Analyse acoustique de la voix pour la détection des émotions du locuteur

Leila Kerkeni

▶ **To cite this version:**

Leila Kerkeni. Analyse acoustique de la voix pour la détection des émotions du locuteur. Vision par ordinateur et reconnaissance de formes [cs.CV]. Le Mans Université; Université de Sousse (Tunisie), 2020. Français. NNT : 2020LEMA1003 . tel-02925116

HAL Id: tel-02925116
https://theses.hal.science/tel-02925116

Submitted on 28 Aug 2020

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

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

THESE DE DOCTORAT EN COTUTELLE DE L'UNIVERSITE DE SOUSSE ET LE MANS UNIVERSITE

ECOLE DOCTORALE N° 602
Sciences pour l'Ingénieur
Spécialité : Acoustique

Par
Leila KERKENI

Analyse Acoustique de la Voix pour la Détection des Émotions Du Locuteur

Thèse présentée et soutenue à Le Mans Université, le « 20 janvier 2020 »
Unité de recherche : LAUM UMR CNRS 6613 France et LATIS, ENISO Tunisie.
Thèse N° : 2020LEMA1003

Rapporteurs avant soutenance :

Najet AROUS OUNI Professeur des universités (HDR) Université de Tunis (ISI)
Lionel PREVOST Professeur des universités (HDR) Paris (ESIEA)

Composition du Jury :

Examinateurs :	Jean-François DIOURIS	Professeur des universités Université de Nantes (Polytech Nantes)
	Fabien RINGEVAL	Maitre de conférences Université Grenoble Alpes (LIG)
Dir. de thèse :	Kosai RAOOF	Professeur Le Mans Université (LAUM)
Co-dir. de thèse :	Mohamed Ali MAHJOUB	Maitre de conférences Université de Sousse (LATIS)
Co-encadrant :	Youssef SERRESTOU	PRAG Le Mans Université (LAUM)
Co-encadrant :	Catherine CLEDER	Maitre de conférences Le Mans Université (CREN)

Résumé

L'objectif de cette thèse est de proposer un système de reconnaissance automatique des émotions (RAE) par analyse de la voix pour une application dans un contexte pédagogique d'orchestration de classe. Ce système s'appuie sur l'extraction, par démodulation en amplitude et en fréquence, de nouvelles caractéristiques de la voix considérée comme un signal multi-composantes modulé en amplitude et en fréquence (AM-FM), non-stationnaire et issue d'un système non-linéaire. Cette démodulation est basée sur l'utilisation conjointe de la décomposition en modes empiriques (EMD) et de l'opérateur d'énergie de Teager-Kaiser (TKEO). Dans ce système, le modèle discret (ou catégoriel) a été retenu pour représenter les six émotions de base (la tristesse, la colère, la joie, le dégoût, la peur et la surprise) et l'émotion dite neutre. La reconnaissance automatique a été optimisée par la recherche de la meilleure combinaison de caractéristiques, la sélection des plus pertinentes et par la comparaison de différentes approches de classification. Deux bases de données émotionnelles de référence, l'une en allemand et l'autre en espagnol, ont servi à entrainer et évaluer ce système. Une nouvelle base de données en français, plus appropriée pour le contexte pédagogique a été construite, testée et validée.

Mots clés : reconnaissance automatique des émotions (RAE), décomposition en modes empiriques (EMD), modulation en amplitude et en fréquence (AM-FM), opérateur d'énergie de Teager-Kaiser (TKEO), RNN, SVM, Recursive Feature Elimination (RFE).

Abstract

The aim of this thesis is to propose a speech emotion recognition (SER) system for application in classroom. This system has been built up using novel features based on the amplitude (AM) and frequency (FM) modulation model of speech signal. This model is based on the joint use of empirical mode decomposition (EMD) and the Teager-Kaiser energy operator (TKEO). In this system, the discrete (or categorical) emotion theory was chosen to represent the six basic emotions (sadness, anger, joy, disgust, fear and surprise) and neutral emotion. Automatic recognition has been optimized by finding the best features combination, selecting the most relevant ones and comparing different classification approaches. Two reference emotional databases, in German and Spanish, were used to train and evaluate this system. A new database in French, more appropriate for the educational context was built, tested and validated.

Keywords : Automatic emotion recognition (AER), Empirical mode decomposition (EMD), Amplitude modulation-Frequency modulation (AM-FM), Teager Kaiser energy operator (TKEO), RNN, SVM, Recursive feature elimination (RFE).

Remerciements

J'aimerais tout d'abord remercier mes directeurs de thèse, Kosai RAOOF, Professeur à l'Université du Mans, et Mohamed Ali MAHJOUB, maitre de conférence à l'Université de Sousse, pour leurs confiances qu'ils m'ont accordé en acceptant d'encadrer ce travail doctoral, pour leurs multiples conseils et pour toutes les heures qu'ils ont consacrées à diriger cette recherche.

Je tiens à remercier grandement mon encadrant, Youssef SERRESTOU, pour toute son aide et son expertise sur des domaines qui m'étaient a priori étrangers et surtout pour son soutien et ses exigences même dans les moments difficiles. Je suis ravie d'avoir travaillé en sa compagnie car outre son appui scientifique, il a toujours été là pour me soutenir et me conseiller au cours de ces années de thèse.

Je tiens à remercier mon encadrante, Catherine CLEDER, pour toutes nos discussions et ses conseils qui m'ont accompagné tout au long de ma thèse. Je le remercie pour son accueil chaleureux à chaque fois que j'ai sollicité son aide.

Je tiens également à remercier les personnels de l'ENSIM pour l'accueil et les conditions de travail privilégiées qui m'ont été offertes.

Je remercie également Monsieur PREVOST et Madame AROUS, m'ont fait l'honneur d'être rapporteurs de ma thèse, ils ont pris le temps d'examiner mon manuscrit et de discuter avec moi. Je tiens à remercier Monsieurs Jean-François DIOURIS et Fabien RINGEVAL pour avoir accepté de participer à mon jury de thèse et pour le temps qu'ils ont consacré à ma soutenance. Je remercie toutes les personnes avec qui j'ai partagé mes études et notamment ces années de thèse.

Mes remerciements vont aussi à ma famille et mes amis qui, avec cette question récurrente, « quand est-ce que tu la soutiens cette thèse ? », bien qu'angoissante en période fréquente de doutes, m'ont permis de ne jamais dévier de mon objectif final. Heureusement, je suis bien

entourée et encouragée.

Mes derniers remerciements vont à mon cher époux Mouez qui a tout fait pour m'aider, qui m'a soutenu et surtout supporté durant ces dernières années. Tu es un homme d'une immense valeur. Merci mon loulou Jad, tu es ma source de motivation et ma raison d'être. Quand tu souris, je sens que la fatigue meurt pour donner vie à l'énergie et finir ce que j'ai commencé. Je vous aime.

Ces remerciements ne peuvent s'achever, sans une pensée pour ma première fan : ma mère. Sa présence et ses encouragements sont pour moi les piliers fondateurs de ce que je suis et de ce que je fais.

Mon Papa... J'ai longtemps fait les choses pour que tu sois fier de moi et pour que tu me le dises. J'en suis heureuse et je ne t'oublierais pas, même si, aujourd'hui, tu n'es plus à mes côtés. Je t'aime ! Repose en paix !

Liste des figures

1.1 Les six émotions primaires 13
1.2 Un exemple de modèle catégoriel (ou discret). 14
1.3 Un exemple de modèle dimensionnel : Circumplex de Russell 15

2.1 Architecture d'un système de reconnaissance des émotions 22
2.2 Les trois paramètres caractéristiques de la voix 24
2.3 Principe de l'apprentissage automatique 33

3.1 Les étapes nécessaires pour la constitution d'un corpus émotionnel. 45
3.2 Photo de l'installation de locuteur lors de la collecte de notre corpus 47
3.3 Taux de reconnaissance par émotion : résultat sur un échantillon 49
3.4 Taux de reconnaissance par émotion : résultat global brut 50
3.5 Comparaison des taux de reconnaissance sur corpus biaisé et non biaisé 52
3.6 Les taux de reconnaissance et de confusion des différentes émotions sur la base entière. 54

4.1 Banc de filtres sur une échelle de Mel. 62
4.2 Étapes de l'extraction des coefficients MFCC. 62
4.3 Étapes de l'extraction des coefficients MSF 64
4.4 Exemple d'un signal vocal décomposé en IMFs en utilisant l'algorithme EMD . . 65
4.5 Le schéma fonctionnel du lissage de l'opérateur d'énergie. 68
4.6 Étapes de l'extraction des coefficients SMFCC 69
4.7 Étapes de l'extraction des coefficients ECC et EFCC 70
4.8 Étapes de l'extraction des coefficients MAF 73
4.9 Moyenne de l'énergie $E(i, j)$ pour les 7 émotions. 75
4.10 Estimation de la densité de probabilité de $\overline{\Phi}_3$ pour les 3 émotions (tristesse, neutre et colère) 75

4.11 Étapes de l'extraction des coefficients MFF 76
4.12 Estimation de la fréquence moyenne de tous les locuteurs allemands. 77
4.13 Exemple d'un signal d'un fichier pour l'émotion "colère" et sa représentation temps-fréquence. 78
4.14 Exemple d'un signal d'un fichier pour l'émotion "neutre" et sa représentation temps-fréquence. 79
4.15 Estimation de la densité de probabilité de $\overline{F}$ de l'IMF10 pour les 3 émotions (tristesse, neutre et colère). 79
4.16 Estimation de la densité de probabilité de $\overline{F}^w$ de l'IMF10 pour les 3 émotions (tristesse, neutre et colère). 80
4.17 Estimation de la densité de probabilité de $\overline{F}$ de l'IMF10 pour les 3 émotions (tristesse, neutre et colère), pour 3 différentes tailles de trame. 80
4.18 Estimation de la densité de probabilité de $\overline{F}^w$ de l'IMF10 pour les 3 émotions (tristesse, neutre et colère), pour 3 différentes tailles de trame. 81
4.19 Estimation de la densité de probabilité de $\overline{F}$ de l'IMF10 d'un homme et d'une femme pour les 3 émotions (tristesse, neutre et colère). 81
4.20 Estimation de la densité de probabilité de $\overline{F}^w$ de l'IMF10 d'un homme et d'une femme pour les 3 émotions (tristesse, neutre et colère). 82

5.1 Schéma d'un réseau de neurones récurrents à une unité reliant l'entrée et la sortie du réseau. A droite la version « dépliée » de la structure 87
5.2 Long Short-Term Memory (LSTM) avec c et $\tilde{c}$ sont respectivement la mémoire et le nouveau contenu de la mémoire. 88
5.3 Notre modèle de réseau LSTM 89
5.4 Comparaison des 3 différentes méthodes de sélection (RFE, PCA, LDA) sur la base allemande avec la combinaison des caratéristiques SMFCC, MFF et MAF. . 98
5.5 Nombre de classifications correctes en fonction du nombre de caractéristiques sélectionnées pour la combinaison SMFCC+MFF+MAF de la base allemande en utilisant la méthode de sélection RFE-SVM. 98
5.6 Comparaison des taux de reconnaissance des émotions du corpus "French Student Emotionnal database" basée sur le classifieur SVM et l'oreille humaine. 102

A.1 Banc de mesure 116
A.2 Spectre de la réponse impulsionnelle. 116
A.3 Zoom sur la plage 60 Hz-1200 Hz. 117

A.4 Schéma de la chaine d'enregistrement 118

Liste des tableaux

1.1 Les émotions primaires selon différents auteurs 12
1.2 Tableau comparatif des différentes catégories de corpus 19

2.1 Résumé des effets des émotions sur la parole traduit de l'anglais. 25
2.2 Types d'émotions associées aux paramètres acoustiques. 25
2.3 Tableau récapitulatif des études sur la reconnaissance automatiques des émotions 38

3.1 Exemples des phrases et scénarios établis en fonction des différentes émotions. . . 46
3.2 Nombre de locuteurs et d'enregistrements. 48
3.3 Nombre d'enregistrements par émotion et par genre. 48
3.4 Score et intervalle de confiance en fonction de chaque émotion 51
3.5 Décomposition des taux de reconnaissance avec et sans biais sémantique. 53
3.7 Taux de reconnaissance par locuteur pour la base entière. 56
3.8 Taux de reconnaissance par testeur pour la base avec biais sémantique. 57
3.9 Taux de reconnaissance par testeur pour la base sans biais sémantique 57

4.1 Description du nombre de caractéristiques extraites. 82

5.1 Récapitulatif des résultats des 3 classifieurs avec les différentes caractéristiques sur la base allemande. 90
5.2 Récapitulatif des résultats des 3 classifieurs avec les différentes caractéristiques sur la base espagnole. 91
5.3 Récapitulatif des résultats des 3 classifieurs avec les meilleures combinaisons sur la base allemande. 92
5.4 Récapitulatif des résultats des 3 classifieurs avec les meilleures combinaisons sur la base espagnole. 92

5.5 Matrice de confusion pour la combinaison des nouvelles caractéristiques (MAF et MFF) avec les SMFCC, ECC et EFCC en utilisant le classifieur SVM sur le corpus allemand. 93

5.6 Matrice de confusion pour la combinaison des nouvelles caractéristiques (MAF et MFF) avec les SMFCC et EFCC en utilisant le classifieur SVM sur le corpus espagnol. 93

5.7 Résultats sans et avec normalisation par locuteur obtenus sur la base allemande . 94

5.8 Résultats sans et avec normalisation par locuteur obtenus sur la base espagnole . 95

5.9 Résultats avec normalisation par locuteur pour les meilleures combinaisons obtenus sur la base allemande 96

5.10 Résultats avec normalisation par locuteur pour les meilleures combinaisons obtenus sur la base espagnole ; 97

5.11 Comparaison des résultats sans et avec sélection de caractéristiques en utilisant le classifieur RNN sur la base allemande (avec SN). 99

5.12 Comparaison des résultats sans et avec sélection de caractéristiques en utilisant le classifieur SVM sur la base espagnole (avec SN). 100

5.13 Récapitulatif des meilleurs résultats obtenus sur le corpus en français : résultats avec normalisation par locuteur et avec sélection. 102

A.1 Tableau comparatif des salles envisagées 114

A.2 Tableau détaillé contenant les informations sur les locuteurs. 120

A.3 Les hypothèses pour un risque de 1% et de 5% des différents taux de reconnaissance des émotions pour le corpus avec et sans biais sémantique. 123

A.4 L'intervalle de confiance à 95% et à 99% des différents taux de reconnaissance des émotions pour le corpus avec et sans biais sémantique. 124

A.5 Taux de reconnaissance par testeur pour la base avec biais sémantique. 125

A.6 Taux de reconnaissance par testeur pour la base sans biais sémantique. 127

A.6 Taux de reconnaissance par testeur pour la base sans biais sémantique 128

B.1 Récapitulatif des résultats des trois classifieurs avec les différentes combinaisons sur la base allemande. 129

B.2 Récapitulatif des résultats des trois classifieurs avec les différentes combinaisons sur la base espagnole. 130

B.3 Comparaison des résultats sans et avec sélection de caractéristiques en utilisant le classifieur RL sur la base allemande (avec SN). 131

B.4 Comparaison des résultats sans et avec sélection de caractéristiques en utilisant le classifieur SVM sur la base allemande (avec SN). 131

B.5 Comparaison des résultats sans et avec sélection de caractéristiques en utilisant le classifieur RL sur la base espagnole (avec SN). 132

B.6 Comparaison des résultats sans et avec sélection de caractéristiques en utilisant le classifieur RNN sur la base espagnole (avec SN). 132

Table des matières

Résumé i

Abstract ii

Remerciements iii

Introduction générale 1

Liste des publications 5

I Etat de l'art 7

1 Théorie des émotions 9

1.1 Introduction . . . 9

1.2 Définition de l'émotion . . . 10

1.3 Classification des émotions . . . 11

1.3.1 Émotions primaires . . . 11

1.3.2 Émotions secondaires . . . 13

1.4 Représentation des émotions . . . 13

1.4.1 Approche catégorielle . . . 14

1.4.2 Approche dimensionnelle . . . 14

1.5 Canaux de communication émotionnelle . . . 15

1.5.1 Les expressions faciales . . . 16

1.5.2 Les signaux physiologiques . . . 16

1.5.3 La voix . . . 17

1.6 Corpus émotionnels de parole . . . 17

1.6.1 Corpus naturel . . . 17

1.6.2 Corpus induit . . . 18

1.6.3 Corpus acté . . . 18
1.7 Conclusion . . . 19

2 Reconnaissance automatique des émotions à partir de la voix 20
2.1 Introduction . . . 20
2.2 Voix et émotion . . . 22
2.3 Extraction de caractéristiques . . . 26
2.3.1 Descripteurs prosodiques . . . 27
2.3.2 Descripteurs spectraux . . . 27
2.4 Sélection de caractéristiques . . . 29
2.4.1 Stratégies de recherche . . . 29
2.4.1.1 Stratégie de recherche exhaustive . . . 29
2.4.1.2 Stratégie de recherche heuristique . . . 30
2.4.1.3 Stratégie de recherche aléatoire . . . 30
2.4.2 Critères d'évaluation . . . 31
2.4.2.1 Méthodes filtres (filter) . . . 31
2.4.2.2 Méthodes enveloppes (wrapper) . . . 31
2.4.2.3 Méthodes intégrées (embedded) . . . 31
2.5 Apprentissage automatique . . . 32
2.5.1 Comprendre l'apprentissage automatique . . . 32
2.5.2 Types d'apprentissage automatique . . . 33
2.5.3 Algorithmes existants pour la classification . . . 34
2.6 Conclusion . . . 36

II Système de reconnaissance des émotions à partir de la voix 39

3 Les corpus émotionnels, base de données et collecte d'informations 41
3.1 Introduction . . . 41
3.2 Corpus utilisés dans le cadre de ces travaux . . . 42
3.2.1 Berlin database . . . 42
3.2.2 Spanish database . . . 43
3.3 Constitution de notre corpus émotionnel : le corpus "French Student Emotionnal database" : . . . 44
3.3.1 Construction . . . 45
3.3.1.1 Élaboration des énoncés et scénarios . . . 45

3.3.1.2 Salle et matériel d'enregistrement . . . 45
3.3.1.3 Méthode d'enregistrement . . . 47
3.3.1.4 Résultats . . . 47
3.3.2 Test . . . 48
3.3.2.1 Méthode de test . . . 49
3.3.2.2 Résultats . . . 50
3.3.3 Validation . . . 50
3.3.3.1 Questions . . . 51
3.3.3.2 Analyse des résultats . . . 51
3.4 Conclusion . . . 58

4 Extraction de caractéristiques **59**
4.1 Introduction . . . 59
4.2 Extraction des caractéristiques basée sur un banc de filtres . . . 61
4.2.1 Les coefficients cepstraux sur l'échelle Mel (MFCC) . . . 61
4.2.2 Les caractéristiques spectrales de modulation (MSF) . . . 63
4.3 Démodulation par décomposition en modes empiriques . . . 64
4.3.1 Décomposition en modes empiriques (EMD) . . . 64
4.3.1.1 Fonction Mode intrinsèque (IMF) . . . 64
4.3.1.2 Algorithme . . . 65
4.3.2 Modélisation de la voix par un signal modulé en amplitude et en fréquence 66
4.3.3 Opérateur d'énergie de Teager-Kaiser (TKEO) . . . 66
4.4 Extraction des caractéristiques basée sur la démodulation par EMD-TKEO . . . 68
4.4.1 Les coefficients cepstraux . . . 68
4.4.1.1 SMFCC . . . 68
4.4.1.2 Les coefficients cepstraux d'énergie (ECC) et les coefficients cepstraux d'énergie pondérée en fréquence (EFCC) . . . 69
4.4.2 Les caractéristiques de modulation AM-FM . . . 72
4.4.2.1 Les caractéristiques de modulation d'amplitude (MAF) . . . 72
4.4.2.2 Les caractéristiques de modulation de fréquence (MFF) . . . 74
4.5 Conclusion . . . 81

5 Sélection de caractéristiques et classification **84**
5.1 Introduction . . . 84
5.2 Classification . . . 85

5.2.1 Les algorithmes utilisés . . . 85
5.2.1.1 Régression linéaire . . . 85
5.2.1.2 Machines à support de vecteurs . . . 85
5.2.1.3 Réseaux de neurones récurrents . . . 87
5.2.2 Résultats et discussions . . . 89
5.3 Fusion de caractéristiques . . . 91
5.4 Normalisation par locuteur . . . 93
5.5 Sélection de caractéristiques . . . 95
5.5.1 Recursive Feature Elimination . . . 97
5.5.2 Résultats et discussions . . . 99
5.6 Comparaison avec les systèmes existants . . . 100
5.7 Évaluation du corpus “French Student Emotionnal database” . . . 101
5.8 Conclusion . . . 102

Conclusion générale **104**

A Notre corpus émotionnel **106**
A.1 Énoncés et mises en situation . . . 106
A.2 Phrases contenant un biais sémantique . . . 112
A.3 Enregistrement . . . 113
A.3.1 Choix de la salle d’enregistrement . . . 113
A.3.1.1 Temps de Réverbération . . . 113
A.3.1.2 Réponse impulsionnelle . . . 115
A.3.2 Choix du matériel . . . 117
A.3.3 Procédure d’enregistrement . . . 118
A.3.4 Résultat . . . 119
A.4 Validation du corpus . . . 121
A.4.1 Étude statistique des tests de reconnaissance des émotions . . . 121
A.4.1.1 Étude 1 : Impact du biais sémantique sur la reconnaissance de l’émotion par l’humain . . . 121
A.4.1.2 Étude 2 : Intervalle de confiance la reconnaissance de l’émotion par l’humain . . . 122
A.4.2 Impact du testeur . . . 125

B Résultats détaillés des expériences du système de RAE **129**

Introduction générale

Contexte général

La voix est un signal généré par un processus psychoacoustique complexe développé à la suite de milliers d'années d'évolution humaine. Il comporte une multitude d'informations sur le locuteur telles que son âge, son identité, son genre, ses émotions et troubles physiologiques ressentis lors de l'expression orale. L'extraction de ces informations a donné naissance à plusieurs domaines de recherche sur la parole notamment la reconnaissance des émotions à partir de la voix. C'est un champ de recherche pluridisciplinaire, qui s'appuie sur les travaux en psychologie, physiologie, neurologie, traitement de la parole, traitement du signal, réalité virtuelle, etc. Les émotions ou « états émotionnels » sont fondamentaux pour les humains dans la mesure où elles nous imprègnent de façon consciente et inconsciente dans les domaines les plus variés de notre vie. Elles influencent nos perceptions, nos comportements, nos états mentaux et nos activités quotidiennes telles que la communication, l'apprentissage et la prise de décision. L'importance des émotions dans le processus d'apprentissage est connue depuis longtemps . Elles font l'objet de différentes études comme par exemple la reconnaissance des états émotionnels de l'apprenant, l'influence des moyens d'interaction sur les émotions de l'apprenant, ou l'induction d'états émotionnels permettant de maximiser la performance de l'apprenant dans son activité d'apprentissage, telle que la mémorisation, la compréhension. De la même manière, elle peut être un vecteur d'échec et de perte de confiance quand elle est inadaptée, dérégulée, ou ignorée. A l'école, la classe est une génératrice d'émotions positives et négatives. Les émotions « positives » ou agréables permettent de faciliter l'apprentissage contrairement aux émotions « négatives » ou désagréables qui entraveraient plutôt ce dernier .

Ainsi l'émotion représente un catalyseur puissant, non seulement du bien-être à l'école, mais également de la réussite scolaire. Dans ce contexte, il est essentiel d'outiller les lieux d'apprentissage tels que les classes ou des outils dans le cas de relation homme machine.

La reconnaissance automatique des émotions (RAE) à partir de la voix a suscité un intérêt considérable au cours des dernières années. Cependant, malgré les progrès réalisés dans ce domaine, la reconnaissance des émotions à partir de la voix fait toujours face à de nombreux défis. En particulier, en tant que tâche d'apprentissage automatique, les systèmes de RAE de hautes performances exigent des caractéristiques efficaces. Les caractéristiques acoustiques utilisées pour la reconnaissance des émotions peuvent être regroupées en deux catégories : prosodiques et spectrales. Il a été démontré que les caractéristiques prosodiques transmettent des informations très importantes à propos du locuteur. Bien qu'aucun accord n'a été trouvé sur l'ensemble optimal à utiliser, les caractéristiques prosodiques constituent le type de caractéristiques le plus utilisé dans les systèmes de RAE et ont été étudiées de manière approfondie dans des travaux antérieurs. Les caractéristiques spectrales jouent également un rôle important dans le système de RAE, car elles sont liées au contenu fréquentiel du signal vocal, et fournissent ainsi des informations complémentaires aux caractéristiques prosodiques. Cependant, la voix étant un signal non stationnaire et issu d'un système non linéaire, trouver les caractéristiques les plus performantes reste un défi. La plupart des méthodes utilisées pour obtenir des caractéristiques, basées sur des techniques de décomposition du signal dans une base prédéfinie, sont limitées dans leurs analyses par l'hypothèse de stationnarité et de linéarité. La décomposition en mode empirique (EMD) permet la projection du signal dans une base adaptative. Le couplage de cette méthode avec l'opérateur d'estimation d'énergie TKEO permet de faire une analyse temps-fréquence sans être contraint par les hypothèses de stationnarité et de linéarité.

Un autre défi, et non des moindres, de la reconnaissance automatique des émotions réside dans la construction d'un corpus émotionnel. Il existe un grand nombre de corpus (principalement en allemand ou en anglais), mais peu d'entre eux sont accessibles gratuitement à l'ensemble de la communauté scientifique. Ainsi nous avons crée un corpus en français dans le contexte pédagogique. Dans cette thèse, un corpus en allemand et un corpus en espagnol, sont aussi exploités.

Contribution

La présente thèse apporte principalement les contributions suivantes :

1. Une méthode d'extraction des caractéristiques basée sur la démodulation AM-FM (modulation d'amplitude et modulation de fréquence) de la voix par le couplage EMD-TKEO.
2. Un corpus de discours émotionnel en langue française dans un contexte pédagogique, appelé

"French student emotional database". Ce corpus testé et validé, sera mis à disposition de la communauté.

3. Un système de RAE utilisant les caractéristiques proposées et les caractéristiques spectrales et cepstrales existantes dans la littérature. Trois classifieurs (régression linéaire, machines à vecteurs de support et réseau de neurones réccurents) sont utilisés pour l'apprentissage automatique.

4. Une méthode de combinaison des nouvelles caractéristiques et des caractéristiques cepstrales avec la sélection des plus pertinentes afin d'augmenter les performances.

Structure du document

Ce document est organisé en deux grandes parties. La première partie présente l'état de l'art et les différentes connaissances utiles à la bonne compréhension des travaux de cette thèse. Dans le premier chapitre nous commençons par présenter les notions nécessaires concernant les émotions telles que leurs définitions, leurs types, leurs représentations et les différents canaux permettant de communiquer ces émotions. Ensuite, nous décrivons les différents types de corpus émotionnels. Le chapitre 2 fournit une revue de la littérature pour le système de RAE. Nous expliquons le lien entre la voix et l'émotion. Nous présentons ensuite les composants de ce système et nous décrivons toutes les étapes nécessaires à ce système pour passer d'un signal vocal à une émotion bien précise, c'est à dire l'extraction des caractéristiques qui permettent de modéliser le mieux possible les classes d'émotions, les méthodes de sélection et les algorithmes d'apprentissage automatique utilisés pour un système de RAE.

Dans la deuxième partie, nous développons le travail réalisé durant cette thèse, à savoir la construction d'un corpus émotionnel, l'extraction des nouvelles caractéristiques, l'élaboration d'un système de RAE et enfin l'amélioration de la performance de ce système. Nous présentons ainsi dans le troisième chapitre les deux corpus que nous avons utilisés pour valider notre système. Puis nous détaillons la méthode de construction d'un nouveau corpus émotionnel en français et nous présentons les résultats de tests et de validation de ce corpus. Dans le quatrième chapitre, une description détaillée des nouvelles caractéristiques proposées est donnée. Les caractéristiques de comparaison sont également introduites. Les résultats de la simulation sont présentés et discutés dans le chapitre 5. La classification des émotions discrètes est effectuée. Nous proposons également, suite aux résultats obtenus, d'évaluer différentes combinaisons des caractéristiques afin d'augmenter les performances. Nous nous intéressons aussi à enrichir le système de RAE avec

la normalisation de données et la sélection des meilleures caractéristiques. Enfin, nous présentons un résumé des points clés de la thèse et quelques perspectives pour des travaux futurs.

Liste des publications

Articles de revues

— Leila Kerkeni, Youssef Serrestou, Kosai Raoof, Mohamed Mbarki, Mohamed Ali Mahjoub and Catherine Cleder, "Automatic Speech Emotion Recognition using an Optimal Combination of Features based on EMD-TKEO", Speech Communication(2019), Elsevier, DOI : https://doi.org/10.1016/j.specom.2019.09.002.

Chapitre d'ouvrage

— Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub and Catherine Cleder, "Automatic Speech Emotion Recognition Using Machine Learning", Social Media and Machine Learning book, IntechOpen, March 2019. DOI : 10.5772/intechopen.84856.

Conférences internationales

— Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof and Mohamed Ali Mahjoub, "A Review on Speech Emotion Recognition : Case of Pedagogical Interaction in Classroom", The 3rd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), May 2017, Fez (Morocco).

— Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Mohamed Ali Mahjoub, Kosai Raoof, and Catherine Cleder, "Speech Emotion Recognition : Recurrent Neural Networks compared to SVM and Linear Regression", The 26th International Conference on Artificial Neural Networks (ICANN), September 2017, Sardinia (Italy).

— Leila Kerkeni1, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof and Mohamed Ali Mahjoub, "Speech Emotion Recognition : Methods and Cases Study", The 10th International Conference on Agents and Artificial Intelligence (ICAART), January 2018, Madeira (Portugal).

Présentations scientifiques sans comité de lecture

— Leila Kerkeni, "Reconnaissance acoustique des émotions par réseaux de neurones récurrents : Application à l'interaction pédagogique en classe", journée de l'AS "Visage, geste, action et comportement " de la GDR 720 ISIS (Information, Signal, Image et Vision), Novembre 2016, Télécom ParisTech, Paris.

— Leila Kerkeni, "Détections des états émotionnels-Application à l'interaction pédagogique en classe", Journée des Doctorants, Avril 2017, Nantes.

Première partie

Etat de l'art

Chapitre 1

Théorie des émotions

Contents

1.1 Introduction **9**

1.2 Définition de l'émotion **10**

1.3 Classification des émotions **11**

1.3.1 Émotions primaires 11

1.3.2 Émotions secondaires 13

1.4 Représentation des émotions **13**

1.4.1 Approche catégorielle 14

1.4.2 Approche dimensionnelle 14

1.5 Canaux de communication émotionnelle **15**

1.5.1 Les expressions faciales 16

1.5.2 Les signaux physiologiques 16

1.5.3 La voix 17

1.6 Corpus émotionnels de parole **17**

1.6.1 Corpus naturel 17

1.6.2 Corpus induit 18

1.6.3 Corpus acté 18

1.7 Conclusion **19**

1.1 Introduction

L'objectif de ce travail est de réaliser un système qui identifie des émotions exprimées vocalement. Pour parvenir à cela, il est nécessaire de comprendre le phénomène émotionnel avant de

pouvoir construire ce système. Nous débutons donc ce manuscrit par la présentation des notions nécessaires concernant les émotions telles que leurs définitions, leurs types, leurs représentations et les différents canaux permettant de communiquer ces émotions. Ensuite, nous présentons les différents types de corpus émotionnels.

1.2 Définition de l'émotion

Qu'est-ce qu'une émotion ?

La définition d'une émotion est un sujet de controverse sur lequel se sont penchés des psychologues, des neurologues, des philosophes et aussi des chercheurs en informatique. C'est l'un des concepts les plus difficiles à définir en psychologie. Le mot "émotion", qui signifie remuer ou mouvement vers l'extérieur, remonte à l'année 1579. Son origine vient du mot français "se mouvoir" et du mot latin "emovere, emotum"(enlever, secouer). Le dictionnaire définit l'émotion comme un état affectif intense caractérisé par des troubles physiques et mentaux, ce qui est perçu comme le résultat d'un phénomène extérieur (appelé stimulus : une image, une odeur, un son...) [1]. Définir une émotion est une réelle question scientifique qui donne même son titre à l'article théorique le plus célèbre du psychologue et philosophe américain William James. En 1884, James a définit l'émotion comme une réaction physiologique : lorsqu'on est dans la forêt et qu'un ours apparaît, nos os tremblent à cause de l'ours et on éprouve de la peur parce qu'on sent nos os trembler (et non pas à cause de l'ours). Dans [2], Ekman et Davidson définissent l'émotion comme une réaction aiguë et transitoire, provoquée par un stimulus spécifique et caractérisée par un ensemble cohérent de réponses cognitives, physiologiques et comportementales. Lecomte définit aussi l'émotion comme "une réaction de l'organisme à un événement extérieur, et qui comporte des aspects physiologiques, cognitifs et comportementaux "[3]. D'après l'étymologie, les émotions provoquent des états internes qui engagent le corps et le cerveau. Ces états internes peuvent être positifs comme la joie ou négatifs comme la colère. Le terme "émotion "peut aussi désigner un état extrêmement complexe associé à une grande variété de changements mentaux, physiologiques et physiques. Plusieurs définitions ont été données à l'émotion, une liste non exhaustive de celles ci est donnée dans [4]. Malgré ces divergences, la majorité des scientifiques s'accordent à dire qu'une émotion présente cinq composantes :

— une expression (faciale, vocale ou encore posturale), qui permet d'interagir avec l'environnement et de communiquer l'émotion,

— une motivation (la tendance à l'action),

— une réaction corporelle (un changement du rythme cardiaque, un changement de la température du corps, une variation de la fréquence respiratoire) ,

— un sentiment, une prise de conscience de son état émotionnel, ce qui permet notamment de le verbaliser ou de le réguler,

— une évaluation cognitive qui détermine la nature et l'intensité de la réaction émotionnelle.

Nous pouvons donc considérer l'émotion comme une réaction physiologique de durée brève et transitoire à une situation donnée. Les émotions ont différentes répercussions physiques. Par exemple, la peur peut déclencher un cri, des tremblements, une accélération du pouls, la tristesse peut provoquer les larmes ou la joie génère quant à elle un sourire.

1.3 Classification des émotions

Il existe une très grande variété d'émotions avec leurs nuances, leurs combinaisons, leurs variantes. Tous les psychologues et philosophes ne sont pas d'accord quant à la classification mais tout le monde semble admettre qu'il existe des émotions principales avec des nuances de celles-ci. Les émotions peuvent alors être divisées en deux classes, à savoir les émotions primaires et les émotions secondaires.

1.3.1 Émotions primaires

Les émotions primaires, dites également "émotions de base " [5], sont universelles et innées. Non seulement elles sont innées mais en plus de ça elles sont automatiques, elles sont inconscientes et elles ont un déclenchement rapide, telles des réflexes. Izard, Plutchik, Ekmanet Tomkins ont développé la théorie des émotions de bases [6]. Le Tableau 1.1 présente une liste des émotions primaires selon différents auteurs. En 1960, Arnold a défini 11 émotions primaires (la colère, l'aversion, le courage, le dégoût, le désespoir, le désir, la peur, la haine, l'espoir, l'amour et la tristesse). En 1972, Ekman, Friesen et Ellsworth [7] ont conçu une liste de six émotions primaires dont chacune correspond à une expression du visage. Les expressions sont les mêmes entre plusieurs personnes d'âge, de culture ou de sexe différents [8]. Ces émotions telles qu'illustre la Figure 1.1 sont : la tristesse, la colère, la joie, le dégoût, la peur et la surprise. Charles Darwin [9] publie aussi dans son ouvrage intitulé "L'expression des émotions chez l'Homme et les animaux"

1. Cette photo a été réalisée par les étudiants 5A à l'ENSIM dans le cadre d'un projet que nous avons proposé. Nous les remercions d'avoir aimablement autorisé l'inclusion de cette photo.

