrerons que les modèles d'ancrage peuvent aussi servir de méthode de combinaison efficace de classificateurs.

8.4.1 Définition de l'espace d'ancrage de fusion

Soit r le nombre de classificateurs à combiner dans un problème à C-classes d'émotion. L'espace d'ancrage de fusion, de dimension égale à $C \times r$, sera engendré par les C modèles de classes de chacun des r classificateurs à combiner. Soit $\{\lambda_1^1, \cdots, \lambda_C^1, \cdots, \lambda_1^r, \cdots, \lambda_C^r\}$ l'ensemble des modèles de référence de l'espace de fusion où λ_i^k est le modèle GMM de la classe i du k-ième classificateur.

Soit $\mathbf{X}$ une séquence de T vecteurs de traits acoustiques de dimension d représentant un énoncé. La projection de $\mathbf{X}$ dans l'espace d'ancrage de fusion est définie comme suit :

$$\Re^d \longrightarrow \Re^{C \times r}$$

$$\mathbf{X} \longrightarrow L(\mathbf{X}) = \left[l_1^1 \cdots l_C^1, \cdots, l_1^r \cdots l_C^r \right]^T$$

où $l_i^k(\mathbf{X}) = \frac{1}{T} \log P\left(\mathbf{X} \middle| \lambda_i^k\right)$, représente la moyenne des logarithmes de la probabilité de vraisemblance de l'énoncé $\mathbf{X}$ contre le modèle de la classe i du k-ième classificateur.

8.4.2 Normalisation des scores

La normalisation des scores avec WCCN est une étape importante dans la modélisation d'un système d'ancrage. WCCN permet de réduire la variance intraclasse des scores. WCCN nécessite l'estimation de la matrice **R** issue de la décomposition de *Cholesky* de la matrice de covariance intraclasse des scores dans l'espace d'ancrage. La matrice **R** est de taille $(C \times r, C \times r)$, et nécessite donc un apprentissage de $Cr \times (Cr-1)/2$ paramètres. Pour le cas de fusion de cinq classificateurs d'un problème à cinq classes d'émotion, le nombre de paramètres de la matrice de covariance à estimer s'élève à 300 éléments, un nombre qui peut se révéler élevé pour la taille d'un corpus d'émotion. Pour cette raison, nous proposons une nouvelle forme de matrice **R**, qui permettrait de réduire considérablement le nombre d'éléments non nuls de la matrice à estimer. L'idée est de substituer l'apprentissage de la matrice de covariance associée à tout l'espace d'ancrage de fusion par l'apprentissage de sous matrices, de dimension peu élevées, qui correspondront aux sous espaces d'ancrage associés à chacun des sous-classificateurs (classificateurs individuels). Chaque sous-matrice est estimée indépendamment des autres sous-matrices en utilisant les données d'apprentissage associées au sous-classificateur sous-jacent. Ainsi le nombre de paramètres à estimer est réduit à C(C-1)/2 (10 éléments pour notre cas de figure au lieu de 300) tout en continuant de bénéficier du même nombre de données d'apprentissage que celui utilisé pour l'estimation de la matrice de covariance intégrale dans la première méthode. La nouvelle matrice creuse de covariance intraclasse à estimer est constituée de r sous-matrices non-nulles de dimension $(C \times C)$ chacune, alors que la valeur des éléments restants est mis à zéro. La matrice creuse est formulée comme suit:

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}^1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{R}^2 & \cdots & \mathbf{O} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{R}^r \end{bmatrix} \tag{8.8}$$

où $\mathbf{R}^i$ représente, la matrice de dimension (C,C) de la décomposition de *Cholesky* de la matrice de la covariance intraclasse associée au i-ième classificateur fusionné, et **O** la matrice zéro de dimension (C,C). La matrice de covariance ainsi calculée ne contiendra donc que l'information sur la corrélation entre variables engendrées par le même sous-classificateur alors que l'information sur la corrélation qui existe entre les variables appartenant à différents classificateurs sera absente. Notre hypothèse motivant cette proposition est que le bénéfice (impact) de l'information perdue sera compensé par l'apport d'une estimation plus robuste des autres paramètres et par conséquent une meilleure capacité de généralisation pour les données de test.

8.4.3 Résultats expérimentaux des modèles d'ancrage de fusion

Le Tableau 8.2 montre les performances de reconnaissance des émotions obtenues sur les données de test après fusion au moyen des modèles d'ancrage en fonction de la méthode de normalisation utilisée. D'abord, nous observons qu'il y a dégradation considérable des performances quand le modèle d'ancrage de fusion est utilisé sans normalisation préalable des scores et ceci indépendamment de la mesure de similarité utilisée (cosinus ou euclidienne). Ce comportement des modèles d'ancrage utilisé dans un contexte de fusion contraste avec celui observé quand le modèle d'ancrage est utilisé comme simple classificateur, où la métrique cosinus offrait une amélioration même en absence de toute normalisation. Par ailleurs, nous observons que la normalisation WCCN utilisant le premier modèle de matrice de covariance intraclasse (méthode standard) remédie au problème de dégradation des performances sans pour autant améliorer les performances par rapport aux résultats du meilleur classificateur simple. Finalement nous constatons que la deuxième méthode que nous avons proposée pour le calcul de la matrice **R** (le système Matrice-creuse-Euclid.) dépasse en performance la première méthode et permet d'améliorer les résultats de 5,64 % en terme UAR par rapport au meilleur classificateur fusionné.

Tableau 8.2 Résultats UAR de la combinaison de classificateurs au moyen des modèles d'ancrage en fonction des méthodes de normalisation

Systèmes	**UAR**	**WAR**
Sans-WCCN Euclid.	28,54 %	24,25 %
Sans-WCCN Cosinus	28,79 %	23,70 %
Matrice-intégrale-Euclid.	44,50 %	42,40 %
Matrice-creuse-Euclid.	**47,01 %**	42,91 %

8.4.4 Modèles d'ancrage versus autres méthodes de combinaison

Nous allons maintenant comparer l'efficacité des modèles d'ancrage utilisés dans le contexte d'une fusion de classificateurs avec d'autres méthodes de combinaison. Les méthodes de combinaison peuvent être des méthodes démunies d'étape d'apprentissage à l'instar des règles de la *somme* et du *maximum*. C'est à cette catégorie qu'appartient la méthode de fusion basée sur les modèles d'ancrage. Le deuxième groupe de méthodes que nous allons évaluer reposent sur une phase d'apprentissage pour entraîner les paramètres de fusion. Dans ce groupe on retrouve des méthodes plus complexes telles que les réseaux de neurones, SVM, la régression logistique et les forêts d'arbres décisionnels. La plupart de ces méthodes sont sensibles au problème de la distribution biaisée des classes. Par conséquent, nous aurons recours aux mêmes méthodes d'échantillonnage vues au CHAPITRE 6 pour mitiger l'impact des données non équilibrées sur les performances de classification après combinaison.

La Figure 8.5 montre les résultats UAR sur les données de test du corpus FAU AIBO Emotion obtenus pour chacune des méthodes de combinaison impliquant un apprentissage en fonction de la méthode d'échantillonnage utilisée. Nous constatons d'abord que les techniques d'échantillonnage permettent en général d'améliorer les performances par rapport aux résultats obtenus avant échantillonnage. Notons également que la méthode naïve bayésienne est la méthode la moins sensible face au problème de distribution biaisée. Par ailleurs, le sous-échantillonnage est la méthode la plus performante indépendamment de la méthode de combinaison utilisée.

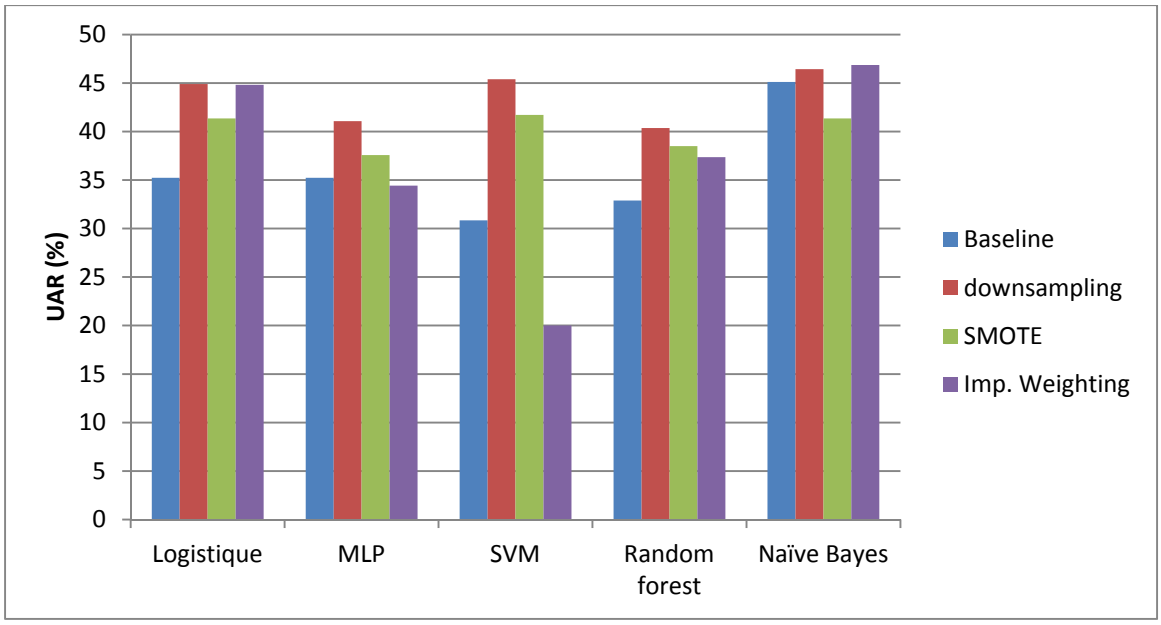

Figure 8.5 Résultats de classification au moyen de méthodes impliquant un apprentissage, en fonction de la méthode d'échantillonnage. Les résultats UAR sont obtenus en utilisant les données de test du corpus FAU AIBO Emotion

La méthode de pondération d'importance donne des performances similaires à la méthode de sous-échantillonnage pour naïve bayésienne et la régression logistique. Finalement, les résultats montrent que la méthode naïve bayésienne offre également les meilleures performances de reconnaissance dans cette catégorie de méthodes de combinaison de classificateurs. À noter que des méthodes d'ensemble telles que *Bagging* et *Boosting* ont été expérimentées. Cependant, aucune amélioration significative n'a été observée à l'issue de l'utilisation de ces méthodes d'ensemble.

Les meilleurs résultats obtenus pour chacune des méthodes de combinaison sont récapitulés au Tableau 8.3. Nous constatons que la règle *somme* donnent de meilleurs résultats que la règle du *maximum*, les méthodes MLP, régression logistique et les forêts d'arbres décisionnels. Rappelons que la règle *somme* effectue une somme sur des logarithmes de probabilités de vraisemblance ce qui correspond à l'utilisation de la règle du *produit* sur les probabilités de vraisemblance.

Les résultats montrent également que les meilleures performances de reconnaissance des émotions sont obtenues à travers une fusion avec les modèles d'ancrage. Ces résultats mettent en évidence la puissance des modèles d'ancrage comme plateforme pour la combinaison de classificateurs au niveau des scores de décision. Soulignons que les performances obtenues avec les modèles d'ancrage représentent les résultats de l'état de l'art, dépassant ainsi les meilleures performances obtenues dans (Le *et al.* 2013). Combinant plusieurs systèmes hybrides de type DBN-HMM, Le et Provost ont obtenu 45,60 % et 46.6 % en terme UAR, quand les systèmes hybrides sont évalués respectivement en mode indépendant et en mode dépendant du locuteur. Des gains relatifs de 3,75 % et 2,05 % sont donc obtenus par les modèles d'ancrage par rapport aux DBN-HMM, des systèmes plus complexes et coûteux en temps de calcul.

Tableau 8.3 Tableau récapitulatif des meilleurs résultats obtenus pour chacune des méthodes de combinaison expérimentées sur les données de test du corpus FAU AIBO Emotion

Méthodes de fusion des cinq systèmes	**UAR (%)**	**WAR (%)**
Logistique	44,24	41,49
MLP	40,60	45,89
Forêts d'arbres décisionnels	38,94	40,04
Naïve bayésienne	46,56	38,55
SVM (linéaire)	46,38	42,76
Règle de la *somme*	45,11	44,97
Règle du *maximum*	43,53	41,96
MA-EUC	47,08	42,91
MA-EUC (3 systèmes : AMCC+PLP_MT+MFCC_MT)	**47,31**	38,46
[1]Combinaison de HMM-DBN (Le et al. 2013)	46,36	-
[2]Combinaison de HMM-DBN (Le et al. 2013)	45,60	-
1 et 2 se rapportent respectivement aux résultats d'expériences basées ou non sur une normalisation spécifique au locuteur.		