Auteurs	Émotions primaires (langue d'origine)
Ekman et al. (1982)	anger, disgust, fear, joy, sadness, surprise
Izard (1971)	anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
Plutchik (1980)	acceptance, anger, anticipation, disgust, fear, joy, sadness, surprise
Tomkins (1984)	anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Arnold (1960)	anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Fridja (1986)	desire, happiness, interest, surprise, wonder, sorrow
Gray (1982)	rage, terror, anxiety, joy
Mower (1960)	pain, pleasure
James (1884)	fear, grief, love, rage, anger, disgust
McDougall (1926)	elation, fear, subjection, tender-emotion, wonder
Weiner and Graham (1984)	happiness, sadness
Panksepp (1982)	expectancy, fear, rage, panic

Table 1.1: Les émotions primaires selon différents auteurs [6].
(Nous avons laissé les émotions dans la langue d'origine pour respecter au mieux le sens des auteurs)

que l'espèce humaine présente six émotions fondamentales et universelles, ce sont les émotions primaires : peur, joie, tristesse, dégoût, surprise et colère. Gray (1982) a proposé quatre émotions primaires : l'espoir, la panique, la peur et la rage. En psychologie, il n'y a pas le même nombre d'émotions de base, mais seules 6 émotions de base sont communes aux divers auteurs. En 1976, Ekman a mis au point un outil (le Facial Action Coding System (FACS)) répertoriant les 46 composantes de base des expressions du visage humain (froncement de sourcils, clignement de l'œil, mouvement des narines, serrement des lèvres,...) [10]. Cet outil lui a permis d'identifier, chez l'adulte, les photographies de visages exprimant les six émotions de base et la neutralité (aucune émotion). Ces sept émotions sont aujourd'hui mondialement connues. En outre, les bébés, de plus de 3 mois, sont capables de discriminer les expressions de visage des six émotions de base et la neutralité quand on les leur montre dans des films ou sur des photographies [11].

Dans notre travail, nous avons retenu cette catégorisation. Pour information et par souci d'exhaustivité, nous présentons ci-après la décomposition en émotions secondaires, mais qui est trop "complexe "à exploiter.

FIGURE 1.1: Les six émotions primaires[1]

1.3.2 Émotions secondaires

Il y a des émotions primaires, et il y a aussi des émotions secondaires qui jouent un rôle important dans la vie. Ces émotions dites également complexes sont issues des émotions primaires et résultent d'un mélange de ces dernières : la colère peut par exemple donner l'agressivité, la haine ; à partir de la peur, la culpabilité, l'angoisse... Selon Ekman, ces émotions secondaires sont des mélanges des émotions de base, il les nomme parfois des émotions mixtes. Par exemple, pour lui la honte est une émotion mixte, puisque c'est un mélange de peur et de colère. Parmi ces émotions secondaires, on trouve aussi la jalousie, la culpabilité, l'embarras et l'envie... Contrairement aux émotions primaires, les émotions secondaires ne sont pas innées. Elles n'apparaissent qu'entre un an et quatre ans. Elles ne sont pas systématiquement automatiques, leur déclenchement n'est pas forcement rapide et leur action peut durer dans le temps.

1.4 Représentation des émotions

En psychologie, plusieurs modèles ont été conçus pour décrire l'ensemble des émotions. Parmi ceux-ci, nous pouvons présenter les deux principales familles, les modèles dits catégoriels et les modèles dits dimensionnels. Les figures 1.2 et 1.3 donnent deux exemples d'illustration de ces familles de modèles.

1.4.1 Approche catégorielle

L'approche catégorielle (ou discrète), est basée sur un ensemble d'émotions primaires, qui sont innées et communes chez tous les peuples. Ce modèle permet de représenter une infinité d'émotions obtenues en combinant les émotions primaires. Il est utilisé par la plupart des études qui s'intéressent à la reconnaissance vocale des émotions [12]. Un exemple de modèle discret est celui de Plutchik qui est illustré dans la Figure 1.2. Ce modèle est composé de 8 émotions de base faites de 4 paires opposées deux à deux (joie-tristesse, anticipation-surprise, colère-peur et dégoût-confiance) et de multiple variations. Ces émotions et leurs variations sont représentées par des couleurs et des teintes différentes.

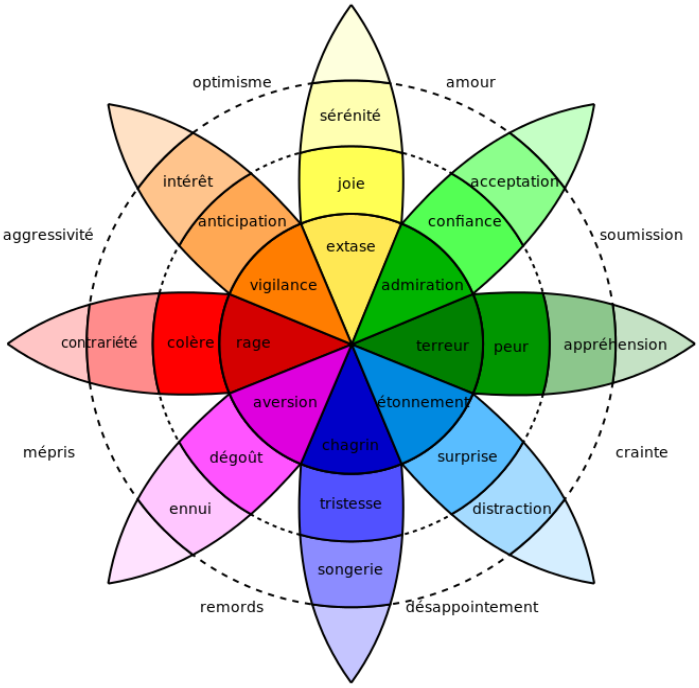


Figure 1.2: Un exemple de modèle catégoriel (ou discret) : La roue des émotions de Plutchik [13].

1.4.2 Approche dimensionnelle

Dans l'approche dimensionnelle (ou continue), les émotions sont situées dans un espace de deux ou trois dimensions. Les deux principales dimensions sont la valence[2] et l'activation[3] (ou arousal)

2. La valence, telle que définie en psychologie, est la qualité intrinsèquement agréable ou désagréable d'un stimulus ou d'une situation.
3. L'activation correspond au degré de l'expression corporelle ou gestuelle qui se traduit par des réactions physiologiques (accélération du rythme cardiaque, transpiration, etc.)

[14]. En effet, la notion de valence permet de représenter la manière dont se sent une personne quand , par exemple, elle regarde une image. Elle permet de distinguer deux types d'émotions, positives et négatives ou agréables et désagréables. La notion d'activation permet quant à elle de représenter le niveau d'excitation corporel. Un exemple de ce modèle est celui de Russell avec les dimensions valence et activation (ou encore valence-arousal) représenté par le circumplex de Russel sur la Figure 1.3.

Dans cet espace bidimensionnel, nous pouvons distinguer les émotions selon 4 quadrants : les émotions à valence positive et faible activation (par exemple le contentement et la relaxation) et les émotions à valence positive et forte activation (par exemple l'excitation et la joie), les émotions à valence négative et faible activation (par exemple la tristesse et l'ennui), les émotions à valence négative et forte activation (par exemple la peur et la colère).

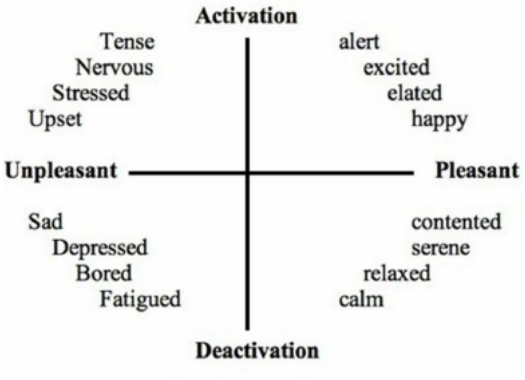


FIGURE 1.3: Un exemple de modèle dimensionnel : circumplex de Russell [15]

Dans ce travail, nous avons retenu l'approche catégorielle avec un ensemble de 7 émotions de base : la tristesse, la colère, le dégoût, la peur, la surprise, la joie et le neutre. Ce choix se justifie par le fait que l'approche dimensionnelle a été critiqué par plusieurs théoriciens tel que Izard [16] et Ekman [17], qui ont montré que la représentation des émotions dans un espace bidimensionnel ou tridimensionnel impliquait une perte d'informations. Par ailleurs, certaines émotions peuvent se trouver en dehors de l'espace à deux ou trois dimensions (par exemple, la surprise).

1.5 Canaux de communication émotionnelle

L'être humain exprime quotidiennement un grand nombre d'émotions. Les émotions sont perceptibles pour autrui visuellement, par l'ouïe et aussi à travers les gestes. Cette section présente les différentes modalités permettant de communiquer des émotions : l'expression faciale, les signaux physiologiques et la voix. L'expression de l'émotion par la voix sera abordée plus en détail dans

la seconde partie de cette étude bibliographique puisqu'elle va être utilisée dans ces travaux de thèse.

1.5.1 Les expressions faciales

Notre visage véhicule des informations riches (telles que l'identité individuelle, les expressions de communication verbale et non verbale, les émotions) à travers la direction du regard et les expressions faciales. Comme nous l'avons montré, dans la section 1.3.1, les émotions primaires possèdent des expressions faciales spécifiques, en termes de positions des sourcils, forme de la bouche... De plus, ces expressions faciales semblent être universelles selon Ekman [8]. L'expression faciale est la modalité d'expression non-verbale des émotions la plus étudiée. Cependant, elle a des limites dans son efficacité à communiquer des émotions. Dans les travaux de [18], les auteurs ont montré que les femmes semblent être meilleures que les hommes pour la reconnaissance des émotions à partir d'expressions faciales . En outre, bien que certaines émotions semblent avoir une expression faciale spécifique, Russell [19] a montré qu'il est rare que l'émotion contenue dans une expression faciale puisse être reconnue par 100 % des individus. Cela peut être justifié par le fait que : certaines expressions faciales sont communes à plusieurs émotions. Par exemple, la peur et la surprise se manifestent sur les visages des hommes par des sourcils levés. De plus, l'expression faciale n'est pas le seul moyen pour exprimer l'émotion, il existe d'autres modalités.

1.5.2 Les signaux physiologiques

L'expression des émotions est étroitement liée au système nerveux végétatif. Selon les émotions, nous pouvons observer des augmentations ou des diminutions de la fréquence cardiaque, de la sudation, du débit sanguin cutané (rougissement ou pâleur), variations de température et la conductivité électrique de la peau, du volume respiratoire, etc. Plusieurs indices physiologiques sont utilisés pour caractériser et mesurer une émotion [20]. Nous trouvons par exemple les signaux électroencéphalogrammes (EEG), le signal du volume respiratoire (VR), le rythme cardiaque (Heart Rate : HR), les variations de pression sanguine (Blood Volume Pulse : BVP), les variations de température de la peau (Skin Temperature : SKT), etc... Utiliser des signaux physiologiques pour reconnaître les émotions est également utile pour les personnes souffrant de maladie physique ou mentale, présentant ainsi des problèmes d'expressions faciales ou du timbre de la voix.

1.5.3 La voix

La voix est la forme la plus naturelle de communication humaine. Elle est considérée depuis longtemps comme une mesure de l'émotion et comme un réflecteur de la personnalité en raison de son potentiel pour exploiter les différences individuelles dans les états émotionnels et les dispositions de la personnalité [21]. La reconnaissance des émotions à partir de la voix est le cœur de notre travail. Nous avons donc dédié un chapitre aux travaux relatifs à ce domaine (voir chapitre 2).

Les premiers travaux sur la RAE ont exploité les expressions faciales (Ekman, Friesen Ellsworth, 1972 ; Izard 1995 ; Emde 1993). Ces travaux utilisent des types statiques d'expressions faciales qui ne prennent pas en compte la dynamique des états émotionnels et leurs variabilité. Pour prendre en compte ces aspects dynamiques, les chercheurs se sont intéressés aux signaux physiologiques qui donnent des bons résultats mais imposent un certain matériel et contraignent les utilisateurs : les dispositifs proposés sont relativement invassifs. Finalement travailler sur la voix avère à des dispositifs légers, peu couteux et non invassifs pour les utilisateurs. C'est cette approche qui est la plus répondue dans les travaux actuels.

1.6 Corpus émotionnels de parole

Nous distinguons essentiellement trois catégories de corpus émotionnels utilisées dans le domaine de la reconnaissance automatique des émotions : les émotions simulées, les émotions induites et les émotions naturelles. Une comparaison de différents corpus est illustré dans le Tableau 1.2.

1.6.1 Corpus naturel

Les corpus naturels ou réalistes sont obtenus à partir des enregistrements non contrôlés c'est à dire spontanées et naturelles. Les contextes dans lesquels sont recueillis ce type de données sont très variés (émission de TV, centre d'appels, interview de consommateurs). L'inconvénient est que ces données sont très limitées en quantité de données émotionnelles et en nombre de locuteurs, souvent de piètre qualité d'enregistrement (l'environnement d'enregistrement est publique quelque fois bruyant), de courtes durées. Pour ce type de corpus, l'étiquetage en classes d'émotions est particulièrement complexe [22]. En effet, dans une même phrase il peut y avoir

plusieurs émotions, il faut aussi faire abstraction de la sémantique qui peut biaiser l'émotion exprimée par le locuteur, etc.

1.6.2 Corpus induit

Une deuxième catégorie de corpus, appelée corpus induit, est utilisée dans le domaine de la reconnaissance automatique des émotions. Les émotions de cette catégorie sont induites en utilisant plusieurs techniques, par exemple l'exposition du sujet à des tâches difficiles à faire en peu du temps pour induire le stress, présentation des images, de films ou de jeux permettant d'induire des émotions. Les émotions induites sont souvent de faible intensité. De plus, les mêmes protocoles d'induction n'induisent pas nécessairement les mêmes états émotionnels chez les individus [22].

1.6.3 Corpus acté

Les corpus actés (appelés aussi simulés), contiennent des émotions produites par des acteurs professionnels ou semi professionnels en se basant sur le nom de la classe d'émotion et/ou des scénarios typiques. Ce type de corpus représente à l'heure actuelle la majorité des données utilisées dans les études mennées sur la reconnaissance automatique des émotion (RAE) [23], [24]. Ces corpus présentent plusieurs avantages. Ils permettent d'obtenir un grand nombre de données très prototypiques. Ces données sont faciles à collecter. L'enregistrement en laboratoire permet de faciliter la contrôlabilité et l'exploitation du corpus (pas de problèmes concernant les droits de diffusion, haute résolution). Les performances obtenues avec des données simulées sont largement meilleures à celles obtenues avec des données réalistes [25].

En conclusion de cette partie, la méthode la plus fréquemment utilisée pour étudier les expressions vocales émotionnelles consiste à enregistrer des acteurs qui simulent l'émotion. Bien que cette utilisation suscite aussi des critiques car les expressions ne correspondent pas exactement aux expressions véritables. Nous avons choisi, dans un premier temps, par souci d'efficacité, de travailler avec des corpus émotionnels actés pour la détection des émotions.

Corpus	Réaliste	Qualité d'enregistrement	Reconnaissance par l'oreille humaine
naturel	plus réaliste	mauvaise qualité	n'est pas garantie
induit	moins réaliste	bonne qualité	moins garantie
acté	hors contexte	très bonne qualité	plus garantie

TABLE 1.2: Tableau comparatif des différentes catégories de corpus

1.7 Conclusion

Dans ce chapitre, nous avons abordé les généralités qui donnent le cadre de notre travail. Il est à noter que nous avons retenu une représentation catégorielle des émotions en s'appuyant sur des corpus émotionnels actés.

Pour approfondir la compréhension de notre travail, nous allons, au chapitre 2, détailler ce que dans la voix caractérise les émotions et présenter les fondamentaux d'un système de RAE.

Chapitre 2

Reconnaissance automatique des émotions à partir de la voix

Contents

2.1 Introduction **20**

2.2 Voix et émotion **22**

2.3 Extraction de caractéristiques **26**

2.3.1 Descripteurs prosodiques 27

2.3.2 Descripteurs spectraux 27

2.4 Sélection de caractéristiques **29**

2.4.1 Stratégies de recherche 29

2.4.2 Critères d'évaluation 31

2.5 Apprentissage automatique **32**

2.5.1 Comprendre l'apprentissage automatique 32

2.5.2 Types d'apprentissage automatique 33

2.5.3 Algorithmes existants pour la classification 34

2.6 Conclusion **36**

2.1 Introduction

Depuis 1858, Herbert Spencer a examiné, dans son intéressant essai sur la musique, le caractère que prend la voix humaine sous l'influence de l'émotion [26]. Il a montré que la voix est le siège de nombreuses manifestations émotionnelles et physiologiques. Elle se modifie suivant les

circonstances, dans sa qualité et dans sa force, c'est-à-dire dans son timbre, dans sa hauteur, dans sa sonorité et dans son étendue. Personne ne peut écouter un homme s'adressant avec colère à un autre, ou un orateur éloquent, ou quelqu'un exprimant de l'étonnement, sans être touché. L'influence est perçue, ressentie par l'auditeur et peut conduire à modifier sa réaction. M. Spencer est un des premiers à mettre en relation la fréquence fondamentale et les émotions. Pour lui, il existe une relation entre la hauteur de la voix et certains états de la sensibilité. Par exemple, une personne qui souffre légèrement, ou qui se plaint doucement de mauvais traitement, parle presque toujours d'une voix que l'on dit haut perchée. Après lui, Charles Darwin en 1872 a essayé de citer et compléter de son analyse et de ses observations les travaux de Herbert Spencer [26].

Au XX^{e} siècle, les recherches sur la voix et l'émotion se sont développées jusqu'à proposer aujourd'hui une reconnaissance automatique de l'émotion (RAE). De nos jours, la RAE à partir de la voix est un domaine de recherche particulièrement dynamique qui couvre un large champ d'applications. Parmi les applications on trouve par exemple : les systèmes embarqués pour contrôler l'état du conducteur de voiture, les systèmes de traduction automatique, les centres d'appels et les cockpits d'avions. Un système de RAE est utilisé aussi pour les applications nécessitant une interaction naturelle homme-machine, telles que les films Web, l'e-learning et les didacticiels. Il peut également être utilisé comme outil de diagnostic pour les thérapeutes.

Nous commençons tout d'abord par donner l'architecture type d'un système de RAE (voir Figure 2.1) puis les différentes étapes nécessaires à sa mise en œuvre. Ce système prend en entrée un signal vocal et produit en sortie l'émotion véhiculée par la voix du locuteur. Les étapes sont listées ci-dessous et chaque étape sera explicitée dans un chapitre la concernant :

a) Choix d'un corpus de données : le corpus émotionnel a pour objectif l'entrainement des systèmes. Pour le construire, on peut demander à des acteurs de jouer des émotions particulières, construire des systèmes pour induire des émotions ou utiliser des données plus ou moins naturelles et spontanées ;

b) Extraction de caractéristiques : le signal de la parole contient un grand nombre de paramètres qui reflètent les caractéristiques émotionnelles. L'un des principaux points de la reconnaissance des émotions est le choix des caractéristiques à utiliser ;

c) Sélection de caractéristiques : les caractéristiques extraites à l'étape précédente peuvent contenir des informations redondantes et/ou non pertinentes susceptibles d'affecter la précision du système de RAE. Une phase de sélection des meilleures caractéristiques est

donc nécessaire ;

d) Classification : elle suit souvent une phase de sélection des meilleurs caractéristiques. A partir de la création d'un modèle caractérisant chaque classe (dite phase apprentissage), la phase de classification consiste à associer une classe à un segment de parole en comparant ses caractéristiques aux modèles de chaque classe. De nombreux algorithmes peuvent être utilisés ou mélangés.

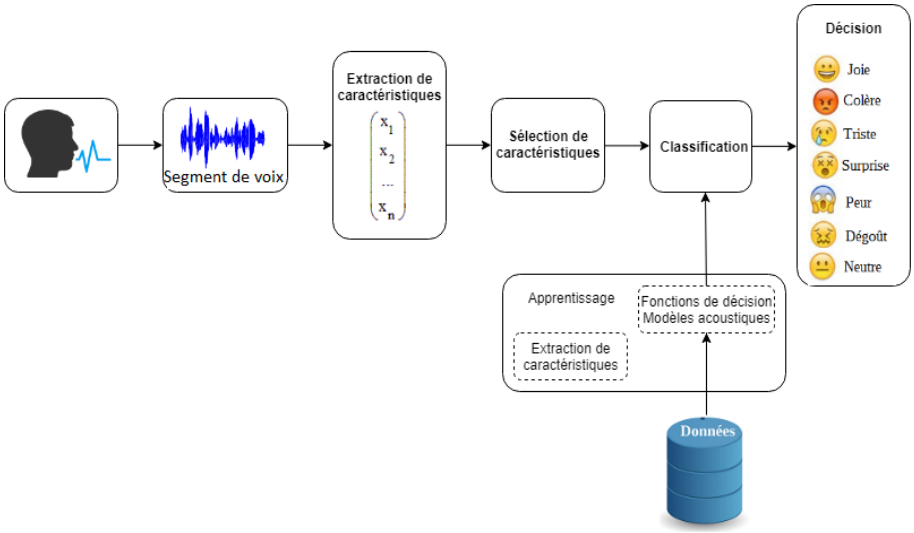


FIGURE 2.1: Architecture d'un système de reconnaissance des émotions

Pour réussir le développement d'un système de RAE, trois grands problèmes doivent être résolus : la construction d'une base de données de parole émotionnelle consistante, l'extraction de caractéristiques efficaces et la conception de classifieurs fiables à l'aide d'algorithmes d'apprentissage automatique.

Dans ce chapitre, et sur la base de la littérature scientifique, nous expliquons le lien entre la voix et l'émotion. Ensuite, nous décrivons chacun des composants de système de RAE. Premièrement, nous présentons l'ensemble des descripteurs acoustiques qui permettent de modéliser le mieux possible les données pour la tâche de reconnaissance. Deuxièmement, nous donnons une description des méthodes de sélection de caractéristiques utilisées pour la classification. Enfin, nous présentons les algorithmes d'apprentissage automatique.

2.2 Voix et émotion

Le signal vocal est un signal réel, continu, d'énergie finie, et peut être supposé stationnaire sur une courte durée (25 à 30 ms) au cours du temps [27]. Il transporte une quantité importante d'informations comme le message linguistique, la langue adoptée, l'identité du locuteur, ainsi que

ses émotions. Un système de RAE consiste à récupérer seulement l'information paralinguistique représentant l'état émotionnel du locuteur indépendamment des autres informations. Dans un tel système, l'analyse acoustique consiste à extraire du signal vocal un ensemble de paramètres discriminants et pertinents pour la tâche de classification. Le nombre de ces paramètres doit rester raisonnable, afin de réduire le temps de calcul dans la tâche d'apprentissage et améliorer ses résultats.

Le chemin de la voix est divisé entre trois parties (voir Figure 2.2) : les poumons, le larynx et les cordes vocales, les résonateurs (la gorge, la bouche, les fosses nasales). Les poumons produisent un flux d'air pour permettre la vibration des cordes vocales : plus le flux est faible, moins les cordes vocales vibrent et moins la voix s'entend. Donc la voix humaine est une émission de sons produits par la bouche et résultant de la vibration des cordes vocales et du choc de la glotte sur les muscles du larynx. Elle se caractérise par trois paramètres acoustiques : l'intensité, la fréquence et le timbre (voir Figure 2.2) [28].

— **L'intensité** de la voix (appellée aussi énergie) permet de distinguer un son fort d'un son faible. Elle dépend de l'amplitude de la vibration : plus elle est importante, plus le son est fort ; plus l'amplitude est faible, plus le son est faible. L'amplitude se crée par la variation de pression de l'air exercée sur ces cordes vocales [29].

— **La fréquence fondamentale (F0)** correspond à la fréquence de vibration des cordes vocales [13], c'est à dire au nombre d'ouvertures/fermetures par seconde des cordes vocales. Selon ce nombre de vibrations de l'air, on peut trouver des sons graves et des sons aigus. La fréquence fondamentale correspond d'un point de vue perceptif à la hauteur de la voix [29]. Dans le langage technique, le terme de ton/tonalité est aussi utilisé pour désigner la hauteur d'un son [29].

— **Le timbre** de la voix correspond aux sons sombres et sons clairs. L'air circule dans les résonateurs (gorge, bouche, fosses nasales) et va prendre sa couleur, son timbre et ses harmoniques. C'est ce qui permet par exemple de reconnaître qu'une note est jouée au hautbois ou au piano. C'est aussi grâce au timbre qu'on reconnait les voix de nos proches, tout comme les empreintes digitales, le timbre est unique à chacun et ne varie pas. Le timbre de la voix est lié au spectre de la voix [30].

Les même paramètres acoustiques caractérisant la voix mais sous des appellations différentes, dans la suite de ce travail, Nous utiliserons les termes : intensité pour désigner l'amplitude

et l'énergie, fréquence fondamentale pour la hauteur et la tonalité, et timbre pour désigner le spectre.

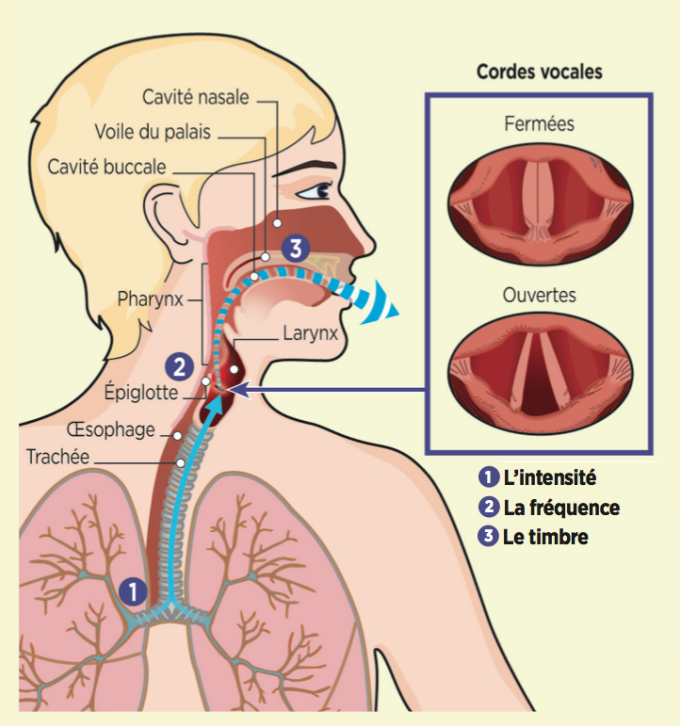


FIGURE 2.2: Les trois paramètres caractéristiques de la voix [31]

L'émotion joue un rôle prépondérant dans l'expression de la voix [26] : elle peut modifier la fréquence fondamentale (aigu, grave) , l'intensité et les harmoniques (mélange de fréquences vocales) de notre voix. Par exemple, lorsqu'un individu est angoissé sa voix devient cassée car certaines remontées acides vont imprégner les cordes vocales. Par contre, s'il est joyeux sa voix s'adoucit et devient plus harmonieuse tandis que la peur provoque quant à elle la contraction violente des cordes vocales et la voix devient peu harmonieuse. Plusieurs recherches sont faites dans le domaine des relations entre la voix et les émotions. Dans [32], l'auteur a proposé un résumé des caractéristiques prosodiques relevées par les auteurs de plusieurs études. Le tableau 2.1 présente une description des effets les plus communément associés avec les cinq émotions (colère, joie, tristesse, peur et dégoût) sur la voix.

Une autre étude (Scherer et Oshinsky, 1977) résume les résultats obtenus pour les types d'émotions associées aux paramètres acoustiques (voir Tableau 2.2) [28].

Dans [33], les auteurs montrent aussi qu'il existe des paramètres vocaux qui sont influencés par les

1. Le débit de la voix (appelé également tempo/rythme) est la vitesse à laquelle on parle, le nombre de mots par minute. Il peut donc être lent, normal ou rapide.
3. l'articulation désigne l'aisance à parler clairement. Elle exprime la manière dont nous parlons pour nous faire comprendre.

	Colère	Joie	Tristesse	Peur	Dégoût
Débit [1]	légèrement plus rapide	plus rapide ou plus lent	légèrement plus lent	beaucoup plus rapide	vraiment beaucoup plus lent
Moyenne de F0	vraiment beaucoup plus haut	beaucoup plus haut	légèrement plus bas	vraiment beaucoup plus haut	vraiment beaucoup plus bas
Amplitude de F0	beaucoup plus large	beaucoup plus large	légèrement plus restreinte	beaucoup plus large	légèrement plus restreinte
Intensité	plus forte	plus forte	plus faible	normale	plus faible
Variations de F0	abruptes, sur les syllabes accentuées	régulières, inflexions montantes	inflexions descendantes	normales	grandes, inflexions terminales descendantes
Articulation [3]	tendue	normale	relâchée	précise	normale

TABLE 2.1: Résumé des effets des émotions sur la parole traduit de l'anglais. [32]

Paramètres acoustiques	Direction de l'effet	Émotions
Variation de l'intensité	Faible Forte	Bonheur, plaisir, activité Peur
Variation de F0	Faible Forte	Dégoût, colère, peur ennui Bonheur, plaisir, activité, surprise
Niveau de F0	Bas Elevé	Ennui, plaisir, tristesse Surprise, puissance, colère, peur, activité
Débit	Lent Rapide	Tristesse, ennui, dégoût Activité, surprise, bonheur, plaisir, puissance, peur, colère
Enveloppe	Ronde Aiguë	Dégoût, tristesse, peur, ennui, puissance Plaisir, bonheur, surprise, activité

TABLE 2.2: Types d'émotions associées aux paramètres acoustiques. [28]

émotions. Ces paramètres dépendent largement de facteurs physiologiques tels que la fréquence cardiaque, la tension musculaire et le flux sanguin, qui varient eux-mêmes avec l'état émotionnel d'une personne. La prosodie (fréquence, contour sonore et rythme) et la qualité de la vocalisation non verbale (timbre et changements spectraux) doivent transmettre l'état émotionnel d'une personne. Ces indices sont la fréquence fondamentale, le timbre, le niveau de pression sonore, le rythme (durée des segments et des pauses). D'après certains de ces paramètres, la tristesse serait, par exemple, traduite par une hauteur plus grave, une variation faible du niveau de pression sonore et un discours lent, alors que la colère et le bonheur seraient exprimés par une variation forte du niveau de pression sonore. Aussi, une augmentation de la fréquence fondamentale traduit une émotion avec un fort arousal, c'est-à-dire une forte intensité émotionnelle, telle que la colère.

Au contraire, une diminution de la fréquence fondamentale peut représenter une émotion telle que la tristesse.

En 1997, Picard [27] montre que les modifications corporelles/physiologiques, qui accompagnent certains états émotionnels, influencent fortement le mode de production du message oral du locuteur . Par exemple, dans le cas de la peur, les modifications physiologiques typiques sont l'augmentation du pouls, la pression du sang et la sécheresse de la bouche. Elle se manifeste par un débit plus rapide, une voix plus forte et plus aigüe, au contraire de l'ennui et de la tristesse qui sont corrélés avec un abaissement du rythme cardiaque et se manifestent par un débit plus lent, une voix plus grave et moins intense.

Enfin, nous pouvons conclure que l'état émotionnel induit la modification de la voix [26] par le biais du cerveau qui contrôle l'appareil de production de la parole (contrôle la respiration et contracte plus ou moins les cordes vocales) [28]. Ces modifications dépendent de chaque locuteur. Pour cela, nous constatons que chaque individu a sa propre voix qui n'est pas toujours la même : elle varie suivant la situation et le contexte.

Dans la section suivante, nous entrerons dans les détails de ce qu'on appelle descripteurs acoustiques utilisés dans le domaine de traitement du signal pour étudier leur intérêt dans la reconnaissance automatique des émotions.

2.3 Extraction de caractéristiques

Le signal audio contient un grand nombre de paramètres reflétant les caractéristiques émotionnelles. L'un des principaux points de la reconnaissance des émotions est de savoir quelles caractéristiques (ou descripteurs) doivent être utilisées. L'extraction de caractéristiques (ou feature extraction en anglais) consiste à extraire un ensemble de vecteurs représentatifs du signal vocal. Il y a deux grands types d'informations dans le signal vocal. La partie sémantique du discours contient des informations linguistiques dans la mesure où les énoncés sont faits selon les règles de prononciation de la langue. L'information paralinguistique, par contre, fait référence aux messages implicites tels que l'état émotionnel du locuteur. Pour la reconnaissance des émotions vocales, l'identification des caractéristiques paralinguistiques représentant l'état émotionnel du locuteur est une première étape importante.

Plusieurs études [34], [35], [34], [35] ont montré qu'un être humain est capable de comprendre les

émotions transmises par des énoncés dans une langue étrangère que celles dans sa langue natale. Il peut être conclu de ces études que l'information paralinguistique est plus déterminante ou suffisamment pertinente dans la reconnaissance de l'état émotionnel d'un être humain à partir de la voix. C'est pour cette raison que nous nous intéresserons, dans ce qui suit, particulièrement à la description de l'information paralinguistique et plus spécifiquement à l'information spectrale et à l'information prosodique. Ces deux groupes de caractéristiques (prosodique et spectral) sont les caractéristiques acoustiques les plus largement utilisés pour la reconnaissance des émotions [36].

2.3.1 Descripteurs prosodiques

D'après [36], la prosodie est, définie comme la phonologie suprasegmentale des sons de la parole, c'est à dire que les domaines d'interprétation des traits prosodiques sont au-delà de la limite de l'unité du phone [34]. Le locuteur peut donner à l'énoncé un ton d'une question ou de déclaration à travers la prosodie [34]. Les caractéristiques prosodiques modélisant l'accent, le rythme, l'intonation, la mélodie de la voix [37], sont appropriés pour la modélisation de l'émotion du locuteur [38]. Chacun de ces phénomènes prosodiques se manifeste par des variations au niveau de la fréquence fondamentale (F0), de l'intensité et/ou de débit de parole [13]. F0 est étroitement lié au comportement glottal et vocal et a été reconnu comme un indicateur d'émotion efficace [36]. L'intensité de la parole est mesurée en tant que puissance de parole à court terme. Elle a un impact sur le volume perceptuel de la parole (faible ou fort) et indique le niveau d'émotion de l'excitation. Le débit de parole est généralement calculé de manière quantitative en syllabes en mots ou en segments par unité de temps et mesure la rapidité avec laquelle la parole est produite [34].

2.3.2 Descripteurs spectraux

Une grande partie de descripteurs repose sur l'analyse spectrale de la voix, parmi lesquels se trouvent les descripteurs spectraux et les descripteurs dites cepstraux.

Les descripteurs spectraux (les formants, énergie en bande de Bark, centroïde spectral, LPC - *Linear Prediction Coding*...) décrivant les cordes vocales et le comportement des pistes vocales [37]. Ils cherchent à représenter le timbre d'un signal, c'est-à-dire la répartition de l'énergie en fonction des fréquences [13]. Les formants, appelés aussi harmoniques, sont les résonances de la fréquence fondamentale produites par le conduit vocal. Ils sont variables en fonction de

la position des différents articulateurs (lèvres, langue, ...). Les descripteurs formantiques sont largement utilisés pour la reconnaissance de l'émotion [38], [39]. L'énergie par bandes de Bark [13] est basée sur une échelle perceptive qui est l'échelle de Tonie[1] [38]. Le spectre du signal est découpé en différentes bandes de fréquences déterminées par cette échelle. L'énergie contenue dans ces bandes de fréquences particulières est un descripteur qui a été utilisé pour l'étude des manifestations acoustiques des émotions (colère, peur, joie, ennui et tristesse) dans [41] et des émotions (rire et acclamation) dans [42]. Les coefficients LPC (analyse par prédiction linéaire ou Linear Prediction Coding en anglais) permettent d'obtenir les formants ainsi de transmettre les informations d'enveloppe spectrale relatives au conduit vocal [13]. Le centroïde spectral est une mesure utilisée dans le traitement du signal numérique pour caractériser un spectre. Il indique où se trouve le "centre de masse" du spectre. C'est un bon indicateur de brillance "brightness" du son, il est largement utilisé dans le traitement de l'audio et de la musique comme une mesure du "timbre" de la voix [38].

Une famille de descripteurs utilisés pour la reconnaissance acoustique des émotions se base sur les descripteurs cepstraux (MFCC - Mel Frequency Cepstral coefficients , LPCC - *Linear Prediction Cepstral Coefficients*, PLP - *Perceptual Linear Prediction*...). Ces descripteurs sont dérivés du spectre du signal [13]. Les MFCC (ou Mel-Frequency Cepstral Coefficients) sont des coefficients cepstraux obtenus au moyen d'une transformation en cosinus discrète appliquée au logarithme du spectre de puissance d'un signal. Les bandes de fréquence de ce spectre sont espacées logarithmiquement selon l'échelle de Mel qui est une approximation de la résolution fréquentielle du système auditif humain. Ces coefficients sont utilisés de manière standard comme vecteur de traits caractéristiques dans les systèmes de la reconnaissance du locuteur et de la parole [43], [34], [44]. Les coefficients PLP (Perceptual Linear Prediction) correspondent à une amélioration des LPC. Ils sont calculés par une transformée de Fourier inverse appliquée à la racine cubique du spectre de puissance suivie d'une analyse par prédcition linéaire (LPC) [45] [46]. Les coefficients LPCC (ou Linear Prediction Cepstral Coefficients) sont un autre type de coefficients cepstraux utilisés pour la reconnaissance des émotions. Ce sont des coefficients calculés à partir de l'analyse LPC du signal [13].

1. L'échelle de Tonie, ou échelle de hauteur tonale, est une échelle traduisant la perception relative de deux sons de hauteurs différentes : elle permet de distinguer un son grave d'un son aigu. Cette échelle a été proposée en 1937 par Stevens, Volkman et Newman [40].

2.4 Sélection de caractéristiques

Du signal audio sont extraites des caractéristiques exprimées sous forme de vecteurs. Cet ensemble de caractéristiques permet de distinguer une forme appartenant à une classe par rapport aux formes des autres classes. L'utilisation de toutes les caractéristiques disponibles pour la classification peut ne pas offrir les meilleures performances de reconnaissance en raison de l'existence de caractéristiques redondantes ou non pertinentes d'où l'importance de la phase de sélection des caractéristiques. L'objectif principal de la sélection de caractéristiques en machine learning (ML), tel qu'indiqué par Aha et Bankert [47], est de trouver un sous-ensemble de caractéristiques les plus pertinents parmi celles de l'ensemble de départ. Les éléments redondants ou n'apportant pas d'information utile pour que le système prenne une décision sont écartés. La sélection de caractéristiques permet d'optimiser le système de reconnaissance en le simplifiant, en réduisant son temps de calcul et en améliorant ses résultats.

Il existe de nombreuses méthodes consistant à réduire le nombre de caractéristiques. Ces méthodes sont appliquées dans plusieurs domaines tels que la classification, l'apprentissage automatique (machine learning), la modélisation et l'analyse exploratoire de données (Data Mining). Nous nous intéressons à la sélection de caractéristiques pour la classification. Dans cette section, nous décrivons ces méthodes d'une manière générale et particulièrement la méthode Recursive Feature Elimination (RFE), utilisée dans notre travail, sera décrite plus en détails dans le chapitre 5 (section 5.5). Ces techniques de sélection de caractéristiques utilisent différentes stratégies de recherche et se répartissent en trois catégories : les méthodes filtres (filter), les méthodes enveloppes (wrapper) et les méthodes intégrées (embedded).

2.4.1 Stratégies de recherche

La génération des sous-ensembles découle de la stratégie de recherche choisie. Selon Liu et Yu [48], cette dernière peut être exhaustive, heuristique ou aléatoire.

2.4.1.1 Stratégie de recherche exhaustive

Lors d'une recherche exhaustive, aussi appelée recherche complète, tous les sous-ensembles possibles sont analysés afin de trouver le plus optimal. Le nombre de combinaisons évolue exponentiellement en fonction du nombre de caractéristiques. Par conséquent, cette méthode peut s'avérer extrêmement longue au point de rendre ce type de recherche impossible lorsque le

nombre de caractéristiques est trop important. Sa complexité est $O(2^N)$. L'avantage de cette recherche toutefois est qu'elle est extrêmement précise.

2.4.1.2 Stratégie de recherche heuristique

La recherche heuristique, également appelée recherche séquentielle, est plus rapide que la recherche exhaustive car elle ne parcourt pas l'ensemble des caractéristiques pour générer des sous-ensembles. Ce type de recherche peut être divisé en trois sous-types :

— la recherche séquentielle ascendante (sequential forward feature selection ou SFFS) qui part d'un ensemble vide et y ajoute progressivement des caractéristiques ;

— la recherche séquentielle descendante (sequential backward feature elimination ou SBFE) qui correspond à la méthode inverse de la SFFS. L'ensemble de départ est l'ensemble total des caractéristiques auquel on retirera progressivement les caractéristiques les moins pertinentes ;

— la recherche bidirectionnelle (bidirectional selection ou stepwise) qui mélange les méthodes SFFS et SBFE. Le sous-ensemble de départ se voit ajouter et retirer des caractéristiques au fur et à mesure.

La complexité de la recherche heuristique est $O(N^2)$. L'avantage de cette méthode est qu'elle est simple à implémenter et très rapide en temps d'exécution. Son inconvénient est qu'elle est sensible aux changements de l'ensemble de données.

2.4.1.3 Stratégie de recherche aléatoire

Aussi appelée recherche stochastique ou non déterministe, cette stratégie de recherche concerne les cas où l'ensemble de caractéristiques contient des milliers de données, comme dans le cas de traitement du langage naturel par exemple. Dans un cas comme celui-ci, il n'est pas possible d'effectuer une recherche sur l'intégralité de l'espace de caractéristiques. Une recherche stochastique permet de se concentrer sur un échantillon de cet espace, en écartant l'optimalité de la solution au profit de l'efficacité de la recherche. La complexité de cet algorithme s'exprime en $O(Nlog(N))$. Bien que cette méthode puisse manquer de précision, elle permet tout de même de laisser à l'utilisateur le choix entre précision et vitesse afin d'éviter de tomber dans des optimums locaux.

2.4.2 Critères d'évaluation

Les méthodes utlisées pour évaluer un sous-ensemble de caractéristiques dans les algorithmes de sélection peuvent etre classées en trois catégories : la catégorie filtre (Filter), la catégorie enveloppante (Wrapper) et la catégorie intégrée (Embedded) [48].

2.4.2.1 Méthodes filtres (filter)

Les méthodes correspondant au modèle filtre sont indépendantes de l'algorithme d'apprentissage et ne dépendent donc que des données. Ce modèle est réalisée comme un pré-traitement précédant la phase d'apprentissage de l'algorithme, c'est-à-dire que la sélection se fait sans tenir compte de son influence sur les performances du système. Les méthodes de ce modèle utilisent généralement une approche heuristique en guise de stratégie de recherche. En général, les filtres utilisés pour évaluer la pertinence des données comprennent des critères de mesures statistiques, la pertinence maximale et la redondance minimale (Minimum Redundancy Maximum Relevance ou mRMR), ou encore l'algorithme Relief. Ces filtres attribuent un poids aux données qui peuvent alors être classées.

2.4.2.2 Méthodes enveloppes (wrapper)

Les méthodes du modèle "filter" interviennent en amont de la phase d'apprentissage. Elles peuvent donc avoir une influence sur les classifieurs qui seront utilisés lors de cette dernière mais ne mesurent pas cet impact. Par opposition, l'approche Wrapper utilise les performances de l'algorithme d'apprentissage comme critère d'évaluation. Cette évaluation se fait sur un sous-ensemble de caractéristiques par le biais d'un classifieur qui estime sa pertinence. De ce fait, l'algorithme d'apprentissage sera toujours effectué sur des sous-ensembles pertinents.

2.4.2.3 Méthodes intégrées (embedded)

Contrairement aux méthodes filter et wrapper, les méthodes intégrées ne séparent pas la phase d'apprentissage de la sélection de caractéristiques. Par conséquent les classifieurs utilisés par ces méthodes n'ont pas besoin de tout recommencer à chaque nouvel ensemble de caractéristiques. Ces méthodes sont donc beaucoup plus rapides que celles des modèles précédents. Parmi les

méthodes intégrées existant dans la littérature on peut trouver : les algorithmes de type SVM ou encore l'algorithme de Gram-Schmidt.

En conclusion, la recherche de l'ensemble des caractéristiques les plus pertinentes joue un rôle majeur dans le domaine de la reconnaissance des émotions. Plusieurs études ont utilisé les méthodes de sélection de caractéristiques afin de déterminer le sous-ensemble optimal. Dans les travaux [49], Oudeyer a fait la sélection des six meilleurs caractéristiques à partir d'un ensemble de 200 caractéristiques. Vogt [25] est parti de 1280 caractéristiques pour en sélectionner les 90 meilleures. Nous présentons ici la liste de quelques méthodes de sélection de caractéristiques proposées dans la littérature de la reconnaissance des émotions. Ces méthodes sont fondées sur les différentes stratégies de recherche définies précédemment ainsi que différents critères d'évaluation :

- SFS (Sequential Forward Selection) ou (sélection séquentielle croissante) ([50], [36], [51])
- SBS (Sequential Backward Selection) ou (sélection séquentielle arrière) ([52], [53], [50])
- Algorithme RELIEF-F ([34], [4])
- Algorithme génétique ([49], [54], [55])

2.5 Apprentissage automatique

2.5.1 Comprendre l'apprentissage automatique

L'apprentissage automatique (ML pour machine learning en anglais, littéralement "l'apprentissage machine ") est un champ d'étude de l'intelligence artificielle qui est consacrée à l'analyse des données. Le but de ML est de créer de la connaissance (modèle) de manière automatique à partir de données brutes. Ce modèle peut alors être exploité sur de nouvelles données pour prendre des décisions. Comme le modèle est construit à partir des données disponibles, donc plus nous disposons de données, plus le modèle construit est précis et permettra ainsi de prendre de bonne décisions. L'apprentissage automatique comporte généralement deux phases. La première phase dite d'entraînement (ou d'apprentissage) du modèle, consiste à déterminer un modèle à partir des données d'entrée. Selon la taille de ces données dépend la précision du modèle et le temps de génération de ce modèle. La seconde phase correspond à la phase d'évaluation (ou de test) : le modèle étant entrainé, de nouvelles données sont utilisées afin d'obtenir le résultat correspondant à la tâche souhaitée. Les données peuvent être de types différents, tels que des graphes, des

FIGURE 2.3: Principe de l'apprentissage automatique

courbes, des arbres ou des vecteurs d'attributs (ou caractéristiques). Les données sont souvent notées sous forme d'un vecteur de caractéristiques. Les données brutes (les graphes, les courbes, les arbres) sont inexploitables. Donc, il faut procéder à un pré-traitement des données afin d'extraire les caractéristiques (features en anglais) pour la prise de décision. Il existe deux grandes familles de données : les données labélisées et les données non labélisées. Le modèle prend un vecteur de caractéristiques en entrée, et renvoie une décision (voir Figure 2.3). En fonction de la nature des données d'entrée (labélisés ou pas), il existe deux types d'apprentissage : si les données sont labélisés, nous parlons d'apprentissage supervisé. Si les données sont non-labélisés, nous parlons d'apprentissage non supervisé. En fonction de la nature de la sortie (catégorisée ou pas), il existe deux types de décisions. Si les sorties sont catégorisées, nous disons que le modèle effectue une classification, si non nous disons que le modèle effectue une régression.

2.5.2 Types d'apprentissage automatique

Les algorithmes d'apprentissage peuvent se catégoriser selon le mode d'apprentissage qu'ils emploient. Parmi les méthodes d'apprentissage automatique, deux sont les plus largement adoptées :

Apprentissage supervisé

Dans l'apprentissage supervisé, les données d'entrée sont étiquetées avec les sorties souhaitées. L'algorithme puisse « apprendre » à faire des prédictions sur des données non étiquetées. Par exemple, avec un apprentissage supervisé, un algorithme peut être alimenté avec des photos étiquetées de chats et des photos étiquetées de chiens. En étant formé sur ces données, l'algorithme devrait être capable d'identifier plus tard des photos non marquées de chats comme chat et les photos de chiens non étiquetées comme chien. Dans ce cas, nous parlons de classification puisque nous affectons une catégorie « chat» ou «chien » à une image. Dans le cas où nous prédisons une valeur, par exemple un algorithme qui prend en entrée les caractéristiques d'un véhicule, et essayera de prédire le prix du véhicule, l'apprentissage supervisé est appelé régression.

Apprentissage non supervisé

Dans l'apprentissage non supervisé, les données d'entrée ne sont pas étiquetées. En effet, l'objectif de l'algorithme d'apprentissage est de trouver tout seule la similarité (les points communs) parmi ses données d'entrée. Par exemple, les photos non étiquetées de chats et les photos non marquées de chiens peuvent être utilisées comme données d'entrée pour l'algorithme afin de trouver des similitudes et de regrouper les photos de chats ensemble et les photos de chiens ensemble. Dans ce cas, nous parlons de regroupement (ou clustering en anglais).

2.5.3 Algorithmes existants pour la classification

Pour la RAE, sans être exhaustif, les méthodes de classification utilisées sont la régression linéaire (RL), les machines à vecteurs de support (SVM), les K plus proches voisins (KNN), les modèles de Markov cachés (HMM), les modèles de mélange gaussien (GMM), les réseaux de neurones, les réseaux de neurones récurrents (RNN). Dans ce qui suit, nous décrivons brièvement chacun de ces algorithmes.

Régression Linéaire (RL)

La régression linéaire [56] entre dans la catégorie de la classification de sous-espace la plus proche. Pour chaque classe, on cherche une fonction de décision linéaire. Il s'agit alors de trouver une frontière de décision linéaire par minimisation des moindres carrés. Elle est très efficace, simple et rapide, mais peut être trop simpliste pour certains problèmes .

Machines à vecteurs de support (SVM)

SVM [57] [58] est un algorithme de classification linéaire qui sépare les classes de l'espace en utilisant un hyperplan optimal, c'est à dire qui maximise la marge entre les échantillons et l'hyperplan séparateur. SVM est un classifieur simple et optimal en apprentissage automatique. L'avantage de SVM par rapport aux autres classifieurs est que, pour des données d'apprentissage limitées, il montre une meilleure performance de classification [59].

K plus proches voisins (KNN)

Le KNN [60] est considéré parmi les algorithmes d'apprentissage automatique les plus basiques. Elle ne nécessite pas d'apprentissage mais simplement le stockage des données d'apprentissage. Cette méthode consiste à associer à l'élément à classer, la classe la plus représentée parmi ses k plus proches voisins selon une distance à définir. L'inconvénient majeur de la méthode des plus proches voisins est qu'elle peut être lourde pour des grandes bases de données, puisqu'elle implique le calcul des distances entre l'élément à classer et chacun des éléments de l'ensemble.

Modèles de Markov cachés (HMM)

HMM [61] est un modèle statistique qui permet d'analyses les séquences de vecteurs. Pour les tâches de classification, un modèle est estimé pour chaque type de signal. Ainsi, il faudrait prendre en compte autant de modèles que de types de signaux à reconnaître. Lors de la classification, un signal est pris et la probabilité pour chaque signal donné par le modèle est calculée. La sortie du classifieur est basée sur la probabilité maximale que le modèle ait généré ce signal. Un modèle de Markov caché comporte deux éléments de base : un processus de Markov et un ensemble de distributions de probabilité de sortie.

Modèles de mélange gaussien (GMM)

GMM [62] est un modèle probabiliste qui fait une analyse discriminante sur les densité de probabilité des descripteurs. Il suppose que tous les données sont générés à partir d'une somme de plusieurs gaussiennes. Il s'agit après de déterminer la moyenne, la variance et l'amplitude de chaque gaussienne. Ces paramètres sont estimées par maximisation de la vraisemblance pour approcher le plus possible la distribution recherchée.

Réseaux de neurones

Les réseaux de neurones [63], dont les méthodes d'apprentissage profond ("deep learning" en anglais), sont des modèles inspirés du fonctionnement du cerveau humain. Ils sont composés d'un ensemble de neurones connectés entre eux qui permettent de propager les signaux d'un neurone à un autre. Les réseaux de neurones se reposent sur des modèles intrinsèques de systèmes de neurones biologiques du traitement de l'information, qui ont conduit au développement de

systèmes informatiques plus intelligents, applicables dans de nombreux problématiques statistiques et d'analyse des données. Le succès croissant des réseaux de neurones par rapport aux autres algorithmes de classification peut s'attribuer à leur puissance, leur capacité d'apprentissage, leur polyvalence et à leur simplicité d'utilisation. De ce fait, ils sont largement utilisés dans de nombreux problèmes de classification (reconnaissance de formes, ciblage marketing, traitement de signal,. . .).

En fait, il n'y a pas de réponse définitive sur le choix de l'algorithme d'apprentissage, car chaque technique a ses avantages et ses limites, mais aucune ne peut fournir la meilleure performance de reconnaissance dans toutes les situations. Pour cette raison, nous avons choisi dans ce travail de comparer les performances des classifieurs les plus souvent utilisés, dans la communauté scientifique, pour la détection des émotions (RL, SVM et RNN). Par conséquent, le critère de sélection devrait être basé sur la tâche concrète. À titre d'exemple, SVM est une technique puissante capable d'avoir une très bonne performance de classification par rapport à d'autres classifieurs (RNN), en particulier pour des données d'entraînement limitées.

De plus, la nature de données à modéliser influence également la sélection de l'algorithme pour le système de RAE. Par exemple, le RNN [64] est un choix courant si nous traitons des séries temporelles et des données séquentielles. Dans la présente étude, les algorithmes RL, SVM et RNN sont utilisés pour construire des systèmes de reconnaissance des émotions, car ils ont été démontré qu'ils donnaient des bonnes performances dans de nombreux travaux sur les systèmes RAE [65],[59], [66]. Dans le chapitre 5 (section 5.2), nous explorons plus en détail ces algorithmes.

2.6 Conclusion

Le signal vocal est l'un des signaux les plus complexes à caractériser. Il présente une grande variabilité ce qui rend difficile la tâche des systèmes de RAE. Cette complexité est due à la présence de plusieurs facteurs, comme les conditions d'enregistrement. Les problématiques qui se posent pour toutes les études sur la reconnaissance automatique des émotions à partir de la voix : Quel corpus des données émotionnelles ? Quelles caractéristiques extraire ? Quelle méthode pour choisir les meilleures caractéristiques ? Quels sont les algorithmes les plus appropriés pour la classification ?
Le Tableau 2.3 donne une liste non exhaustive des études rapportant les expériences de reconnaissance automatique des émotions. Le taux de reconnaissance de chaque étude est présenté

avec le type de corpus, les émotions, les caractéristiques et les classifieurs utilisés. Après un état de l'art sur les différentes étapes nécessaires à la mise en œuvre d'un système de RAE, nous décrivons dans le chapitre 3 notre corpus en langue française ainsi les deux corpus employés dans nos travaux.

Référence Auteur	Style de corpus	Émotions	Type de descripteurs	classifieur	Taux de détection
He L and al.(2011) [67]	naturel	neutral, angry, anxious, dysphoric and happy	spectrogram features	GMM,KNN	53%,53%
Sheikhan M and al.(2012) [68]	naturel	anger, happiness and neutral	MFCC, formant and pitch	GMM, C5.0, MLP, MODULAR NEURAL-SVM	65.9%, 56.3%, 68.3%, 76%
Hamidi M, Mansorizadeh M (2012) [69]	naturel	anger, disgust, fear, sadness and happiness	energy, rate, pitch and MFCC	NN	78%
Monorama Sand al. (2015) [70]	acté	anger, happiness, disgust, fear, sadness, surprise, and neutral	MFCC	SVM	82.14 %
Sih-Huei C and al. (2016) [71]	acté	anger, boredom, disgust, fear, joy, neutral and sadness.	prosodic features and some general features	multiple kernel Gaussian process (GP)	77.74%
Weishan Z and al. (2017) [58]	acté	anger, fear, joy, neutral, sadness, and surprise	Pitch frequency, short-term energy, short-term zero-crossing rate, formant, and MFCC	SVM, DBN	84.54%, 94.6%
Pravena and Govind (2017) [72]	acté	neutral, anger, happy and sad	Instantaneous F0 and strength of excitation parameters	GMM	89.20%
Zhen-Tao L and al.(2018) [73]	acté	neutral, angry, surprise, happy, fear and sadness	Prosodic features, quality characteristics, spectrum characteristics	ELM decision tree, SVM decision tree, BPNN, KNN	89.6%, 87.2%, 82.3%, 80.7%

Table 2.3: Tableau récapitulatif des études sur la reconnaissance automatiques des émotions : référence de l'auteur, type de corpus (acté, induit, naturel), les émotions, type de descripteurs (Spectraux, Prosodique (Fréquence Fondamentale, Energie, Débit)...), modèle d'apprentissage (GMM : ≪ Gaussian Mixture Model≫ , KR : ≪ Kernel Regression ≫, kNN : k Nearest-Neighbors, SVM : Support Vector Machine, HMM : Hidden Markov Model, NN : Neural Networks, decision trees, Adaboost, DBN : deep belief networks, ELM : extreme learning machine, etc), et finalement le taux de reconnaissance. (Nous avons laissé les émotions et les descripteurs dans la langue d'origine pour respecter au mieux le sens des auteurs

Deuxième partie

Système de reconnaissance des émotions à partir de la voix

Chapitre 3

Les corpus émotionnels, base de données et collecte d'informations

Contents

3.1 Introduction **41**

3.2 Corpus utilisés dans le cadre de ces travaux **42**

3.2.1 Berlin database 42

3.2.2 Spanish database 43

3.3 Constitution de notre corpus émotionnel : le corpus "French Student Emotionnal database" : **44**

3.3.1 Construction 45

3.3.2 Test 48

3.3.3 Validation 50

3.4 Conclusion **58**

3.1 Introduction

Les performances de tout système de reconnaissance automatique des émotions (tout au long de ce manuscrit nous abrégeons par RAE), reposent sur la qualité des corpus émotionnels utilisés. Ces corpus permettent à la fois l'entraînement, le test et la validation d'un tel système. Dans la littérature, pour les deux modèles descriptifs des émotions, discret et continu (voir section 1.4), des corpus sont disponibles [74]. Au niveau de la construction de ces corpus, nous distinguons trois types : corpus d'émotions actées, corpus d'émotions induites et corpus d'émotions naturelles. Bien qu'il existe un grand nombre de corpus, peu d'entre eux sont gratuits et accessibles à

la communauté de recherche. Il faut noter également qu'il n'existe ni standards de corpus, ni tests normalisés pour la RAE [36], ce qui rend difficile la comparaison des performances. Pour valider notre système, nous avons fait appel dans un premier temps à un corpus largement utilisé dans la littérature, et qui peut être considéré comme une référence pour mieux comparer les résultats obtenus. Nous avons également utilisé un autre corpus, moins répandu dans la littérature de la RAE, mais qui offre un grand nombre d'enregistrements. Enfin, nous avons également créé notre propre corpus en français. Ceci a été motivé par le besoin d'avoir un corpus en français et dans le contexte applicatif initial de cette thèse, qui est un contexte pédagogique.

Dans ce chapitre, nous décrivons tout d'abord les corpus que nous avons utilisés pour valider notre approche. Puis pour présenter notre corpus, nous détaillerons la méthode de construction, les résultats de test pour la validation de ce corpus et l'analyse statistique de ces résultats.

3.2 Corpus utilisés dans le cadre de ces travaux

L'étude [12] donne une liste détaillée des corpus de parole émotionnelle utilisés dans différents travaux de recherche. L'analyse de cette liste de corpus (principalement en allemand ou en anglais) indique que les émotions sont principalement décrites par une approche catégorielle, et que les émotions simulées prévalent davantage par rapport aux émotions spontanées ou induites. Ce qui explique le fait que la plupart des travaux sur la reconnaissance des émotions ont été réalisés sur des données «simulées» enregistrées par des acteurs avec un nombre bien défini de classes émotionnelles.

Une étude plus récente [24] sur les corpus émotionnels existants est menée en 2018. Dans cette thèse, deux bases de données en libre accés pour les chercheurs, la base de données Berlin [75] et la base de données Espagnole [76], sont utilisées pour évaluer les performances de notre système. Les résultats obtenus en utilisant ces BD seront exposés dans le 5ème chapitre .

3.2.1 Berlin database

Berlin database [75] est un corpus d'émotions acté en langue allemande. Ce corpus constitue une référence et a été utilisé dans plusieurs travaux de recherche sur la reconnaissance des émotions à partir de la voix. Ce corpus est public . L'utilisation de ce corpus permet donc la comparaison des performances avec d'autres études. 10 acteurs (5 hommes et 5 femmes) ont prononcé chacun 10 phrases de la vie quotidienne en allemand (5 courtes et 5 longues, d'une durée moyenne comprise

entre 1,5 et 4 secondes). Les phrases sont choisies pour être sémantiquement neutres et peuvent donc être facilement interprétées dans les 7 émotions simulées. La parole est enregistrée avec une précision de 16 bits et un taux d'échantillonnage de 48 kHz (sous-échantillonné ultérieurement à 16 kHz) dans une chambre anéchoïque. La base de données brute contient environ 800 phrases (7 émotions $\times$ 10 phrases $\times$ 10 acteurs + quelques secondes versions). Les fichiers de parole sont ensuite évalués par un test de perception subjective avec 20 auditeurs afin de garantir la reconnaissabilité et la naturalité des émotions. Seules les expressions ayant un taux de reconnaissance des émotions supérieur à 80% et considérées comme naturelles par plus de 60% des auditeurs sont retenues. Le nombre final d'énoncés pour les sept catégories d'émotions de la base de données Berlin est le suivant : colère (127), ennui (81), dégoût (46), peur (69), joie (71), neutre (79) et tristesse (62), c'est à dire 535 fichiers en total. Cette base de données a été choisie pour les raisons suivantes : i) la qualité de son enregistrement est très bien ii) elle est populaire pour la reconnaissance des émotions et recommandée dans la littérature [77].

3.2.2 Spanish database

Le corpus émotionnel acté en espagnol appelé INTER1SP [76], auquel nous avons le droit d'accès (gratuit pour un usage universitaire et pour la recherche), a été crée dans le cadre du projet Interface financé par l'Union européenne. Elle contient 6041 enregistrements de locuteurs espagnols professionnels, un homme et une femme, dans une salle à bruit réduit. Les enregistrements ont été réalisés à l'aide d'un microphone électrodynamique AKG 320. Six émotions de base ou principales plus le neutre (colère, ennui, dégoût, peur, bonheur, tristesse et neutre/normal) ont été enregistrées, ainsi que quatre variations supplémentaires du neutre (neutre/doux, neutre/fort, neutre/lent et neutre/rapide). La base de données contient deux sessions d'enregistrement de discours émotionnels. Le matériel textuel se compose de 184 éléments incluant des phrases phonétiquement équilibrées, des chiffres, des mots isolés, ainsi que des paragraphes, pour un total de 3h 59min de parole enregistrée pour le locuteur homme et 3h 53min pour la locutrice. Les fichiers sont stockés sous forme de signaux (extension de fichier .l16). Les signaux vocaux ont d'abord été enregistrés à 32 kHz, puis ré-échantillonnés à 16 kHz. Cette base de données est intéressante dans notre cas, car elle contient plus de données. Bien que INTER1SP propose des échantillons représentant onze émotions, nous n'avons retenus que les sept émotions principales afin de faire la comparaison avec la base de données Berlin détaillée ci-dessus. Les autres sont uniquement des variétés du neutre.

3.3 Constitution de notre corpus émotionnel : le corpus "French Student Emotionnal database" :

Ce travail sur la RAE vise une application dans un contexte pédagogique. Les émotions jouent un rôle important dans les tâches d'apprentissage [78], [79]. La connaissance de l'état émotionnel d'un apprenant, s'avère être un élément clé, dans le choix de la méthode pédagogique la plus pertinente, pour aider l'apprenant à accomplir sa tâche d'apprentissage dans des meilleures conditions [80]. L'objectif à moyen terme est d'utiliser notre système pour détecter l'état émotionnel d'un apprenant afin de faire remonter à l'enseignant des indicateurs utiles à l'animation de sa classe. Pour être conforme à cet objectif, nous avons construit un corpus, lié à un contexte pédagogique, avec les conditions suivantes :

- un nombre suffisant de locuteurs (au moins vingt),
- les locuteurs sont des étudiants de genres différents (femme et homme),
- des paroles neutres et émotionnelles en lien avec le contexte pédagogique,
- un environnement acoustique bien défini,
- des conditions d'enregistrement (matériel) similaires pour tous les enregistrements,
- possibilité d'extension et d'enregistrement libre plus tard (en utilisant une application mobile déjà réalisée).

La figure 3.1 illustre les étapes de constitution de notre corpus émotionnel (construction, test et validation). La première étape fut l'écriture des phrases qui sont prononcées par les volontaires. Pour aider ces derniers à produire l'émotion désirée, ou plutôt les stimuler à acter naturellement celle-ci, juste avant l'enregistrement ils sont invités à lire des scénarios spécialement écrits pour leur permettre de se plonger dans l'émotion à acter. La deuxième étape était le choix du matériel et du méthode d'enregistrement. La troisième étape est l'étude et le choix de la salle d'enregistrement. L'étape qui clot la construction est l'invitation d'un échantillon de volontaires (désignés dans la suite par locuteurs[1]) choisis et la réalisation des enregistrements. Les dernières étapes sont dédiées au test et à la validation de ce corpus par un autre échantillon différents de volontaires (désignés dans la suite par testeurs[2]) invités à écouter les enregistrements et identifier les émotions actées. Nous détaillerons dans la suite ces différentes étapes.

1. Un locuteur est une personne qui parle, qui énonce une phrase avec une émotion précise.
2. Un testeur est une personne qui écoute des enregistrements différents et les classer selon l'émotion reconnue.

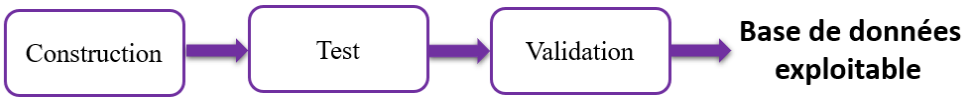


FIGURE 3.1: Les étapes nécessaires pour la constitution d'un corpus émotionnel.

3.3.1 Construction

3.3.1.1 Élaboration des énoncés et scénarios

Notre objectif a été d'écrire au moins dix phrases par émotion dans un contexte pédagogique universitaire. Pour ce faire, nous avons organisé des "focus group", c'est à dire des réunions de travail avec des étudiants (le public cible de l'application de nos corpus) pour déterminer le vocabulaire et les phrases souvent utilisées dans leurs vies d'étudiant. En comparaison avec des poèmes et des phrases qui ne rélèvent pas de la vie quotidienne , l'utilisation de la communication quotidienne est la meilleure solution [75], car il s'agit de la forme naturelle de la parole sous une excitation émotionnelle. De plus, les étudiants peuvent immédiatement les méméoriser pour mieux les acter. Lors de la construction de la base de données, la priorité a été donnée au caractère naturel (du matériel de langage). Au total, 79 phrases ont été construites de manière à pouvoir être prononcées avec les émotions ciblées. De plus, ce sont des énoncés qui, tant par le choix de leurs mots que par leur construction syntaxique, peuvent être utilisés au quotidien. Pour mettre les personnes dans un état proche de l'émotion à interpréter lors de la phase d'enregistrement, nous avons imaginé des scénarios. Nous avons travaillé d'une manière empirique et itérative pour construire le dictionnaire des phrases et l'ensemble des scénarios permettant d'acter ce corpus. Le tableau 3.1 montre quelques exemples de phrases et scénarios établis pour les différentes émotions. La liste complète des phrases et des textes introductifs de mise en situation se trouvent dans Annexe A.

3.3.1.2 Salle et matériel d'enregistrement

Les enregistrements ont eu lieu dans une salle légèrement isolée, avec un faible bruit de fond, et qui comporte des sièges en mousse pouvant absorber certains bruits non désirés. L'étude de cette salle d'enregistrement est détaillée en Annexe A (section A.3). Le corpus a été enregistré en utilisant le logiciel d'enregistrement de son numérique "Audacity" sur deux canaux. Les enregistrements ont été pris avec une fréquence d'échantillonnage de 44 kHz, puis réduits à 16 kHz. Les deux micros utilisés sont de type Rode NT1-A Complete Vocal Bundle monté sur un

Émotion	Scénarios	Exemple de phrase
Tristesse	Toute la semaine, tu as un cours du 8h à 18h. La fatigue se fait ressentir de plus en plus pour toute la classe. Et tu te rends compte que demain matin encore tu ne pourra pas dormir.	Demain, on a encore un cours à 8h.
Colère	Quand l'enseignant prend beaucoup de temps pour corriger les examens et ça commence à t'énerver.	Ça fait trois mois qu'on a passé le contrôle et on n'a pas encore reçu les notes.
Dégoût	Tu arrives à ta place et tu trouves des mouchoirs usagés sous ta chaise. Tu sais que tu dois les retirer toi-même sinon personne d'autre ne le fera ! Tu prends ton courage à deux-mains et tu le fais.	Il faut que j'aille me laver les mains après !
Peur	Tu as eu une bonne note au dernier examen mais ton semestre est mal parti. Tu es vraiment inquiète parce que tu ne sais pas si tu vas le valider.	J'espère que ça va compenser.
Surprise	Depuis plusieurs cours, tu as des difficultés à comprendre les explications de l'enseignant en particulier, mais aujourd'hui c'est bien différent.	Aujourd'hui, le cours s'est vraiment bien passé !
Joie	Tu es en cours, mais l'enseignant n'est encore arrivé. Tu pars voir l'administration pour avoir des informations. Et là, tu reviens en criant.	L'enseignant est malade, donc le cours est annulé !
Neutre	Pas de scénarios.	C'est quoi le prochain cours ?

Table 3.1: Exemples des phrases et scénarios établis en fonction des différentes émotions.

pied de micros et deux filtres anti-réflexions de type LD Systems RF1 permettant d'effectuer des enregistrements.

Les émotions sont actées par 32 locuteurs (8 femmes et 24 hommes), dont l'âge varie entre 22 et 27 ans. 7 catégories d'émotion : tristesse, colère, dégoût, peur, surprise, joie et neutre sont actées. Cette répartition (homme/femme) est représentative de la population estudiantine de l'école d'ingénieurs partenaire de ce projet. En effet dans [81], les statistiques montrent que les femmes sont minoritaires en classes préparatoires aux grandes écoles, en IUT et surtout en écoles d'ingénieurs, où elles représentent environ un quart des effectifs (28% en 2017 par rapport aux hommes 72%). L'enregistrement des énoncés de notre corpus est réalisé, en demandant à chacun des étudiants de prononcer une phrase, en simulant un des états émotionnels.

3.3.1.3 Méthode d'enregistrement

Pendant une séance d'enregistrement, le participant tient devant les deux microphones pour pouvoir utiliser le langage du corps à leur guise, uniquement gêné par la nécessité de parler en direction du microphone à une distance d'environ 10 cm (voir Figure 3.2). Il doit ensuite s'imaginer dans une situation qui lui est proposée (mauvaise note, arrivé à l'école en forme,...). Ces situations sont situées dans le contexte de classe (voir Annexe A, section A.1). Elles visent à induire chez le participant des émotions comme la tristesse, la colère, le dégoût, la peur ... Une seule session d'enregistrement a eu lieu avec chaque participant sous la supervision de quatre personnes, deux d'entre eux donnant des instructions et un retour d'information, deux surveillant l'enregistrement. Chaque session durait environ dix minutes. Le texte de chaque énoncé a été donné à l'avance aux locuteurs pour le préparer.

FIGURE 3.2: Photo de l'installation de locuteur lors de la collecte de notre corpus

3.3.1.4 Résultats

Les tableaux 3.2 et 3.3 présentent le nombre de locuteurs, le nombre d'enregistrements par émotion et par genre, et le nombre total d'enregistrements contenus dans le corpus émotionnel. 32 locuteurs et locutrices ont enregistré leurs voix pour avoir en total 502 enregistrements. Un tableau plus détaillé, comprenant les informations sur les locuteurs (genre, culture, nombre d'enregistrements par émotion...), est annexé à la présente étude (voir Annexe A, Tableau A.2).

Genre	Nombre de locuteurs	Nombre d'enregistrements
Homme	24	363
Femme	8	139
Nombre total	32	502

Table 3.2: Nombre de locuteurs et d'enregistrements.

Émotion	Tristesse	Colère	Dégoût	Peur	Surprise	Joie	Neutre
Nombre d'enregistrements par des hommes	52	53	51	52	52	52	51
Nombre d'enregistrements par des femmes	20	20	20	19	20	20	20
Nombre d'enregistrements total	72	73	71	71	72	72	71

Table 3.3: Nombre d'enregistrements par émotion et par genre.

3.3.2 Test

La perception des émotions est la capacité de l'être humain à identifier et reconnaitre les émotions d'autrui. Dans [22], Scherer a étudié les performances humaines pour la reconnaissance des émotions à partir de la voix et de l'expression faciale. Il a montré que, lors des études menées dans des pays occidentaux et non occidentaux, la reconnaissance de six émotions dans la voix atteint une précision de 55% à 65%. L'émotion colère est la plus reconnue avec un taux d'environ 80%, puis la tristesse autour de 70%, la peur et la joie autour de 60%. Enfin, l'émotion le moins reconnue est le dégoût avec un score de 31%. De même, dans l'étude [82], les auteurs ont effectué des tests d'évaluation humaine avec 4 auditeurs anglophones non formés pour distinguer 4 émotions jouées par des acteurs professionnels : Colère, Joie, Tristesse et Neutre. Ils ont montré que 68.3% des énoncés ont été correctement identifiés et que la Joie est l'émotion la plus difficile à distinguer des autres émotions. Il y a principalement des confusions entre les émotions Neutre/Tristesse et Colère/Joie.

Les résultats de ces études mentionnées ci-dessous, nous servirons de base pour valider les enregistrements de notre corpus. Pour y parvenir dessus et afin de déterminer le taux de reconnaissance de chaque émotion, chaque enregistrement a été évalué par des testeurs afin de déterminer l'émotion associée. Dans ce qui suit nous présentons nos méthodes de test et de validation.