8.5 Conclusion

Dans le dernier chapitre de cette thèse nous avons proposé de nouveaux types de descripteurs plus discriminants pour la reconnaissance des émotions quand expérimenté sur le corpus

FAU AIBO : i) les traits MFCC et PLP basés sur une estimation de spectre de type multifenêtre et ii) les coefficients cepstraux de modulation d'amplitude AMCC. Les performances obtenues avec ces traits, et particulièrement avec AMCC, dépassent les résultats obtenus avec des traits MFCC- ou PLP-conventionnels qui pourtant sont considérés parmi les plus performants. Nous avons montré également qu'il existe une complémentarité d'information entre les différents traits, ce qui a permis d'améliorer davantage les performances à travers leurs combinaisons. Finalement, nous avons montré que les modèles d'ancrage constituent également une méthode puissante de combinaison des classificateurs moyennant une normalisation qui réduirait la variance intraclasse des émotions. Nous avons donc proposé une nouvelle forme de matrice de covariance intraclasse pour la normalisation WCCN adaptée aux modèles d'ancrage utilisés dans un contexte de fusion. Cette méthode permet de pallier au problème de sur-apprentissage des éléments de la matrice de covariance de dimension élevée en présence de peu de données d'apprentissage. Cette nouvelle configuration des modèles d'ancrage a permis non seulement de dépasser les performances de classification des autres méthodes de combinaison mais aussi de dépasser les résultats de l'état de l'art obtenus par des systèmes beaucoup plus complexes.

CONCLUSION GÉNÉRALE

La reconnaissance automatique des émotions à partir de la parole est une tâche difficile particulièrement lorsqu'il s'agit d'expressions affectives spontanées issues du monde réel loin des simulations de laboratoires. Les émotions véhiculées par la parole spontanée sont souvent subtiles, parfois mixtes, caractérisées par une grande variabilité et leurs collecte et annotation constituent des tâches onéreuses et difficiles. Les énoncés analysés sont généralement de courtes durées et la distribution de leurs classes d'émotion est largement biaisée. C'est dans ce contexte que s'inscrivait notre objectif de proposer une méthodologie capable d'améliorer les performances des systèmes de RAE traitant des expressions affectives réelles.

La méthodologie que nous avons présentée a été motivée principalement par les connaissances a priori sur les modèles théoriques des émotions en psychologie. Nous avons proposé de combiner deux modèles traditionnels d'émotion qui sont généralement en compétition : le modèle dimensionnel et le modèle discret. L'idée est d'intégrer les concepts du modèle d'émotion dimensionnel dans la conception d'un classificateur basé sur des classes d'émotions discrètes. Deux concepts ont été dégagés du modèle dimensionnel : d'une part, l'existence d'un espace multidimensionnel dans lequel les émotions catégoriques peuvent être projetées et, d'autre part, l'existence d'une relation de proximité entre ces catégories d'émotion relativement à chacune de ces dimensions. Le premier concept s'est traduit dans notre méthodologie par l'extraction de traits de haut niveau destinés à jouer un rôle similaire à celui incarné par les dimensions du modèle théorique. Le second a motivé l'adoption d'une approche basée sur la similarité pour la représentation et la classification des émotions. Nous avons montré que les scores de vraisemblance générés par les modèles GMM constituent de puissants traits de similarité pour la RAE et répond bien à la contrainte relative à la taille limitée des énoncés.

Nous avons proposé une nouvelle méthode de classification, intitulé le *plus proche patron de similarité pondéré*, basée sur ces scores de vraisemblance. Cette méthode est conçue autour d'un nouveau vecteur caractérisant un énoncé à travers la description de son patron de classes

voisines. Les classes au sein d'un patron sont ordonnées selon leurs degrés de proximité estimés sur la base des scores de vraisemblance. Contrairement à la règle de décision de *Bayes*, les rangs de tous les scores influencent la décision de classification. Deux types de modèles ont été proposés et expérimentés : linéaire et non linéaire (pour investiguer l'interaction entre rangs de classe à l'intérieur d'un patron de voisinage). Dans la règle de décision de cette méthode, tous les rangs participent dans le processus de prise de décision de classification mais avec différents poids contrairement à la règle de décision *Bayes* où seul le rang de classe maximisant la probabilité des scores est prépondérant. Les résultats expérimentaux sur le corpus FAU AIBO Emotion montrent une amélioration relative de 5,1 % obtenue par le système WOC-NN à double interaction par rapport au système GMM-*Bayes*.

Nous avons proposé également les modèles d'ancrage basés sur les mesures de similarité euclidienne et cosinus comme méthode de classification et d'analyse du contenu émotionnel des énoncés. Dans les modèles d'ancrage, les émotions sont projetées dans un espace multidimensionnel de dimension égale au nombre de classes d'émotion du problème de classification. Chaque dimension est engendrée par un modèle de classe d'émotion et mesure par conséquent le degré de similarité d'un énoncé avec les données de cette classe. Les dimensions engendrées par les modèles de classes peuvent être vues comme des dimensions alternatives aux axes du modèle théorique dimensionnel (tels que les axes *valence* et *excitation*). L'évaluation des modèles d'ancrage pour la classification des cinq classes d'émotion montre un gain relatif de 6 % par rapport au modèle GMM-Bayes.

Nous avons montré qu'il était possible d'appliquer avec succès les modèles d'ancrage dans un contexte de classification binaire à travers une extension de l'espace d'ancrage avec des modèles générés avec des données de classes externes au problème de classification. Une amélioration relative moyenne de 2,3 % a été obtenue par les modèles d'ancrage sur les dix expériences de classification binaire par rapport aux systèmes GMM-Bayes.

Nous avons également analysé et comparé les modèles d'ancrage selon les propriétés géométriques de leurs frontières de décision. Nous avons analysé et répondu à quelques questions telles que la raison : de la chute des performances des modèles d'ancrage quand la distance euclidienne est utilisée, de la similarité des performances avec le modèle GMM-*Bayes*. Nous avons proposé de normaliser les scores avec WCCN pour à la fois atténuer ce problème et aussi améliorer les résultats de classification obtenus avec la similarité cosinus. Nous avons également montré que des modèles d'ancrage basés sur de simples métriques possèdent d'intéressantes propriétés telles que leur insensibilité au problème de distribution biaisée des classes, légèreté du système dorsal (absence de phase d'entrainement) et les bonnes performances de classification obtenues qui font de ces modèles des solutions adéquates au problème de reconnaissance des émotions spontanées en comparaison avec des systèmes dorsaux plus complexes. Une autre caractéristique intéressante des modèles d'ancrage réside également dans leur capacité de servir d'outil d'aide à l'analyse du contenu émotionnel des énoncés qu'on pourrait exploiter dans les domaines de recherche en psychologie par exemple.

Nous avons montré que les modèles d'ancrage peuvent servir également comme méthode puissante de combinaison de classificateurs moyennant une normalisation de la variance intraclasse. Nous avons proposé par conséquent une nouvelle forme de matrice de covariance intraclasse pour la normalisation WCCN adaptée aux modèles d'ancrage destinés à la fusion. Cette méthode a permis de pallier au problème de sur-apprentissage du nombre élevé d'éléments de la matrice de covariance en présence de peu de données d'entrainement. Les résultats de combinaison de classificateurs obtenus avec cette méthode ont dépassé toutes les autres méthodes que nous avons expérimentées telles que la régression logistique, SVM, MLP et les forêts d'arbres décisionnels.

Sur le plan des traits caractéristiques acoustiques, nous avons proposé deux nouveaux types de descripteurs pour la reconnaissance des émotions : i) les traits MFCC et PLP basés sur une estimation de spectre de type multifenêtre et ii) les coefficients cepstraux de modulation d'amplitude (AMCC). Les performances obtenues avec ces traits, et avec AMCC en

particulier, dépassent nettement les résultats obtenus avec des traits MFCC- ou PLP-conventionnels. Par ailleurs, la combinaison des traits proposés au moyen des modèles d'ancrage a permis de dépasser les résultats de l'état de l'art basés sur la combinaison de plusieurs systèmes DBN-HMM avec un gain relatif de 3.75 % quand testé sur FAU AIBO Emotion.

RECOMMANDATIONS

Malgré l'amélioration des performances de classification apportée par l'architecture des modèles d'ancrage expérimentés dans cette thèse, nous pensons que ces travaux ne représentent qu'une contribution initiale exploitant cette approche. D'autres types d'architectures basées sur les traits de similarité susceptibles d'améliorer le pouvoir de discrimination entre classes d'émotion restent à explorer. Nous pouvons penser par exemple à un nouveau modèle d'ancrage intégrant la modélisation de l'information temporelle véhiculée par les énoncés. Dans les modèles d'ancrage actuels la similarité est mesurée au niveau de l'énoncé en additionnant les scores du logarithme de probabilité de vraisemblance de chacune des trames. L'information temporelle pourra être exploitée dans le nouveau modèle en mesurant la similarité au niveau des trames dans l'espace d'ancrage.

D'autres propositions de recherche peuvent s'orienter vers la conception de modèles d'ancrage basés sur une architecture multi-étage. Au lieu de créer un modèle d'ancrage global pour classer toutes les classes en une seule étape, plusieurs modèles d'ancrage seront créés et répartis sur plusieurs étages. A chaque étage, un modèle d'ancrage est utilisé pour discriminer entre deux grandes catégories d'émotion, en commençant d'abord par celles qui sont généralement facilement discernables avant de raffiner la classification dans les étapes ultérieures. Par exemple, les émotions peuvent être séparées selon la dimension *excitation* (*arousal*) suivie de la dimension *valence*.

Par ailleurs, un important mécanisme permettant de réduire le taux d'erreur de classification est celui du rejet. Le rejet est un moyen naturel et efficace utilisé dans divers domaines d'application afin de détecter les instances ambiguës, c.-à-d. les instances reconnues avec un faible niveau de confidence et de différer le processus de prise de décision en confiant la tâche à une tierce partie plus experte (machine ou humaine). Bien qu'il ne soit pas encore exploité en RAE, le rejet est particulièrement utile et indispensable pour la reconnaissance des émotions où la confusion demeure une caractéristique inhérente à l'émotion. D'ailleurs le rejet est toujours appliqué au cours de l'étape de préparation du corpus d'émotion avant

même la phase d'apprentissage des modèles en écartant du corpus les énoncés pour lesquels il y a absence d'accord entre annotateurs. Il existe donc une importante partie des données déjà étiquetées mais qui est rejetée non pas parce qu'elle est corrompue (à cause des mauvaises conditions d'enregistrement) mais tout simplement parce qu'on ne partage pas la même perception du type d'émotion véhiculée dans le discours. Si le rejet est réalisé manuellement durant la phase de constitution du corpus, doter le système de RAE d'une compétence de rejet devient dès lors indispensable pour la phase la classification automatique où seule la machine est présente. Il est intéressant de noter qu'il est possible d'exploiter les modèles d'ancrage pour implanter un rejet plus *discriminant* dans le sens où la partie des données rejetées (ambigües) contiendra un maximum de données qui seraient mal classifiées si le principe de rejet n'était pas appliqué et vice-versa, c.-à-d. qui contiendra un minimum de données qui seraient autrement bien classifiées. Le rejet à travers les modèles d'ancrage peut être appliqué en se basant sur les données situées dans la région de décision délimitée par l'intersection des frontières des discriminants linéaires des modèles d'ancrage basés sur les différentes distances euclidienne et/ou cosinus et/ou en combinaison avec GMM-*Bayes*. Un tel rejet, caractérisé par une incertitude renforcée par plusieurs classificateurs en comparaison à un classificateur unique, augmentera la probabilité de mal classifier les données situées dans cette zone si le rejet n'était pas appliqué.