3.3.2.1 Méthode de test

Afin d'assurer la qualité de nos enregistrements, nous avons fait une première évaluation[3] de la base de données (BD). 5 personnes ont pris part à un test de reconnaissance des émotions. 18.4% de la BD a été testée, cela revient donc à un total de 93 enregistrements. Ces personnes n'ont ni pris part aux enregistrements, ni participé à l'élaboration des énoncés. Elles ont pu écouter chacun des enregistrements puis les ont classé par rapport à l'émotion perçue. La Figure 3.3 présente les taux de reconnaissance par émotion. Les résultats de reconnaissance sur un

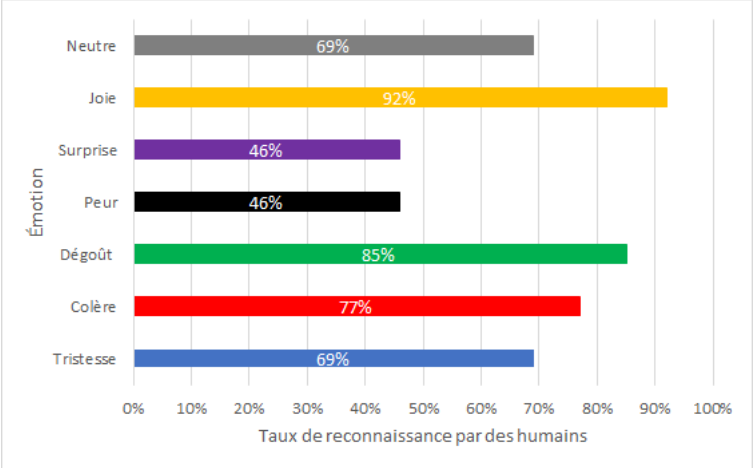


FIGURE 3.3: Taux de reconnaissance par émotion : résultat sur un échantillon

échantillon montrent que les émotions colère, dégoût et joie sont les émotions les plus reconnues. Les émotions les moins reconnues sont la surprise et la peur avec un score de 46%.

Une évaluation plus exhaustive a été réalisée auprès de 106 personnes (hommes et femmes) en utilisant une interface Web. Pour ce faire, nous avons tout d'abord créé des ensembles à faire écouter aux testeurs. Ces ensembles sont conçus en tenant compte des phrases, des locuteurs et des émotions actées. Ainsi chaque ensemble est construit de telle manière qu'il contient :

— 2 enregistrements de chaque émotion ;

— 14 phrases différentes ;

— 14 locuteurs différents dont 4 à 6 locutrices.

Pendant une séance de test, le testeur s'enregistre en précisant son âge, son genre, sa culture et sa langue maternelle, puis écoute l'enregistrement et choisi l'émotion que le locuteur voulait exprimer. Chaque testeur a écouté 14 enregistrements, pour une durée d'à peu prés 10 minutes.

3. La construction de cette base de données a été faite dans le cadre d'un projet des étudiants de 4ème année de l'ENSIM. Ce test partiel a été fait dans l'optique de valider leurs projet, nous le citons comme première évaluation de notre corpus.

3.3.2.2 Résultats

La Figure 3.4 montre le taux de reconnaissance par émotion sur la base entière. Par exemple, le neutre est l'émotion la plus reconnue par nos testeurs avec un score de 86%. Par contre, l'émotion tristesse est reconnue seulement à 51% . 70% des énoncés ont été correctement identifiés ce qui est en corrélation avec l'étude précédemment mentionnée (un taux de reconnaissance des émotions de 68,3%).

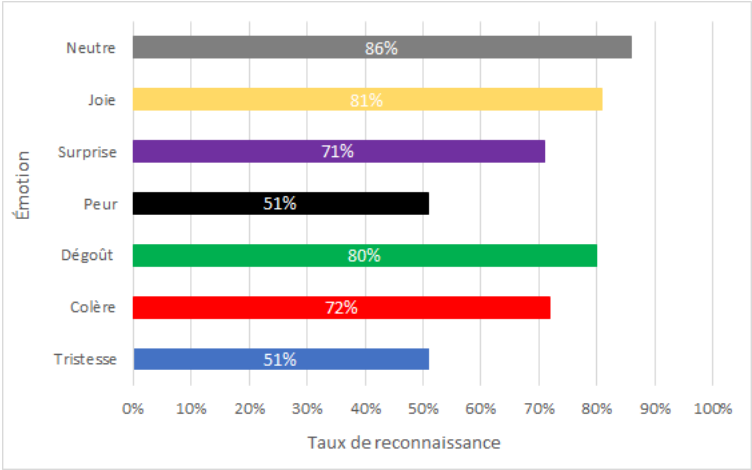


FIGURE 3.4: Taux de reconnaissance par émotion : résultat global brut

La question qui se pose est : les résultats obtenus sont ils statistiquement significatifs ?

Dans tout système de reconnaissance automatique, il est important d'avoir une indication de la qualité d'un résultat obtenu. Pour traiter cette problématique, nous avons utilisé, pour notre modèle, une mesure d'évaluation qui se base sur un intervalle de confiance permettant de valider statistiquement ces résultats. La formulation pour calculer cet intervalle est donnée dans Annexe A (voir section A.4.1.2). Nous avons calculé l'intervalle de confiance à 95% et à 99% des différents taux de reconnaissance des émotions. Les résultats sont présentés dans le Tableau 3.4. Le taux de reconnaissance de l'émotion neutre atteint 85,57% avec un intervalle de confiance à 95% de $\pm 5\%$.

3.3.3 Validation

Dans la section précédente, les enregistrements du corpus ont été évalués par un test de perception avec 106 testeurs afin de garantir la reconnaissabilité des émotions. Dans ce qui suit, nous allons procéder à la validation de ce corpus. La validation, dans notre cas, est une opération destinée à démontrer si le corpus construit est fiable ou non, c'est à dire les émotions actées

Émotion	$IC_{95\%}$	$IC_{99\%}$	TR
Colère	[65,94% ; 78,34%]	[63,98% ; 80,30%]	72,14%
Joie	[75,60% ; 86,35%]	[70,44% ; 91,51%]	80,98%
Peur	[43,57% ; 57,43%]	[36,92% ; 64,08%]	50,50%
Surprise	[64,71% ; 77,29%]	[58,67% ; 83,33%]	71,00%
Dégoût	[74,08% ; 85,31%]	[68,69% ; 90,71%]	79,70%
Tristesse	[44,00% ; 58,07%]	[37,24% ; 64,82%]	51,03%
Neutre	[80,71% ; 90,43%]	[76,05% ; 95,09%]	85,57%

TABLE 3.4: Score et intervalle de confiance en fonction de chaque émotion (TR : taux de reconnaissance ; IC : intervalle de confiance).

sont-elles réalistes ? La validation est la réponse à la question suivante : Avons-nous construit le bon corpus ?

3.3.3.1 Questions

Les énoncés utilisés dans notre corpus sont réalistes et s'inscrivent dans la vie estudiantine. Ils contiennent des mots susceptibles d'indiquer l'émotion actée par un locuteur, nous nous sommes posés la question de l'influence de ces indications, qu'on appellera biais sémantique, sur la reconnaissance de l'émotion lors d'un test. Voici quelques exemples de phrases contenant un biais sémantique :

— **Je suis vraiment déçu**, comment on va faire le TP, ça n'arrête pas de bugger.

— **Ça m'énerve**, j'ai rien compris.

La liste complète de toutes les phrases contenant un biais sémantique se trouve dans Annexe A (section A.2). Plusieurs autres questions se posent ici : Quel est l'impact du locuteur sur le taux de reconnaissance : est-il vraiment incapable d'exprimer cette émotion ou bien il ne joue pas bien le jeu ? Quel est l'impact du testeur : est-il incapable à reconnaitre l'émotion jouée ou bien il n'est pas concentré ?

3.3.3.2 Analyse des résultats

Impact du biais sémantique

Certaines phrases de notre corpus contiennent un biais sémantique. Dans ce cas deux hypothèses s'opposent : ou bien la présence d'indication sémantique de l'état émotionnel est sans effet, ou bien elle modifie le taux de reconnaissance. Pour établir la validité de l'une ou l'autre de ces deux

hypothèses, nous faisons recours aux tests statistiques d'hypothèses. Dans Annexe A (section A.4.1), nous décrivons la formulation du problème avec les hypothèses de travail.

Les résultats de ces tests statistiques d'hypothèses montrent que le biais sémantique diminue significativement les taux de reconnaissance des émotions (voir Annexe A, Tableau A.4) :

— dégoût (au seuil de 1%, $IC_{99\%} = [57{,}15\% ; 88{,}94\%]$ avec bias, $IC_{99\%} = [80{,}12\% ; 97{,}93\%]$ sans biais),

— colère (au seuil de 5%, $IC_{95\%} = [57{,}22\% ; 74{,}41\%]$ avec biais, $IC_{95\%} = [72{,}55\% ; 89{,}35\%]$ sans biais)

— neutre (au seuil de 5%, $IC_{95\%} = [74{,}33\% ; 88{,}38\%]$ avec biais, $IC_{95\%} = [85{,}21\% ; 97{,}92\%]$ sans biais).

La figure 3.5 montre les taux de reconnaissance sur la base entière, sur la base sans les enregistrements contenant le biais sémantique et celle sans biais sémantique. Ces taux sont répartis selon le nombre de testeurs dans le Tableau 3.5. 64 testeurs ont participé à la validation du corpus contenant les énoncés biaisés avec un score de 68% et 42 testeurs pour valider le corpus sans biais sémantique avec un score de 73%. Nous pouvons conclure à partir des résultats obtenus que le biais sémantique diminue aussi le taux global de reconnaissance, le corpus non biaisé est reconnu avec 5% de plus que le corpus biaisé.

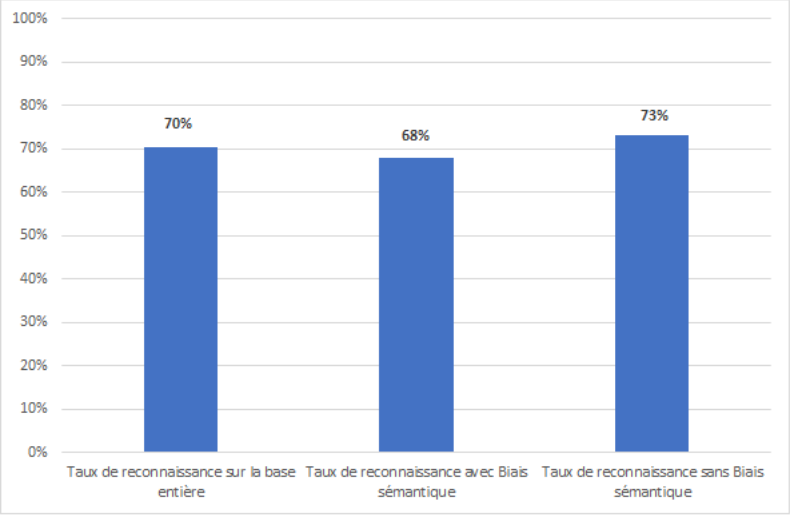


Figure 3.5: Comparaison des taux de reconnaissance sur corpus biaisé et non biaisé

Résultats par émotion

Intéressons nous maintenant aux performances par émotion. Les taux de reconnaissance des différentes émotions sont présentés ci-dessous (voir Figure 3.6). Chaque émotion est mise en

	Nb de testeurs	TR	$IC_{95\%}$	$IC_{99\%}$
Avec biais sémantique	64	68%	[65% ; 71%]	[64% ; 72%]
Sans biais sémantique	42	73%	[69% ; 77%]	[68% ; 78%]
Total	106	70%	[68% ; 72%]	[67% ; 73%]

TABLE 3.5: Décomposition des taux de reconnaissance avec et sans biais sémantique (Nb : nombre ; TR : taux de reconnaissance ; IC : intervalle de confiance).

regard avec les émotions avec lesquelles elle est confondue. Les principales confusions ont lieu entre le dégoût et la tristesse d'une part et surprise et joie d'autre part. Dans la Figure 3.6, le taux de reconnaissance de la tristesse est de 51% environ. Cette émotion a été confondue principalement avec le dégoût et, avec moins de confusion, la colère, le neutre et la peur.

Le tableau 3.6 représente la matrice de confusion pour la base entière. Une matrice de confusion permet de donner une idée pour chaque classe de modèle, les vraies classifications versus les classifications prédites. La colonne la plus à gauche représente les émotions reconnues. La colonne indique pour une émotion, le nombre de prédictions correctes pour cette émotion et le nombre d'échantillons confondues avec une autre émotion. Les valeurs correctes sont organisées dans une ligne diagonale allant du haut à gauche au bas à droite de la matrice. Par exemple, le nombre total d'échantillons pour l'émotion "Peur" dans l'ensemble de données est la somme des valeurs de la colonne "Peur". Nous voyons sur ce tableau que la peur est confondue avec la joie dans 7 cas sur les 200. Le tableau 3.6 montre que la surprise est le plus souvent confondue avec la joie et que l'émotion tristesse est le plus souvent confondue avec l'émotion dégoût. Nous constatons aussi que le neutre est l'émotion la plus reconnue : 172 reconnus parmi 201 enregistrements.

Ces données suggèrent que les analyses statistiques basées sur le pourcentage de reconnaissance des émotions doivent être vérifiées par une mesure de précision appelée taux de réussite non biaisé (ou unbiased hit rate (H_u) en anglais) [83], afin de vérifier le biais de jugement. Le H_u est le taux des émotions correctement identifiées multiplié par le taux de jugement correct des émotions. Par exemple, dans le tableau 3.6, pour l'émotion colère, $H_u = \frac{145}{201} \times \frac{145}{191} = .547$. Le calcul de H_u pour chacune des émotions donne une meilleure indication de la précision relative de la reconnaissance.

Résultats par locuteur

Dans cette partie, nous allons présenter les scores par locuteur. Pour chaque locuteur, le taux de reconnaissance de chaque émotion ainsi que le taux global sont calculés. Les résultats sont

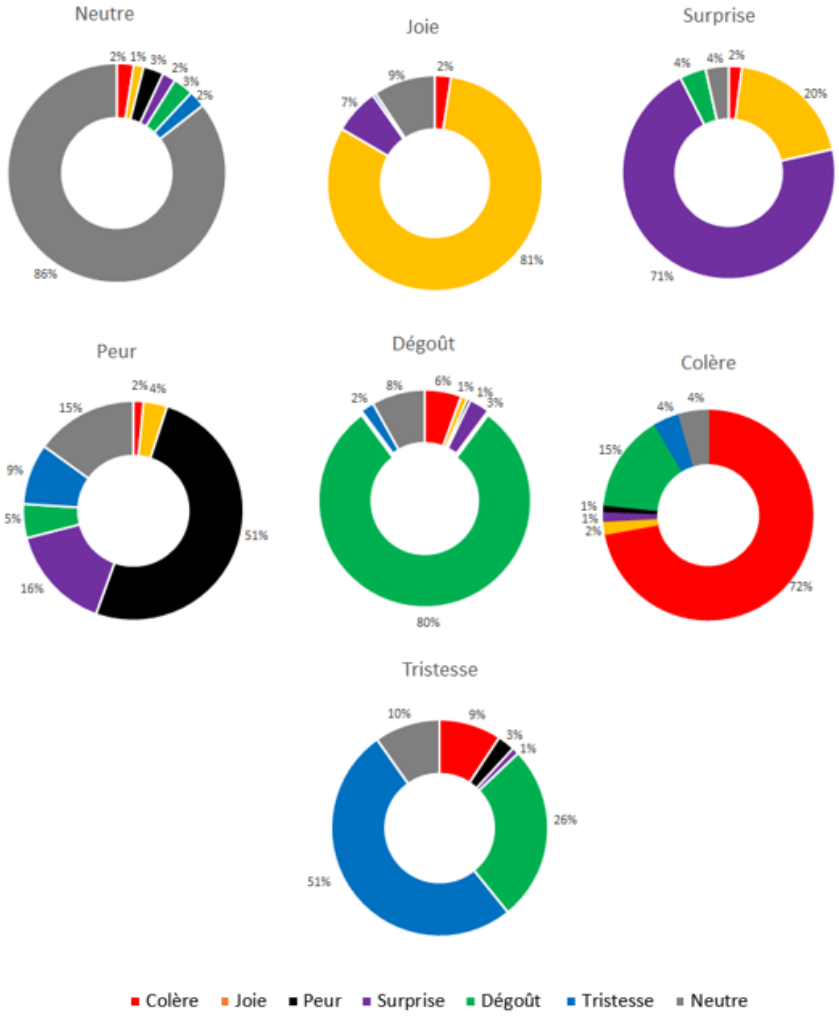


FIGURE 3.6: Les taux de reconnaissance et de confusion des différentes émotions sur la base entière.

Prédite / Exprimée	Colère	Joie	Peur	Surprise	Dégoût	Tristesse	Neutre	Total	H_u
Colère	**145**	5	3	4	11	18	5	191	.547
Joie	4	**166**	7	39	2	0	3	221	.608
Peur	3	0	**101**	0	1	5	6	116	.439
Surprise	2	14	31	**142**	6	2	4	201	.501
Dégoût	30	0	10	8	**157**	51	6	262	.477
Tristesse	8	1	18	0	4	**99**	5	135	.374
Neutre	9	19	30	7	16	19	**172**	272	.541
Total	201	205	200	200	197	194	201	1398	–

TABLE 3.6: Comparaison entre les émotions reconnues et les émotions exprimées avec l'indicateur (H_u) : matrice de confusion de la base entière.

présentés par ordre décroissant sur le taux de reconnaissance global : le locuteur le plus reconnu vers le locuteur le moins reconnu. Le Tableau 3.7 présente le taux de reconnaissance par locuteur

pour la base entière. Nous constatons que les émotions exprimées par 2/3 des locuteurs sont reconnues à plus de 65%. La question qui se pose est la suivante : faut-il retirer les enregistrements des locuteurs les moins reconnus ?

Locuteur	Genre	Nationalité	Taux de reconnaissance par émotion (%)							Taux global (%)
			Colère	Joie	Peur	Surprise	Dégoût	Tristesse	Neutre	
1	H	Française	100	100	100	50	100	100	67	90,32
2	H	Espagnole	100	100	67	100	86	75	100	89,28
3	H	Française	100	100	100	100	50	75	83	86,36
4	H	Française	100	100	100	89	67	*	*	84,61
5	F	Française	100	87	60	74	88	80	95	82,29
6	H	Française	67	100	67	88	75	50	100	80,00
7	H	Française	63	92	71	75	86	40	100	78,48
8	F	Française	93	100	62	55	91	74	60	77,64
9	F	Française	82	100	64	74	94	22	80	77,63
10	H	Française	100	60	50	83	100	57	86	75,75
11	F	Française	100	86	53	80	71	33	81	75,30
12	H	Française	100	100	67	57	75	0	83	75,00
13	H	Française	50	80	43	100	80	67	100	72,41
14	H	Française	83	0	75	50	71	50	86	69,69
15	H	Française	67	100	0	33	86	67	100	68,18
16	H	Française	50	58	58	92	93	47	69	65,97
17	F	Espagnole	75	71	22	91	50	67	90	66,23
18	H	Française	100	67	67	67	100	29	60	65,71
19	H	Française	100	25	71	20	100	71	100	65,62
20	H	Française	100	50	57	50	100	25	100	65,38
21	H	Française	*	100	100	25	67	25	50	65,38
22	H	Française	38	100	25	75	100	20	86	63,15
23	H	Française	33	75	60	100	50	33	100	62,06
24	H	Française	100	80	100	50	33	33	*	61,90
25	H	Française	83	57	20	80	100	75	40	61,76
26	H	Française	33	50	33	100	50	71	100	61,76
27	F	Française	45	91	20	50	71	57	100	60,71
28	H	Française	20	60	17	100	83	100	100	57,89
29	H	Française	71	100	29	100	75	29	50	56,25
30	H	Française	75	0	0	38	100	0	100	53,33
31	H	Française	0	43	25	80	100	17	100	50,00
32	H	Française	100	50	25	29	50	33	100	47,05

Table 3.7: Taux de reconnaissance par locuteur pour la base entière (* signifie que les testeurs n'ont pas écouté l'émotion actée par ce locuteur.)

Résultats par testeur

Après l'analyse du corpus selon la "qualité" des locuteurs, regardons les taux des différents testeurs. Dans les tableaux 3.8 et 3.9, nous présentons quelques scores selon le genre et la culture. Nous constatons que 56 testeurs sur les 64 reconnaissent des émotions avec un score supérieur ou égal à 50 % lorsque le corpus contient des énoncés avec un biais sémantique. Lorsque le corpus ne contient pas de biais sémantique, 41 testeurs sur les 42 reconnaissent des émotions avec un score supérieur ou égal à 50% (voir annexe A, tableaux A.5 et A.6).

Testeur	Genre	Culture	Taux de reconnaissance
1	H	Berbère	100%
2	H	Française	92,85%
3	H	Arabe	78,57%
4	H	Camerounais	71,42%
5	F	Française	50,00%
6	H	Française	42,85%

TABLE 3.8: Taux de reconnaissance par testeur pour la base avec biais sémantique.

Testeur	Genre	Culture	Taux de reconnaissance
1	F	Française	92,85%
2	H	Bresilienne	85,71%
3	F	Française	78,57%
4	H	Américaine	64,28%
5	F	Asiatique	64,28%
6	F	Française	42,85%

TABLE 3.9: Taux de reconnaissance par testeur pour la base sans biais sémantique

En se référant aux études présentées ci dessus sur la perception des émotions (voir section 3.3.2) et aux résultats des tests, nous avons donc, procédé au filtrage des données, pour ne garder que les énoncés correctement simulés. Pour la base de données Berlin, seuls les enregistrements ayant un taux de reconnaissance des émotions supérieur à 80% et considérés comme naturels par plus de 60% des auditeurs sont retenus. 503 enregistrements parmi 800 enregistrements sont retenus.

Dans notre travail, nous avons retenu seulement les enregistrements ayant un taux de reconnaissance supérieur à 50%. Après la suppression des enregistrements qui ne respectent pas ce critère de validation et la suppression du biais sémantique, le nombre final d'enregistrements retenus par émotion est le suivant : tristesse (55), colère (67), dégoût (67), peur (51), surprise (62), joie (62) et neutre (70). 434 enregistrements parmi 502 sont retenus, pour obtenir un taux de reconnaissance de 88%.

3.4 Conclusion

Pour analyser la voix émotionnelle de locuteurs, les émotions sont collectées en bases de données. Il faut d'abord rappeler que la plupart des corpus émotionnels aujourd'hui sont des corpus actés, c'est à dire construits à partir de données prototypiques sur peu de locuteurs (moins de 10). Nous avons présenté dans ce chapitre des corpus actés utilisés dans nos expériences. Nous avons également collecté un corpus émotionnel acté dans le contexte pédagogique. Notre corpus en français que nous avons dénommé "French Student Emotionnal database" contient plus de 500 énoncés différents. Cette base sera mise à disposition de la communauté sur internet via les sites web du laboratoire CREN[4] et du laboratoire LAUM[5] .

Nous décrivons dans le chapitre suivant notre méthodologie d'extraction des caractéristiques existantes extraites à partir des corpus.

4. http://cren.univ-nantes.fr/
5. http://laum.univ-lemans.fr/fr/index.html/

Chapitre 4

Extraction de caractéristiques

Contents

4.1 Introduction 59

4.2 Extraction des caractéristiques basée sur un banc de filtres 61

4.2.1 Les coefficients cepstraux sur l'échelle Mel (MFCC) 61

4.2.2 Les caractéristiques spectrales de modulation (MSF) 63

4.3 Démodulation par décomposition en modes empiriques 64

4.3.1 Décomposition en modes empiriques (EMD) 64

4.3.2 Modélisation de la voix par un signal modulé en amplitude et en fréquence 66

4.3.3 Opérateur d'énergie de Teager-Kaiser (TKEO) 66

4.4 Extraction des caractéristiques basée sur la démodulation par EMD-TKEO 68

4.4.1 Les coefficients cepstraux 68

4.4.2 Les caractéristiques de modulation AM-FM 72

4.5 Conclusion 81

4.1 Introduction

L'extraction d'un ensemble de caractéristiques pertinent pour un système de RAE était et reste encore un grand défi à relever, ce qui a ouvert la voie à plusieurs travaux de recherche [36],[84], [85]. La plupart de ces caractéristiques sont regroupées en deux catégories : prosodique et spectrale [36]. De nombreuses méthodes ont été utilisées pour extraire des caractéristiques de la voix, dont la plupart sont basées sur l'analyse de Fourier à court terme. La nécessité d'une telle analyse à court terme se justifie par l'hypothèse de stationnarité de la parole pendant des

courts intervalles du temps (de 25 à 30 ms). Cependant, la question se pose sur la pertinence des caractéristiques extraites avec cette hypothèse. En effet, des informations contenues dans des intervalles plus longs peuvent être pertinentes pour la reconnaissance des émotions [86].

Étant donné que le signal vocal est naturellement non stationnaire[1]. La transformée de Fourier à court temps (STFT pour Short-Time Fourier Transform en anglais) ne permet pas de suivre l'évolution du contenu fréquentiel du signal, toutes ses composantes étant réparties sur toute l'échelle de temps. Cette méthode est alors limitée par le principe d'incertitude fondamentale, selon lequel le temps et la fréquence ne peuvent pas être résolus simultanément avec la même précision et elle n'est pas appropriée pour extraire de caractéristiques d'un signal issue d'un système non linéaire. Une reconnaissance précise des émotions à partir de la voix reste difficile en raison de la variabilité, de la complexité et des changements non linéaires des caractéristiques des émotions dans la voix. Cependant, l'information sur l'émotion est plutôt située sur le long terme (syllabe, mot et phrase) et non pas au niveau phonème, d'où vient la nécessité de faire une analyse à long terme (plus de 100 ms).

La décomposition en modes empiriques (EMD), proposée par Huang et al. [87], apporte une réponse aux limites évoquées précédemment. Cette méthode décompose le signal, d'une façon adaptative, en somme de composantes oscillantes, appelées IMF (pour intrinsic mode function en anglais). Des études ont montré que la représentation temps-fréquence du signal est extrêmement importante pour l'analyse du signal non stationnaire et/ou généré par un système non-linéaire [88], [89]. Cette représentation est donnée par l'utilisation conjointe de la décomposition en modes empiriques pour estimer les IMF du signal et de l'opérateur d'énergie de Teager-Kaiser (TKEO) pour démoduler les différentes IMF en estimant la fréquence instantanée (notée FI) et l'amplitude instantanée (notée AI). Dans ce chapitre, nous proposons un ensemble de nouvelles caractéristiques basées sur cette représentation, capturant les informations ignorées par les méthodes classiques. Dans la section suivante (4.2), nous présentons les caractéristiques les plus utilisés et qui nous serviront de référence pour la comparaison avec nos nouvelles caractéristiques. Nous allons présenter seulement les caractéristiques les plus utilisées pour la RAE à savoir MFCC. Ensuite, la méthode de démodulation du signal est décrite dans la section 4.3. Enfin, les nouvelles caractéristiques proposées sont détaillées dans la section 4.4.2.

1. Un signal aléatoire est dit non stationnaire si ses propriétés statistiques sont variantes au cours du temps. Exemples : le signal de la parole, la musique, un chirp, etc.

4.2 Extraction des caractéristiques basée sur un banc de filtres

Dans notre travail, nous proposons un ensemble de caractéristiques unifiant les deux catégories (spectrale et prosodique) et qui sont basées sur la méthode EMD combinée avec TKEO. Ces caractéristiques sont inspirées de précédents travaux sur les MFCC et les MSF. Dans cette section, nous allons présenter ces deux caractéristiques à des fins de comparaison.

4.2.1 Les coefficients cepstraux sur l'échelle Mel (MFCC)

Les coeffcients cepstraux sur l'échelle Mel (MFCC pour Mel-Frequency Cepstral Coefficients en anglais), introduites en [90], sont les paramètres acoustiques les plus utilisés dans les systèmes de RAE [44] car ils prennent en compte la sensibilité de perception fréquentielle de l'oreille [91]. Le processus de calcul de la MFCC est présenté dans la Figure 4.2. Le signal d'entrée est découpé en une séquence de trames de durées relativement courtes sur lesquelles le signal peut en général être considéré comme stationnaire. Chaque segment (trame) dure $25ms$ avec un chevauchement de 10 ms. La segmentation du signal en trames produit des discontinuités aux frontières des trames. Pour réduire cet effet, les trames sont multipliées par une fenêtre de Hamming. Ensuite, la transformée de Fourier (FFT) est appliquée pour passer du domaine temporel vers le domaine spectral (ou fréquentiel), on obtient donc le spectre $X_k(m)$. La simulation de l'oreille humaine, qui se base sur une échelle fréquentielle spécifique, nécessite le passage à l'échelle de Mel (voir Figure 4.1). Cette dernière est composée d'un banc de 12 à 24 filtres triangulaires, dont les bandes sont plus fines pour les basses fréquences que pour les hautes fréquences. Le but de l'analyse par un banc de filtres est de trouver l'énergie du signal dans des bandes de fréquences déterminées. Dans notre travail, 12 filtres triangulaires (espacés linéairement jusqu'à 1000 Hz, puis logarithmiquement au-delà) sont utilisés (voir Figure 4.1). Chaque filtre fournit un coefficient qui donne l'énergie du signal dans la bande couverte par le filtre. La formule de conversion de l'échelle linéaire en échelle de Mel est la suivante :

$$mel(f) = 259 \log_{10}(1 + \frac{f}{700}) \tag{4.1}$$

En général, le signal vocal résulte de la convolution de la source par le conduit vocal. Dans le domaine spectral, cette convolution devient un produit qui rend difficile la déconvolution de ce signal. Ce problème peut être surmonté en passant dans le domaine log-spectral : c'est le principe de l'analyse cepstrale. Par définition, le cepstre d'un signal est une transformation de ce signal du domaine temporel vers un autre domaine analogue au domaine temporel. En pratique, le cepstre

réel d'un signal est obtenu par la transformée de Fourier inverse appliquée au logarithme de la transformée de Fourier du signal. Le calcul de la FFT et de la FFT inverse est laborieux ce qui rend ce type de calcul des coefficients cepstraux inutilisable en reconnaissance de la parole [92]. En revanche, la transformée en cosinus discrète (DCT) appliquée sur le logarithme des énergies sortant d'un banc de filtres permet d'obtenir les coefficients cepstraux (MFCC).

Nous travaillons avec les 12 premiers coefficients qui sont les coefficients les plus efficaces. Cela peut s'expliquer par le fait que la plupart des informations utiles pour la discrimination des émotions concernant la prosodie se trouvent dans la partie des basses fréquences de l'échelle Mel. Nous savons que le signal de parole varie au cours du temps. Ces variations peuvent elles aussi être propres à un individu. Il est donc utile de modéliser l'évolution temporelle de chacun des coefficients. Cela est fait par le calcul des statistiques (moyenne, variance, coefficient de variation[2], kurtosis et asymétrie) pour chaque coefficient, pour un total de 60 caractéristiques.

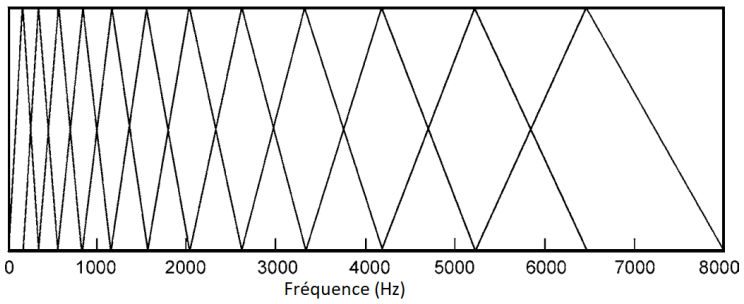


Figure 4.1: 12 Banc de filtres sur une échelle Mel.

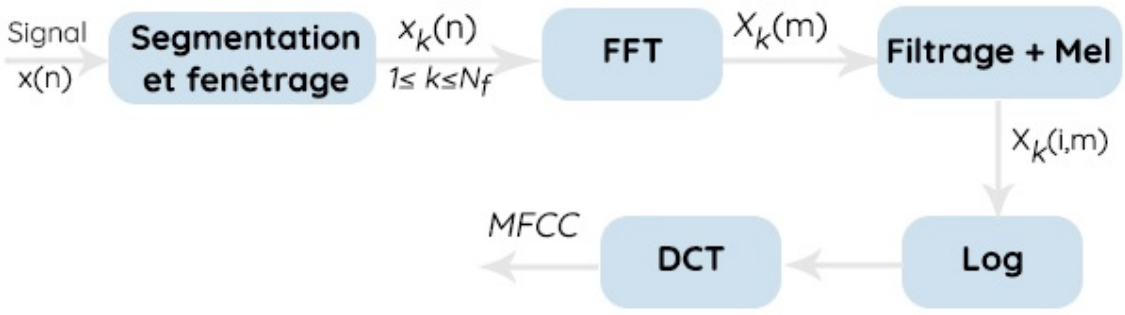


Figure 4.2: Étapes de l'extraction des coefficients MFCC.

Dans [92] et en se basant sur des résultats de recherche existants, l'auteur a montré que les MFCC donnent des meilleurs résultats dans les systèmes de RAE, ce qui montre plus généralement l'intérêt d'un pré-traitement par banc de filtres, d'une échelle fréquentielle non linéaire, et de la

2. Le coefficient de variation se calcule comme le ratio de l'écart-type rapporté à la moyenne.

représentation cepstrale.

4.2.2 Les caractéristiques spectrales de modulation (MSF)

En supposant qu'un signal vocal soit un processus stationnaire à court terme, les caractéristiques (MFCC) décrites dans la section précédente sont extraites sur une fenêtre courte (d'une durée de 25 à 30 ms). Ces caractéristiques transmettent uniquement le spectre à court terme du signal vocal et omettent des informations importantes sur le comportement temporel. De plus, des études en neurosciences [86] ont montré l'existence de champs récepteurs spectro-temporels (ST) dans le cortex auditif humain. Ces champs peuvent s'étendre sur des plages temporelles de plusieurs centaines de millisecondes et répondre à des modulations dans le domaine temps-fréquence. Les caractéristiques spectrales de modulation à long terme (MSF pour long-term modulation spectral features en anglais), introduites dans [86], sont spécifiquement extraites pour résoudre les problèmes de caractéristiques spectrales à court terme et pour mieux modéliser la nature de la perception auditive humaine. Ces caractéristiques sont obtenues en émulant le traitement spectro-temporel (ST) effectué dans le système auditif humain et prend en compte la fréquence acoustique régulière conjointement avec la fréquence de modulation.

Les étapes de calcul des caractéristiques MSF sont illustrées dans la Figure 4.3. Pour obtenir la représentation ST, le signal d'entrée est d'abord découpé en une séquence de trames de durée de 256 ms avec un chevauchement de 64 ms. Cette durée est relativement longue par rapport aux valeurs typiques utilisées dans le traitement de la parole traditionnel (environ 20 à 30 ms). Ensuite, le signal vocal est décomposé par un banc de filtres auditifs (19 filtres au total). Les enveloppes de Hilbert des sorties en bande critique sont calculées pour former les signaux de modulation. Un banc de filtres de modulation est en outre appliqué aux enveloppes de Hilbert pour effectuer une analyse de fréquence. Les contenus spectraux des signaux de modulation sont appelés spectres de modulation et les caractéristiques proposées s'appellent ainsi des caractéristiques spectrales de modulation (MSF). Enfin, la représentation ST est formée en mesurant l'énergie des signaux d'enveloppe décomposés, en fonction de la fréquence acoustique et de la fréquence de modulation régulière. La moyenne des énergies sur toutes les fenêtres dans chaque bande spectrale fournit une caractéristique. Dans notre expérience, un banc de filtres auditifs avec N = 19 filtres et un banc de filtres de modulation avec $M = 5$ filtres sont utilisés. Au total, 95 (19×5) coefficients sont calculés à partir de la représentation ST.

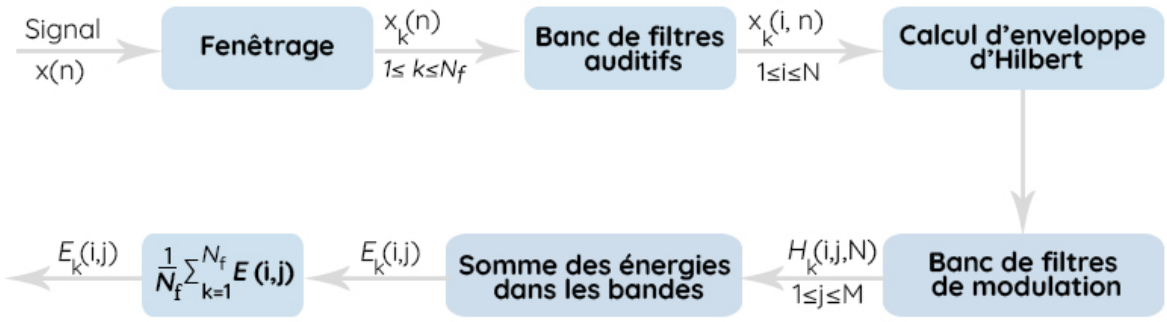


Figure 4.3: Étapes de l'extraction des coefficients MSF

4.3 Démodulation par décomposition en modes empiriques

4.3.1 Décomposition en modes empiriques (EMD)

La décomposition en modes empiriques (EMD pour Empirical Mode Decomposition en anglais) est une méthode de décomposition de tout signal non stationnaire ou issu d'un système non linéaire, en une somme de composantes appelées fonctions en mode intrinsèque (IMF pour intrinsic mode function en anglais). Ces fonctions sont des signaux modulés en amplitude et en fréquence (AM-FM). Chaque IMF contient localement des oscillations plus lentes (basses fréquences) que celle extraite précédemment (haute fréquence) [88] (voir exemple Figure 4.4). L'extraction des IMFs est non linéaire, mais l'ajout de tous les IMF permet une reconstruction linéaire du signal d'origine sans perte ni distorsion des informations initiales.