ANNEXE I

CORPUS DE PAROLE ÉMOTIONNELLE

Dans cet annexe nous allons présenter les différents corpus de parole émotionnelle exploités aussi bien dans le domaine de la psychologie que celui l'informatique affective. Un recensement des corpus les plus utilisés peut être trouvé dans (Douglas-Cowie *et al.* 2007; El Ayadi *et al.* 2007; Ververidis et Kotropoulos, 2006) ou encore sur le site web de l'association AAAC (Association for the Advancement of Affective Computing, ex-HUMAINE) où la liste des corpus est continuellement mise à jour. Beaucoup de ces corpus sont privés et ne sont pas disponibles à la communauté. En plus du critère de disponibilité, ces corpus peuvent être caractérisés par le type d'émotion véhiculée qui peut être actée, naturelle ou induite, la taille du corpus, la langue dans laquelle il a été collecté et le contenu émotionnel des énoncés (nombre et types de classes d'émotion). Les corpus peuvent se distinguer également par la méthode d'annotation qui peut être de type catégorique (discrète) ou dimensionnelle (continue). Enfin, certains corpus peuvent contenir plusieurs modalités telles la vidéo et le gestuel à côté de la parole

Tableau-A I-1 Liste des corpus de parole émotionnelle disponibles ainsi que leurs descriptions
Tiré du site web de l'association AAAA (2015)

Identifier	**Modalities**	**Emotional content**	**Emotion elicitation methods**	**Size**	**Language**
HUMAINE Database from www.emotion-research.net/download/pilot-db/	Audiovisual + gesture	Categorical and continue	Naturalistic and induced material	50 clips ranging from 5 seconds to 3 minutes	English and some French and Hebrew
Belfast Naturalistic Database (Douglas-Cowie et al 2000, 2003)	Audio- visual	Wide range	Natural	125 subjects; 31 male, 94 female	English
RECOLA - REmote COLlaborative and Affective interactions - Database (Ringeval *et al.* 2013);	Multimodal: audio, video, EDA, ECG	(i) Discrete (5 categories: agreement, dominance, engagement, performance and rapport. (ii) Continuous (arousal and valence).	Natural	34 subjects; 14 male, 20 female	French
Geneva Airport Lost Luggage Study (Scherer & Ceschi 1997; 2000)	Audio-visual	Anger, good humor, indifference, stress, sadness	Natural	109 subjects	
Chung (Chung 2000)	Audio-visual	Joy, neutrality, sadness (distress)	Natural: (television interviews)	77 subjects; 6 1 Korean speakers, 6 Americans	English and Korean
SMARTKOM	Audio-visual and gestures	Joy, gratification, anger, irritation, helplessness, pondering, reflecting, surprise, neutral	Human machine in WOZ scenario	224 speakers; 4/5 minute sessions	German
Amir *et al.* (Amir et al, 2000)	Audio + physiological(EMG,GSR, Heart Rate, Temperature, Speech)	Anger, disgust, fear, joy, neutrality, sadness	Induced	140 subjects 60 Hebrew speakers 1 Russian speakers	Hebrew Russian
SALAS database	Audio-visual	Wide range of emotions/emotion related states but not very intense	Induced: subjects talk to artificial listener	Pilot study of 20 subjects	English
ORESTEIA database (McMahon *et al.* 2003)	Audio + physiological (some visual data)	Stress, irritation, shock	Induced	29 subjects, 90min sessions per subject	English
Belfast Boredom database (Cowie *et al.* 2003)	Audio-visual	Boredom	Induced	12 subjects: 30 minutes each	English
XM2VTSDB multi-modal face database	Audio-visual	None	n/a	295 subjects Video	English

ISLE project corpora	Audio-visual+ gesture	None	n/a		
Polzin (Polzin, 2000)	Audio- visual	Anger, sadness, neutrality	Acted	Segment numbers 1586 angry, 1076 sad, 2991 neutral	English
Banse and Scherer (Banse and Scherer 1996)	Audio- visual	Anger (hot), anger (cold), anxiety, boredom, contempt, disgust, elation, fear (panic), happiness, interest, pride, sadness, shame	Acted	12 (6 male, 6 female)	German
TALKAPILLAR (Beller, 2005)	Speech	neutral, happiness, question, positive and negative surprised, angry, fear, disgust, indignation, sad ,bore	Contextualised acting	1 actor reading 26 semantically neutral sentences for each emotion (each repeated 3 times in different activation level : low,middle,high)	French
Reading-Leeds database (Greasley *et al.* 1995; Roach *et al.* 1998, Stibbard 2001)	Speech	Range of full blown emotions	Natural	Around 4 ½ hours material	English
France *et al.* (France *et al.* 2000)	Speech	Depression, suicidal state, neutrality	Natural	115 subjects: 48 females 67 males.	English
Campbell CREST database, ongoing (Campbell 2002; see also Douglas-Cowie *et al.* 2003)	Speech	Wide range of emotional states and emotion-related attitudes	Natural	Target - 1000 hrs over 5 years	English Japanese Chinese
Capital Bank Service and Stock Exchange Customer Service (as used by Devillers & Vasilescu 2004)	Speech	Mainly negative - fear, anger, stress	Natural: call center human-human interactions		English
SYMPAFLY (as used by Batliner *et al.* 2004b)	Speech	Joyful, neutral, emphatic, surprised, ironic, helpless, touchy, angry, panic	Human machine dialogue system	110 dialogues, 29.200	German
DARPA Communicator corpus (as used by Ang *et al.* 2002) See Walker *et al.* 2001	Speech	Frustration, annoyance	Human machine dialogue system	average length about 2.75 words 13187	English
AIBO (Erlangen database) (Batliner *et al.* 2004a)	Speech	Joyful, surprised, emphatic, helpless, touchy (irritated), angry, motherese, bored, reprimanding, neutral	Human machine: interaction with robot	51 german children, 51.393 words English (Birmingham): 30 children, 5.822 words	German
Fernandez *et al.* (Fernandez *et al.* 2000, 2003)	Speech	Stress	Induced:	4 subjects	English
Tolkmitt and Scherer (Tolkmitt and Scherer, 1986)	Speech	Stress (both cognitive & emotional)	Induced:	60 (33 male, 27 female)	German

Iriondo *et al.* (Iriondo *et al.* 2000)	Speech	Desire, disgust, fury, fear, joy, surprise, sadness	Contextualised acting:	8 subjects reading paragraph length passages (20-40mmsec each)	Spanish
Mozziconacci (Mozziconacci, 1998)	Speech	Anger, boredom, fear, disgust, guilt, happiness, haughtiness, indignation, joy, rage, sadness, worry, neutrality	Contextualised acting	3 subjects reading 8 semantically neutral sentences (each repeated 3 times)	Dutch
McGilloway (McGilloway, 1997; Cowie and Douglas-Cowie, 1996)	Speech	Anger, fear, happiness, sadness, neutrality	Contextualised acting:	40 subjects reading 5 passages each	English
Belfast structured Database An extension of McGilloway database above (Douglas-Cowie *et al.* 2000)	Speech	Anger, fear, happiness, sadness, neutrality	Contextualised acting:	50 subjects reading 20 passages	English
Danish Emotional Speech Database (Engberg *et al.* 1997)	Speech	Anger, happiness sadness, surprise neutrality	Acted	4 subjects read 2 words, 9 sentences & 2 passages	Danish
Groningen ELRA corpus number S0020	Speech	Database only partially oriented to emotion	Acted	238 subjects reading 2 short texts	Dutch
Berlin database (Kienast & Sendlmeier 2000; Paeschke & Sendlmeier 2000)	Speech	Anger- hot, boredom, disgust, fear-panic, happiness, sadness-sorrow, neutrality	Acted	10 subjects (5 male, 5 female) reading 10 sentences each	German
Pereira (Pereira, 2000)	Speech	Anger (hot), anger (cold), happiness, sadness, neutrality	Acted	2 subjects reading 2 utterances each	English
van Bezooijen (van Bezooijen, 1984)	Speech	Anger, contempt disgust, fear, interest joy, sadness shame, surprise, neutrality	Acted	8 (4 male, 4 female) reading 4 phrases	Dutch
Abelin (Abelin 2000)	Speech	Anger, disgust, dominance, fear, joy, sadness, shyness, surprise	Acted	1 subject	Swedish
Yacoub et al (2003) (data from LDC)	Speech	15 emotions : Neutral, hot anger, cold anger, happy, sadness, disgust, panic, anxiety, despair, elation, interest, shame, boredom, pride, contempt	Acted	2433 utterances from 8 actors	English

ANNEXE II

LE CORPUS FAU AIBO EMOTION

Description générale du corpus

Emotion FAU AIBO (Steidl, 2009) est un corpus d'enregistrements d'émotions spontanées de 51 enfants allemands, âgés de 10 à 13 ans, interagissant avec un chien robot de compagnie. FAU AIBO Emotion est constitué de 8,9 heures de parole et 48 401 mots répartis sur 13 642 tours de parole. Notons que l'état émotionnel de l'enfant peut même changer à l'intérieur d'un tour de parole. De longues pauses peuvent se produire entre les mots. Le corpus a été recueilli dans deux écoles différentes OHM (13 garçons et 13 filles), et MONT (8 garçons et 17 filles). La partie du corpus OHM a été utilisée pour l'entrainement des modèles alors que la partie MONT a été utilisée pour le test de performances. Dans ce qui suit nous décrirons en plus de détails le robot AIBO, le scénario d'induction des émotions, la segmentation et la transcription enregistrements ainsi que leurs annotations. Pour plus de détails nous référons le lecteur à la thèse de Steidl (2009).

Description du robot AIBO

Le robot AIBO (ERS-210A) de Sony est conçu comme un petit chien autonome qui peut prendre ses propres décisions. La voix et les mouvements du robot AIBO peuvent être contrôlés à travers une carte réseau sans fil. AIBO possède une capacité de reconnaissance de la voix d'environ 50 mots. Divers capteurs permettent à AIBO de réagir à des stimuli externes et signaler des émotions telles que la joie, la colère ou l'anxiété. Par exemple, AIBO peut manifester son attention envers l'enfant à travers les mouvements d'oreilles, les lumières LED clignotantes ou encore en tournant la tête vers lui. En revanche, AIBO peut ignorer l'enfant en tournant ou en s'éloignant. La joie peut être signalée en laissant AIBO déplacer sa queue alors que la sollicitation (*begging* en anglais) peut être signalée en mettant AIBO en position assise avec pattes allongées.

Description du scénario d'enregistrement

Dans le scénario de la collection de la parole émotionnelle, les enfants jouaient avec le robot AIBO. Les enfants étaient instruits de parler à AIBO comme s'ils parleraient à un vrai chien (i.e., réprimander AIBO s'il désobéit, et le louer s'il exécute les commandes).
Les expériences sont menées selon le paradigme magicien d'Oz (WOZ, *Wizard-of-OZ* en anglais) où AIBO était entièrement contrôlé à distance par l'expérimentateur à l'insu des enfants. Le comportement autonome d'AIBO ainsi que la fonction de reconnaissance vocale ont été désactivées. Les actions d'AIBO sont effectuées selon un ordre préétabli fixe, indépendamment des ordres donnés par l'enfant, afin d'être en mesure de comparer le comportement des enfants envers AIBO. Pour évoquer les émotions, les enfants ont été légèrement mis sous la pression du temps. Au cours de certaines étapes prédéfinies de l'expérience, AIBO désobéissait dans le but d'induire la colère et dansait afin d'évoquer la joie. Aussi, les enfants ont été instruits qu'un des trois mangeoires placés sur le tapis contenait du poison et qu'ils devaient s'assurer qu'AIBO ne s'y approchait en aucun cas. Néanmoins, AIBO s'y approchait dans le but d'induire de légères formes de peur ou de panique.

Le signal de la parole a été enregistré à un taux d'échantillonnage de 48 kHz et une quantification de 16 bits, en utilisant un casque sans fil de la série Shure UT 14/20 UHF et un microphone WH20TQG en combinaison avec un enregistreur DAT Tascam DA-P1. Les enregistrements vocaux ont été par la suite sous-échantillonnés à 16 kHz.

Description, segmentation et transcription des enregistrements

Le corpus FAU AIBO Emotion est constitué de 8,9 heures de parole et 48 401 mots répartis sur 13 642 tours de parole. En moyenne, un tour de parole est de longueur de 3,5 mots. Les enregistrements ont été segmentés automatiquement en tours de parole en utilisant un seuil de pause d'une seconde (l'enfant est supposé être en attente d'une réaction d'AIBO au cours de la pause qui suit l'énoncé).

Ni le tour de parole ni le mot ne représente l'unité d'analyse émotionnelle optimale, mais une unité intermédiaire syntaxiquement et sémantiquement significative. Par conséquent, les tours de parole sont manuellement segmentés en *segments* (*chunk* en anglais) sur la base de critères syntaxiques-prosodique. « *Get up * and go to the left* » et « *AIBO * get up* » sont deux exemples de tours de paroles segmentés en deux segments chacun (* représente l'endroit de segmentation). La taille moyenne d'un segment est de 2,66 en termes du nombre de mots et de 1,7 seconde en durée. Le corpus est composé de 18 216 segments dont 9959 sont utilisés pour l'entrainement et 8 257 pour le test.

Le corpus de la parole a été manuellement transcrit. En outre, les sons non-verbaux tels que les bruits respiratoires, le rire, la toux, le bruit, et les hésitations vocales ou nasales ont été également transcrits.

Annotation des enregistrements

Le mot a été choisi comme unité d'annotation afin de permettre de capturer les changements rapides d'émotion à l'intérieur d'un même tour de parole. Les données sont étiquetées en se basant exclusivement sur la parole (bien que les enregistrements vidéo soient disponibles). Les données ont été étiquetées en dix classes d'émotions : *angry* (en colère), *touchy/irritated* (susceptible/irrité), *joyful* (joyeux), *surprised* (surpris), *bored* (ennuyé), *helpless* (impuissant), *motherese* (mamanais), *reprimanding* (réprimandant), *emphatic* (emphatique), *other* (autre) en plus de classe *neutral* (neutre). Une description plus détaillée des catégories d'émotions est donnée dans le Tableau-A II-1.

Les catégories utilisées ont été déterminées au préalable après inspection des données. Seules les cinq premières catégories décrivent des états émotionnels au sens étroit du terme. Les autres catégories décrivent des états typiques de l'utilisateur (tel que *helpless*) et des modèles comportementaux (*behavioral patterns*, tels que *motherese*, *reprimanding*) qui représentent en fait des états liées aux émotions (*emotion-related*).