4.3.1.1 Fonction Mode intrinsèque (IMF)

Une fonction est appelée fonction de mode intrinsèque lorsqu'elle satisfait les propriétés suivantes :

- Le nombre d'extrema (maximum + minima) dans le signal et le nombre de passages par zéro ne diffèrent pas de plus d'un ;
- La moyenne des enveloppes définies par les maxima locaux et les minima locaux doit être nulle en tout point.

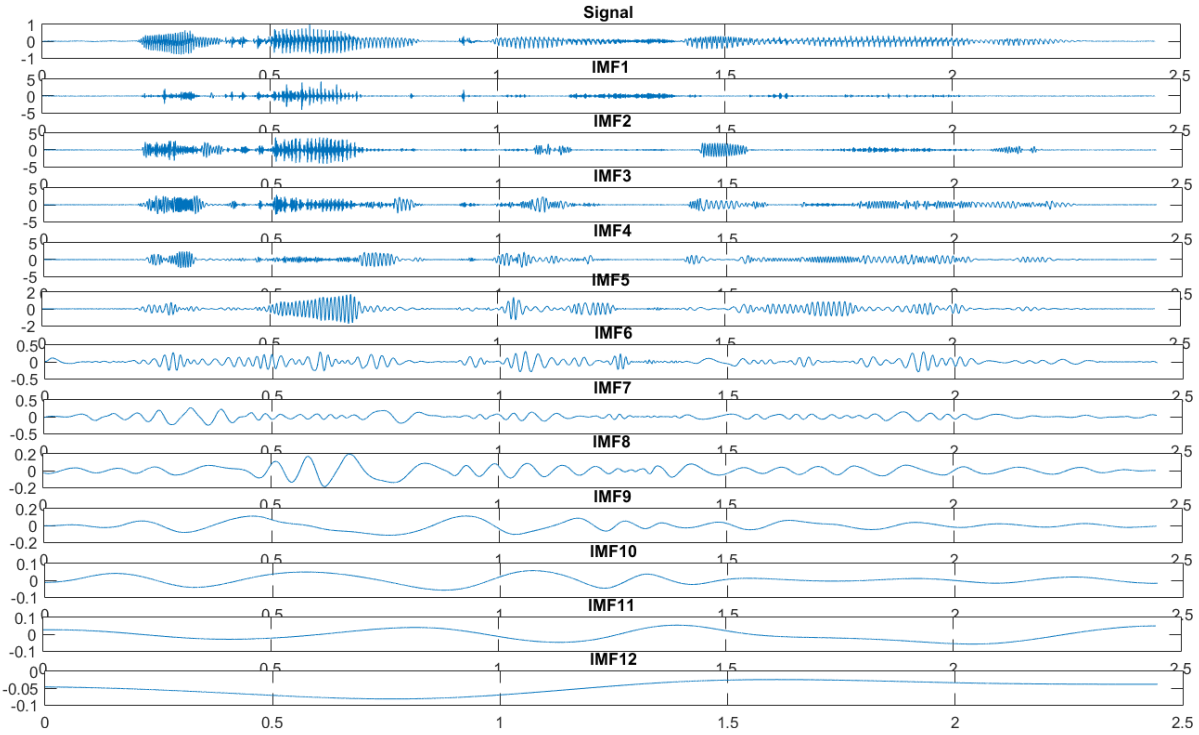


FIGURE 4.4: Exemple d'un signal vocal décomposé en IMFs en utilisant l'algorithme EMD

La procédure pour extraire un IMF, à partir d'un signal donné, est appelée *sifting process* (traduite par processus de tamisage).

4.3.1.2 Algorithme

L'algorithme de l'EMD est décrit comme suit (voir Algorithme 1) :

L'algorithme de la décomposition s'arrête lorsqu'il n'existe plus d'oscillations à extraire dans le résidu : le résidu $res(t)$ est une constante, une fonction monotone ou une fonction avec un seul extrema. Le résultat du processus EMD produit N IMFs $(r_1(t), r_2(t), ..., r_N(t))$ et le résidu $(res_N(t))$. Pour tout signal x(t), l'EMD donne la décomposition suivante [93] :

$$x(t) = \sum_{i=1}^{N} r_i(t) + res_N(t) \tag{4.2}$$

où $r_i(t)$ est lA $i^{\text{ème}}$ IMF du signal, N le nombre d'IMF et $res_N(t)$ est le résidu final (un polynôme de faible degré). Ce résidu est généralement exclu du signal car il ne contient pas le nombre suffisant des extrema pour construire les enveloppes maxima et minima [89].

Algorithme 1 Algorithme de la décomposition par EMD

Entrée : Un signal vocal $x(t)$
Sortie : IMFs $r_1, r_2, ..., r_N$, résidu res

1. Calculer tous les extrema locaux dans le signal $x(t)$: les maxima locaux et les minima locaux ;
2. Construire l'enveloppe supérieure $E_u(t)$ et l'enveloppe inférieure $E_l(t)$, interpolant respectivement les maxima et les minima par spline cubique.
3. Calculer l'enveloppe moyenne par la demi-somme de son enveloppe supérieure et inférieure : $m(t) = \frac{(E_u(t)+E_l(t))}{2}$;
4. Soustraire du signal original $x(t)$ son enveloppe moyenne $m(t)$, puis obtenir des modes $r(t)$ à partir de laquelle la basse fréquence est supprimée : $r(t) = x(t) - m(t)$;
5. Répéter les étapes 1-4 (soustraction de l'enveloppe moyenne) jusqu'à obtenir un IMF $r(t)$ (qui répondent aux deux conditions ci-dessus) : *sifting process* ;
6. Soustraire cet IMF $r(t)$ du signal original $x(t)$: $res(t) = x(t) - r(t)$;
7. Répéter les étapes 1-6 jusqu'à ce qu'il ne reste plus de IMFs dans le signal résiduel $res(t)$ afin d'obtenir tous les IMFs $r_1(t)$, $r_2(t)$,..., $r_N(t)$ du signal $x(t)$.

4.3.2 Modélisation de la voix par un signal modulé en amplitude et en fréquence

Chaque IMF obtenue par la décomposition EMD est considérée comme étant un signal AM-FM monocomposante. Ce signal peut s'écrire sous la forme [94] :

$$r_i(t) = Re\left(a_i(t)\, exp\left[\, j\ \phi_i(t)\right]\right) \tag{4.3}$$

$$\phi_i(t) = 2\pi \int f_i(\tau)\mathrm{d}\tau \tag{4.4}$$

avec Re représente la partie réelle, $\phi_i(t)$ représente la phase instantanée, $a_i(t)$ est l'amplitude instantanée et $f_i(t)$ la fréquence instantanée de la i ème IMF. Cette définition contient à la fois la fonction AM, $a_i(t)$ et la fonction FM, $f_i(t)$. L'intérêt principal de ce type des signaux en traitement du signal est la facilité d'extraction d'informations relatives à des grandeurs comme l'amplitude instantanée $a(t)$ et la fréquence instantanée $f(t)$ à partir d'un signal modulé $r_i(t)$.

4.3.3 Opérateur d'énergie de Teager-Kaiser (TKEO)

Les IMFs obtenues par l'EMD ne transmettent aucune information et ne donnent aucun sens. Il est ainsi nécessaire, pour chaque IMF, d'estimer l'amplitude instantanée (AI) et la fréquence instantanée (FI), ce qui donne une certaine signification physique. Différents opérateurs énergétiques

ont été utilisés pour estimer les grandeurs FI et AI. Maragos et al. [94] ont proposé l'application d'un opérateur TKEO (Teager-Kaiser Energy Operator) afin d'estimer la FI et AI d'un signal mono-composante. Dans [95], [96] et [97], les auteurs montrent que la combinaison de l'EMD et le TKEO a un rôle important pour l'analyse temps-fréquence et en particulier pour la démodulation du signal AM-FM.

L'opérateur TKEO est appliqué sur chacune des N IMFs séparément. TKEO est un opérateur non linéaire qui calcule l'énergie des signaux monocomposants en tant que produit du carré de l'amplitude et de la fréquence du signal. Les caractéristiques instantanées de ces signaux sont ensuite obtenues par l'application de l'algorithme de séparation d'énergie discrète (DESA-2) [94]. Une méthode basée sur l'algorithme EMD de Huang et al. [98] et le TKEO, qui permet d'estimer les grandeurs instantannées d'un signal, est appelée Tranformée de Huang-Teager (THT pour Teager-Huang Transform en anglais) [99]. TKEO est défini pour l'IMF $r_i(t)$ à temps continu comme [94] :

$$\Psi[r_i(t)] = [\dot{r_i}(t)]^2 - r_i(t)\ddot{r_i}(t) \tag{4.5}$$

où $\dot{r_i}$ et $\ddot{r_i}$ sont respectivement la dérivée première et seconde de $r_i(t)$. Dans le cas discret, les dérivées temporelles de l'équation 4.5 peuvent être exprimées comme suit [100] :

$$\Psi[r_i(n)] = r_i^2(n) - r_i(n+1)r_i(n-1) \tag{4.6}$$

où n est l'index de temps discret. L'algorithme (DESA-2) permet de séparer la fréquence instantannée $f(n)$ et l'amplitude instantannée $a(n)$ d'un sigal $r_i(n)$ AM-FM mono-composante [94] :

$$f(n) = \frac{1}{2}\arccos(1 - \frac{\Psi[r_i(n+1) - r_i(n-1)]}{2\Psi[r_i(n)]} \tag{4.7}$$

$$\mid a(n) \mid = \frac{2\Psi[r_i(n)]}{\sqrt{\Psi[r_i(n+1) - r_i(n-1)]}} \tag{4.8}$$

Les auteurs dans [101] ont trouvé que cette approximation induisait une composante d'erreur de haute fréquence. Il faut donc éliminer la composante d'erreur de haute fréquence en passant la sortie de l'opérateur d'énergie à travers un filtre passe-bas approprié, comme indiqué dans la Figure 4.5. Pour cela, nous avons utilisé dans notre travail un filtre de lissage binomial linéaire à sept points avec une réponse impulsionnelle (1, 6, 15, 20, 15, 6, 1) [101].

Figure 4.5: Le schéma fonctionnel du lissage de l'opérateur d'énergie.

4.4 Extraction des caractéristiques basée sur la démodulation par EMD-TKEO

4.4.1 Les coefficients cepstraux

4.4.1.1 SMFCC

Dans le signal vocal, des changements complexes et aléatoires peuvent être accompagnés d'une tendance du signal ("signal trend "en anglais) qui peut entrainer des erreurs dans l'analyse spectrale et par conséquent dans le calcul des coefficients MFCC. Pour éviter ce problème, il faut supprimer cette tendance. Dans cette section, nous calculons les caractéristiques extraites après l'élimination du signal trend $T(n)$. Ces caractéristiques sont appelées SMFCC et introduites dans [97], fournissent une description plus précise de la distribution de l'énergie dans le domaine fréquentiel. Le procédé de reconstruction du signal basé sur EMD est mis en oeuvre pour effectuer l'extraction des caractéristiques SMFCC. Le processus d'extraction de SMFCC est représenté dans la Figure 4.6. Tout d'abord, la méthode EMD est appliquée au signal vocal d'entrée. Deuxièmement, la tendance du signal, selon [95], est calculée à l'aide de la formule ci-dessous (équation 4.10) et ensuite supprimée du signal d'origine à l'aide de la méthode de détection du taux de passage par zéro (Zero Crossing Rate (ZCR)). $T(n)$ est la somme des IMFs $r(n)$ qui respectent la contrainte suivante [95] :

$$\frac{R_{r_i}}{R_{r_1}} < 0.01 \quad (i = 2, ...n) \tag{4.9}$$

avec R représente le taux de passage par zéro

$$T(n) = \sum_i r_i(n) \tag{4.10}$$

Ensuite, le signal final $S(n)$ est obtenu par la soustraction de la tendance du signal $T(n)$ du signal vocal original $x(n)$.

$$S(n) = x(n) - T(n) \tag{4.11}$$

Les coefficients SMFCC sont extraites du signal reconstitué obtenu $S(n)$ en appliquant la FFT et la transformée en cosinus discrète (DCT). Pour chaque coefficient, nous calculons la moyenne, la variance, l'écart type, le Kurtosis et l'asymétrie, pour un total de 60 caractéristiques.

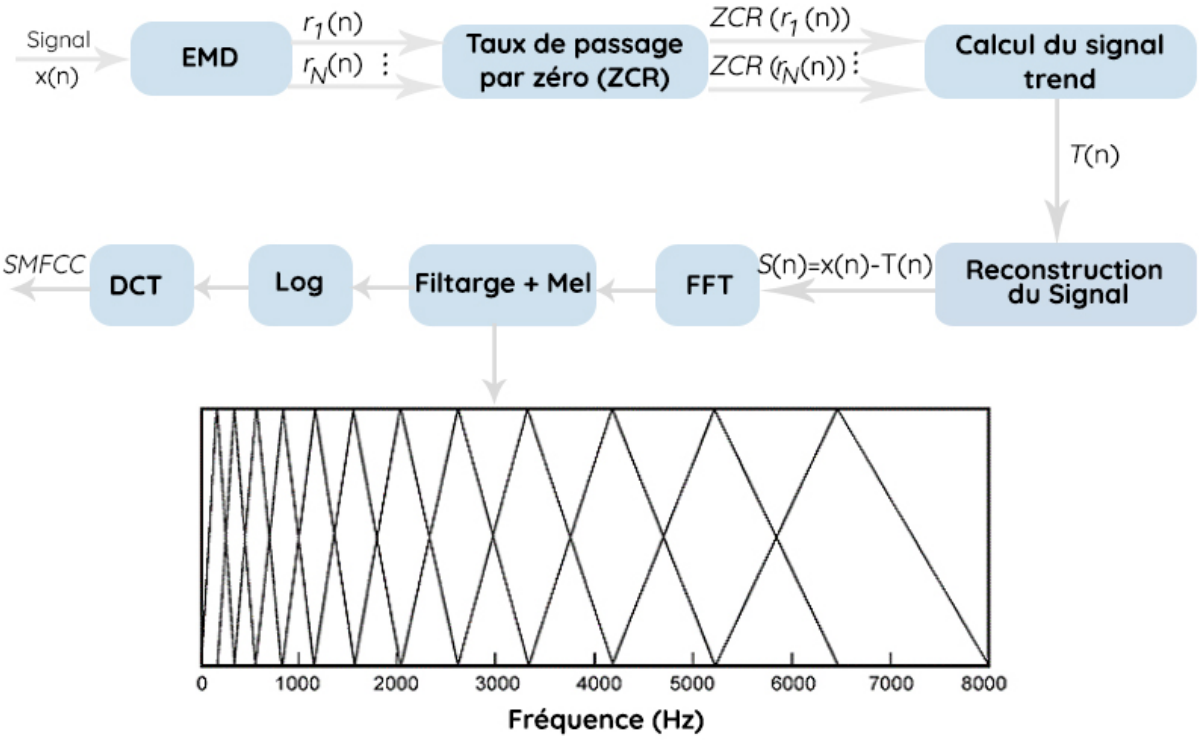


FIGURE 4.6: Étapes de l'extraction des coefficients SMFCC

4.4.1.2 Les coefficients cepstraux d'énergie (ECC) et les coefficients cepstraux d'énergie pondérée en fréquence (EFCC)

Dans [89], les auteurs démontrent que la distribution de l'énergie spectrale varie selon les émotions. Cela signifie que les émotions peuvent affecter la répartition énergétique de la parole entre différentes bandes de fréquences. D'où l'importance d'extraire la fonction ECC qui assure la distribution de l'énergie dans le spectre. En approche standard, l'énergie instantanée de chaque IMF est considérée comme proportionnelle à l'amplitude et n'a rien à voir avec la fréquence instantanée. Selon le modèle physique de fréquence instantanée, l'énergie instantanée dépend aussi de la fréquence instantanée $f(t)$. Sur la base des considérations ci-dessus, l'énergie instantanée pondérée en fréquence (EFCC) permet d'améliorer les caractéristiques ECC présentées ci-dessus [95].

L'implémentation standard du calcul des coefficients ECC et EFCC est montrée à la Figure 4.7.

La première étape de traitement est la décomposition du signal vocal en IMF avec EMD. L'amplitude instantanée de la $j^{\text{ème}}$ IMF $(a(i,n))$ et sa fréquence instantanée $(f(i,n))$ sont estimées à l'aide de TKEO à la deuxième étape. La troisième étape de traitement tente d'enregistrer l'amplitude et la fréquence instantanées en trames qui se chevauchent (la durée de la trame est de 256 ms et le chevauchement est égal à 64 ms), ce qui donne $a_k(i,n)$ et $f_k(i,n)$, N_f représente le nombre de fenêtre. La quatrième étape du traitement consiste à calculer le spectre marginal de Hilbert (MHS pour Marginal Hilbert Spectrum en anglais) qui offre une mesure de l'amplitude totale de chaque fréquence. Par conséquent, le spectre est décomposé en 12 bandes de fréquence. Ces 12 bandes correspondent à l'échelle Mel entre 0 et 8KHz (voir Figure 4.1) qui couvrent largement la bande du signal échantillonné à 16KHz. La puissance de chaque bande de fréquence est calculée. Ensuite, le logarithme naturel de l'énergie des sous-bandes et le complément de la transformée en cosinus discrète (DCT) sont calculés. Les 12 premiers coefficients fournissent les valeurs ECC et EFCC utilisées dans le processus de classification.

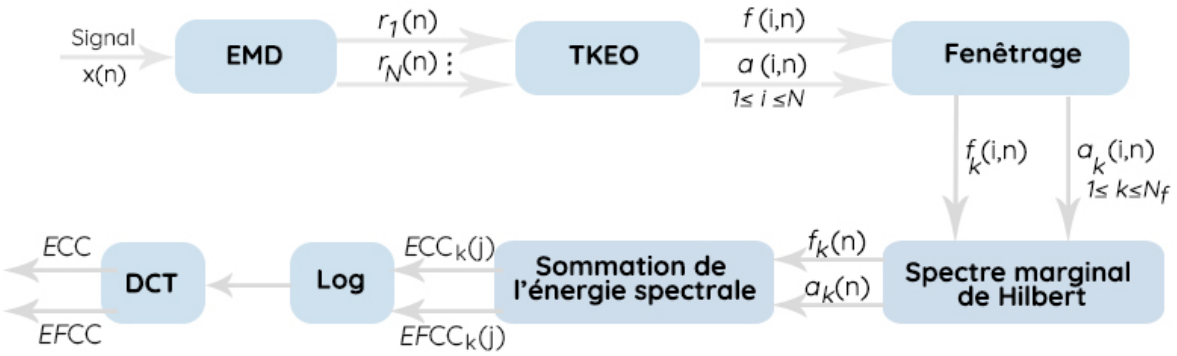


FIGURE 4.7: Étapes de l'extraction des coefficients ECC et EFCC

Pour chaque mode (IMF), la densité spectrale de Hilbert $H(f,n)$ est définie comme le carré de l'amplitude [89] :

$$H(f,n) = \sum_{i=1}^{N} a(i,n)^2 \mathbb{1}_{\{f(i,n)\}}(f) \tag{4.12}$$

où $\mathbb{1}_\Omega$ est la fonction indicatrice de l'ensemble Ω :

$$\mathbb{1}_\Omega(x) = \begin{cases} 1 & \text{if } x \in \Omega \\ 0 & \text{if } x \notin \Omega \end{cases}$$

$H(f,n)$ donne une valeur locale de l'énergie dans une représentation temps-fréquence et on peut extraire la densité de probabilité conjointe $p(f,a)$ de la fréquence et de l'amplitude instantanées,

pour l'ensemble des modes $i = 1...N$. Ceci permet d'estimer le spectre marginal de Hilbert (MHS pour marginal Hilbert spectrum en anglais) qui est définie comme suit [89] :

$$h_{j,k}(f) = \sum_{n=1}^{L_f} H(f,n)\mathbb{1}_{B_j}(f) \tag{4.13}$$

où $h_{j,k}(f)$ représente la densité d'énergie pour une fréquence f pendant la $k^{\text{ème}}$ trame et dans la $j^{\text{ème}}$ bande et L_f représente le nombre d'échantillons par fenêtre. Les fréquences instantanées sont divisées en 12 bandes de fréquence qui se chevauchent (voir figure 4.1).

Les coefficients cepstraux d'énergie (ECC)

Les ECC sont donnés en temps continu par [89] :

$$ECC(B_j, \Pi_k) = \int_{f \in B_j} h_{j,k}(f)\mathrm{d}f,\ t \in \Pi_k\ ,\ j = 1, ..., 12 \tag{4.14}$$

où B_j représente une sous-bande particulière et Π_k représente une trame vocale particulière. Dans ce travail, ces caractéristiques sont calculées, en utilisant un signal vocal en temps discret, comme suit :

$$ECC_k(j) = \sum_{i=1}^{N} \frac{1}{L_f} \sum_{n=1}^{L_f} a_k^2(i,n)\mathbb{1}_{B_j}(f_k(i,n)),\ j = 1, ..., 12 \tag{4.15}$$

Les coefficients cepstraux d'énergie pondérée en fréquence (EFCC)

Les EFCC sont donnés en temps continu par [89] :

$$EFCC(B_j, \Pi_k) = \int_{f \in B_j} f h_{j,k}(f)\mathrm{d}f,\ j = 1, ..., 12 \tag{4.16}$$

où B_j représente une sous-bande particulière et Π_k représente une trame vocale particulière. Dans le cas discret, EFCC est calculée comme suit :

$$EFCC_k(j) = \sum_{i=1}^{N} \frac{1}{L_f} \sum_{n=1}^{L_f} f_k(i,n) a_k^2(i,n)\mathbb{1}_{B_j}(f_k(i,n)),\ j = 1, ..., 12 \tag{4.17}$$

où i représente un IMF particulier, j représente une bande fréquentielle et k représente une trame particulière.

Dans notre travail, nous extrayons les 12 premiers coefficients ECC et EFCC où les signaux vocaux sont échantillonnés à 16 KHz. Pour chaque coefficient, nous calculons la moyenne, la variance, le coefficient de variation, le Kurtosis et l'asymétrie. Chaque vecteur de ECC et EFCC est composé de 60 caractéristiques.

4.4.2 Les caractéristiques de modulation AM-FM

4.4.2.1 Les caractéristiques de modulation d'amplitude (MAF)

Les caractéristiques spectrales de modulation (MSF) présentées dans 4.3 sont extraites spécifiquement pour résoudre les problèmes de caractéristiques spectrales à court terme (MFCC) et pour mieux modéliser la nature de la perception auditive humaine. La méthode est basée sur l'émulation du traitement Spectro-temporal (ST) effectué dans le système auditif humain et prend en compte la fréquence acoustique régulière conjointement à la fréquence de modulation. Ces caractéristiques sont basées sur la décomposition du signal de parole par un banc de filtres auditifs et le calcul de l'enveloppe de Hilbert sur chaque bande. L'enveloppe de Hilbert est basée sur la transformée de Fourier (FT) et certaines des limitations de FT lui sont donc associées.

Pour cette raison, nous proposons dans ce travail une nouvelle méthode d'extraction de ces caractéristiques basée sur l'EMD utilisée conjointement avec TKEO. Les étapes d'extraction de ces caractéristiques sont décrites dans la Figure 4.8. Après l'application de TKEO sur des IMFs décomposées au moyen d'EMD, un groupe de filtres de modulation est appliqué à l'amplitude instantanée. Nous utilisons un banc de filtres de modulation avec M=8 filtres, de type passe-bas, dont les fréquences centrales allant de 4 à 512 Hz. Ces filtres sont, de type passe-bande de second ordre avec un facteur de qualité de 2, comme suggéré dans [86]. En appliquant le banc de filtres de modulation à chaque sortie $a_k(i,n)$, M sorties $a_k(i,j,n)$ sont générées avec j dénote le $j^{\text{ème}}$ filtre de modulation ($1 \leq j \leq M$). Les expériences ont montré que le banc de filtres a instauré un bon équilibre entre la performance et la complexité du modèle. Les contenus spectraux des signaux de modulation sont appelés spectres de modulation, et les caractéristiques proposées sont appelées caractéristiques de modulation d'amplitude (MAF pour modulation amplitude features en anglais). L'énergie, prise sur toutes les trames de la bande spectrale, fournit une caractéristique ($E(i,j)$). L'énergie dans chaque bande spectrale est définie comme :

$$E(i,j) = \sum_{k=1}^{N_f} E_k(i,j) \tag{4.18}$$

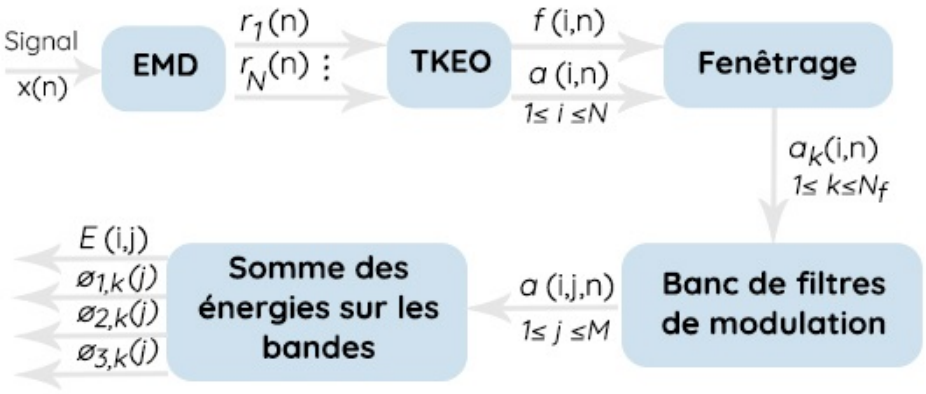


FIGURE 4.8: Étapes de l'extraction des coefficients MAF

où N_f est le nombre de trames pour $1 \leq j \leq 8$ et $E_k(i,j)$ est l'énergie sur les bandes définie par :

$$E_k(i,j) = \sum_{n=1}^{L_f} a_k^2(i,j,n) \tag{4.19}$$

Pour chaque trame k, $E(i,j)$ est normalisée en énergie unitaire avant un calcul ultérieur :

$$\sum_{i,j} E_k(i,j) = 1$$

Trois mesures spectrales Φ_1, Φ_2 et Φ_3 sont ensuite calculées par fenêtre [86]. La première mesure est **l'énergie spectrale** $\Phi_{1,k}(j)$ (spectral energy en anglais) , définie comme la moyenne des échantillons d'énergie appartenant à la $j^{\text{ème}}$ bande de modulation ($1 \leq j \leq 8$) :

$$\Phi_{1,k}(j) = \frac{\sum_{i=1}^{N} E_k(i,j)}{N} \tag{4.20}$$

La planeïté spectrale $\Phi_{2,k}(j)$ (spectral flatness en anglais) est le rapport entre la moyenne géométrique et la moyenne arithmétique de l'énergie spectrale ($\Phi_{1,k}(j)$). $\Phi_{2,k}(j)$ est donc défini comme suit :

$$\Phi_{2,k}(j) = \frac{\sqrt[N]{\prod_{i=1}^{N} E_k(i,j)}}{\Phi_{1,k}(j)} \tag{4.21}$$

Dans notre travail, $\Phi_{2,k}(j)$ est exprimée sur une échelle logarithmique comme suit :

$$\log \Phi_{2,k}(j) = \frac{1}{N} \sum_{i=1}^{N} \log E_k(i,j) - \log \Phi_{1,k}(j)$$

La dernière mesure utilisée est **le centroïde spectral** $\Phi_{3,k}(j)$ (spectral centroïd en anglais) qui fournit une mesure du "centre de masse" du spectre dans chaque bande de modulation. Pour la $j^{\text{ème}}$ bande de modulation, $\Phi_{3,k}(j)$ est défini comme suit :

$$\Phi_{3,k}(j) = \frac{\sum_{i=1}^{N} f(i)E_k(i,j)}{\sum_{i=1}^{N} E_k(i,j)} \qquad \text{où} \quad f(i) = i. \tag{4.22}$$

Dans [86], les auteurs expérimentent deux types similaires de mesure de fréquence $f(i)$. Dans notre cas, nous avons choisi la méthode la plus simple : l'indice du filtre de bande critique, c'est-à-dire, $f(i) = i$. Nous avons constaté une corrélation considérable entre deux bandes de modulation adjacentes. Afin de réduire la forte corrélation pour la planéité spectrale et le centroïde spectral, $\Phi_{2,k}(j)$ et $\Phi_{3,k}(j)$ ne sont calculés que pour $j \in \{1, 3, 5, 7, 8\}$. La moyenne de l'énergie dans chaque bande spectrale est calculée. Parallèlement à cette statistique, la moyenne et la variance de l'énergie spectrale, de la planeïté spectrale et du centroïde spectral sont évaluées et utilisées comme des coefficients.

La Figure 4.9 montre la moyenne de l'énergie $E(i,j)$ pour les 7 émotions (colère, ennui, dégoût, peur, joie, tristesse et neutre) de la base de données en allemand. Chaque $E(i,j)$ représente la moyenne sur toutes les fenêtres de tous les locuteurs pour une émotion. Cette figure confirme les effets des émotions sur l'intensité présentées dans le tableau 2.1. Nous constatons que la répartition de l'énergie sur le plan, fréquence de modulation et IMF, est similaire pour certaines émotions qui pourraient devenir des paires confondues, comme la colère et la joie, la peur et la tristesse. Mais il est très distinct pour d'autres émotions, comme la colère et le neutre, elles pourraient être bien discriminées les unes des autres.

La Figure 4.10 montre que les émotions tristesse et neutre ont significativement une énergie à basse fréquence que l'émotion colère, ce qui est à peu près la même chose dans des travaux antérieurs [86]. Pour la tristesse et le neutre, la dispersion est plus grande. Cependant, les émotions, plus expressives telles que la colère, présentent des courbes plus serrées, ce qui est à peu près la même chose dans la figure décrite précédemment.

4.4.2.2 Les caractéristiques de modulation de fréquence (MFF)

Des études sur la perception de la parole ont montré que les informations perceptuelles les plus importantes se trouvent dans les basses fréquences de modulation [102]. Dans cette section, nous allons illustrer comment l'analyse de ces fréquences de modulation peut être formulée et

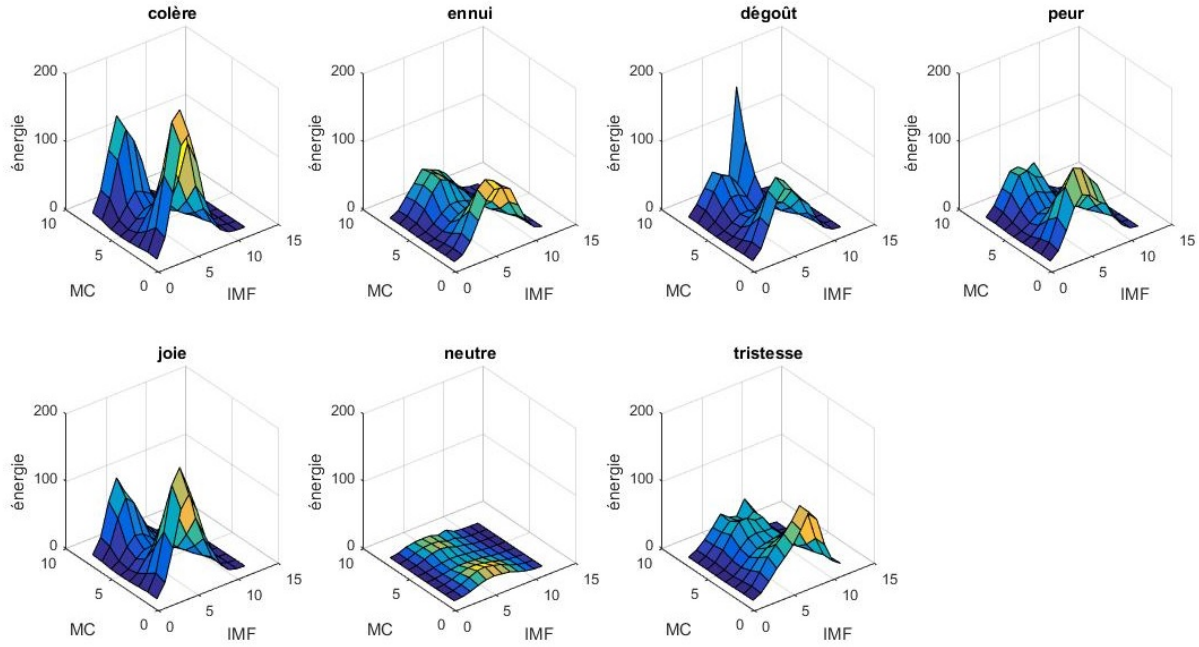


FIGURE 4.9: Moyenne de l'énergie $E(i,j)$ pour les 7 émotions ; Pour chaque émotion, la moyenne de $E_k(i,j)$ est calculée sur toutes les fenêtres pour tous les locuteurs de cette émotion ; "MC" désigne les bandes de fréquence de modulation.

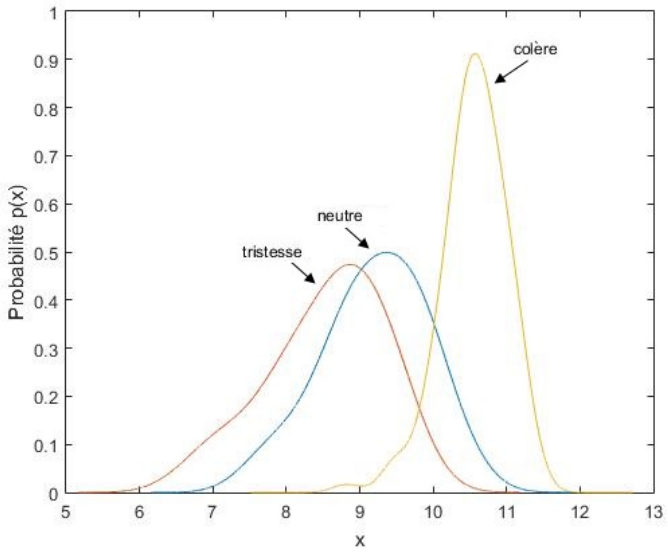


FIGURE 4.10: Estimation de la densité de probabilité de $\overline{\Phi}_3$ pour les 3 émotions (tristesse, neutre et colère)

appliquée au problème du RAE. Le modèle d'un signal AM-FM est utilisé pour cette analyse. Les étapes de calcul basé sur la décomposition AM-FM du signal d'entrée sont illustrées par la Figure 4.11. Le signal vocal est décomposé en plusieurs IMF par l'EMD. L'amplitude et la fréquence instantanées de chaque IMF sont estimées à l'aide du TKEO. Ensuite, ils sont utilisés pour obtenir des estimations à long terme (256 ms) de la fréquence instantanée et de

la largeur de bande moyenne sur une trame. Les vecteurs acoustiques sont constitués de ces descripteurs accompagnés de leurs statistiques : moyenne, minimum, maximum et l'écart type. La reconnaissance est sensiblement améliorée par la présence de ces données statistiques [26]. La

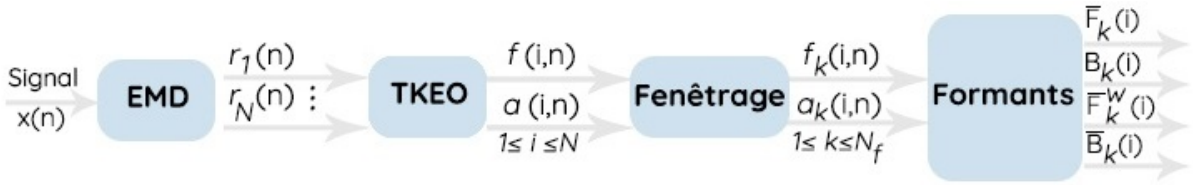


FIGURE 4.11: Étapes de l'extraction des coefficients MFF

fréquence instantanée pondérée en amplitude moyenne (F_w) et la bande passante instantanée pondérée en amplitude moyenne (B_w) sont décrites en temps continu dans [89] par :

$$F_w = \frac{\int_{t_0}^{t_0+T} f(t)a^2(t)\mathrm{d}t}{\int_{t_0}^{t_0+T} a^2(t)\mathrm{d}t} \tag{4.23}$$

$$B_w^2 = \frac{\int_{t_0}^{t_0+T} [\{\dot{a}(t)/2\pi\}^2 + \{f(t) - F_w\}^2 a^2(t)]\mathrm{d}t}{\int_{t_0}^{t_0+T} a^2(t)\mathrm{d}t} \tag{4.24}$$

où t_0 et T représentent le début et la durée de la trame analysée, $f(t)$ et $a(t)$ représentent respectivement la fréquence et l'amplitude instantanées de chaque signal AM-FM. Dans ce travail, nous estimons à long terme, en utilisant l'amplitude et la fréquence instantanée, la fréquence instantanée moyenne ($\overline{F}$), la largeur de bande instantanée moyenne (B), la fréquence instantanée moyenne pondérée en amplitude ($\overline{F}_k^w$) et la bande passante instantanée pondérée en amplitude moyenne (B_k^w) :

$$\overline{F}_k(i) = \frac{1}{L_f}\sum_{n=1}^{L_f} f_i(k,n) \tag{4.25}$$

$$B_k(i) = \sqrt{\frac{1}{L_f}\sum_{n=1}^{L_f}(f_k(i,n) - \overline{F}_k(i))} \tag{4.26}$$

$$\overline{F}_k^w(i) = \frac{\sum_{n=1}^{L_f} f_k(i,n)a_k^2(i,n)}{\sum_{n=1}^{L_f} a_k^2(i,n)} \tag{4.27}$$

$$B_k^w(i) = \sqrt{\frac{\sum_{n=1}^{L_f}\{\dot{a}_k(i,n)/2\pi\}^2 + \{f_k(i,n) - \overline{F}_k^w(i)\}^2 a_k^2(i,n)}{\sum_{n=1}^{L_f} a_k^2(i,n)}} \tag{4.28}$$

où i représente un IMF particulier, L_f représente un nombre d'échantillons par fenêtre, f_k et a_k

représentent respectivement la fréquence instantanée et l'amplitude instantanée dans une trame vocale particulière k. $\dot{a}(n)$ est définie par :

$$\dot{a}(n) = a(n+1) - a(n) \tag{4.29}$$

Dans ce travail, pour chaque IMF la moyenne, le maximum et le minimum de B, B^w, $\overline{F}$, $\overline{F}^w$) ont été calculés pour chaque caractéristique et ont été utilisés pour la classification. La figure 4.12 montre les valeurs moyennes de la fréquence instantanée ($\overline{F}$), pour l'ensemble de 535 énoncés, correspondant à chaque fichier de la base de données en allemand. Nous remarquons que la fréquence fondamentale se situe entre IMF10 - IMF12. Nous pouvons voir l'utilité de l'analyse

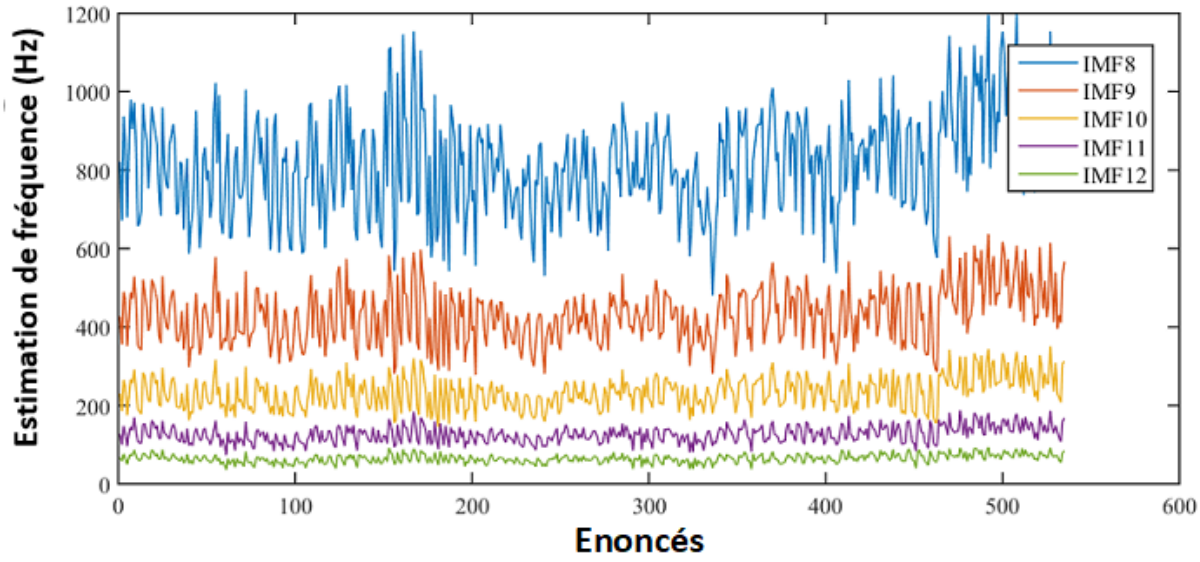
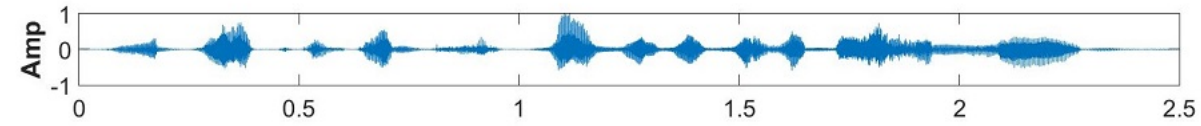
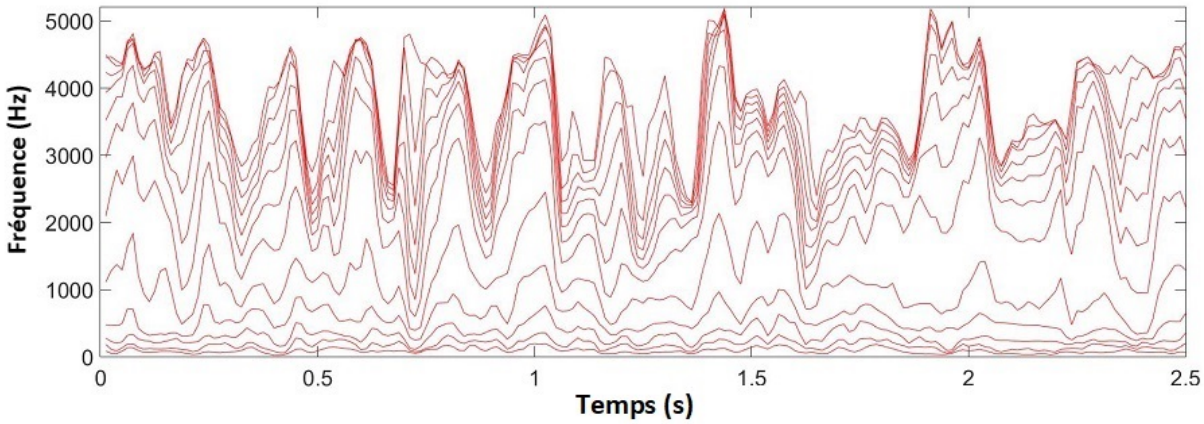


FIGURE 4.12: Estimation de la fréquence moyenne de tous les locuteurs allemands.

AM-FM à partir de la représentation temps-fréquence d'un signal vocal présentée aux figures 4.13 et 4.14. Pour chaque composante IMF, une estimation de fréquence à long terme $\overline{F}_k(i)$ est obtenue à chaque trame en utilisant l'équation 4.25. La durée de la trame analysée est de 256 ms avec un chevauchement de 128 ms. Les figures ci-dessous indiquent que l'analyse AM-FM pourrait fournir des informations utiles pour discriminer les émotions dans le signal vocal. Elles identifient des régions avec des zones denses dans le but de suivre les fréquences des formants et leurs largeurs de bande [89]. Nous pouvons remarquer sur les figures 4.13 et 4.14 que les zones denses dans la représentation temps-fréquence sont différentes entre le signal vocal de l'émotion "colère" et le signal vocal de l'émotion "neutre", cette différence est due à l'émotion. La densité de probabilité estimée (Figure 4.15) confirme cette hypothèse. La Figure 4.12 montre que l'IMF10 s'approche le mieux de la fréquence fondamentale qui se situe à l'environ de 200 Hz. Dans ce qui suit, nous estimons la densité de probabilité de cette IMF. Nous constatons à



(a)



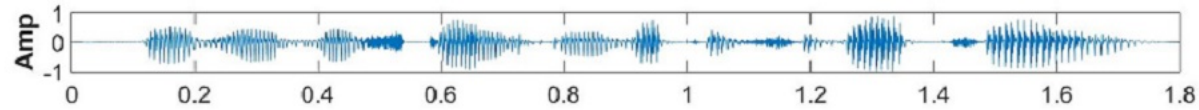
(b)

FIGURE 4.13: (4.13a) Signal d'un fichier pour l'émotion "colère" ; (4.13b) La représentation temps-fréquence de (4.13a) en utilisant 12 IMFs. La taille de trame est de 256 ms avec un chevauchement de 128 ms est utilisée.

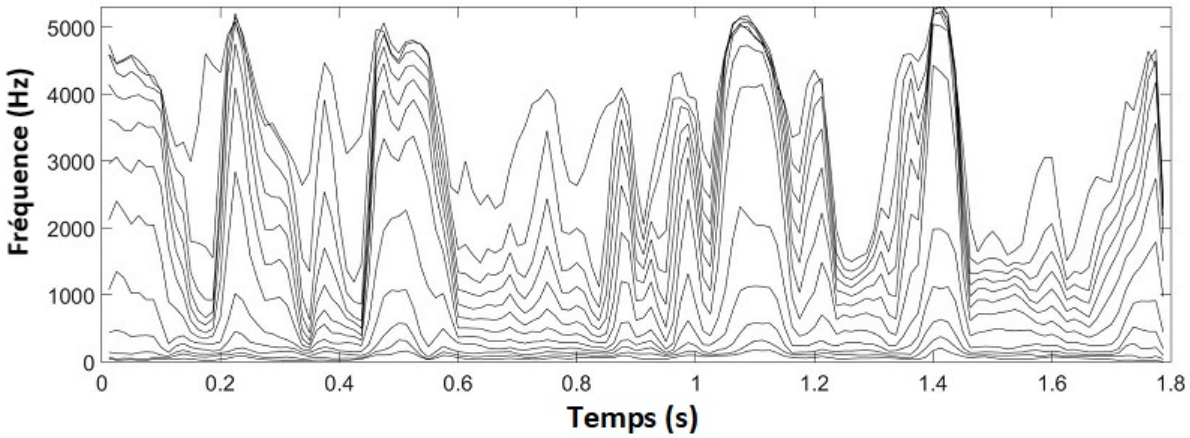
partir de la Figure 4.15 que la courbe de l'émotion "colère" se déplace vers des hautes fréquences (entre 200 et 400 Hz). Cependant, les émotions "neutre" et "tristesse" vers les basses fréquences. Ces résultats confirment les conclusions des recherches qui démontrent que la moyenne de F0 d'une personne en colère est plus haute que celle d'une personne triste tristesse (plus bas) (voir tableau 2.1). Ces résultats sont aussi pratiquement identiques à ceux de la figure 4.10.

La Figure 4.16 montre l'estimation de la densité de probabilité de la fréquence instantanée moyenne pondérée en amplitude $\overline{F}^w$. Nous pouvons constater que la fréquence instantanée pondérée en amplitude présente une meilleure capacité discriminatoire. Ces résultats sont aussi pratiquement identiques à ceux du tableau 2.1. Nous avons testé 3 différentes tailles de trame de court terme à long terme (25 ms, 256 ms et 512 ms). Les résultats montrent que la fréquence instantanée moyenne et la fréquence instantanée pondérée en amplitude sont plus discriminante sur une trame de 256 ms (voir Figures 4.17 et 4.18).

Plusieurs études phonétiques portent sur les différences qui existent entre hommes et femmes. Dans [29] et [103], les auteurs ont montré qu'il y a une différence significative entre hommes et femmes concernant la variation de la fréquence fondamentale chez les locuteurs allemands (en moyenne vers 100 Hz chez l'homme et 200 Hz chez la femme). Une illustration de cette différence est visible ci-dessous avec les figures 4.19 et 4.20.



(a)



(b)

FIGURE 4.14: (4.14a) Signal d'un fichier pour l'émotion "neutre" ; (4.14b) La représentation temps-fréquence de (4.14a) en utilisant 12 IMFs. La taille de trame est de 256 ms avec un chevauchement de 128 ms est utilisée.

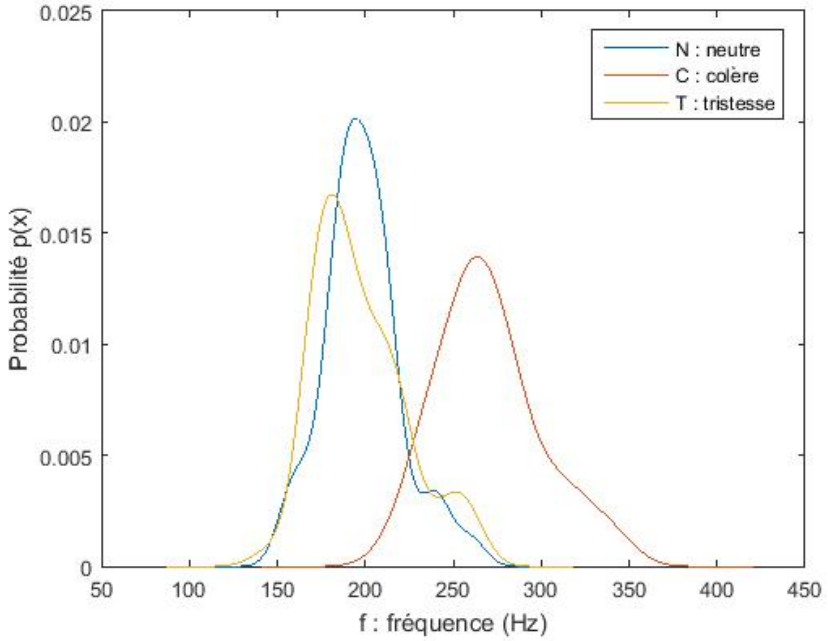


FIGURE 4.15: Estimation de la densité de probabilité de $\overline{F}$ de l'IMF10 pour les 3 émotions (tristesse, neutre et colère).

Comme nous pouvons le voir sur les figures 4.19 et 4.20, il y a une différence entre les plages de variation de la voix d'un homme et celles d'une femme.

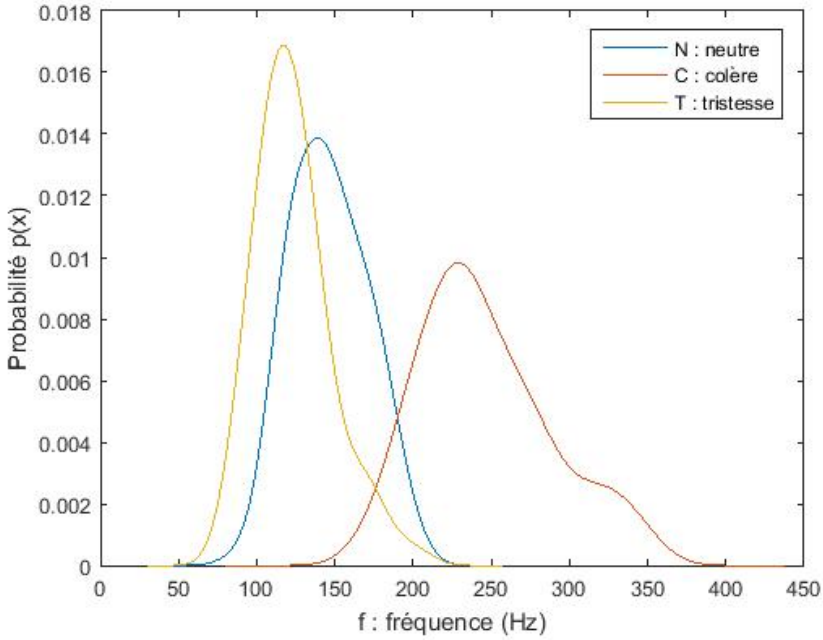
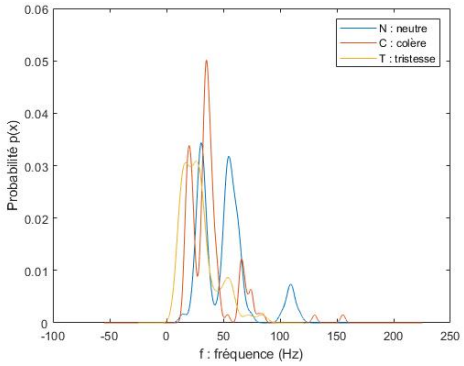
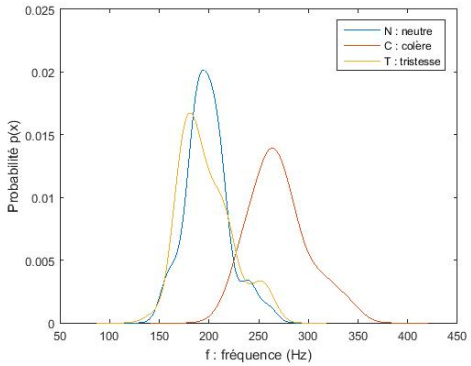


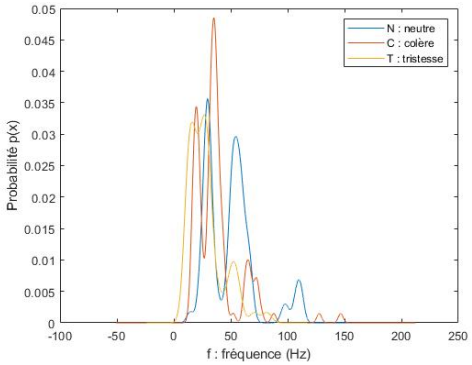
FIGURE 4.16: Estimation de la densité de probabilité de $\overline{F}^w$ de l'IMF10 pour les 3 émotions (tristesse, neutre et colère).



(a) Taille de la trame est de 25 ms.

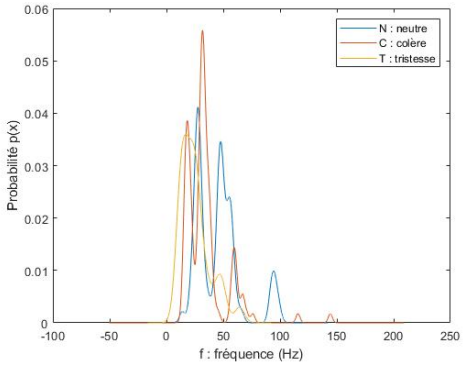


(b) Taille de la trame est de 256 ms.

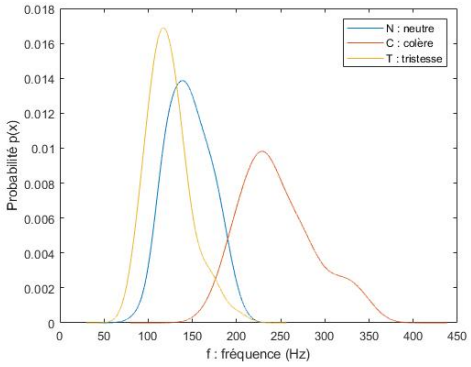


(c) Taille de la trame est de 512 ms.

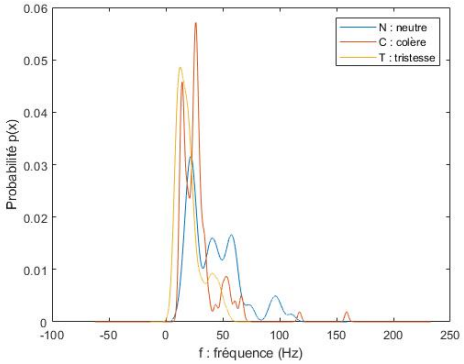
FIGURE 4.17: Estimation de la densité de probabilité de $\overline{F}$ de l'IMF10 pour les 3 émotions (tristesse, neutre et colère), pour 3 différentes tailles de trame.



(a) Taille de la trame est de 25 ms.

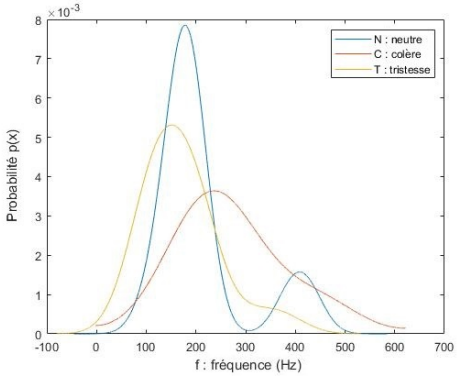


(b) Taille de la trame est de 256 ms.

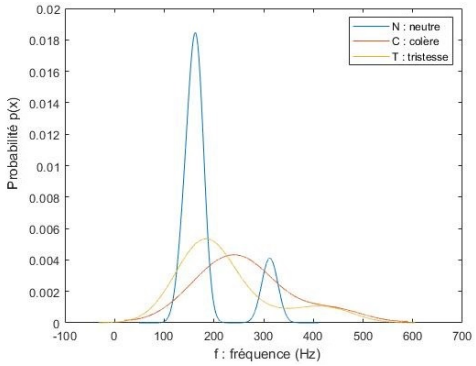


(c) Taille de la trame est de 512 ms.

FIGURE 4.18: Estimation de la densité de probabilité de $\overline{F}^w$ de l'IMF10 pour les 3 émotions (tristesse, neutre et colère), pour 3 différentes tailles de trame.



(a) Représentation de $\overline{F}$ d'une femme

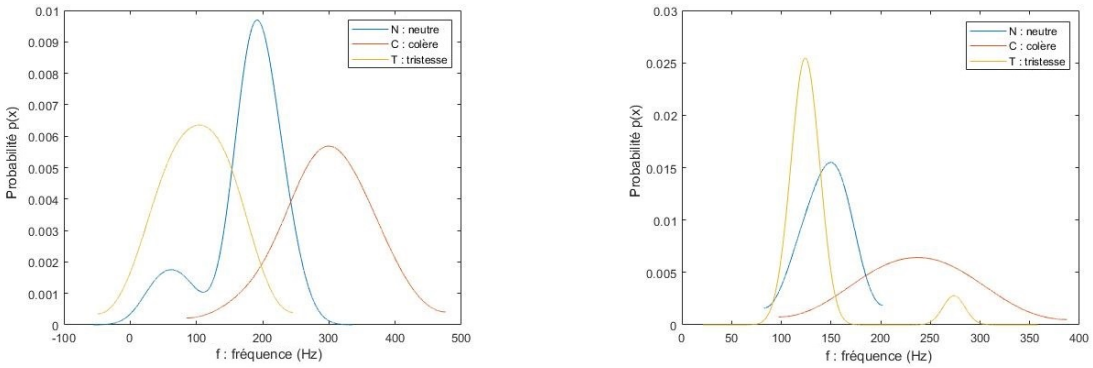


(b) Représentation de $\overline{F}$ d'un homme

FIGURE 4.19: Estimation de la densité de probabilité de $\overline{F}$ de l'IMF10 d'un homme et d'une femme pour les 3 émotions (tristesse, neutre et colère).

4.5 Conclusion

Dans ce chapitre, nous avons proposé un ensemble de caractéristiques basées sur la décomposition EMD pour l'analyse de la voix, appelées les caractéristiques de modulation d'amplitude (MAF) et les caractéristiques de modulation de fréquence (MFF). Les caractéristiques les plus couram-



(a) Représentation de $\overline{F}^w$ d'une femme

(b) Représentation de $\overline{F}^w$ d'un homme

FIGURE 4.20: Estimation de la densité de probabilité de $\overline{F}^w$ de l'IMF10 d'un homme et d'une femme pour les 3 émotions (tristesse, neutre et colère).

ment utilisées dans les systèmes de RAE (à savoir les MFCC, les caractéristiques de modulation spectrales...) extraites à des fins de comparaison et de complémentarité ont également été décrites. Le tableau ci-dessous décrit le nombre de caractéristiques extraites dans notre travail pour les deux bases de données.

Caractéristique	Nombre de caractéristiques	
	Base allemande	Base espagnole
MFCC	60	60
MSF	95	95
SMFCC	60	60
ECC	60	60
EFCC	60	60
MFF	96	72
MAF	132	108

TABLE 4.1: Description du nombre de caractéristiques extraites.

Nous avons fait plusieurs tests pour déterminer le nombre optimal des IMFs permettant de reconstruire le signal sans perte d'informations importantes pour chaque base de données. Nous avons vu qu'après un certain nombre d'itérations, les IMF ne fournissent pas d'informations. C'est pourquoi le signal dans notre travail est reconstruit à partir des 9 premiers modes (9 IMFs) pour la base espagnole et des 12 premiers modes (12 IMFs) pour la base allemande. Nous pensons que cette différence est due au fait que les deux bases n'ont pas la même langue. Ces IMFs ont ensuite été utilisés pour l'extraction de caractéristiques afin de reconnaître les émotions. La différence entre le nombre de caractéristiques (MAF, MFF) pour les deux bases de données est liée au nombre des IMFs (respectivement 12 et 9 IMFs pour les bases de données allemande et espagnole).

Dans le chapitre qui suit, nous allons décrire les méthodes utilisées pour la classification et

nous allons présenter le système de reconnaissance automatique des émotions que nous avons développé. Les résultats de la reconnaissance avec différents types de caractéristiques sont présentés.

Chapitre 5

Sélection de caractéristiques et classification

Contents

5.1 Introduction 84
5.2 Classification 85
5.2.1 Les algorithmes utilisés 85
5.2.2 Résultats et discussions 89
5.3 Fusion de caractéristiques 91
5.4 Normalisation par locuteur 93
5.5 Sélection de caractéristiques 95
5.5.1 Recursive Feature Elimination 97
5.5.2 Résultats et discussions 99
5.6 Comparaison avec les systèmes existants 100
5.7 Évaluation du corpus "French Student Emotionnal database" . . . 101
5.8 Conclusion 102

5.1 Introduction

Dans le présent chapitre, nous nous intéressons au développement d'un système de RAE à partir de la voix, expérimenté sur les corpus émotionnels décrits dans le chapitre 3. Ce système a été développé sous MATLAB et PYTHON. Dans ce qui suit, nous décrirons la méthodologie suivie pour concevoir ce système. Nous présenterons les modèles utilisés pour la classification notamment la méthode de régression linéaire (RL), les machines à support de vecteurs (SVM)

et les réseaux de neurones récurrents (RNN). Ensuite, nous allons donné une description de la méthode de sélection de caractéristiques (RFE pour recursive feature elimination en anglais) utilisée. Nous présentons au fur et à mesure les résultats des expérimentations.

5.2 Classification

Dans cette partie, nous allons décrire brièvement les méthodes d'apprentissage utilisées dans notre travail et nous présentons par la suite les résultats de classification de chaque caractéristique pour les 3 classifieurs et sur les 2 bases (allemande et espagnole).

5.2.1 Les algorithmes utilisés

Dans le chapitre 2, nous avons dressé un état de l'art sur les algorithmes de classification utilisés dans la littérature. Dans cette section seules les méthodes utilisées dans cette thèse seront présentées.

5.2.1.1 Régression linéaire

La régression linéaire (RL) pour la classification est basée sur la recherche de séparations linéaires entre les différentes classes, par minimisation des moindres carrés. Ce qui revient à chercher les hyperplans qui séparent les différentes classes à partir des données d'apprentissage. Ce modèle, au delà de la simplicité, s'avère efficace pour certains problèmes de classification où la séparation entre différentes classes peut être linéaire. Mais dans la plupart des cas cette propriété n'est pas assurée. L'algorithme utilisé est décrit ci-dessous (voir Algorithme 2). Le lecteur intéressé peut se référer aux références [104] et [56]. La RL est un algorithme simple et efficace pour la classification des données. Le seul problème est que la RL décrit des relations linéaires entre les données sauf que beaucoup des relations entre classes et caractéristiques sont pas linéaires.

5.2.1.2 Machines à support de vecteurs

Les machines à support de vecteurs (SVM pour Support Vector Machines en anglais) sont une classe d'algorithmes d'apprentissage initialement construits pour la classification binaire (à deux classes), après étendues au cas multiclasse. Ils reposent sur l'idée qu'il existe un hyperplan dans l'espace des caractéristiques pour lequel la distance entre les vecteurs de poids propres à chaque

Algorithme 2 Régression linéaire

Entrées : Modèles de classes $X_i \in \mathbb{R}^{q \times p_i}$, $i = 1, 2, ..., N$ et un vecteur de test $y \in \mathbb{R}^{q \times 1}$
(N : nombre de classes, p_i : nombre de données d'entrainement de la $i$ème classe et q : nombre de caractéristiques)
Sorties : Classe de y

1. Calculer l'estimateur $\hat{\beta}_i \in \mathbb{R}^{p_i \times 1}$ pour que $\hat{y}_i$ soit proche du vrai vecteur y au sens des moindres carrés : $\hat{\beta}_i = (X_i^T X_i)^{(-1)} X_i^T y, \quad i = 1, 2, ..., N$
2. Pour chaque classe, calculer la prédiction $\hat{y}_i$ associé à $\hat{\beta}_i$ de la façon suivante : $\hat{y}_i = X_i \hat{\beta}_i, \quad i = 1, 2, ..., N$;
3. Calculer la distance entre le vecteur des observations et le vecteur des prédictions : $d_i(y) = \mid y - \hat{y}_i \mid, \quad i = 1, 2, ..., N$;
4. Choisir la classe qui a la distance minimale $d_i(y)$

donnée les plus proches de cet hyperplan est maximale. Ces vecteurs sont nommés vecteurs de support et la distance séparant l'hyperplan de ces points est appelée « marge ». Dans le cas où les données ne sont pas linéairement séparables, l'espace de représentation des données d'entrées est projeté dans un espace de plus grande dimension, dans lequel il est probable qu'il existe une séparation linéaire. Ceci est réalisé en utilisant une fonction noyau[1] [36]. Plusieurs fonctions noyaux sont utilisés avec les SVM : linéaire, gaussien, polynomial, etc. Le noyau polynomial, utilisé dans toutes les expériences de classification SVM suivantes, est défini comme suit [36] :

$$K(x_i, x_j) = (\langle x_i, x_j \rangle + c)^d \tag{5.1}$$

où d est le degré de polynôme, x_i et x_j sont des vecteurs de caractéristiques calculées à partir d'échantillons d'apprentissage ou de test, et $c \geq 0$ est un paramètre libre. Ce choix de noyau nous a permis de gagner en précision.

Il existe plusieurs manières d'effectuer la classification multiclasses (par exemple one-versus-one, one-vesus-all). Ici, nous introduisons seulement la méthode un contre un (one-versus-one en anglais) qui est largement utilisée par la communauté [36]. Cette méthode construit $C(C-1)/2$ classifieurs binaires, formés chacun à l'aide de deux (une paire) des classes C [36]. La stratégie de vote est utilisée pour attribuer à un échantillon de test, analysé par chaque classifieur, la classe avec le plus de votes. La librairie Matlab implémentant SVM, utilisé dans notre travail, est disponible en accès libre dans [105].

Le choix du noyau et des autres paramètres est souvent un problème pratique et critique lors de la mise en œuvre des SVM. Nous avons utilisé la méthode de validation croisée pour le choix de noyau et des autres paramètres. Les meilleurs résultats sont obtenus en utilisant une SVM avec

1. Une fonction noyau (Kernel) transforme un produit scalaire dans un espace de grande dimension.

un noyau polynomial de degré 2 et de paramètre lambda=0.001, la constante de marge douce C =1.0 a été choisie. Le modèle peut reconnaître 7 classes différentes, avec une classification à la one-vs-one.

Les SVM sont beaucoup plus efficaces que d'autres algorithmes [36] (par exemple RNN) quand on ne dispose que de peu de données d'entraînement. Cependant, quand les données sont trop nombreuses, les SVM ont tendance à baisser en performance [106], [107].

5.2.1.3 Réseaux de neurones récurrents

Un réseau de neurones récurrents (RNN pour Recurrent Neural Network en anglais) est un réseau de neurones qui est capable de traiter une entrée de longueur variable (donnée séquentielle) telle que les séries chronologiques, l'audio, la vidéo, la parole, les textes, les données météorologiques, etc.,. Ceci est effectué en renvoyant la sortie d'une couche de réseau neuronal à l'instant t à l'entrée de la même couche de réseau à l'instant $t+1$. La figure 5.1 présente le schéma d'un réseau de neurones récurrents à une unité reliant l'entrée et la sortie du réseau :

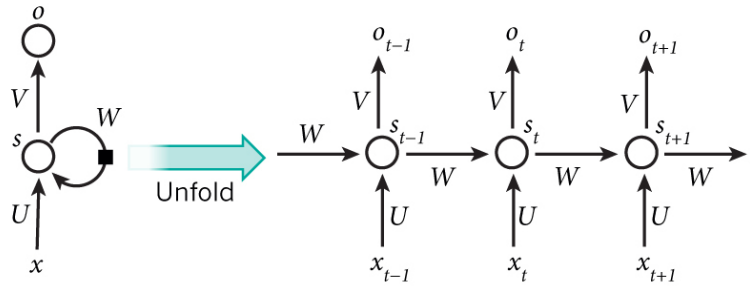


FIGURE 5.1: Schéma d'un réseau de neurones récurrents à une unité reliant l'entrée et la sortie du réseau. A droite la version « dépliée » de la structure [108].

Contrairement aux réseaux de neurones traditionnels qui utilisent des paramètres différents à chaque couche, le RNN partage les mêmes paramètres (U, V et W) à toutes les étapes. La formule et les variables d'état cachées sont comme suit :

$$s_t = f(Ux_t + Ws_{t-1}) \tag{5.2}$$

$$o_t = Vs_t \tag{5.3}$$

Avec :

— x_t, s_t et o_t sont respectivement le vecteur d'entrée, vecteur de la couche cachée et le vecteur de sortie à l'instant t ;

— U, V, W sont les matrices et vecteur (paramètres) ;

— f est la fonction d'activation.

Long short-term memory

Un réseau de neurones récurrents à mémoire court-terme et long terme (LSTM pour Long short-term memory en anglais) est une sorte de réseau de neurones récurrents [109]. Grâce à leur mémoire interne, les LSTM sont capables de se rappeler des informations importantes sur l'entrée reçue, ce qui leurs permet d'être très précis dans la prédiction de ce qui va suivre.

Un réseau LSTM est composée d'une mémoire (c) et de trois portes. La porte d'entrée "IN" (input) doit choisir les informations pertinentes qui seront transmises à la mémoire. La sortie "OUT" (output) protège le réseau du contenu de sa mémoire. La porte d'oubli f (forget) permet à l'unité de remettre à zéro le contenu de sa mémoire.

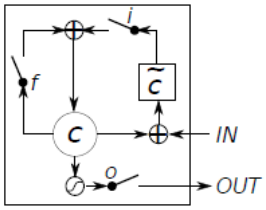


FIGURE 5.2: Long Short-Term Memory (LSTM) avec c et $\tilde{c}$ sont respectivement la mémoire et le nouveau contenu de la mémoire.

Le modèle que nous avons utilisé, illustré dans la figure 5.3, est formé d'une succession de 2 couches LSTM suivies d'une couche dense[2] (avec une fonction d'activation de type "tanh"). La succession de plusieurs couches permet au modèle d'apprendre des représentations temporelles de niveau supérieur. L'entrée est une séquence de vecteurs (sous forme d'une matrice). Chaque vecteur de descripteurs (vecteur à n éléments avec n correspond au nombre de caractéristiques) est utilisé séparément pour entrainer ce modèle. La couche de sortie est un vecteur de 7 unités, de façon à ce que chaque neurone correspond à une émotion à prédire. On normalise la prédiction via une fonction "softmax" pour obtenir une distribution de probabilité. Les RNN permettent de mieux comprendre une séquence et son contexte par rapport aux autres algorithmes. C'est la raison pour laquelle ils sont préférés pour ce type de données. Le seul inconvénient d'un algorithme de réseaux neuronaux est que son apprentissage peut prendre beaucoup de temps, plus

2. Une couche dense : est une couche de projection linéaire (couche de neurones cachés complètement connectée). Dense parce que tous les neurones de couche précédente seront connectés à tous les neurones de la couche suivante [110].