Tableau-A II-1 Description des classes d'émotion du corpus FAU AIBO Emotion
Tirée et traduit de Steidl (2009)

État utilisateur	Description
Joyful (J)	L'enfant jouit de l'action d'AIBO et/ou remarque que quelque chose est drôle.
Surprised (S)	L'enfant est (positivement) surpris, car de toute évidence elle/il ne s'attendait pas à ce qu'AIBO réagisse de sorte.
Motherese (M)	L'enfant s'adresse à AIBO de la manière que les mères/parents s'adressent à leurs bébés (également appelé «*infant-directed speech*») - soit parce qu'AIBO se comporte bien ou parce que l'enfant veut qu'AIBO obéisse; c'est l'équivalent positif de réprimander.
Neutral (N)	État utilisateur par défaut, n'appartenant à aucune une autre catégorie.
Other (O)	État non neutre et n'appartenant à aucune autre catégorie, i.e., autres émotions non essentielles.
Bored (B)	L'enfant n'est (momentanément) pas intéressé à interagir avec AIBO.
Emphatic (E)	L'enfant parle d'une façon accentuée, prononcée, parfois hyper-articulée, mais sans montrer aucune émotion
Helpless (H)	L'enfant hésite, semble ne pas savoir quoi dire par la suite à AIBO; peut être marqué par des diffluences et/ou des pauses remplies.
Touchy (T) (irritated)	L'enfant est légèrement irrité; pré-étape de la colère.
Reprimanding (R)	L'enfant est reprochant, réprimandant, «remue le doigt»; c'est l'équivalent négatif de mamanais (*motherese*).
Angry (A)	L'enfant est clairement en colère, contrarié, parle à haute voix.

Les données ont été étiquetées par cinq annotateurs (des étudiants en linguistiques de niveau avancé). Seuls les mots ayant obtenu un vote majoritaire (c'est-à-dire qu'au moins trois annotateurs se sont accordés pour attribuer la même classe d'émotion à un même mot), ont été sélectionnés pour les expérimentations. Au total, 48 401 mots ont reçu un vote majoritaire alors que les 4 707 mots restants avaient obtenu un agrément de deux annotateurs au plus. Les différents taux d'accord entre les cinq annotateurs pour chacune des classes sont donnés au Tableau-A II-2.

Tableau-A II-2 Taux d'accord entre les cinq annotateurs pour chacune des catégories d'émotions
Tirée et traduit de Steidl (2009)

État utilisateur	**Fréquence**		**Accord des cinq annotateurs**			
			0,4	**0,6**	**0,8**	**1,0**
Neutral (N)	39 975	82,6 %	2,0 %	26,1 %	40,3 %	31,6 %
Emphatic (E)	2 807	5,8 %	9,9 %	68,1 %	19,6 %	2,4 %
Motherese (M)	1 311	2,7 %	3,9 %	56,1 %	38,8 %	1,2 %
Reprimanding (R)	463	1,0 %	33,0 %	51,8 %	11,4 %	3,7 %
Touchy (T)	419	0,9 %	46,3 %	48,4 %	5,0 %	0,2 %
Angry (A)	134	0,3 %	37,3 %	53,7 %	9,0 %	0,0 %
Joyful (J)	109	0,2 %	7,3 %	74,3 %	18,3 %	0,0 %
Bored (B)	16	0,0 %	31,3 %	50,0 %	18,8 %	0,0 %
Other (O)	10	0,0 %	70,0 %	30,0 %	0,0 %	0,0 %
Hesitant (H)	4	0,0 %	25,0 %	75,0 %	0,0 %	0,0 %
Surprised (S)	0	0,0 %				
–	3 153	6,5 %				
Tous	48 401	100,0 %	9,6 %	28,2 %	35,7 %	26,3 %

Le Tableau-A II-3 présente la matrice de confusion entre les classes d'émotion lors de l'opération étiquetage des mots par les cinq annotateurs. Nous constatons que la plus part des classes d'émotion sont confondues avec la classe neutre.

Tel que l'on pourrait l'observer (voir Tableau-A II-2), la fréquence de certains états émotionnels dans le corpus est très rare et ne peut suffire pour un apprentissage machine des modèles d'émotions. C'est pourquoi les classes moins fréquentes sont regroupées dans des classes de famille d'émotions plus large pour former au total cinq classes d'émotion : *Neutral* (**N**), *Emphatic* (**E**), *Anger* (**A**) regroupant les états *angry*, *touchy* et *repremanding*, la classe *Positive* (**P**) regroupant *motherese* et *joyful* et enfin la classe *Rest* (**R**) qui regroupe toutes les autres classes d'émotion restantes.

Tableau-A II-3 Matrice de confusion d'étiquetage en pourcentage des cinq annotateurs du corpus FAU AIBO Emotion. Les entrées des catégories rares (*helpless*, *bored*, *surprised*, and *other*) ne sont pas affichées
Tiré et traduit de Steidl (2009)

Emotion	**A**	**T**	**R**	**E**	**N**	**M**	**J**	**Mots**
Angry (**A**)	43,3	13,0	12,9	12,1	18,1	0,1	0,0	134
Touchy (**T**)	4,5	42,9	11,7	13,7	23,5	1,0	0,1	419
Reprimanding (**R**)	3,8	15,7	45,8	14,0	18,2	1,3	0,1	463
Emphatic (**E**)	1,3	5,8	6,7	53,6	29,9	1,2	0,5	2 807
Neutral (**N**)	0,4	2,2	1,5	13,9	77,8	2,7	0,5	39 975
Motherese (**M**)	0,0	0,8	1,4	4,9	30,4	61,1	0,9	1 311
Joyful (**J**)	0,1	0,6	1,1	7,3	32,4	2,0	54,2	109

Nous présentons dans le Tableau-A II-4 la nouvelle matrice de confusion de l'opération d'étiquetage pour les nouvelles classes d'émotion obtenues après fusion des classes plus rares. Nous avons calculé cette matrice après compilation des valeurs de la matrice de confusion du Tableau-A II-3.

Tableau-A II-4 Matrice de confusion d'étiquetage en pourcentage des cinq annotateurs du corpus FAU AIBO Emotion après regroupement des petites catégories d'émotions en des classes plus grandes (compilées à partir des valeurs du Tableau-A II-3)

	A	**E**	**N**	**P**
A	64,3	13,9	20,7	1,1
E	13,9	54,1	30,2	1,7
N	4,1	14,0	78,6	3,2
P	2,2	5,1	30,7	62,0

Enfin, les valeurs de la dernière colonne du Tableau-A II-3, met clairement en évidence que la distribution des cinq classes d'émotion est fortement déséquilibrée. Le pourcentage de chacune des cinq classes est comme suit: **A** (8,8 %), **E** (21 %), **N** (56,1 %), **P** (6,8 %),

R (7,2 %) pour les données d'apprentissage et **A** (7,4 %), **E** (18,3 %), **N** (65,1 %) , **P** (2,6 %), **R** (6,6 %) pour les données de test.

ANNEXE III

MÉTHODES D'ESTIMATION DES PARAMÈTRES

Algorithme Estimation-Maximisation (EM)

Introduit initialement par Baum (Baum, 1972; Baum et Petrie, 1966; Dempster *et al.* 1977), l'algorithme EM permet de déterminer, suivant un processus itératif, les paramètres du modèle λ en maximisant dans l'espace des paramètres λ la fonction de la vraisemblance $p(\mathbf{X}|\lambda)$ ou de manière équivalente le logarithme de la vraisemblance $\log p(\mathbf{X}|\lambda)$ de l'ensemble des observations **X** conditionné sur l'ensemble des paramètres λ. Soit $Y = \{y_1,...,y_n,...,y_N\}$ les variables manquantes supposées connues afin de simplifier le problème. Les valeurs y_n peuvent représenter la composante gaussienne qui se réalise pour la donnée observable $\mathbf{x}_n$ dans le cas des GMM. Soit $Q(\lambda,\hat{\lambda})$ une fonction auxiliaire incluant les paramètres $\lambda = \{w_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}$ du modèle courant et leurs valeurs estimées $\hat{\lambda} = \{\hat{w}_m, \hat{\boldsymbol{\mu}}_m, \hat{\boldsymbol{\Sigma}}_m\}$ à l'itération *t*. Elle est définie comme étant l'espérance mathématique du logarithme de la vraisemblance jointe des variables observées et des variables cachées :

$$Q(\lambda,\hat{\lambda}) = \sum_{Y} P\left(Y\middle|\mathbf{X},\hat{\lambda}\right)\log p\left(\mathbf{X},Y\middle|\lambda\right) \tag{1}$$

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

$$Q(\lambda,\hat{\lambda}) \geq Q(\lambda,\lambda) \quad \Rightarrow \quad \log P(\mathbf{X}\middle|\hat{\lambda}) \geq \log P(\mathbf{X}|\lambda) \tag{2}$$

C'est-à-dire que nous avons trouvé un nouveau modèle $\hat{\lambda}$, plus probable que λ, ayant probablement généré la séquence d'observations. En se basant sur cette procédure, si nous procédons au remplacement de λ par $\hat{\lambda}$ d'une manière itérative, et que nous répétions le

calcul de ré-estimation, nous pouvons alors améliorer la probabilité que **X** soit observée à partir du modèle λ, et ce, jusqu'à ce qu'un point limite soit atteint.

Algorithme EM

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

Aucune donnée $\mathbf{x}_n$ n'est associée exclusivement à une gaussienne unique, mais plutôt sera considérée comme étant générée par chacune des gaussiennes avec une certaine vraisemblance. Les valeurs des paramètres λ sont données par les formules suivantes :

La probabilité a posteriori pour la composante gaussienne j :

$$\hat{P}\left(j \,\middle|\, \mathbf{x}_n, \hat{\lambda}\right) = \frac{w_j\, b_j(\mathbf{x}_n)}{\sum_{m=1}^{M} w_m b_m(\mathbf{x}_n)} \tag{3}$$

où

$$b_j(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\left|\mathbf{\Sigma}_j\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T \mathbf{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right) \tag{4}$$

La pondération d'une gaussienne :

$$\hat{w}_j = \frac{1}{N}\sum_{n=1}^{N} \hat{P}\left(j \,\middle|\, \mathbf{x}_n, \hat{\lambda}\right) \tag{5}$$

La moyenne :

$$\hat{\boldsymbol{\mu}}_j = \frac{\sum_{n=1}^{N} \hat{P}\left(j \middle| \mathbf{x}_n, \hat{\lambda}\right) \mathbf{x}_n}{\sum_{n=1}^{N} \hat{P}\left(j \middle| \mathbf{x}_n, \hat{\lambda}\right)} \tag{6}$$

La covariance :

$$\hat{\boldsymbol{\Sigma}}_j = \frac{\sum_{n=1}^{N} \hat{P}\left(j \middle| \mathbf{x}_n, \hat{\lambda}\right) \left(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_j\right)\left(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_j\right)^T}{\sum_{n=1}^{N} \hat{P}\left(j \middle| \mathbf{x}_n, \hat{\lambda}\right)} \tag{7}$$

Adaptation MAP

L'adaptation MAP (Reynolds *et al.* 2000 ; Huang *et al.* 2001) permet d'ajuster les paramètres du modèle préentrainé (UBM) de manière à ce que de nouvelles données modifient les paramètres du modèle, guidé par la connaissance a priori. En utilisant les données observées **X**, l'estimation MAP peut être formulée par :

$$\hat{\lambda} = \arg\max_{\lambda} \left[p(\lambda|\mathbf{X})\right] = \arg\max_{\lambda} \left[p(\mathbf{X}|\lambda)p(\lambda)\right] \tag{8}$$

En absence de l'information a priori, c'est-à-dire aucune connaissance sur le modèle λ et si $p(\lambda)$ possède une distribution uniforme, alors l'estimation MAP devient identique à l'estimation ML).

Nous pouvons utiliser l'algorithme EM pour estimer les paramètres du GMM de la même façon que nous l'avons fait pour la méthode ML. La *Q*-fonction correspondante est définie par la formule suivante (Huang *et al.* 2001) :

$$\underset{MAP}{Q}(\lambda, \hat{\lambda}) = Q(\lambda, \hat{\lambda}) + \log p(\lambda) \tag{9}$$

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

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

$$P(j|\mathbf{x}_n) = \frac{w_j b_j(\mathbf{x}_n)}{\sum_{m=1}^{M} w_m b_m(\mathbf{x}_n)} \tag{10}$$

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

$$s_j = \sum_{n=1}^{N} P(j|\mathbf{x}_n) \tag{11}$$

$$E_j(\mathbf{x}) = \frac{1}{s_j} \sum_{n=1}^{N} P(j|\mathbf{x}_n)\mathbf{x}_n \tag{12}$$

$$E_j(\mathbf{x}^2) = \frac{1}{s_j} \sum_{n=1}^{N} P(j|\mathbf{x}_n)\, diag(\mathbf{x}_n \mathbf{x}_n{}^T) \tag{13}$$

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

$$\hat{w}_j = \left[\alpha_j^w s_j / N + \left(1-\alpha_j^w\right) w_j\right]\gamma \tag{14}$$

$$\hat{\boldsymbol{\mu}}_j = \alpha_j^m \, E_j(\mathbf{x}) + \left(1-\alpha_j^m\right)\boldsymbol{\mu}_j \tag{15}$$

$$\hat{\boldsymbol{\sigma}}_j^2 = \alpha_j^v \, E_j\left(\mathbf{x}^2\right) + \left(1-\alpha_j^v\right)\left(\boldsymbol{\sigma}_j^2 + \boldsymbol{\mu}_j^2\right) - \hat{\boldsymbol{\mu}}_j^2 \tag{16}$$

où γ représente un facteur de pondération assurant que la somme des pondérations des mélanges de gaussiennes adaptés est égale à l'unité.