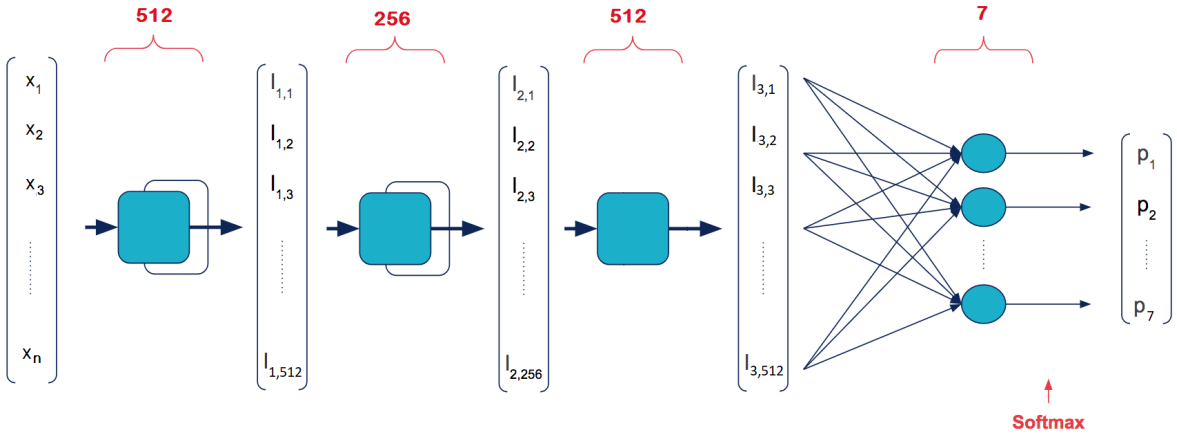


FIGURE 5.3: Notre modèle de réseau LSTM

particulièrement pour des grandes bases de données avec un grand nombre de caractéristiques. Ils ont également plus de paramètres que les autres algorithmes, ce qui signifie que le balayage de paramètres allonge grandement la durée d'apprentissage.

5.2.2 Résultats et discussions

Tous les résultats de la classification sont obtenus en utilisant la méthode de « K-fold cross-validation ». Cette méthode est une pratique couramment utilisée pour évaluer des modèles d'apprentissage automatique [34]. Cela fonctionne en divisant d'abord les données en K sous échantillons. Un de ces K sous échantillons est utilisé pour le test du modèle et les K-1 restants pour l'apprentissage du modèle. Ce processus est répété K fois de sorte que chacun des K sous-échantillons est utilisé exactement une fois pour le test. La validation croisée des K-fold formera et évaluera K modèles et donnera K scores (précision ou erreur). Ensuite la moyenne et l'écart type de ces scores, pour obtenir une statistique des performances du modèle, peuvent être calculées. L'avantage de cette méthode est que toutes les données sont utilisées à la fois pour l'entraînement et le test. La « 10-fold cross validation » est une méthode simple à comprendre et aboutit généralement à une estimation moins biaisée du modèle que d'autres méthodes, comme la simple répartition par train/test.

La normalisation des caractéristiques est une étape indispensable pour un modèle de Machine learning (ML) [36]. Cette normalisation permet d'éviter d'avoir des caractéristiques avec des valeurs numériques élevées, et des caractéristiques avec des valeurs numériques faibles. En effet, il est possible que la variation d'échelle de valeurs entre chaque caractéristique ralentisse le calcul et empêche le classifieur utilisé d'apprendre efficacement à résoudre le problème de

classification. Comme suggéré dans [36], les caractéristiques sont normalisées de manière linéaire entre $[-1, 1]$ avant de faire une classification. Les 3 classifieurs présentés ci dessus sont évalués en utilisant les différentes caractéristiques individuellement. Les résultats sont présentés dans les tableaux suivants. Chacun des tableaux illustre les performances de classification sur les 2 corpus émotionnels. Comme le montrent les tableaux 5.1 et 5.2, le classifieur SVM utilisant les SMFCC donne de meilleurs résultats 73,01% et 95.74% respectivement pour la base allemande et la base espagnole. Nous constatons aussi que les SMFCC donnent de meilleurs scores que les MFCC avec le SVM (pas de différence significative pour les autres classifieurs) pour la base allemande et que pour la base espagnole les SMFCC ont de meilleurs précisions que les MFCC pour les deux classifieurs RL et SVM. Les résultats montrent aussi que les caractéristiques

Caractéristique	Méthode	Taux de reconnaissance (%)							
		Peur	Dégoût	Joie	Ennui	Neutre	Tristesse	Colère	Moy. (σ)
MFCC	RL	62.10	59.58	48.43	61.12	66.71	85.84	87.52	67.92 (6.41)
MSF		45.93	31.61	39.01	76.52	52.36	85.74	79.97	61.69 (8.15)
SMFCC		63.10	65.09	52.99	63.72	63.48	81.50	79.78	67.16 (8.01)
ECC		56.28	36.54	37.29	36.82	49.91	71.35	76.64	54.33 (7.43)
EFCC		55.63	40.48	46.26	29.54	45.63	68.92	84.41	55.28 (5.62)
MAF		55.42	32.98	24.65	78.33	50.49	73.66	80.87	59.05 (6.22)
MFF		47.21	47.16	32.71	55.75	37.57	78.81	76.54	56.03 (3.56)
MFCC	SVM	74.68	62.58	66.49	49.10	60.60	91.68	82.47	70.37 (7.06)
MSF		61.81	54.95	44.04	80.86	66.44	82.63	77.35	67.92 (4.87)
SMFCC		75.37	62.45	76.02	53.95	69.57	87.66	85.16	**73.01** (5.48)
ECC		59.52	48.09	43.20	47.59	49.16	49.85	64.67	53.77 (6.30)
EFCC		58.89	58.88	55.59	52.18	47.56	61.34	66.49	56.41 (7.57)
MAF		66.45	72.21	14.73	73.82	65.32	80.47	69.55	63.96 (7.57)
MFF		66.98	51.83	43.46	41.78	52.64	73.14	72.67	59.43 (6.72)
MFCC	RNN	76.40	62.50	58.90	55.60	53.20	88.50	81.30	68.88 (6.07)
MSF		40.08	40.00	40.04	77.02	72.04	75.07	77.01	64.47 (4.82)
SMFCC		73.50	62.00	52.40	51.90	59.50	83.70	82.80	68.12 (6.15)
ECC		52.50	43.00	42.60	32.50	48.40	54.60	66.30	51.73 (4.51)
EFCC		44.00	51.50	39.80	34.70	46.60	61.20	74.90	51.60 (6.83)
MAF		62.50	30.00	32.60	66.80	57.20	77.10	77.70	61.04 (7.40)
MFF		42.60	36.50	48.50	32.00	51.00	66.90	58.80	53.68 (9.44)

Table 5.1: Récapitulatif des résultats des 3 classifieurs avec les différentes caractéristiques sur la base allemande ; "Moy." désigne le taux de reconnaissance moyen ; σ désigne l'écart type des 10 précisions de validation croisée.

proposées (MAF, MFF) donnent des taux de reconnaissance satisfaisant mais une combinaison de ces caractéristiques avec les autres permet d'obtenir de meilleurs résultats. La différence importante entre les résultats des 2 corpus s'explique par la différence de taille. En effet, la base allemande ne contient que 535 fichiers tandis que la base espagnole en contient 6041 fichiers. Pour conclure, l'algorithme SVM (linéaire) plus performant que RNN ou RL en utilisant chaque

Caractéristique	Méthode	Taux de reconnaissance (%)							
		Peur	Dégoût	Joie	Ennui	Neutre	Tristesse	Colère	Moy. (σ)
MFCC	RL	67.85	61.41	75.97	60.17	95.79	71.89	84.94	77.21 (0.76)
MSF		67.72	44.04	68.78	46.95	89.58	63.10	78.49	69.22 (1.37)
SMFCC		86.54	82.00	85.45	73.07	98.39	82.97	97.00	88.26 (0.95)
ECC		45.02	45.50	23.47	48.49	79.75	30.75	65.97	53.26 (2.62)
EFCC		46.04	44.95	23.94	42.19	80.29	30.06	68.32	52.79 (2.48)
MAF		57.59	50.11	63.01	50.19	81.48	36.08	79.00	63.04 (1.44)
MFF		32.11	40.73	24.33	46.65	81.69	20.83	59.22	49.35 (2.46)
MFCC	SVM	98.22	90.93	88.96	84.92	95.95	84.90	96.36	91.04 (1.04)
MSF		81.04	82.18	80.17	70.13	91.08	78.63	89.08	83.29 (1.15)
SMFCC		95.04	95.58	94.24	90.66	99.01	92.02	98.88	**95.74** (0.94)
ECC		64.43	70.55	61.57	67.71	79.09	52.11	84.71	69.68 (2.15)
EFCC		64.29	69.86	58.01	64.58	77.89	48.88	82.26	67.58 (2.55)
MAF		71.46	76.99	71.04	63.03	83.65	63.73	86.22	74.63 (1.30)
MFF		57.12	57.35	52.33	51.07	68.06	38.38	71.54	57.54 (1.85)
MFCC	RNN	86.04	83.01	85.00	77.70	95.30	80.40	93.60	87.30 (0.92)
MSF		76.00	76.40	76.00	66.10	92.30	82.90	89.00	81.66 (1.62)
SMFCC		83.80	86.50	81.60	83.00	86.80	79.10	93.90	87.80 (1.82)
ECC		55.60	65.40	49.20	65.20	74.50	46.90	77.20	63.93 (2.15)
EFCC		57.40	62.30	42.00	61.10	75.00	42.00	75.40	61.74 (1.25)
MAF		59.60	60.90	56.30	51.70	72.00	57.90	85.30	64.71 (2.46)
MFF		46.60	46.60	35.60	54.00	58.50	24.90	65.10	49.05 (3.49)

TABLE 5.2: Récapitulatif des résultats des 3 classifieurs avec les différentes caractéristiques sur la base espagnole ; "Moy." désigne le taux de reconnaissance moyen ; σ désigne l'écart type des 10 précisions de validation croisée.

caractéristique individuellement pour les 2 bases allemande et espagnole.

5.3 Fusion de caractéristiques

Dans la section précédente, nous avons testé individuellement les différents types de caractéristiques. En raison de l'utilisation généralisée des caractéristiques cepstraux (MFCC, ECC, EFCC et SMFCC) et spectrales (MSF) dans les systèmes de RAE, il est important d'étudier la contribution des caractéristiques (MAF et MFF) proposées en tant que caractéristiques complémentaires. Dans [111], l'auteur a montré que la reconnaissance automatique est optimisée par la recherche de la meilleure combinaison de caractéristiques.

Nous avons testé différentes combinaisons qui sont détaillées en annexe B (voir section B : tableaux B.1 et B.2). Dans le tableau 5.3, nous présentons seulement les combinaisons qui donnent les meilleurs scores. Le tableau 5.3 montre que la combinaison des nouvelles caractéristiques avec SMFCC, ECC et EFCC donne une meilleure performance de 80.56% sur la base allemande. En

outre, la combinaison des nouvelles caractéristiques avec SMFCC et EFCC fournit la meilleure discrimination à 96.24% pour la base espagnole (voir tableau 5.4). Nous constatons des tableaux

Caractéristique	Taux de reconnaissance (%)		
	RL	SVM	RNN
MFCC+MSF	73.20	76.60	63.67
SMFCC+ECC	65.47	76.60	72.95
SMFCC+MAF	68.11	76.98	72.59
SMFCC+MFF	62.45	76.03	68.88
SMFCC+MFF+MAF	59.43	76.60	73.81
EFCC+MFF+MAF+SMFCC	54.52	79.24	76.50
SMFCC+ECC+EFCC+MAF+MFF	35.47	**80.56**	77.09

Table 5.3: Récapitulatif des résultats des 3 classifieurs avec les meilleures combinaisons sur la base allemande.

Caractéristique	Taux de reconnaissance (%)		
	RL	SVM	RNN
MFCC+MSF	84.25	91.83	91.36
SMFCC+ECC	90.49	96.20	89.54
SMFCC+MAF	91.25	95.69	89.97
SMFCC+MFF	89.83	95.76	90.35
SMFCC+MFF+MAF	92.38	96.00	90.52
EFCC+MFF+MAF+SMFCC	92.99	**96.24**	90.72
SMFCC+ECC+EFCC+MAF+MFF	93.04	95.81	90.65

Table 5.4: Récapitulatif des résultats des 3 classifieurs avec les meilleures combinaisons sur la base espagnole.

précédents qu'après la combinaison de caractéristiques proposées avec les autres, le pouvoir de discrimination de caractéristiques est considérablement accru. En regardant les tableaux 5.3 et 5.4, nous pouvons voir aussi que le taux de reconnaissance de la combinaison des caractéristiques, basées sur la transformée de Fourier (MFCC et MSF), est considérablement augmentée par rapport aux résultats de chacune individuellement.

Le tableau 5.5 montre la matrice de confusion pour la meilleure performance de reconnaissance obtenue (80.56%) en combinant les nouvelles caractéristiques (MAF et MFF) avec les SMFCC, ECC et EFCC sur le corpus allemand. La colonne la plus à gauche représente les émotions exprimées (réelles). La colonne indique pour une classe, le nombre de prédictions correctes pour cette classe et le nombre d'échantillons confondus avec une autre classe. Les valeurs correctes sont organisées dans une ligne diagonale allant du haut à gauche au bas à droite de la matrice. Par exemple, le nombre total d'échantillons pour l'émotion "Tristesse" dans l'ensemble de données est la somme des valeurs de la colonne "Tristesse" (11 échantillons). Le tableau 5.5 montre que la joie est le plus souvent confondue avec la colère et que la tristesse est le plus souvent confondue

avec l'ennui, ce qui est à peu près les mêmes résultats obtenus dans la figure 4.9 du chapitre 4.

Le tableau 5.6 montre la matrice de confusion pour la meilleure performance de reconnaissance

Prédite / Exprimée	Peur	Dégoût	Joie	Ennui	Neutre	Tristesse	Colère
Peur	**3**	0	0	0	0	0	0
Dégoût	0	**6**	0	0	0	1	0
Joie	0	0	**4**	0	0	0	2
Ennui	1	0	0	**6**	0	1	0
Neutre	1	0	0	0	**7**	0	0
Tristesse	0	0	0	0	0	**9**	0
Colère	0	0	4	0	0	0	**8**
précision	**60%**	**100%**	**50%**	**100%**	**100%**	**81.81%**	**80%**

TABLE 5.5: Matrice de confusion pour la combinaison des nouvelles caractéristiques (MAF et MFF) avec les SMFCC, ECC et EFCC en utilisant le classifieur SVM sur le corpus allemand.

obtenue (96.24%) en combinant les nouvelles caractéristiques (MAF et MFF) avec les SMFCC et EFCC sur le corpus espagnole.

Prédite / Exprimée	Peur	Dégoût	Joie	Surprise	Neutre	Tristesse	Colère
Peur	**82**	2	1	0	0	0	0
Dégoût	2	**74**	0	0	1	0	0
Joie	1	0	**54**	2	0	0	2
Surprise	0	0	1	**67**	0	0	3
Neutre	0	0	0	0	**196**	0	0
Tristesse	0	0	0	0	0	**70**	0
Colère	1	0	3	3	0	0	**66**
précision	**95.34%**	**96.10%**	**91.52%**	**93.05%**	**99.49%**	**100%**	**92.95%**

TABLE 5.6: Matrice de confusion pour la combinaison des nouvelles caractéristiques (MAF et MFF) avec les SMFCC et EFCC en utilisant le classifieur SVM sur le corpus espagnol.

5.4 Normalisation par locuteur

L'effet de la normalisation du locuteur sur la reconnaissance est étudié dans cette section. La normalisation du locuteur (SN pour speaker normalization en anglais) est utile pour compenser les variations dues à la diversité des locuteurs plutôt qu'à un changement d'état émotionnel. Trois schémas de SN différents sont définis dans [36], nous avons utilisé dans notre travail le schéma qui a donné les meilleurs résultats. Les caractéristiques de chaque locuteur sont normalisées avec une moyenne nulle et un écart type de 1. Soit $f_{u,v}(1 \leq n \leq N_{u,v})$ représente la $u^{\text{ème}}$ caractéristique du locuteur v avec $N_{u,v}$ est la taille de l'échantillon. La nouvelle caractéristique, après SN, $f_{u,v}^{SN}$

est donnée par :

$$f_{u,v}^{SN} = \frac{f_{u,v}(n) - \overline{f_{u,v}}}{\sqrt{\frac{1}{N_{u,v}-1}\sum_{m=1}^{N_{u,v}}(f_{u,v}(m) - \overline{f_{u,v}})^2}} \tag{5.4}$$

avec $\overline{f_{u,v}} = \frac{1}{N_{u,v}}\sum_{n=1}^{N_{u,v}} f_{u,v}(n)$.

Les résultats obtenus sans et avec normalisation par locuteur sont les suivants :

Méthode	Caractéristique	Sans normalisation Moyenne (σ)	Avec normalisation Moyenne (σ)
RL	MFCC	67.92 (6.41)	71.69 (5.89)
	MSF	61.69 (8.15)	67.16 (6.42)
	SMFCC	67.16 (8.01)	75.28 (7.67)
	ECC	54.33 (7.43)	54.71 (6.65)
	EFCC	55.28 (5.62)	57.92 (6.35)
	MAF	59.05 (6.22)	59.24 (4.37)
	MFF	56.03 (3.56)	63.96 (4.74)
SVM	MFCC	70.37 (7.06)	73.77 (7.82)
	MSF	67.92 (4.87)	72.45 (7.23)
	SMFCC	**73.01** (5.48)	75.66 (7.35)
	ECC	53.77 (6.30)	57.16 (8.34)
	EFCC	56.41 (7.57)	56.41 (7.46)
	MAF	63.96 (7.57)	65.47 (5.48)
	MFF	59.43 (6.72)	65.66 (3.30)
RNN	MFCC	68.88 (6.07)	74.17 (6.47)
	MSF	64.47 (4.82)	70.03 (6.62)
	SMFCC	68.12 (6.15)	**76.81** (4.95)
	ECC	51.73 (4.51)	58.37 (7.65)
	EFCC	51.60 (6.83)	57.11 (5.73)
	MAF	61.04 (7.40)	64.04 (6.40)
	MFF	53.68 (9.44)	65.29 (8.57)

TABLE 5.7: Résultats sans et avec normalisation par locuteur obtenus sur la base allemande

D'après les tableaux 5.7 et 5.8, nous pouvons constater que l'application de SN améliore significativement les résultats de reconnaissance pour la base allemande jusqu'à 12% pour le RNN utilisant MFF (voir tableau 5.7), mais ce n'est pas le cas pour la base espagnole. Les résultats montrent que la normalisation par locuteur apporte peu d'amélioration pour la base espagnole (voir tableau 5.8). Les mêmes résultats sont obtenus avec les 3 classifieurs. Ceci peut être expliqué par le nombre de locuteurs dans chaque base de données, où la base allemande contient 10 locuteurs différents, par rapport à la base espagnole qui ne contient que 2 locuteurs.

Nous pouvons remarquer aussi que pour la base espagnole et le classifieur RNN, la normalisation par locuteur permet d'augmenter le pouvoir de discrimination des émotions des caractéristiques

Méthode	Caractéristique	Sans normalisation Moyenne (σ)	Avec normalisation Moyenne (σ)
RL	MFCC	77.21 (0.76)	46.97 (2.23)
	MSF	69.22 (1.37)	45.29 (1.94)
	SMFCC	88.26 (0.95)	83.06 (1.00)
	ECC	53.26 (2.62)	51.67 (1.96)
	EFCC	52.79 (2.48)	51.37 (1.80)
	MAF	63.04 (1.44)	65.36 (1.86)
	MFF	49.35 (2.46)	49.08 (1.57)
SVM	MFCC	91.04 (1.04)	66.85 (1.93)
	MSF	83.29 (1.15)	61.60 (2.01)
	SMFCC	**95.74** (0.94)	**90.41** (1.40)
	ECC	69.68 (2.15)	65.66 (1.85)
	EFCC	67.58 (2.55)	64.42 (1.36)
	MAF	74.63 (1.30)	72.43 (1.16)
	MFF	57.54 (1.85)	55.33 (1.61)
RNN	MFCC	87.30 (0.92)	85.38 (1.37)
	MSF	81.66 (1.62)	79.93 (1.15)
	SMFCC	87.80 (1.82)	85.09 (1.56)
	ECC	63.93 (2.15)	64.67 (1.69)
	EFCC	61.74 (1.25)	62.73 (1.95)
	MAF	64.71 (2.46)	68.25 (1.55)
	MFF	49.05 (3.49)	60.39 (2.52)

TABLE 5.8: Résultats sans et avec normalisation par locuteur obtenus sur la base espagnole

ECC, EFCC, MAF et MFF. Ce n'est pas le cas des caractéristiques MFCC, MSF et SMFCC (voir tableau 5.8).

5.5 Sélection de caractéristiques

L'entrainement d'un algorithme d'apprentissage en utilisant un grand nombre des caractéristiques pour un corpus de taille réduite peut amener à une dégradation des performances du système[111]. A l'heure actuelle, il n'y a pas de consensus sur le meilleur ensemble de descripteurs à prendre en compte pour un système de détection automatique. La pratique la plus courante reste alors d'utiliser l'expérimentation pour découvrir l'ensemble des caractéristiques qui fonctionnent le mieux pour un jeu de données et un modèle donné dans le but d'avoir une classification plus riche. Plusieurs méthodes pour sélectionner les meilleures caractéristiques existent dans la littérature. Nous avons testé 3 différentes méthodes : Analyse en Composant Principal (PCA pour Principal Component Analysis en anglais) permet la projection des données sur une base réduite [112]. La méthode d'analyse discriminante linéaire (LDA pour Linear Discriminant Analysis en anglais) permet de projeter un jeu de données sur un espace de dimension inférieure avec une bonne

Méthode	Caractéristique	Moyenne (avg)	Ecart type (σ)
RL	MFCC+MSF	75.66	4.66
	SMFCC+ECC	72.07	3.64
	SMFCC+MAF	70.00	6.74
	SMFCC+MFF	70.75	5.98
	SMFCC+MFF+MAF	63.58	7.75
	EFCC+MFF+MAF+SMFCC	53.58	7.66
	SMFCC+ECC+EFCC+MAF+MFF	29.05	8.76
SVM	MFCC+MSF	80.00	5.35
	SMFCC+MAF	79.43	3.71
	SMFCC+ECC	78.67	3.77
	SMFCC+MFF	80.18	4.55
	SMFCC+MFF+MAF	83.96	4.00
	EFCC+MFF+MAF+SMFCC	**85.84**	5.05
	SMFCC+ECC+EFCC+MAF+MFF	84.52	3.18
RNN	MFCC+MSF	80.13	3.36
	SMFCC+MAF	77.48	4.49
	SMFCC+ECC	77.44	6.65
	SMFCC+MFF	78.36	4.77
	SMFCC+MFF+MAF	83.14	4.62
	EFCC+MFF+MAF+SMFCC	82.34	4.91
	SMFCC+ECC+EFCC+MAF+MFF	82.97	4.19

Table 5.9: Résultats avec normalisation par locuteur pour les meilleures combinaisons obtenus sur la base allemande

séparabilité de classe afin d'éviter les surajustements et réduire les coûts de calcul [113]. Ces 2 méthodes sont généralement classées comme des techniques de réduction de dimension. La méthode RFE (pour Recursive Feature Elimination en anglais) qui construit un modèle avec toutes les caractéristiques et rejettent récursivement la caractéristique la moins importante jusqu'à l'obtention d'un nombre prédéfini. A partir de la figure 5.4, nous pouvons remarquer qu'il n'y a pas une méthode de sélection qui donne toujours le meilleur résultat avec les 3 classifieurs. Par exemple, la sélection des caractéristiques en utilisant la méthode PCA est meilleure que celles obtenues par les méthodes RFE et LDA avec l'algorithme RL alors que ce n'était pas le cas avec l'algorithme RNN. Nous constatons que le meilleur score de 79% est obtenu en utilisant la méthode RFE avec l'algorithme SVM.

LDA et PCA sont deux techniques de réduction de dimension. Les 3 différentes méthodes permettent de réduire le nombre d'attributs dans le jeu de données, mais une méthode de réduction de dimension le fait en créant de nouvelles combinaisons d'attributs (parfois appelée transformation d'entité), alors que les méthodes de sélection (comme RFE) incluent et excluent des attributs présents dans les données sans les modifier. À cette fin, cette méthode (RFE), décrite plus en détail dans la section suivante, est utilisée dans le reste de notre travail pour sélectionner

Méthode	Caractéristique	Moyenne (avg)	Ecart type (σ)
RL	MFCC+MSF	55.33	2.38
	SMFCC+ECC	86.19	0.96
	SMFCC+MAF	88.17	1.04
	SMFCC+MFF	86.30	1.07
	SMFCC+MFF+MAF	89.40	1.51
	EFCC+MFF+MAF+SMFCC	89.65	1.15
	SMFCC+ECC+EFCC+MAF+MFF	89.95	1.48
SVM	MFCC+MSF	67.20	2.05
	SMFCC+MAF	90.04	1.09
	SMFCC+ECC	90.79	0.93
	SMFCC+MFF	**91.30**	0.85
	SMFCC+MFF+MAF	90.23	1.56
	EFCC+MFF+MAF+SMFCC	90.62	1.73
	SMFCC+ECC+EFCC+MAF+MFF	90.46	1.39
RNN	MFCC+MSF	88.20	1.22
	SMFCC+MAF	87.40	0.92
	SMFCC+ECC	86.95	1.50
	SMFCC+MFF	86.15	1.08
	SMFCC+MFF+MAF	88.18	0.47
	EFCC+MFF+MAF+SMFCC	87.25	1.20
	SMFCC+ECC+EFCC+MAF+MFF	87.25	1.11

TABLE 5.10: Résultats avec normalisation par locuteur pour les meilleures combinaisons obtenus sur la base espagnole ;

le sous-ensemble des meilleures caractéristiques permettant d'améliorer les résultats de classification.

5.5.1 Recursive Feature Elimination

Comme son nom l'indique, RFE correspond à Recursive Feature Elimination en anglais, cette technique supprime récursivement les caractéristiques en construisant un modèle sur les caractéristiques restants. C'est une méthode de sélection de type Wrapper qui est basée sur l'élimination backward [114]. RFE utilise la précision du modèle pour identifier les attributs (et la combinaison d'attributs) qui contribuent le plus à la prédiction de l'attribut cible. Nous avons utilisé la méthode d'élimination récursive de caractéristiques avec réglage automatique du nombre de caractéristiques sélectionnées (RFE avec validation croisée) : il s'agit d'un type de RFE utilisant un système de boucle pour trouver le nombre optimal de caractéristiques. L'algorithme RFE débute avec toutes les caractéristiques, construit un modèle, et supprime la caractéristique de poids faible (calculée à l'aide d'une SVM) c'est à dire la moins importante pour ce modèle. Ensuite, un nouveau modèle est construit avec les caractéristiques restantes, et ainsi de suite jusqu'à ce que le nombre prédéfini de caractéristiques soit atteint.

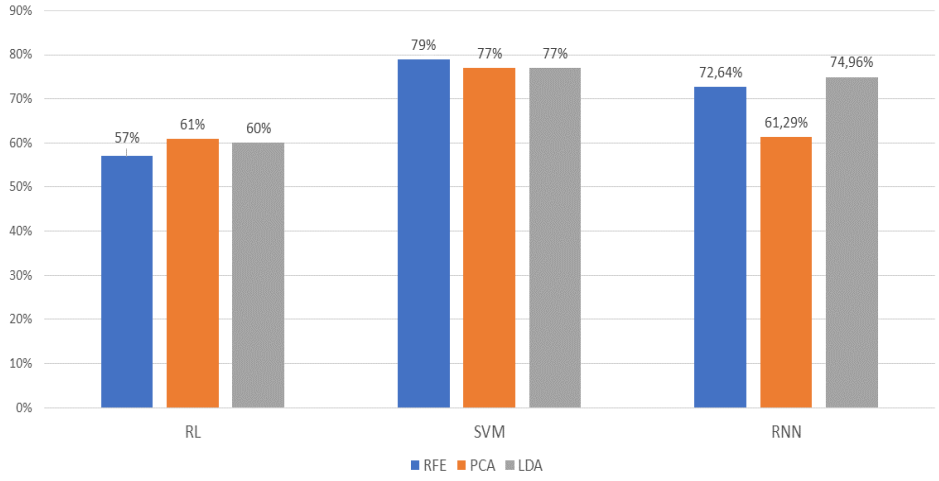


FIGURE 5.4: Comparaison des 3 différentes méthodes de sélection (RFE, PCA, LDA) sur la base allemande avec la combinaison des caratéristiques SMFCC, MFF et MAF.

L'ajout de la sélection par RFE avait pour objectif d'améliorer le taux de classification correcte obtenue. La figure 5.5 illustre l'impact des nombres de caractéristiques sélectionnées sur le taux de reconnaissance. Le premier constat qu'il est possible de faire à partir de la courbe ci-dessus est

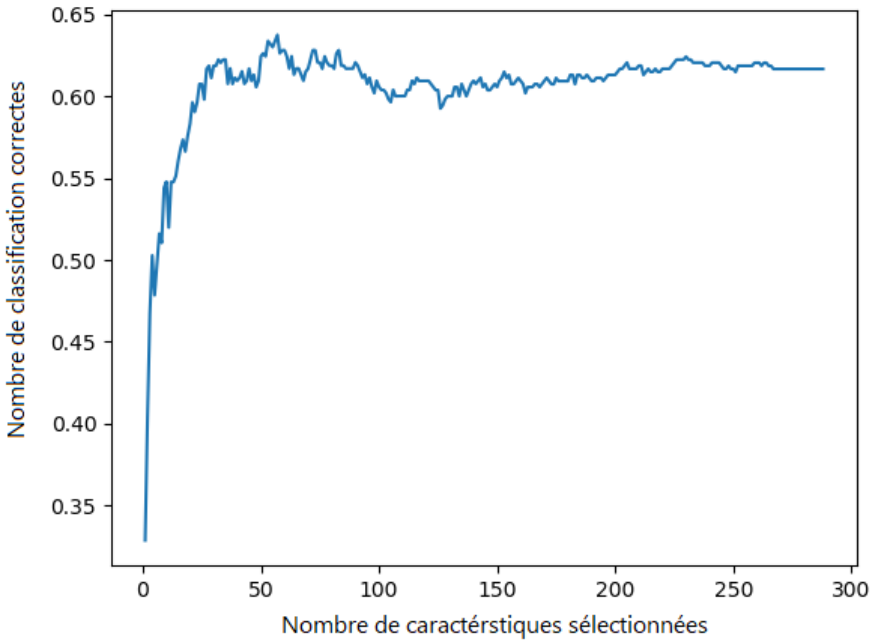


FIGURE 5.5: Nombre de classifications correctes en fonction du nombre de caractéristiques sélectionnées pour la combinaison SMFCC+MFF+MAF de la base allemande en utilisant la méthode de sélection RFE-SVM.

que l'utilisation de l'ensemble des caractéristiques donne un résultat moins bon que la sélection des 30 à 90 meilleures caractéristiques. Il est donc possible de supposer que la combinaison SMFCC, MFF et MAF de la base allemande contient un grand nombre de données redondantes ou non pertinentes. Il est également possible de noter que la variation du taux de bonnes classifications peut être importante pour une différence de sélection d'une seule caractéristique.

5.5.2 Résultats et discussions

Nous présentons ici les meilleurs résultats pour le corpus allemand et le corpus espagnol respectivement avec le classifieur RNN et le classifieur SVM. Ces résultats sont avec normalisation par locuteur. Les précisions de reconnaissance avec et sans sélection de caractéristiques pour les autres classifieurs sont présentés dans B (section B).

Caractéristiques	Résultats sans sélection	Nombre de caractéristiques sélectionnées	Résultats avec sélection
MFCC	74.17%	28 sur 60	77.32%
MSF	70.03%	75 sur 95	71.50%
SMFCC	76.81%	55 sur 60	78.12%
ECC	58.37%	40 sur 60	59.91%
EFCC	57.11%	44 sur 60	59.35%
MAF	64.04%	120 sur 132	65.15%
MFF	65.29%	69 sur 96	65.62%
MFCC+MSF	80.13%	103 sur 155	80.13%
SMFCC+ECC	77.44%	95 sur 120	80.28%
SMFCC+MAF	77.48%	175 sur 192	81.83%
SMFCC+MFF	78.36%	124 sur 156	82.76%
SMFCC+MFF+MAF	83.14%	244 sur 288	85.53%
EFCC+MFF+MAF+SMFCC	82.34%	288 sur 348	85.00%
SMFCC+ECC+EFCC+MAF+MFF	82.97%	328 sur 408	**86.13**%

TABLE 5.11: Comparaison des résultats sans et avec sélection de caractéristiques en utilisant le classifieur RNN sur la base allemande (avec SN).

En conclusion, d'après les résultats obtenus, l'utilisation de la méthode de sélection des caractéristiques RFE permet d'améliorer significativement les performances du système jusqu'à +19% sur le corpus espagnol pour les MFCC et les MSF.

Caractéristiques	Résultats sans sélection	Nombre de caractéristiques sélectionnées	Résultats avec sélection
MFCC	66.85%	51 sur 60	85.82%
MSF	61.60%	85 sur 95	80.01%
SMFCC	90.41%	29 sur 60	89.70%
ECC	65.66%	33 sur 60	65.16%
EFCC	64.42%	48 sur 60	64.73%
MAF	72.43%	58 sur 108	66.93%
MFF	55.33%	49 sur 90	32.93%
MFCC+MSF	67.20%	136 sur 155	86.72%
SMFCC+ECC	90.79%	105 sur 120	90.16%
SMFCC+MAF	90.04%	150 sur 168	90.71%
SMFCC+MFF	91.30%	141 sur 150	90.08%
SMFCC+MFF+MAF	90.23%	199 sur 258	90.77%
EFCC+MFF+MAF+SMFCC	90.62%	247 sur 318	**90.79**%
SMFCC+ECC+EFCC+MAF+MFF	90.46%	280 sur 378	90.19%

TABLE 5.12: Comparaison des résultats sans et avec sélection de caractéristiques en utilisant le classifieur SVM sur la base espagnole (avec SN).

5.6 Comparaison avec les systèmes existants

Dans le chapitre 2 nous avons dressé une liste (voir Tableau 2.3), qui ne vaut pas être exhaustive, des différents travaux sur les systèmes de RAE. Les performances sont difficilement comparables car elles varient en fonction de plusieurs éléments tels que : le type de données, le choix des modèles d'apprentissage, de la manière dont les paramètres sont obtenus, etc. Dans cette section, nous allons comparer, nos résultats, avec ceux de quelques travaux menés sur le corpus allemand. Selon nos connaissances scientifiques actuelles, le corpus espagnol n'est pas utilisée à de telles fins.

Iliou [115] a évalué les sept émotions du corpus allemand par un vecteur de 35 caractéristiques prosodiques dont la hauteur, l'énergie et la durée. Le score de reconnaissance en utilisant les réseaux de neurones était de l'ordre de 51%. Les chercheurs dans [116] ont également mené des expériences en utilisant le corpus allemand où la précision a été de 78,64% pour les femmes et 73,40% pour les hommes. Ce taux de précision a été montré pour les caractéristiques prosodiques et spectrales appliquées avec le classificateur LDA. Alonso et al. [117], Luengo et al. [118] et Cao et al. [119] ont extrait les caractéristiques spectrales, et prosodiques du corpus allemand et ont utilisé les machines à vecteurs de support (SVM) comme classifieur. Alonso et al. [118] ont obtenu une précision de reconnaissance des émotions de 94,9% en utilisant cinq émotions : colère, bonheur, neutralité, ennui et tristesse. Luengo et al. [118] ont rapporté une précision de

reconnaissance des émotions de 78,3% utilisant sept états émotionnels : colère, ennui, dégoût, peur, bonheur, neutralité et tristesse. Cao et al. [119] ont atteint une précision de reconnaissance des émotions de 82,1% en utilisant sept émotions : colère, dégoût, peur, bonheur, neutralité, tristesse et ennui. Wang et al. [120] ont rapporté une précision de reconnaissance des émotions de 88,8% en utilisant des caractéristiques prosodiques dans une classification basée sur SVM. Ils ont utilisé six émotions distinctes du corpus allemand : bonheur, tristesse, colère, ennui, anxiété et neutralité. Le système développé dans ses travaux atteint un taux de reconnaissance de 86,13% sur le corpus allemand. Le classifieur RNN qui donne ce meilleur score en utilisant la combinaison des nouvelles caractéristiques proposées avec les caractéristiques cepstrales (SMFCC, ECC et EFCC).

5.7 Évaluation du corpus "French Student Emotionnal database"

Notre système de RAE appliqué au corpus français que nous avons construit présente un score de 37% avec les caractéristiques MFCC et le classifieur RNN. Dans ce corpus, nous observons deux différences avec les autres corpus (allemand et espagnol) : 1) il y a beaucoup de locuteurs (32 ici, 2 pour le corpus espagnol et 10 pour le corpus allemand) donc la variabilité liée au locuteur est plus forte que celle liée aux émotions et 2) il y a peu d'enregistrements par locuteur. Par conséquence pour améliorer notre taux de reconnaissance, nous utilisons une normalisation du locuteur. Nous obtenons ainsi une augmentation de ±8%.