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

LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES

Abou-Moustafa, K. T., Cheriet, M., & Suen, C. Y. (2004). Classification of time-series data using a generative/discriminative hybrid. Paper presented at the Frontiers in Handwriting Recognition, 2004. IWFHR-9 2004. Ninth International Workshop on.

Abou-Moustafa, Karim T, Mohamed Cheriet et Ching Y Suen. 2004. « Classification of time-series data using a generative/discriminative hybrid ». In Frontiers in Handwriting Recognition, 2004. IWFHR-9 2004. Ninth International Workshop on. p. 51-56. IEEE.

Alam, Md Jahangir, Yazid Attabi, Pierre Dumouchel, Patrick Kenny et Douglas D O'Shaughnessy. 2013. « Amplitude modulation features for emotion recognition from speech ». In INTERSPEECH. p. 2420-2424.

Alam, Md Jahangir, Patrick Kenny et Douglas O'Shaughnessy. 2013. « Low-variance Multitaper Mel-frequency Cepstral Coefficient Features for Speech and Speaker Recognition Systems ». Cognitive Computation, vol. 5, no 4, p. 533-544.

Álvarez, Aitor, Idoia Cearreta, Juan Miguel López, Andoni Arruti, Elena Lazkano, Basilio Sierra et Nestor Garay. 2007. « A comparison using different speech parameters in the automatic emotion recognition using feature subset selection based on evolutionary algorithms ». In Text, Speech and Dialogue. p. 423-430. Springer.

Ang, Jeremy, Rajdip Dhillon, Ashley Krupski, Elizabeth Shriberg et Andreas Stolcke. 2002. « Prosody-based automatic detection of annoyance and frustration in human-computer dialog ». In INTERSPEECH.

Attabi, Yazid, Md Jahangir Alam, Pierre Dumouchel, Patrick Kenny et Douglas O'Shaughnessy. 2013. « Multiple windowed spectral features for emotion recognition ». In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. p. 7527-7531. IEEE.

Attabi, Yazid, et Pierre Dumouchel. 2011. « Weighted Ordered Classes-Nearest Neighbors: A New Framework for Automatic Emotion Recognition from Speech ». In INTERSPEECH. p. 3125-3128.

Attabi, Yazid, et Pierre Dumouchel. 2012a. « Anchor models and WCCN normalization for speaker trait classification ». In Thirteenth Annual Conference of the International Speech Communication Association.

Attabi, Yazid, et Pierre Dumouchel. 2012b. « Emotion Recognition from Children's Speech Using Anchor Models ». In Third Workshop on Child, Computer and Interaction.

Attabi, Yazid, et Pierre Dumouchel. 2012c. « Emotion recognition from speech: WOC-NN and class-interaction ». In Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on. p. 126-131. IEEE.

Attabi, Yazid, et Pierre Dumouchel. 2013. « Anchor models for emotion recognition from speech ». Affective Computing, IEEE Transactions on, vol. 4, no 3, p. 280-290.

Banse, Rainer, et Klaus R Scherer. 1996. « Acoustic profiles in vocal emotion expression ». Journal of personality and social psychology, vol. 70, no 3, p. 614.

Banziger, Tanja , et Klaus R Scherer. 2010. « Introducing the Geneva multimodal emotion portrayal (GEMEP) corpus ». In A Blueprint for Affective Computing: A sourcebook and manual, sous la dir. de Scherer, Klaus R, Tanja Bänziger et Etienne Roesch. p. 271-294. Coll. « Series in Affective Science ». New York: Oxford University Press.

Banziger, Tanja , Stéphane With et Susanne Kaiser. 2010. « The face and voice of emotions: The expression of emotions ». In A Blueprint for Affective Computing: A sourcebook and manual, sous la dir. de Scherer, Klaus R, Tanja Bänziger et Etienne Roesch. p. 85-104. Coll. « Series in Affective Science ». New York: Oxford University Press.

Batista, Luana, Eric Granger et Robert Sabourin. 2010. « A multi-classifier system for off-line signature verification based on dissimilarity representation ». In Multiple Classifier Systems. p. 264-273. Springer.

Baum, Leonard E. 1972. « An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes ». Inequalities, vol. 3, p. 1-8.

Baum, Leonard E, et Ted Petrie. 1966. « Statistical Inference for Probabilistic Functions of Finite State Markov Chains ». The Annals of Mathematical Statistics, vol. 37, no 6, p. 1554-1563.

Belhumeur, Peter N., João P Hespanha et David Kriegman. 1997. « Eigenfaces vs. fisherfaces: Recognition using class specific linear projection ». Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 19, no 7, p. 711-720.

Beritelli, F., S. Casale, A. Russo, S. Serrano et D. Ettorre. 2006. « Speech Emotion Recognition Using MFCCs Extracted from a Mobile Terminal based on ETSI Front End ». In Signal Processing, 2006 8th International Conference on. (16-20 Nov. 2006) Vol. 2, p. 1.

Bhatti, Muhammad Waqas, Yongjin Wang et Ling Guan. 2004. « A neural network approach for human emotion recognition in speech ». In Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on. Vol. 2, p. II-181-4 Vol. 2. IEEE.

Bicego, Manuele, Vittorio Murino et Mário AT Figueiredo. 2004. « Similarity-based classification of sequences using hidden Markov models ». Pattern Recognition, vol. 37, no 12, p. 2281-2291.

Bitouk, Dmitri, Ragini Verma et Ani Nenkova. 2010. « Class-level spectral features for emotion recognition ». Speech communication, vol. 52, no 7, p. 613-625.

Blouet, Raphaël, Chafic Mokbel, Hoda Mokbel, Eduardo Sánchez Soto, Gérard Chollet et Hanna Greige. 2004. « Becars: A free software for speaker verification ». In ODYSSEY04-The Speaker and Language Recognition Workshop. p. 145-148.

Boite, René. 2000. Traitement de la parole. PPUR presses polytechniques.

Boufaden, Narjès, et Pierre Dumouchel. 2008. « Leveraging emotion detection using emotions from yes-no answers ». In INTERSPEECH. p. 241-244.

Breazeal, Cynthia. 2003. « Emotion and sociable humanoid robots ». International Journal of Human-Computer Studies, vol. 59, no 1, p. 119-155.

Breazeal, Cynthia, et Lijin Aryananda. 2002. « Recognition of affective communicative intent in robot-directed speech ». Autonomous robots, vol. 12, no 1, p. 83-104.

Brueckner, Raymond, et Björn Schuller. 2012. « Likability Classification-A Not so Deep Neural Network Approach ». In INTERSPEECH. p. 290-293.

Brümmer, Niko, Albert Strasheim, Valiantsina Hubeika, Pavel Matejka, Lukás Burget et Ondrej Glembek. 2009. « Discriminative acoustic language recognition via channel-compensated GMM statistics ». In Interspeech. p. 2187-2190.

Campbell, William M, Douglas E Sturim et Douglas A Reynolds. 2006. « Support vector machines using GMM supervectors for speaker verification ». Signal Processing Letters, IEEE, vol. 13, no 5, p. 308-311.

Cao, Houwei, Ragini Verma et Ani Nenkova. 2012. « Combining Ranking and Classification to Improve Emotion Recognition in Spontaneous Speech ». In INTERSPEECH. p. 358-361.

Casale, Salvatore, Alessandra Russo et Salvatore Serrano. 2007. « Multistyle classification of speech under stress using feature subset selection based on genetic algorithms ». Speech Communication, vol. 49, no 10, p. 801-810.

Castaldo, Fabio, Daniele Colibro, Emanuele Dalmasso, Pietro Laface et Claudio Vair. 2008. « Stream-based speaker segmentation using speaker factors and eigenvoices ». In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. p. 4133-4136. IEEE.

Castellano, Ginevra, George Caridakis, Antonio Camurri, Kostas Karpouzis, Gualtiero Volpe et Stefanos Kollias. 2010. « Body gesture and facial expression analysis for automatic affect recognition ». In A Blueprint for Affective Computing: A sourcebook and manual, sous la dir. de Scherer, Klaus R, Tanja Bänziger et Etienne Roesch. p. 245-255. Coll. « Series in Affective Science ». New York: Oxford University Press.

Chandrakala, Shanmuganathan, et Chellu Chandra Sekhar. 2009. « Combination of generative models and SVM based classifier for speech emotion recognition ». In Neural Networks, 2009. IJCNN 2009. International Joint Conference on. p. 497-502. IEEE.

Charlet, Delphine, Mikaël Collet et Frédéric Bimbot. 2007. « VZ-norm: an extension of z-norm to the multivariate case for anchor model based speaker verification ». In INTERSPEECH. p. 742-745.

Chauhan, A., S. G. Koolagudi, S. Kafley et K. S. Rao. 2010. « Emotion recognition using LP residual ». In Students' Technology Symposium (TechSym), 2010 IEEE. (3-4 April 2010), p. 255-261.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall et W Philip Kegelmeyer. 2002. « SMOTE: Synthetic Minority Over-sampling Technique ». Journal of Artificial Intelligence Research, vol. 16, p. 321-357.

Chu, Stephen M, Hao Tang et Thomas S Huang. 2009. « Fishervoice and semi-supervised speaker clustering ». In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. p. 4089-4092. IEEE.

Clavel, Chloé, Ioana Vasilescu, Gaël Richard et Laurence Devillers. 2006. « Voiced and unvoiced content of fear-type emotions in the safe corpus ». In Proc. of Speech Prosody. Vol. 2006.

Collet, M., Delphine Charlet et F. Bimbot. 2005. « A Correlation Metric for Speaker Tracking Using Anchor Models ». In Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on. (March 18-23, 2005) Vol. 1, p. 713-716.

Collet, Mikael, Yassine Mami, Delphine Charlet et Frederic Bimbot. 2005. « Probabilistic anchor models approach for speaker verification ». In INTERSPEECH. p. 2005-2008.

Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz et J. G. Taylor. 2001. « Emotion recognition in human-computer interaction ». Signal Processing Magazine, IEEE, vol. 18, no 1, p. 32-80.

Cowie, Roddy, Ellen Douglas-Cowie, Jean-Claude Martin et Laurence Devillers. 2010a. « The essential role of human databases for learning in and validation of affectively competent agents ». In A Blueprint for Affective Computing: A sourcebook and manual, sous la dir. de Scherer, Klaus R, Tanja Bänziger et Etienne Roesch. p. 151-165. Coll. « Series in Affective Science ». New York: Oxford University Press.

Cowie, Roddy, Ellen Douglas-Cowie, Ian Sneddon, Margaret McRorie, Jennifer Hanratty, Edelle McMahon et Gary McKeown. 2010b. « Induction techniques developed to illuminate relationships between signs of emotion and their context, physical and social ». In A Blueprint for Affective Computing: A sourcebook and manual, sous la dir. de Scherer, Klaus R, Tanja Bänziger et Etienne Roesch. p. 295-307. Coll. « Series in Affective Science ». New York: Oxford University Press.

Cummings, K. E., et M. A. Clements. 1995. « Analysis of the glottal excitation of emotionally styled and stressed speech ». J Acoust Soc Am, vol. 98, no 1, p. 88-98.

Davis, Steven, et Paul Mermelstein. 1980. « Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences ». Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 28, no 4, p. 357-366.

Dehak, Najim, Patrick Kenny, Réda Dehak, Pierre Dumouchel et Pierre Ouellet. 2011. « Front-end factor analysis for speaker verification ». Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, no 4, p. 788-798.

Dellaert, Frank, Thomas Polzin et Alex Waibel. 1996. « Recognizing emotion in speech ». In Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on. Vol. 3, p. 1970-1973. IEEE.

Dempster, Arthur P, Nan M Laird et Donald B Rubin. 1977. « Maximum likelihood from incomplete data via the EM algorithm ». Journal of the Royal Statistical Society. Series B (Methodological), p. 1-38.

Devillers, L., L. Lamel et I. Vasilescu. 2003. « Emotion detection in task-oriented spoken dialogues ». In Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on. (6-9 July 2003) Vol. 3, p. III-549-52 vol.3.

Devillers, Laurence, Sarkis Abrilian et Jean-Claude Martin. 2005. « Representing real-life emotions in audiovisual data with non basic emotional patterns and context features ». In Affective Computing and Intelligent Interaction. p. 519-526. Springer.

Devillers, Laurence, Laurence Vidrascu et Omar Layachi. 2010. « Automatic detection of emotion from vocal expression ». In A Blueprint for Affective Computing: A sourcebook and manual, sous la dir. de Scherer, Klaus R, Tanja Bänziger et Etienne

Roesch. p. 132-144. Coll. « Series in Affective Science ». New York: Oxford University Press.

Douglas-Cowie, Ellen, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian et Anton Batliner. 2007. « The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data ». In Affective computing and intelligent interaction. p. 488-500. Springer.

Dumouchel, Pierre, Najim Dehak, Yazid Attabi, Reda Dehak et Narjes Boufaden. 2009. « Cepstral and long-term features for emotion recognition ». In Interspeech. p. 344-347.

El Ayadi, Moataz MH, Mohamed S Kamel et Fakhri Karray. 2007. « Speech emotion recognition using Gaussian mixture vector autoregressive models ». In IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP. p. 957-960.

Fernandez, Raul, et Rosalind W Picard. 2003. « Modeling drivers' speech under stress ». Speech Communication, vol. 40, no 1, p. 145-159.

Friedman, Jerome H, Jon Louis Bentley et Raphael Ari Finkel. 1977. « An algorithm for finding best matches in logarithmic expected time ». ACM Transactions on Mathematical Software (TOMS), vol. 3, no 3, p. 209-226.

Fu, Liqin, Xia Mao et Lijiang Chen. 2008. « Speaker independent emotion recognition based on svm/hmms fusion system ». In Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on. p. 61-65. IEEE.

Fujie, Shinya, Tetsunori Kobayashi, Daizo Yagi et Hideaki Kikuchi. 2004. « Prosody based attitude recognition with feature selection and its application to spoken dialog system as para-linguistic information ». In INTERSPEECH. p. 2841-2844.

Fukunaga, Keinosuke. 1990. Introduction to statistical pattern recognition. Academic press.

Galarneau , Annie, Pascale Tremblay et Pierre Martin. 2009. « Dictionnaire de la parole ». < http://www.phonetique.ulaval.ca/lexique/dico.html >. Consulté le 11-11-2014.

Garcia-Romero, Daniel, et Carol Y Espy-Wilson. 2011. « Analysis of i-vector Length Normalization in Speaker Recognition Systems ». In Interspeech. p. 249-252.

Georgogiannis, A., et V. Digalakis. 2012. « Speech Emotion Recognition using non-linear Teager energy based features in noisy environments ». In Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European. (27-31 Aug. 2012), p. 2045-2049.

Giripunje, Shubhangi, et Narendra Bawane. 2007. « ANFIS based emotions recognision in speech ». In Knowledge-Based Intelligent Information and Engineering Systems. p. 77-84. Springer.

Gold, Ben, et Nelson Morgan. 1999. Speech and Audio Signal Processing: Processing and Perception of Speech and Music. New York, NY, USA: John Wiley & Sons, Inc., 560 p.

Goldberger, Jacob, Geoffrey E Hinton, Sam T Roweis et Ruslan Salakhutdinov. 2004. « Neighbourhood components analysis ». In Advances in neural information processing systems. p. 513-520.

Goronzy, Silke, et Ralf Kompe. 1999. « A combined MAP+ MLLR approach for speaker adaptation ». In Proceedings of the Sony Research Forum. Vol. 99.

Grimm, Michael, et Kristian Kroschel. 2005. « Rule-based emotion classification using acoustic features ». In in Proc. Int. Conf. on Telemedicine and Multimedia Communication.

Hammarberg, Britta, Bernard Fritzell, J Gaufin, Johan Sundberg et Lage Wedin. 1980. « Perceptual and acoustic correlates of abnormal voice qualities ». Acta oto-laryngologica, vol. 90, no 1-6, p. 441-451.

Hansson-Sandsten, Maria, et Johan Sandberg. 2009. « Optimal cepstrum estimation using multiple windows ». In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. p. 3077-3080. IEEE.

Hansson, M., et G. Salomonsson. 1997. « A multiple window method for estimation of peaked spectra ». Signal Processing, IEEE Transactions on, vol. 45, no 3, p. 778-781.

Hatch, A. O., et A. Stolcke. 2006. « Generalized Linear Kernels for One-Versus-All Classification: Application to Speaker Recognition ». In Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on. (14-19 May 2006) Vol. 5, p. V-V.

Hermansky, Hynek. 1990. « Perceptual linear predictive (PLP) analysis of speech ». the Journal of the Acoustical Society of America, vol. 87, no 4, p. 1738-1752.

Hill, Edward, David Han, Pierre Dumouchel, Najim Dehak, Thomas Quatieri, Charles Moehs, Marlene Oscar-Berman, John Giordano, Thomas Simpatico et Kenneth Blum. 2013. « Long Term Suboxone™ Emotional Reactivity As Measured by Automatic Detection in Speech ». PloS one, vol. 8, no 7, p. e69043.

Hoque, Mohammed E, Mohammed Yeasin et Max M Louwerse. 2006. « Robust recognition of emotion from speech ». In Intelligent Virtual Agents. p. 42-53. Springer.

Hu, Hao, Ming-Xing Xu et Wei Wu. 2007a. « Fusion of global statistical and segmental spectral features for speech emotion recognition ». In INTERSPEECH. p. 2269-2272.

Hu, Hao, Ming-Xing Xu et Wei Wu. 2007b. « GMM supervector based SVM with spectral features for speech emotion recognition ». In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. Vol. 4, p. IV-413-IV-416. IEEE.

Hu, Yi, et P. C. Loizou. 2004. « Speech enhancement based on wavelet thresholding the multitaper spectrum ». Speech and Audio Processing, IEEE Transactions on, vol. 12, no 1, p. 59-67.

Huang, Rongqing, et Changxue Ma. 2006. « Toward a speaker-independent real-time affect detection system ». In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Vol. 1, p. 1204-1207. IEEE.

Huang, Xuedong, Alex Acero, Hsiao-Wuen Hon et Raj Foreword By-Reddy. 2001. Spoken language processing: A guide to theory, algorithm, and system development. Prentice Hall PTR.

Hui, Gao, Chen Shanguang et Su Guangchuan. 2007. « Emotion classification of mandarin speech based on TEO nonlinear features ». In Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. SNPD 2007. Eighth ACIS International Conference on. Vol. 3, p. 394-398. IEEE.

Huiqin, Huang, Luo Qi et Zhu Aiqin. 2007. « Speech Emotion Recognition in Web based Service ». In Communications, Circuits and Systems, 2007. ICCCAS 2007. International Conference on. p. 804-806. IEEE.

Hung, LE Xuan, Georges QUÉNOT et Eric CASTELLI. 2004. « Speaker-Dependent Emotion Recognition For Audio Document Indexing ». In International Conference on Electronics, Informations and Communications (ICEIC) 2004, Vol. 1. p. 92-96.

Iliev, Alexander I., Michael S. Scordilis, João P. Papa et Alexandre X. Falcão. 2010. « Spoken emotion recognition through optimum-path forest classification using glottal features ». Computer Speech & Language, vol. 24, no 3, p. 445-460.

Inanoglu, Zeynep, et Ron Caneel. 2005. « Emotive alert: HMM-based emotion detection in voicemail messages ». In Proceedings of the 10th international conference on Intelligent user interfaces. p. 251-253. ACM.

Jain, Anil K, et Douglas Zongker. 1997. « Representation and recognition of handwritten digits using deformable templates ». Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 19, no 12, p. 1386-1390.

Johnstone, Tom, et Klaus R Scherer. 2000. « Vocal communication of emotion ». Handbook of emotions, vol. 2, p. 220-235.

Jones, Christian Martyn, et Marie Jonsson. 2007. « Performance analysis of acoustic emotion recognition for in-car conversational interfaces ». In Universal Access in Human-Computer Interaction. Ambient Interaction. p. 411-420. Springer.

Kao, Yi-hao, et Lin-shan Lee. 2006. « Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language ». In InterSpeech. p. 1814-1817.

Kapoor, Ashish, Winslow Burleson et Rosalind W Picard. 2007. « Automatic prediction of frustration ». International Journal of Human-Computer Studies, vol. 65, no 8, p. 724-736.

Kay, SM. 1988. Modern Spectral Estimation. Coll. « Englewood Cliffs, PTR Prentice Hall ».

Kenny, Patrick. 2010. « Bayesian Speaker Verification with Heavy-Tailed Priors ». In Odyssey. p. 14.

Kim, Eun Ho, Kyung Hak Hyun, Soo Hyun Kim et Yoon Keun Kwak. 2007a. « Speech emotion recognition using eigen-fft in clean and noisy environments ». In Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on. p. 689-694. IEEE.

Kim, Samuel, Panayiotis G Georgiou, Sungbok Lee et Shrikanth Narayanan. 2007b. « Real-time emotion detection system using speech: Multi-modal fusion of different timescale features ». In Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on. p. 48-51. IEEE.

Kim, Yelin, et Emily Mower Provost. 2013. « Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions ». In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. p. 3677-3681. IEEE.

Kinnunen, Tomi, Rahim Saeidi, Filip Sedlák, Kong Aik Lee, Johan Sandberg, Maria Hansson-Sandsten et Haizhou Li. 2012. « Low-variance multitaper MFCC features: a case study in robust speaker verification ». Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no 7, p. 1990-2001.

Kirby, Michael, et Lawrence Sirovich. 1990. « Application of the Karhunen-Loeve procedure for the characterization of human faces ». Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 12, no 1, p. 103-108.

Kockmann, Marcel, Lukáš Burget et Jan Černocký. 2009. « Brno university of technology system for interspeech 2009 emotion challenge ». In Interspeech, ISCA. (Brighton, UK), p. 348-351.

Koolagudi, Shashidhar G, et Sreenivasa Rao Krothapalli. 2012. « Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features ». International Journal of Speech Technology, vol. 15, no 4, p. 495-511.

Koolagudi, ShashidharG, et K. Sreenivasa Rao. 2012. « Emotion recognition from speech using source, system, and prosodic features ». International Journal of Speech Technology, vol. 15, no 2, p. 265-289.

Kreibig, Sylvia D., Gunnar Schaefer et Tobias Brosch. 2010. « Psychophysiological response patterning in emotion: Implications for affective computing ». In A Blueprint for Affective Computing: A sourcebook and manual, sous la dir. de Scherer, Klaus R, Tanja Bänziger et Etienne Roesch. p. 105-130. Coll. « Series in Affective Science ». New York: Oxford University Press.

Kuhn, Roland, Jean-Claude Junqua, Patrick Nguyen et Nancy Niedzielski. 2000. « Rapid speaker adaptation in eigenvoice space ». Speech and Audio Processing, IEEE Transactions on, vol. 8, no 6, p. 695-707.

Kuncheva, Ludmila I. 2004. Combining pattern classifiers: methods and algorithms. John Wiley & Sons.

Kuncheva, Ludmila I, Christopher J Whitaker, Catherine A Shipp et Robert PW Duin. 2000. « Is independence good for combining classifiers? ». In Pattern Recognition, 2000. Proceedings. 15th International Conference on. Vol. 2, p. 168-171. IEEE.

Kwon, Oh-Wook, Kwokleung Chan, Jiucang Hao et Te-Won Lee. 2003. « Emotion recognition by speech signals ». In Proceedings of Eurospeech (Geneva, Switzerland), p. 125-128.

Laukkanen, Anne-Maria, Erkki Vilkman, Paavo Alku et Hanna Oksanen. 1996. « Physical variations related to stress and emotional state: a preliminary study ». Journal of Phonetics, vol. 24, no 3, p. 313-335.

Laver, John. 1980. « The phonetic description of voice quality ». Cambridge Studies in Linguistics London, vol. 31, p. 1-186.

Le, Duc, et Emily Mower Provost. 2013. « Emotion recognition from spontaneous speech using Hidden Markov models with deep belief networks ». In Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. p. 216-221. IEEE.

Lee, Chi-Chun, Emily Mower, Carlos Busso, Sungbok Lee et Shrikanth Narayanan. 2011. « Emotion recognition using a hierarchical binary decision tree approach ». Speech Communication, vol. 53, no 9, p. 1162-1171.

Lee, Chi-Chun, Emily Mower, Carlos Busso, Sungbok Lee et Shrikanth S Narayanan. 2009. « Emotion Recognition Using a Hierarchical Binary Decision Tree Approach ». In Interspeech, ISCA. (Brighton, UK), p. 320-323.

Lee, Chul Min, et Shrikanth Narayanan. 2003. « Emotion recognition using a data-driven fuzzy inference system ». In European Conference on Speech Communication and Technology. p. 157-160.

Lee, Chul Min, et Shrikanth S Narayanan. 2005. « Toward detecting emotions in spoken dialogs ». Speech and Audio Processing, IEEE Transactions on, vol. 13, no 2, p. 293-303.

Lee, Chul Min, Shrikanth S Narayanan et Roberto Pieraccini. 2002. « Combining Acoustic and Language Information for Emotion Recognition ». In Seventh International Conference on Spoken Language Processing.

Lee, Chul Min, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee et Shrikanth Narayanan. 2004. « Emotion recognition based on phoneme classes ». In INTERSPEECH. p. 205-211.

Lefter, Iulia, Leon JM Rothkrantz, Pascal Wiggers et David A Van Leeuwen. 2010. « Emotion recognition from speech by combining databases and fusion of classifiers ». In Text, Speech and Dialogue. p. 353-360. Springer.

Leggetter, Christopher J, et Philip C Woodland. 1995. « Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models ». Computer Speech & Language, vol. 9, no 2, p. 171-185.

Lei, Zhenchun, Yingchun Yang et Zhaohui Wu. 2006. « An UBM-Based Reference Space for Speaker Recognition ». In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Vol. 4, p. 318-321. IEEE.

Li, Longfei, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, Isabel Gonzalez, Enescu Valentin et Hichem Sahli. 2013. « Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition ». In Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. p. 312-317. IEEE.

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

Lin, Jen-Chun, Chung-Hsien Wu et Wen-Li Wei. 2013. « Emotion recognition of conversational affective speech using temporal course modeling ». In INTERSPEECH. (Lyon, France), p. 1336-1340.

Lin, Yi-Lin, et Gang Wei. 2005. « Speech emotion recognition based on HMM and SVM ». In Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on. Vol. 8, p. 4898-4901. IEEE.

Liu, Jia, Chun Chen, Jiajun Bu, Mingyu You et Jianhua Tao. 2007a. « Speech emotion recognition based on a fusion of all-class and pairwise-class feature selection ». In Computational Science–ICCS 2007. p. 168-175. Springer.

Liu, Jia, Chun Chen, Jiajun Bu, Mingyu You et Jianhua Tao. 2007b. « Speech emotion recognition using an enhanced co-training algorithm ». In Multimedia and Expo, 2007 IEEE International Conference on. p. 999-1002. IEEE.

López-Cózar, Ramón, Zoraida Callejas, Martin Kroul, Jan Nouza et Jan Silovský. 2008. « Two-level fusion to improve emotion classification in spoken dialogue systems ». In Text, Speech and Dialogue. p. 617-624. Springer.

Lugger, M., et Yang Bin. 2007. « The Relevance of Voice Quality Features in Speaker Independent Emotion Recognition ». In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. (15-20 April 2007) Vol. 4, p. IV-17-IV-20.

Lugger, M., M. E. Janoir et Yang Bin. 2009. « Combining classifiers with diverse feature sets for robust speaker independent emotion recognition ». In Signal Processing Conference, 2009 17th European. (24-28 Aug. 2009), p. 1225-1229.

Lugger, Marko, et Bin Yang. 2008. « Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters ». In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. p. 4945-4948. IEEE.

Maeireizo, Beatriz, Diane Litman et Rebecca Hwa. 2004. « Co-training for predicting emotions with spoken dialogue data ». In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. p. 28. Association for Computational Linguistics.

Malandrakis, N., A. Potamianos, G. Evangelopoulos et A. Zlatintsi. 2011. « A supervised approach to movie emotion tracking ». In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. (22-27 May 2011), p. 2376-2379.

Mami, Yassine, et Delphine Charlet. 2002. « Speaker identification by location in an optimal space of anchor models ». In International Conference on Spoken Language Processing. Vol. 2, p. 1333.

Mami, Yassine, et Delphine Charlet. 2003. « Speaker identification by anchor models with PCA/LDA post-processing ». In Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on. Vol. 1, p. I-180-I-183 vol. 1. IEEE.

Maragos, P., J. F. Kaiser et T. F. Quatieri. 1993. « On amplitude and frequency demodulation using energy operators ». Signal Processing, IEEE Transactions on, vol. 41, no 4, p. 1532-1550.

Marchi, Erik, Björn Schuller, Anton Batliner, Shimrit Fridenzon, Shahar Tal et Ofer Golan. 2012. « Emotion in the speech of children with autism spectrum conditions: Prosody and everything else ». In Proceedings 3rd Workshop on Child, Computer and Interaction (WOCCI 2012), Satellite Event of INTERSPEECH.

Marsella, Stacy, Jonathan Gratch et Paola Petta. 2010. « Computational models of emotion ». In A Blueprint for Affective Computing: A sourcebook and manual, sous la dir. de

Scherer, Klaus R, Tanja Banziger et Etienne Roesch. p. 21-41. Coll. « Series in Affective Science ». New York: Oxford University Press.

McCoy, E. J., A. T. Walden et D. B. Percival. 1998. « Multitaper spectral estimation of power law processes ». Signal Processing, IEEE Transactions on, vol. 46, no 3, p. 655-668.

McDuff, Daniel, Rana El Kaliouby, Thibaud Senechal, David Demirdjian et Rosalind Picard. 2014. « Automatic measurement of ad preferences from facial responses gathered over the internet ». Image and Vision Computing, vol. 32, no 10, p. 630-640.

McGilloway, Sinéad, Roddy Cowie, Ellen Douglas-Cowie, Stan Gielen, Machiel Westerdijk et Sybert Stroeve. 2000. « Approaching automatic recognition of emotion from voice: a rough benchmark ». In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

Mehrabian, Albert, et Morton Wiener. 1967. « Decoding of inconsistent communications ». Journal of personality and social psychology, vol. 6, no 1, p. 109.

Meng, Hongying, et Nadia Bianchi-Berthouze. 2011. « Naturalistic affective expression classification by a multi-stage approach based on hidden Markov models ». In Affective Computing and Intelligent Interaction. p. 378-387. Springer.

Mitra, V., H. Franco, M. Graciarena et A. Mandal. 2012. « Normalized amplitude modulation features for large vocabulary noise-robust speech recognition ». In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. (25-30 March 2012), p. 4117-4120.

Mower, Emily, Maja J Mataric et Shrikanth Narayanan. 2011. « A framework for automatic human emotion classification using emotion profiles ». Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, no 5, p. 1057-1070.

Mukhopadhyay, S., et G. C. Ray. 1998. « A new interpretation of nonlinear energy operator and its efficacy in spike detection ». Biomedical Engineering, IEEE Transactions on, vol. 45, no 2, p. 180-187.

Nakatsu, Ryohei, Joy Nicholson et Naoko Tosa. 1999. « Emotion recognition and its application to computer agents with spontaneous interactive capabilities ». In Proceedings of the seventh ACM international conference on Multimedia (Part 1). p. 343-351. ACM.

Neiberg, Daniel, Kjell Elenius et Kornel Laskowski. 2006. « Emotion recognition in spontaneous speech using GMMs ». In INTERSPEECH. p. 809-812.

Nicholson, J, K Takahashi et R Nakatsu. 1999. « Emotion recognition in speech using neural networks ». In Neural Information Processing, 1999. Proceedings. ICONIP'99. 6th International Conference on. Vol. 2, p. 495-501. IEEE.

Noda, Tetsuya, Yoshikazu Yano, Shinji Doki et Shigeru Okuma. 2006. « Adaptive emotion recognition in speech by feature selection based on KL-divergence ». In Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on. Vol. 3, p. 1921-1926. IEEE.

Nogueiras, Albino, Asunción Moreno, Antonio Bonafonte et José B Mariño. 2001. « Speech emotion recognition using hidden Markov models ». In INTERSPEECH. p. 2679-2682.

Nwe, Tin Lay, Say Wei Foo et Liyanage C. De Silva. 2003. « Speech emotion recognition using hidden Markov models ». Speech Communication, vol. 41, no 4, p. 603-623.

O'Shaughnessy, Douglas. 2000. « Speech communications ». In. Institute of Electrical and Electronics Engineers.

Ortego-Resa, Carlos, Ignacio Lopez-Moreno, Daniel Ramos et Joaquin Gonzalez-Rodriguez. 2009. « Anchor model fusion for emotion recognition in speech ». In Biometric ID Management and Multimodal Communication. p. 49-56. Springer.

Oudeyer, Pierre-yves. 2002. « Novel useful features and algorithms for the recognition of emotions in human speech ». In Speech Prosody 2002, International Conference.

Pao, Tsang-Long, Yu-Te Chen, Jun-Heng Yeh, Yun-Maw Cheng et Yu-Yuan Lin. 2007a. « A comparative study of different weighting schemes on KNN-based emotion recognition in Mandarin speech ». In Advanced Intelligent Computing Theories and

Applications. With Aspects of Theoretical and Methodological Issues. p. 997-1005. Springer.

Pao, Tsang-Long, Yu-Te Chen, Jun-Heng Yeh et Wen-Yuan Liao. 2005. « Detecting Emotions in Mandarin Speech ». Computational Linguistics and Chinese Language Processing, vol. 10, no 3, p. 347-362.

Pao, Tsang-Long, Charles S Chien, Yu-Te Chen, Jun-Heng Yeh, Yun-Maw Cheng et Wen-Yuan Liao. 2007b. « Combination of multiple classifiers for improving emotion recognition in Mandarin speech ». In Intelligent Information Hiding and Multimedia Signal Processing, 2007. IIHMSP 2007. Third International Conference on. Vol. 1, p. 35-38. IEEE.

Pękalska, Elżbieta, et Robert PW Duin. 2001. « On combining dissimilarity representations ». In Multiple Classifier Systems. p. 359-368. Springer.

Pękalska, Elżbieta, Robert PW Duin et Pavel Paclík. 2006. « Prototype selection for dissimilarity-based classifiers ». Pattern Recognition, vol. 39, no 2, p. 189-208.

Pękalska, Elżbieta, Robert PW Duin et Marina Skurichina. 2002. « A discussion on the classifier projection space for classifier combining ». In Multiple Classifier Systems. p. 137-148. Springer.

Pekalska, Elzbieta, Pavel Paclik et Robert PW Duin. 2002. « A generalized kernel approach to dissimilarity-based classification ». The Journal of Machine Learning Research, vol. 2, p. 175-211.

Petrushin, Valery A. . 2000. « Emotion recognition in speech signal: Experimental study, development, and application ». In International Conference on Spoken Language Processing (ICSLP 2000). p. 222-225.

Pirker, Hannes. 2007. « Mixed feelings about using phoneme-level models in emotion recognition ». In Affective Computing and Intelligent Interaction. p. 772-773. Springer.

Pittermann, A., et J. Pittermann. 2006. « Getting Bored with HTK? Using HMMs for Emotion Recognition from Speech Signals ». In Signal Processing, 2006 8th International Conference on. (16-20 2006) Vol. 1, p. 1.

Planet, S., et I. Iriondo. 2012. « Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition ». In Information Systems and Technologies (CISTI), 2012 7th Iberian Conference on. (20-23 June 2012), p. 1-6.

Potamianos, Alexandros. 1995. « Speech processing applications using an AM-FM modulation model ». Harvard University Cambridge, Massachusetts.

Prieto, GA, RL Parker, DJ Thomson, FL Vernon et RL Graham. 2007. « Reducing the bias of multitaper spectrum estimates ». Geophysical Journal International, vol. 171, no 3, p. 1269-1281.

Prince, Simon JD, et James H Elder. 2007. « Probabilistic linear discriminant analysis for inferences about identity ». In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. p. 1-8. IEEE.

Rao, K Sreenivasa, Shashidhar G Koolagudi et Ramu Reddy Vempada. 2012. « Emotion recognition from speech using global and local prosodic features ». International Journal of Speech Technology, p. 1-18.

Reynolds, Douglas A, Thomas F Quatieri et Robert B Dunn. 2000. « Speaker verification using adapted Gaussian mixture models ». Digital signal processing, vol. 10, no 1, p. 19-41.

Riedel, K. S., et A. Sidorenko. 1995. « Minimum bias multiple taper spectral estimation ». Signal Processing, IEEE Transactions on, vol. 43, no 1, p. 188-195.

Rosenberg, Andrew. 2012. « Classifying Skewed Data: Importance Weighting to Optimize Average Recall ». In INTERSPEECH. p. 2242-2245.

Rotaru, Mihai, et Diane J Litman. 2005. « Using word-level pitch features to better predict student emotions during spoken tutoring dialogues ». In INTERSPEECH. p. 881-884.

Rybka, Jan, et Artur Janicki. 2013. « Comparison of speaker dependent and speaker independent emotion recognition ». Int. J. Appl. Math. Comput. Sci, vol. 23, no 4, p. 797-808.

Sam, Sethserey. 2011. « Vers une adaptation autonome des modèles acoustiques multilingues pour le traitement automatique de la parole ». Université de Grenoble.

Scherer, Klaus R. 1996. « Adding the affective dimension: a new look in speech analysis and synthesis ». In ICSLP. p. 1014-1017.

Scherer, Klaus R. 2000. « Psychological models of emotion ». The neuropsychology of emotion, vol. 137, no 3, p. 137-162.

Scherer, Klaus R. 2003. « Vocal communication of emotion: A review of research paradigms ». Speech communication, vol. 40, no 1, p. 227-256.

Scherer, Klaus R. 2010a. « The component process model: Architecture for a comprehensive computational model of emergent emotion ». In A Blueprint for Affective Computing: A sourcebook and manual, sous la dir. de Scherer, Klaus R, Tanja Banziger et Etienne Roesch. p. 47-70. Coll. « Series in Affective Science ». New York: Oxford University Press.

Scherer, Klaus R. 2010b. « Emotion and emotional competence: conceptual and theoretical issues for modelling agents ». In A Blueprint for Affective Computing: A sourcebook and manual, sous la dir. de Scherer, Klaus R, Tanja Banziger et Etienne Roesch. p. 3-20. Coll. « Series in Affective Science ». New York: Oxford University Press.

Scherer, Klaus R, et Tanja Banziger. 2010. « On the use of actor portrayals in research on emotional expression ». In A Blueprint for Affective Computing: A sourcebook and manual, sous la dir. de Scherer, Klaus R, Tanja Bänziger et Etienne Roesch. p. 166-176. Coll. « Series in Affective Science ». New York: Oxford University Press.

Scherer, Klaus R, Tanja Bänziger et Etienne Roesch. 2010. A Blueprint for Affective Computing: A sourcebook and manual. Oxford University Press.

Scherer, Klaus R, et Heiner Ellgring. 2007. « Multimodal expression of emotion: Affect programs or componential appraisal patterns? ». Emotion, vol. 7, no 1, p. 158.

Schröder, Marc. 2003. « Experimental study of affect bursts ». Speech Communication, vol. 40, no 1–2, p. 99-116.

Schuller, B., G. Rigoll et M. Lang. 2003. « Hidden Markov model-based speech emotion recognition ». In Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on. (6-10 April 2003) Vol. 2, p. II-1-4 vol.2.

Schuller, Björn, Anton Batliner, Stefan Steidl et Dino Seppi. 2011. « Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge ». Speech Communication, vol. 53, no 9, p. 1062-1087.

Schuller, Björn, Ronald Müller, Manfred K Lang et Gerhard Rigoll. 2005. « Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles ». In INTERSPEECH. p. 805-808.

Schuller, Bjorn, Dino Seppi, Anton Batliner, Andreas Maier et Stefan Steidl. 2007a. « Towards more reality in the recognition of emotional speech ». In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. Vol. 4, p. IV-941-IV-944. IEEE.

Schuller, Björn, Stefan Steidl et Anton Batliner. 2009. « The INTERSPEECH 2009 emotion challenge ». In INTERSPEECH. (Brighton, UK) Vol. 2009, p. 312-315.

Schuller, Björn, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth et Gerhard Rigoll. 2010. « Cross-corpus acoustic emotion recognition: variances and strategies ». Affective Computing, IEEE Transactions on, vol. 1, no 2, p. 119-131.

Schuller, Björn, Bogdan Vlasenko, Ricardo Minguez, Gerhard Rigoll et Andreas Wendemuth. 2007b. « Comparing one and two-stage acoustic modeling in the recognition of emotion in speech ». In Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on. p. 596-600. IEEE.

Seppänen, Tapio, Eero Väyrynen et Juhani Toivanen. 2003. « Prosody-based classification of emotions in spoken finnish ». In INTERSPEECH. (Geneva, Switzerland), p. 717-720.

Sethu, Vidhyasaharan, Eliathamby Ambikairajah et Julien Epps. 2007. « Speaker normalisation for speech-based emotion detection ». In Digital Signal Processing, 2007 15th International Conference on. p. 611-614. IEEE.

Shafran, Izhak, Michael Riley et Mehryar Mohri. 2003. « Voice signatures ». In Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on. p. 31-36. IEEE.

Shahin, Ismail Mohd Adnan. 2013. « Gender-dependent emotion recognition based on HMMs and SPHMMs ». International Journal of Speech Technology, vol. 16, no 2, p. 133-141.

Shami, M. T., et M. S. Kamel. 2005. « Segment-based approach to the recognition of emotions in speech ». In Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. (6-8 July 2005), p. 4 pp.

Shami, Mohammad, et Werner Verhelst. 2007. « Automatic classification of expressiveness in speech: a multi-corpus study ». In Speaker classification II. p. 43-56. Springer.

Shin, Donghyuk, et Saejoon Kim. 2009. « Nearest mean classification via one-class SVM ». In 2009 International Joint Conference on Computational Sciences and Optimization. Vol. 1, p. 593-596.

Shum, Stephen, Najim Dehak, Ekapol Chuangsuwanich, Douglas A Reynolds et James R Glass. 2011. « Exploiting Intra-Conversation Variability for Speaker Diarization ». In INTERSPEECH. p. 945-948.

Sim, Kwee-Bo, In-Hun Jang et Chang-Hyun Park. 2007. « The development of interactive feature selection and GA feature selection method for emotion recognition ». In Knowledge-based intelligent information and engineering systems. p. 73-81. Springer.

Steidl, Stefan. 2009. « Automatic Classification of Emotion Related User States in Spontaneous Children's Speech ». University of Erlangen-Nuremberg Germany.

Stuhlsatz, André, Christine Meyer, Florian Eyben, T ZieIke, Günter Meier et Björn Schuller. 2011. « Deep neural networks for acoustic emotion recognition: raising the

benchmarks ». In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. p. 5688-5691. IEEE.

Sturim, Douglas E, Douglas A Reynolds, Elliot Singer et Joseph P Campbell. 2001. « Speaker indexing in large audio databases using anchor models ». In Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on. Vol. 1, p. 429-432. IEEE.

Thomson, D. J. 1982. « Spectrum estimation and harmonic analysis ». Proceedings of the IEEE, vol. 70, no 9, p. 1055-1096.

Thyes, Olivier, Roland Kuhn, Patrick Nguyen et Jean-Claude Junqua. 2000. « Speaker identification and verification using eigenvoices ». In INTERSPEECH. p. 242-245.

Tin Lay, Nwe, Hieu Nguyen Trung et D. K. Limbu. 2013. « Bhattacharyya distance based emotional dissimilarity measure for emotion classification ». In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. (26-31 May 2013), p. 7512-7516.

Ververidis, Dimitrios, et Constantine Kotropoulos. 2004. « Automatic speech classification to five emotional states based on gender information ». In Signal Processing Conference, 2004 12th European. p. 341-344. IEEE.

Ververidis, Dimitrios, et Constantine Kotropoulos. 2005. « Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm ». In Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. p. 1500-1503. IEEE.

Ververidis, Dimitrios, et Constantine Kotropoulos. 2006. « Emotional speech recognition: Resources, features, and methods ». Speech communication, vol. 48, no 9, p. 1162-1181.

Vidrascu, Laurence, et Laurence Devillers. 2005. « Annotation and detection of blended emotions in real Human-Human dialogs recorded in a Call center ». In Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. p. 4 pp.: IEEE.

Vidrascu, Laurence, et Laurence Devillers. 2007. « Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features ». In Proc. Inter. workshop on Paralinguistic Speech between models and data, ParaLing. p. 11-16.

Vidrascu, Laurence, et Laurence Devillers. 2008. « Anger detection performances based on prosodic and acoustic cues in several corpora ». In Programme of the Workshop on Corpora for Research on Emotion and Affect. p. 13.

Vlasenko, Bogdan, Björn Schuller, Andreas Wendemuth et Gerhard Rigoll. 2007. « Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing ». In Affective Computing and Intelligent Interaction. p. 139-147. Springer.

Vogt, Thurid, et Elisabeth André. 2005. « Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition ». In Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. p. 474-477. IEEE.

Vogt, Thurid, et Elisabeth André. 2006. « Improving automatic emotion recognition from speech via gender differentiation ». In Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa. p. 1123-1126.

Wagner, Johannes, Thurid Vogt et Elisabeth André. 2007. « A systematic comparison of different HMM designs for emotion recognition from acted and spontaneous speech ». In Affective Computing and Intelligent Interaction. p. 114-125. Springer.

Wahab, Abdul, Chai Quek et Sussan De. 2007. « Speech emotion recognition using auditory cortex ». In Evolutionary Computation, 2007. CEC 2007. IEEE Congress on. p. 2658-2664. IEEE.

Weiss, Gary M, et Foster Provost. 2001. « The effect of class distribution on classifier learning: an empirical study ». Rutgers Univ.

Williams, Carl E, et Kenneth N Stevens. 1981. « Vocal correlates of emotional states ». Speech evaluation in psychiatry, p. 221-240.

Wöllmer, Martin, Florian Eyben, Björn Schuller, Ellen Douglas-Cowie et Roddy Cowie. 2009. « Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks ». In INTERSPEECH. p. 1595-1598.

Wöllmer, Martin, Angeliki Metallinou, Florian Eyben, Björn Schuller et Shrikanth S Narayanan. 2010. « Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling ». In INTERSPEECH. p. 2362-2365.

Womack, Brian D, et John HL Hansen. 1999. « N-channel hidden Markov models for combined stressed speech classification and recognition ». Speech and Audio Processing, IEEE Transactions on, vol. 7, no 6, p. 668-677.

Wu, Siqing, Tiago H Falk et Wai-Yip Chan. 2011. « Automatic speech emotion recognition using modulation spectral features ». Speech Communication, vol. 53, no 5, p. 768-785.

Xiao, Zhongzhe, Emmanuel Dellandrea, Weibei Dou et Liming Chen. 2010. « Multi-stage classification of emotional speech motivated by a dimensional emotion model ». Multimedia Tools and Applications, vol. 46, no 1, p. 119-145.

Xie, Bo, Ling Chen, Gen-Cai Chen et Chun Chen. 2007. « Feature selection for emotion recognition of mandarin speech ». JOURNAL-ZHEJIANG UNIVERSITY ENGINEERING SCIENCE, vol. 41, no 11, p. 1816.

Yacoub, Sherif M, Steven J Simske, Xiaofan Lin et John Burns. 2003. « Recognition of emotions in interactive voice response systems ». In INTERSPEECH. (Geneva, Switzerland), p. 729-732.

Yan, Rong, Yan Liu, Rong Jin et Alex Hauptmann. 2003. « On predicting rare classes with SVM ensembles in scene classification ». In Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on. Vol. 3, p. III-21-4 vol. 3. IEEE.

Yang, Yingchun, Min Yang et Zhaohui Wu. 2006. « A rank based metric of anchor models for speaker verification ». In Multimedia and Expo, 2006 IEEE International Conference on. p. 1097-1100. IEEE.

Yildirim, Serdar, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Zhigang Deng, Sungbok Lee, Shrikanth Narayanan et Carlos Busso. 2004. « An acoustic study of emotions expressed in speech ». In INTERSPEECH. (Jeju Island, Korea), p. 2193-2196.

You, Mingyu, Chun Chen, Jiajun Bu, Jia Liu et Jianhua Tao. 2006. « Emotion recognition from noisy speech ». In Multimedia and Expo, 2006 IEEE International Conference on. p. 1653-1656. IEEE.

Young, SJ, G Evermann, MJF Gales, D Kershaw, G Moore, JJ Odell, DG Ollason, D Povey, V Valtchev et PC Woodland. 2006. « The HTK book version 3.4 ».

Yu, Chen, Paul M Aoki et Allison Woodruff. 2004. « Detecting user engagement in everyday conversations ». arXiv preprint cs/0410027.

Yu, Feng, Eric Chang, Ying-Qing Xu et Heung-Yeung Shum. 2001. « Emotion detection from speech to enrich multimedia content ». In Advances in Multimedia Information Processing—PCM 2001. p. 550-557. Springer.

Ze-Jing, Chuang, et Wu Chung-Hsien. 2004. « Emotion recognition using acoustic features and textual content ». In Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on. (27-30 June 2004) Vol. 1, p. 53-56 Vol.1.

Zhang, Shiqing. 2008. « Emotion Recognition in Chinese Natural Speech by Combining Prosody and Voice Quality Features ». In Advances in Neural Networks - ISNN 2008, sous la dir. de Sun, Fuchun, Jianwei Zhang, Ying Tan, Jinde Cao et Wen Yu. Vol. 5264, p. 457-464. Coll. « Lecture Notes in Computer Science »: Springer Berlin Heidelberg. < http://dx.doi.org/10.1007/978-3-540-87734-9_52 >.

Zhang, Zixing, Jun Deng, Erik Marchi et Björn Schuller. 2013. « Active learning by label uncertainty for acoustic emotion recognition ». In INTERSPEECH. p. 2856-2860.

Zhang, Zixing, Jun Deng et Bjorn Schuller. 2013. « Co-training succeeds in computational paralinguistics ». In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. p. 8505-8509. IEEE.

Zhao, Xianyu, Yuan Dong, Hao Yang, Jian Zhao et Haila Wang. 2007. « SVM-based speaker verification by location in the space of reference speakers ». In Acoustics, Speech and

Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. Vol. 4, p. IV-281-IV-284. IEEE.

Zhu, Aiqin, et Qi Luo. 2007. « Study on speech emotion recognition system in E-learning ». In Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments. p. 544-552. Springer.