Le tableau 5.13 présente les meilleurs taux de reconnaissance obtenus sur le corpus en français pour les différentes méthodes de classification. Nous constatons que le meilleur score est atteint en utilisant les caractéristiques que nous avons proposé combinées avec les SMFCC. La Figure 5.6 présente les taux de reconnaissance des émotions du corpus "French Student Emotionnal database" basée sur le classifieur SVM et l'oreille humaine. La comparaison des 3 courbes montre une corrélation entre les résultats obtenus par les humains et ceux obtenus par le système de RAE.

Nous pouvons conclure que notre système de RAE est sensible à une grande variabilité des voix des locuteurs. Il nécessite un ensemble d'enregistrements important par locuteur pour être plus performant.

Caractéristique	Taux de reconnaissance en % : Moy.(σ)		
	RL	SVM	RNN
MFCC	44.88 (6.10)	44.88 (5.59)	46.77 (4.42)
SMFCC	45.81 (9.17)	50.93 (6.43)	42.41 (8.24)
MFCC+MSF	45.81 (8.56)	50.46 (6.67)	48.20 (7.18)
SMFCC+ECC	46.51 (8.49)	53.25 (6.14)	47.30 (6.91)
SMFCC+MAF	45.58 (7.92)	52.55 (6.22)	47.82 (5.18)
SMFCC+MFF	43.48 (8.49)	51.39 (9.07)	42.15 (5.95)
SMFCC+MFF+MAF	44.18 (8.90)	**54.41** (7.03)	52.07 (5.52)
EFCC+MFF+MAF+SMFCC	46.51 (8.83)	53.25 (8.16)	48.62 (5.35)
SMFCC+ECC+EFCC+MAF+MFF	31.39 (7.20)	54.18 (9.56)	47.65 (8.93)

Table 5.13: Récapitulatif des meilleurs résultats obtenus sur le corpus en français : résultats avec normalisation par locuteur et avec sélection.

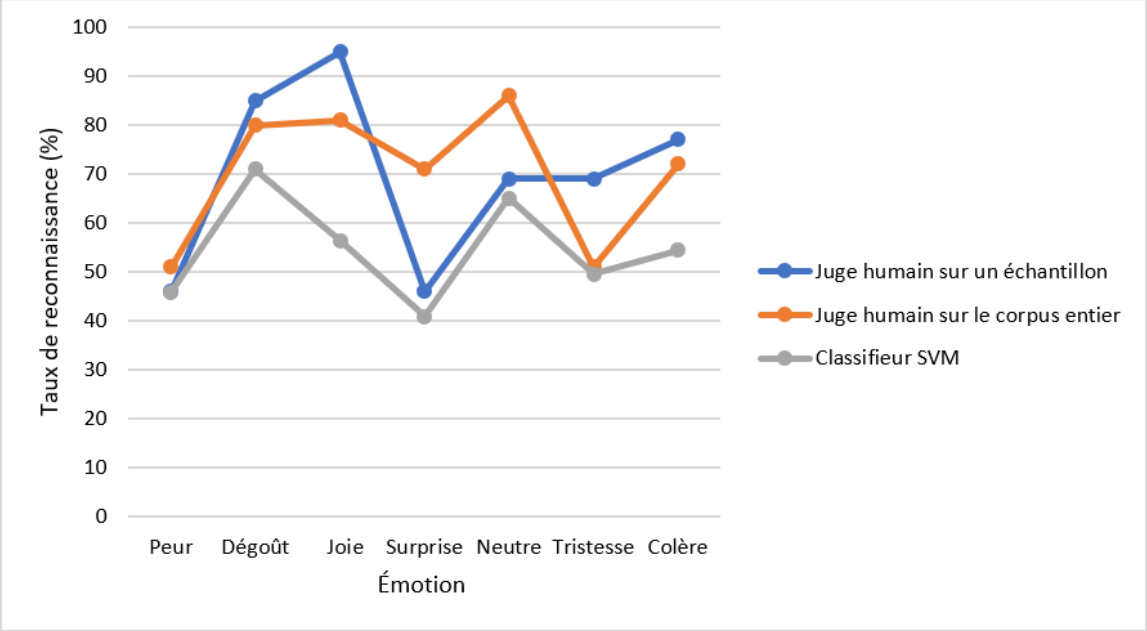


Figure 5.6: Comparaison des taux de reconnaissance des émotions du corpus "French Student Emotionnal database" basée sur le classifieur SVM et l'oreille humaine.

5.8 Conclusion

Dans ce chapitre, nous avons présenté notre système de reconnaissance des émotions à partir de la voix. Ce système a été validé en utilisant deux corpus existants (allemand et espagnol) et le corpus français que nous avons construit. La performance individuelle de chaque caractéristique a été mesurée (voir tableaux 5.1, 5.2 et 5.13). Nous avons également présenté les scores obtenus pour différentes combinaisons de caractéristiques. Sur le corpus allemand, les meilleurs résultats donnent un taux de reconnaissance de 80,56% en utilisant la combinaison des nouvelles caractéristiques (MAF et MFF) avec les SMFCC, ECC et EFCC. C'est par une autre

combinaison de caractéristiques que les meilleurs résultats sont obtenus sur le corpus espagnol : 96.24% en combinant les caractéristiques proposées avec les SMFCC et EFCC. En combinant les nouvelles caractéristiques et les SMFCC, un taux de 54,41% est obtenu sur le corpus français. Ces meilleurs scores sur les 3 corpus sont obtenus en utilsiant le classifieur SVM.

Afin d'améliorer nos résultats, nous avons appliqué successivement la normalisation par locuteur et la méthode de sélection de caractéristiques RFE. Les résultats montrent que la normalisation par locuteur apporte des améliorations. Ainsi sur le corpus allemand le gain est de $\simeq$20% (10 locuteurs et 50 enregistrements par locuteur), sur le corpus français le gain est de $\simeq$8% (32 locuteur et 14 enregistrements par locuteur) et sur le corpus espagnol le gain est de $\simeq$2% (2 locuteurs et 3020 enregistrements par locuteur). Nous constatons que la normalisation par locuteur apporte un gain important lorsqu'il y a une variabilité importante de locuteurs et suffisamment d'enregistrements par locuteur.

Après la sélection des meilleures caractéristiques, c'est le classifieur RNN qui donne le meilleur taux de reconnaissance sur le corpus allemand (86,13% en utilisant la combinaison des caractéristiques proposées avec les SMFCC, ECC et EFCC).

Conclusion générale

Le travail présenté dans cette thèse s'inscrit dans un projet de reconnaissance des émotions à partir de la voix, dans un contexte pédagogique. Le but visé est de détecter l'état émotionnel de chaque apprenant en classe. Pour atteindre cet objectif, nous avons d'abord proposé un système complet et efficace de reconnaissance automatique de l'état émotionnel d'un locuteur à partir d'enregistrements audio. Pour réaliser ce système, nous avons au préalable étudié les notions liées aux émotions telles que leurs définitions, leurs types, leurs représentations et les différents canaux permettant de communiquer ces émotions. Nous avons retenu une approche catégorielle avec un ensemble de 7 émotions de base qui sont la tristesse, la colère, le dégoût, la peur, la surprise, la joie et le neutre en s'appuyant sur des corpus émotionnels actés. Nous avons ensuite identifié les caractéristiques vocales spécifiques à un état émotionnel.

Après avoir étudié les différents composants, les avancés, les défis et problématiques des systèmes de RAE, notre attention s'est portée sur l'extraction des caractéristiques les plus pertinentes du signal vocal. Nous avons ainsi analysé finement les systèmes de RAE développés dans la littérature en utilisant les caractéristiques qui ont prouvé leurs efficacités (comme les MFCC, les caractéristiques de modulation spectrales...). Ces caractéristiques sont améliorées et complétées par de nouvelles caractéristiques. Dans cette étude, nous avons montré la pertinence de l'extraction d'un ensemble de caractéristiques par démodulation AM-FM de la voix en utilisant le couplage EMD-TKEO. Ce couplage permet une meilleure estimation des caractéristiques les plus pertinentes à court et à long terme, sans être restreint par l'hypothèse de stationnarité de la voix. Pour obtenir de bonnes combinaisons de caractéristiques, une étude plus fine est également menée. La sélection de caractéristiques permet d'éliminer la redondance et les caractéristiques les moins pertinentes et ainsi améliorer les performances du système de RAE. De plus, nous avons montré que la normalisation par locuteur, pour réduire la variabilité liée au locuteur, améliore la reconnaissance pour les corpus de taille réduite et avec un grand nombre de locuteurs. Alors que sur un corpus plus important, elle n'apporte pas d'amélioration significative. Nous concluons

que, lorsqu'il y a une variabilité importante de locuteurs et suffisamment d'enregistrements par locuteur, l'application de la normalisation avant la reconnaissance est bénéfique. Les résultats de simulation montrent que la combinaison des caractéristiques proposées avec les caractéristiques cepstrales existantes donnent des meilleures performances. Notre système atteint un taux de reconnaissance de 96,24% sur le corpus espagnol.

L'approche présentée a été validée par expérimentation sur deux corpus émotionnels actés, largement utilisés par la communauté scientifique, le corpus en allemand et le corpus en espagnol. Pour répondre à notre besoin d'application dans une situation pédagogique, un corpus en français a été construit, testé et validé. En construisant nous-même ce corpus, nous avons pu avoir une connaissance complète du contenu émotionnel, linguistique et paralinguistique du corpus utilisé dans notre système de RAE. Ce corpus est constitué de 502 enregistrements prononcés par 32 locuteurs (8 femmes et 24 hommes).

Travaux futurs

Le travail mené dans cette thèse est une première étape pour le projet mené au laboratoire LAUM, en collaboration avec les laboratoires LATIS et CREN, afin de réaliser un système de RAE intelligent capable de détecter l'état émotionnel courant de l'apprenant. Dans ce contexte, d'autres travaux sont à prévoir pour permettre la réelle utilisation de notre système en classe.

— Un premier aspect est l'enrichissement de notre corpus en collectant plus d'enregistrements pour chaque locuteur. La fusion de classifieurs peut également apporter des améliorations significatives. Notamment, tester d'autres méthodes de sélection de caractéristiques permet d'optimiser l'ensemble des caractéristiques existantes à ce jour et extraire des caractéristiques prosodiques à partir du couplage EMD-TKEO.

— Une autre perspective pour cet axe de recherche est l'étude plus poussée de catégorisation des émotions en trois classes (positive, négative et neutre). Pour assurer une bonne gestion de classe, l'enseignant aura besoin de savoir si l'apprenant est de bonne humeur, de mauvaise humeur ou indifférent et rendre le système adapté aux conditions réelles.

— Dans nos travaux, la reconnaissance est effectuée sur un seul canal de communication (canal audio). Plusieurs travaux de recherche portent sur d'autres modalités (les mouvements corporels, les expressions faciales ou les réactions physiologiques). Il est intéressant, dans la situation que nous cherchons, d'utiliser d'autres sources d'informations c'est à dire tester une reconnaissance multicanal à la place d'un seul canal.

Annexe A

Notre corpus émotionnel

A.1 Énoncés et mises en situation

Tristesse :

La tristesse est une réaction douloureuse que l'on ressent en présence de quelque chose de désagréable. En général quand on est triste, on a du chagrin, de la peine ou du regret.

1. Tu viens de passer une semaine intensive de DS et tu penses les avoir tous ratés. Tu dis avec déception : **"Je n'ai validé aucune matière"**.
2. La note du dernier DS est tombée. Tu savais que tu ne l'avais pas bien réussi mais tu ne pensais pas que c'était si grave. Tu as eu 1,5/20 : **"Je ne pensais pas rater à ce point là"**.
3. On est la veille du DS d'une matière que tu n'aimes pas trop. Le prof avait dit qu'un exercice de TD tomberait au DS. Mais tu as beau chercher dans toutes tes notes, il manque une correction et soudain tu te rappelle que tu t'étais endormi pendant ce TD et que tu n'as malheureusement pas la correction : **"Pourquoi je me suis endormi pendant ce TD..."**.
4. Celà fait une semaine que tu fais du 8h-18h. La fatigue se fait ressentir de plus en plus pour toute la classe. Et tu te rends compte que demain matin encore tu ne pourra pas dormir : **"Demain, on a encore un cours à 8h"**.
5. Hier tu es resté plus longtemps à l'école pour réviser un cours et tu es probablement parti trop vite car tu as oublié tous les polycopies de ce cours. Ce matin tu es retourné dans la salle mais il n'y a rien : **"J'ai perdu mes cours"**.

6. Tu as passé un TOEIC blanc car tu passes le TOEIC la semaine prochaine, mais tu as eu un résultat désastreux : **“Tu as vu mon score ? Impossible que j’ai mon TOEIC la semaine prochaine”**.
7. Lorsque vous venez de faire un test difficile et que vous commentez avec vos collègues : **“Oh non, je n’ai pas réussi le test”**.
8. Tu es en cours de physique quantique et on te dit qu’à partir de la notion d’exponentielle de matrice, on peut trouver la solution formelle de l’équation de Schrödinger. Tu n’as pas compris un traître mot de tout ceci. Et tu dis : **“Ça me saoule, je n’ai encore rien compris !”**.
9. Tu arrives en cours en pensant que c’est un cours que tu aimes mais en faite non : **“Ah mince, J’aime pas ce cours”**.
10. Tu es en TD sur un exercice difficile. Le prof n’est pas dans la communication avec les élèves. Il galère et n’arrive pas à expliquer. Il ne fait que parler à son tableau, et tu en parles avec ton voisin : **“On ne comprend rien là !”**.
11. Tu es en TP d’informatique et tu as besoin de la WiFi, mais pas moyen, elle n’arrête pas de bugger : **“Je suis vraiment déçu, comment on va faire le TP, ça n’arrête pas de bugger ?”**.

Colère :

La colère est définie comme un sentiment impulsif de protestation, d’irritation, d’exaspération ou de frustration envers quelqu’un ou quelque chose.

1. En groupe de projet, chacun a sa tâche mais certains n’arrivent jamais à l’heure donc leur tâche n’avancent pas à la vitesse escomptée et le projet prend du retard : **“Merde, on va avoir une mauvaise note à cause de toi”**.
2. **“Ça me soûle”**.
3. Lorsque vous essayez de résoudre un exercice, mais cela ne fonctionne pas : **“Ça marche pas”**.
4. **“Ça m’énerve, j’ai rien compris”**.
5. Quand l’enseignant prend beaucoup de temps pour corriger les examens et ça commence à t’énerver : **“Ça fait trois mois qu’on a fait le contrôle et on a toujours pas les notes”**.
6. Ça t’agace vraiment que ce soit toujours ton groupe qui finit le vendredi après-midi : **“C’est toujours notre groupe qui termine vendredi après-midi”**.

7. Tu viens d'apprendre que le professeur qui tu devais avoir à 8h n'est pas là : **"Purée, je me suis levé pour rien"**.
8. Ça fait 30 minutes que tes camarades parlent derrière toi et que tu as du mal à entendre ce que le prof dit : **"Vous pouvez vous taire ? Merci"**.
9. Tu vois la fin du cours arrivé et vous êtes toujours au TD1 : **"Ça fait une heure qu'on est sur la même chose, on avance pas"**.
10. Le contrôle c'est demain et tu essaies désespérément de comprendre une partie très importante du cours que le prof n'arrive toujours pas à t'expliquer : **"Mais ce n'est pas possible, je n'ai toujours pas compris cette partie !"**.
11. Dans la salle quand vous avez un néon qui grésille : **"Ça n'arrête pas de grésiller ! Ça m'énerve ! Je ne peux pas faire attention en classe"**.

Dégoût :

Le dégoût est un rejet instinctif ou culturel assez violent de quelque chose dont on n'a pas le goût (ni à manger, ni à sentir, ni à toucher, quelque chose qui dégoûte).

1. Aujourd'hui il fait très chaud et des odeurs de transpiration émanent de la pièce où tu vas passer tes trois prochaines heures de cours : **"Cette odeur me dégoute !"**.
2. Tu essaies d'écouter le CM du prof mais tu entends la personne derrière toi mâcher son chewing-gum la bouche ouverte, les bruits de salive te dégoûtent au plus haut point, tu te retournes et tu demandes d'un ton dégoûté : **"Est-ce que tu peux arrêter de mâcher ton chewing-gum la bouche ouverte, s'il te plaît ?"**.
3. Aujourd'hui il fait encore très chaud et le prof se démène pour vous faire comprendre une notion importante. Il bouge tellement que des gouttes de sueur lui coulent sur le visage et une auréole se forme dans tout son dos : **"Tu as vu comment le prof transpire ?"**.
4. Tu arrives à ta place et ta table est toute grasse car la personne avant toi avait mangé sur la table sans la nettoyer : **"Tu as vu l'état de ma table ? C'est tout gras, comment je vais poser mes cahiers ?"**.
5. Tu arrives à ta place mais des mouchoirs usagés sont posés sous ta chaise. Tu sais que tu vas devoir les retirer toi-même sinon personne d'autre ne le fera ! Tu prends ton courage à deux-mains et tu le fais : **"Beurk, ça me dégoute, il faut que j'aille me laver les mains après !"**.
6. Tu viens à vélo tous les matins, mais aujourd'hui tu as du retard car tu as déraillé sur la route. Tu as dû t'arrêter pour remettre ta chaîne, et tu t'es mis de la graisse partout

sur les mains. En arrivant à l'école, comme tu avais accumulé du retard, tu es donc allé directement en cours sans te laver les mains : **"J'ai déraillé ce matin, je m'en suis mis partout, c'est dégueu !"**.

7. En arrivant en classe tu te rends compte que les tables sont trop avancées. Tu décides de reculer ta tables. Au moment de la prendre, tu sens sous ton doigt un chewing-gum encore frais : **"Quelqu'un a posé son chewing-gum sous la table au cours d'avant, c'est dégueu !"**.
8. Hier soir, il y avait une soirée à l'école. Certains n'ont pas pu se laver Soirée heir, et des gens n'ont pas eu le temps de se laver ce matin avant d'aller en cours et ils puent encore l'alcool : **"Wow, quelle puanteur. Quelle odeur d'alcool !"**.
9. Tu essaie d'expliquer un truc à ton voisin et il baille toutes les 5 minutes. Il baille tellement la bouche grande ouverte et sans mettre sa main que tu arrives à voir toute l'intérieur de sa gorge. Tu lui demande : **"Est-ce que tu peux mettre ta main quand tu bailles, ça me dégoute ?"**.
10. Tu es au premier rang en CM et le prof explique quelque chose. D'un coup un énorme postillon arrive en pleins milieu de la page de ton cahier. Tu es écoeuré de la situation et tu dis : **"Regarde le postillon du prof sur mon cahier ? ! Ça m'dégoute !"**.

Peur :

La peur se produit face à une situation de danger éventuelle, quand on se sent menacé, physiquement ou psychologiquement. Dans le domaine pédagogique, un étudiant va ressentir la peur pour ce qui concerne son avenir scolaire donc il présentera de l'agitation et du stress.

1. **"J'espère qu'on aura le temps de terminer le projet"**.
2. Vous avez eu une bonne note au dernier examen mais votre semestre est mal partie. Tu es vraiment inquiète parce que tu ne sais pas si tu vas le valider : **"J'espère que ça va compenser"**.
3. **"J'ai vraiment peur de rater mon semestre"**.
4. Tu vas à un examen mais tu n'as pas eu le temps de réviser. Inquiet, tu dis : **"Je vais rater mon examen**.
5. Pour te rassurer, tu demandes à un camarade s'il a commencé ses révisions : **"Est ce que t'as commencé les révisions ?"**.

6. Tu n'as pas encore commencé à rédiger ton rapport qui est à rendre dans trois jours mais tu demandes à un camarade s'il à commencé le sien pour te rassurer : **"Est ce que t'as commencé à rédiger ton rapport ?"**.
7. Tu n'as pas eu le temps de faire l'exercice et tu as peur que le prof t'interroge : **"j'espère qui ne va pas m'interroger**.
8. Tu es en train de travailler sur un projet, vous voyez que le temps ne sera pas suffisant : **"On aura jamais assez de temps"**.
9. Un contrôle surprise peut avoir lieux au prochain cours, mais tu te rends compte que tu a oublié de réviser : **"J'ai oublié de réviser le cours"**.
10. C'est le jour de ta soutenance de stage ouvrier et tu es le suivant à passé. **"Purée, c'est bientôt à moi, j'espère que ça va aller"**.
11. Ça fait 45 minutes que tu travaillais sur ton programme LABVIEW et tout à coup, toutes tes fenêtres se ferment : **"Mince ! J'espère que ça a enregistré"**.
12. Tu es en TP et jusque là tout se passe bien, mais tu sens une odeur de brûlé : **"Euh, tu ne sens pas une odeur de cramé là ?"**.

Surprise :

La surprise est une émotion ressentie lorsque vous faites face à une situation inattendue. Elle peut être bonne ou mauvaise.

1. La note du DS, que tu pensais avoir complètement raté, est tombée et tu t'aperçois à l'instant, que tu as eu plus de 10/20 et tu dis : **"Incroyable ! J'ai eu la moyenne"**.
2. Quand tu découvres pour la première fois les propriétés étonnantes du mélange Maïzena-eau et tu dis avec surprise : **"C'est incroyable je ne pensais pas que c'était possible !"**.
3. Depuis plusieurs cours, tu as des difficultés à comprendre les explications d'un prof en particulier, mais aujourd'hui c'est bien différent : **"Aujourd'hui, le cours s'est vraiment bien passé !"**.
4. Quand tu arrives dans la salle et qu'il fait très froid dehors. tu t'aperçois avec surprise que la salle est bien chauffée : **"Le chauffage fonctionne !"**.
5. Quand le cours est très intéressant et tu te rends compte qu'il est déjà terminé : **"C'est déjà fini ? Le cours est passé trop vite !"**.
6. Le prof distribue le polycopié de cours et tu constantes sa longueur. **"Quoi ! Tout ça !"**.

7. Un camarade vient de vous dire qu'il y a un exercice à rendre : **"Quoi ! Mais je ne savais pas !"**.
8. Tu aperçois un étudiant, qui est très souvent absent en cours : **"Incroyable, il est là !"**.
9. Vous venez de recevoir la note de vos dernier DS et tu le dis spontanément à tes camarades : **"Ah ! Les notes du dernier examen sont tombés !"**.
10. Tu attends les notes d'un DS depuis tellement longtemps que tu avais même oublié que tu avais fait ce DS. Tu es donc surpris de recevoir la note : **"Ça y est on a reçu les notes de maths, c'est ouf !"**.
11. Alarme incendie se met à retentir d'un coup, ça surprend tout le monde : **"Ahh ? !"**.

Joie

La joie est une émotion vive qui procure un sentiment de plénitude.

1. Vous êtes fatigué, avez eu plusieurs cours dans la journée et apprenez que l'enseignant n'est pas venu : **"Trop cool, le prof n'est pas là !"**
2. Tu es depuis plusieurs jours sur un problème que personne ne réussit et là tu viens enfin de comprendre un point important du cours qui te permet de résoudre le problème, et tu t'exclames : **"Génial, j'ai réussi !"**.
3. Cela fait plusieurs heures que ton programme bloque, ne compile pas et affiche plein d'erreurs, mais d'un coup : **"Ça compile, enfin !"**.
4. En projet éolienne, le prof est venu voir ton groupe plusieurs fois pour dire qu'il ne trouvait pas votre projet innovant et d'un coup tu lances : **"J'ai une idée !"**.
5. Quand quelqu'un vous dit de résoudre le problème d'une certaine manière et vous réussissez : **"Super ! Ça y est ça marche"**.
6. Après avoir postulé à de nombreuses propositions de stage, tu reçois enfin une réponse favorable à ta demande : **"J'ai trouvé un stage !"**.
7. La scolarité vient d'envoyer un mail à toute la promo pour dire que le seul cours que tu as demain est annulé. Tu es le premier à voir le mail, tu annonce donc à toute la classe : **"Il n'y a pas cours demain !"**.
8. Ton téléphone vibre, tu reçois une bonne nouvelle par message et tu dis à ton voisin en montrant le message : **"Trop marrant, t'as vu ce qu'il m'a répondu ?"**.
9. Tu es dans ton cours préféré et là le prof vous énonce le théorème le plus important du cours et il se trouve que tu le connais déjà parfaitement, tu dis à ton voisin **"J'ai trop bien compris cette partie du cours, je peux t'expliquer si tu veux !"**.

10. Ça fait trois semaines, que t'as tous les jours cours de mécanique, tu ne comprends toujours rien aux tenseurs, et là le prof t'as dit le truc qui te débloque et tu lui dis : **"Merci, j'ai enfin compris !"**.
11. Tu es en cours, mais le prof n'est toujours pas là. Tu pars voir l'administration pour avoir des informations. Et là, tu reviens en criant : **"Le prof est malade, donc le cours est annulé !"**.
12. Ton voisin te raconte une blague très marrante, et tu lui réponds **"Elle est bonne celle là !"**.
13. Tu regardes le calendrier et tu remarques que les vacances arrive à grand pas. **"Super ! C'est bientôt les vacances !"**.

Neutre :

L'émotion neutre se traduit par l'absence des autres émotions. Ce sont les situations les plus basiques. Des questions ou affirmations les plus simples.

1. **"Passe-moi ta gomme."**
2. **"Tu peux m'aider ?"**
3. **"Il est quelle heure ?"**
4. **"On est dans quelle salle au prochain cours ?"**
5. **"C'est quoi le prochain cours ?"**
6. **"Il faut que je passe à la scolarité."**
7. **"Est-ce que tu peux dire au prof que je serai absent au prochain cours ?"**
8. **"Tu peux épeler l'adresse URL ?"**
9. **"Tu as reçu mon mail ?"**
10. **"Le prof fait quel exercice ?"**
11. **"Fais un copier coller."**

A.2 Phrases contenant un biais sémantique

1. **"Génial** , j'ai réussi !"
2. **"Incroyable** , il est là !".
3. **"Super !** Ça y est ça marche."

4. **“Trop cool**, le prof n’est pas là !”
5. **“Ah mince**, J’aime pas ce cours”.
6. **“Ça m’énerve**, j’ai rien compris”.
7. **“Incroyable** ! J’ai eu la moyenne”.
8. **“Purée**, je me suis levé pour rien”.
9. **“Oh non**, je n’ai pas réussi le test”.
10. **“Super** ! C’est bientôt les vacances !”
11. **“Mince !** J’espère que ça a enregistré”.
12. **“Ça me soule**, je n’ai encore rien compris !”.
13. **“Trop marrant**, t’as vu ce qu’il m’a répondu ?”
14. **“Ah !** Les notes du dernier examen sont tombés !”
15. “Ça y est on a reçu les notes de maths, **c’est ouf !”**
16. **“Purée**, c’est bientôt à moi, j’espère que ça va aller”.
17. **“Merde**, on va avoir une mauvaise note à cause de toi”.
18. **“C’est incroyable** je ne pensais pas que c’était possible !”.
19. “J’ai déraillé ce matin, je m’en suis mis partout, **c’est dégueu !”**
20. “Regarde le postillon du prof sur mon cahier ? ! **Ça me dégoute !”**.
21. **“Beurk**, ça me dégoute, il faut que j’aille me laver les mains après !”.
22. “Est-ce que tu peux mettre ta main quand tu bailles,**ça me dégoute ?”**.
23. **“Je suis vraiment déçu**, comment on va faire le TP, ça n’arrête pas de bugger ?”
24. “Quelqu’un a posé son chewing-gum sous la table au cours d’avant, **c’est dégueu !”**

A.3 Enregistrement

A.3.1 Choix de la salle d’enregistrement

A.3.1.1 Temps de Réverbération

Le temps de Réverbération (TR) est le paramètre acoustique de loin le plus déterminant. C’est donc la salle avec le plus faible TR qui sera retenue pour les enregistrements. En cas de TR équivalents pour deux salles, alors il faudra prendre en compte d’autres paramètres acoustiques

Comparaison des temps de réverbération des salles envisagées								
Aquarium C06			Box C06			Salle réunion RdC		
Freq. (Hz)	TR20 (s)	TR(30) (s)	Freq. (Hz)	TR20 (s)	TR30 (s)	Freq. (Hz)	TR20 (s)	TR30 (s)
250	0,18	0,18	250	**0,23**	**0,23**	250	0,46	0,54
315	0,16	0,17	315	**0,20**	**0,28**	315	0,54	0,63
400	0,32	0,33	400	**0,12**	**0,25**	400	0,38	0,47
500	0,23	0,29	500	**0,13**	**0,16**	500	0,45	0,48
630	0,37	0,39	630	**0,07**	**0,12**	630	0,38	0,39
800	0,36	0,43	800	**0,11**	**0,13**	800	0,41	0,40
1000	0,58	0,51	1000	**0,07**	**0,11**	1000	0,31	0,35
1250	0,33	0,38	1250	**0,07**	**0,08**	1250	0,38	0,36
1600	0,20	0,29	1600	**0,05**	**0,07**	1600	0,30	0,34
2000	0,17	0,22	2000	**0,06**	**0,07**	2000	0,34	0,33

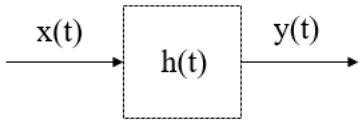
Table A.1: Tableau comparatif des salles envisagées

(tels que la clarté, l'intelligibilité, etc.). Le TR est défini comme le temps nécessaire pour que le niveau sonore d'une impulsion dans la pièce diminue de 60 dB, d'où son nom complet "TR60". En pratique on utilisera le TR30 et le TR20 afin de pouvoir faire des mesures sans générer une impulsion trop forte, qui serait dangereuse pour notre audition. La mesure des TR30 et TR20 est faite à l'aide d'un sonomètre. Une impulsion sonore est générée en faisant éclater un ballon de baudruche, et la décroissance par bande d'octave est calculée par le sonomètre, qui indique ensuite les valeurs des TR30 et TR20 pour chaque bande d'octave. Étant donné que le but est d'enregistrer des voix, les fréquences prises en compte pour la comparaison des salles sont axées autour de celles de la voix humaine. Le tableau suivant montre les mesures obtenues.

Nous pouvons conclure d'après le tableau que le Box C06 est le meilleur choix de salle : en effet, son temps de réverbération est beaucoup plus bas que celui des autres salles étudiées. Cela se traduit par un confort de parole et d'écoute, ce qui correspond à l'ambiance sonore recherchée pour les enregistrements que nous devons effectuer. Mais par faute de place, il ne serait pas possible de faire les enregistrements de façon rapide et efficace dans les boxs en C06. L'aquarium de la salle C06 est donc le meilleur compromis que nous avons pu trouver, en terme de qualité acoustique, confort physique, et accessibilité. Le seul désavantage de cette salle est que 3 de ses murs sont en verre, qui est un matériau peu absorbant.

Une fois que la salle est choisie, il sera maintenant nécessaire de connaître expérimentalement sa réponse impulsionnelle afin de savoir si cette salle était vraiment propice à des enregistrements de type vocaux. Pour cela, nous avons étudié différentes méthodes de mesure de réponse impulsionnelle afin de déterminer celle qui correspondrait le mieux à notre cas. La mesure de la réponse impulsionnelle d'une salle se fait en utilisant une excitation $x(t)$ (dirac, bruit blanc,

sinus glissant) qui est diffusée par un haut-parleur. Nous captons la réponse à cette excitation $y(t)$ avec un microphone de mesure.



$$y(t) = h(t) * x(t)$$

A.3.1.2 Réponse impulsionnelle

La réponse impulsionnelle d'un système est le signal h(t), qu'il faut extraire du signal y(t) par déconvolution. Les deux méthodes les plus utilisés sont la MLS (Maximum Length Signal) et l'ESS (Exponentially Swept Sine). Afin d'avoir la meilleure réponse possible, nous avons opté pour la méthode ESS développé par le chercheur Angelo Farina en 2000 [121]. Elle consiste à faire un balayage de fréquences en utilisant un signal sinusoïdal dont la fréquence varie de façon exponentielle en fonction du temps. En effet, la méthode ESS présente de nombreux avantages par rapport à la méthode MLS :

— Elle est moins sensible aux effets non linéaires de la source,

— Il y a plus d'énergie à basse fréquences, et vu que les fréquences de la voix se trouvent en moyenne autour de 200Hz, cela permet une meilleure précision,

— Un seul balayage assez long est suffisant, ce qui permet d'ignorer les petites variations temporelles du système.

La diffusion du signal d'excitation et l'acquisition de la réponse se fait via le logiciel gratuit Audacity, avec le plugin AURORA, plugin créé par Angelo Farina afin de réaliser des mesures de certains paramètres acoustiques, dont la réponse impulsionnelle.

Le logiciel ne pouvant pas diffuser et enregistrer en même temps, nous avons utilisé deux ordinateurs pour cette manipulation. L'enregistrement a été fait avec deux microphones cardioïdes (captent le son à l'avant et sur les côtés et rejettent le son venant de l'arrière) placés à $1m$ de la source. Dans le cas de l'ESS, l'excitation $x(t)$ un signal sinusoïdal dont la fréquence varie de façon exponentielle en fonction du temps, commençant à f1 et finissant à f2, de période T=10s. Ici, Fs=44,1kHz / f1=20Hz / f2=20kHz.
Le signal x(t) est généré grâce au plugin AURORA. La déconvolution de la réponse impulsionnelle revient à faire la convolution du signal mesuré $y(t)$, avec le filtre inverse $z(t)$ (inversement

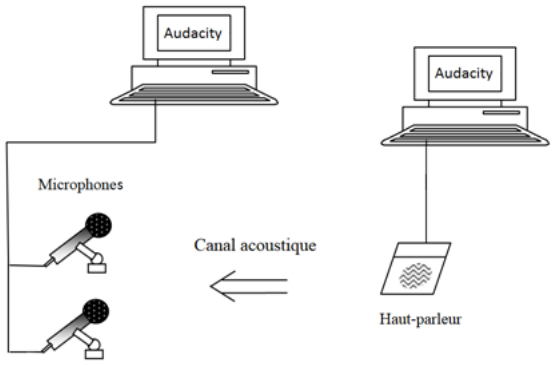


FIGURE A.1: Banc de mesure

$$x(t) = \sin\left[\frac{2\cdot\pi\cdot f_1\cdot T}{\ln\left(\frac{f_2}{f_1}\right)}\cdot\left(e^{\frac{t}{T}\cdot\ln\left(\frac{f_2}{f_1}\right)}-1\right)\right]$$

temporel du signal $x(t)$). Le plugin AURORA permet de réaliser la convolution de deux signaux $z(t)$ et $y(t)$. On obtient la réponse impulsionnelle $h(t)$ mais elle n'est pas exploitable si l'on veut tracer son spectre en fréquence. Afin de réaliser la convolution du signal mesuré y(t) et du filtre inverse z(t), et de tracer le spectre de la réponse impulsionnelle h(t) en fréquentiel, nous avons utilisé le logiciel Scilab.

Le spectre suivant est obtenu :

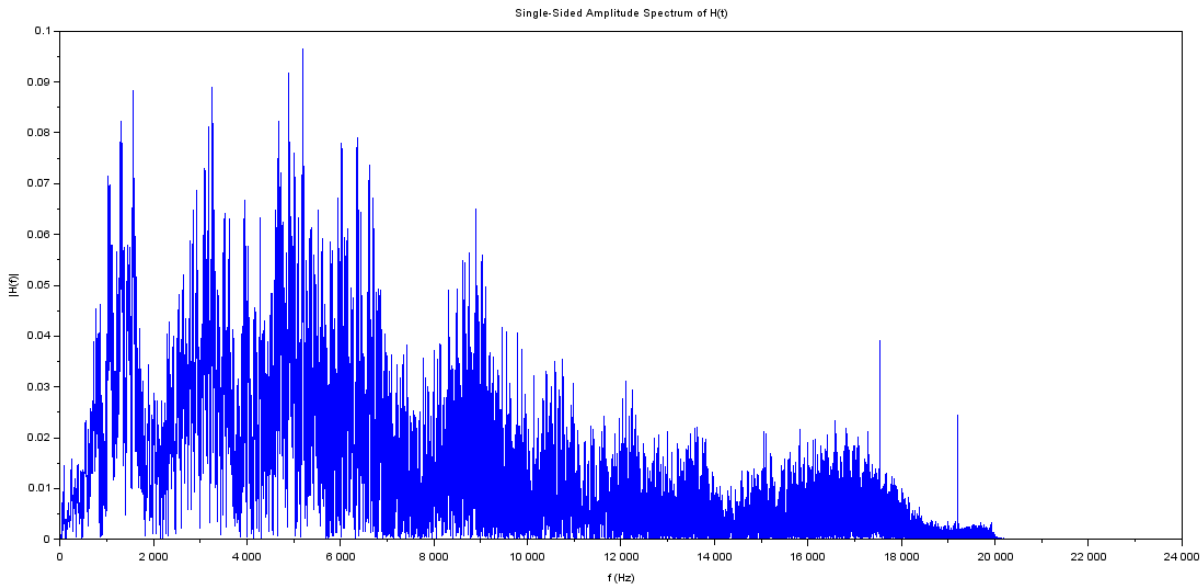


FIGURE A.2: Spectre de la réponse impulsionnelle.

Le spectre indique une bonne réponse de la salle dans la plage où se situe les fréquences de la

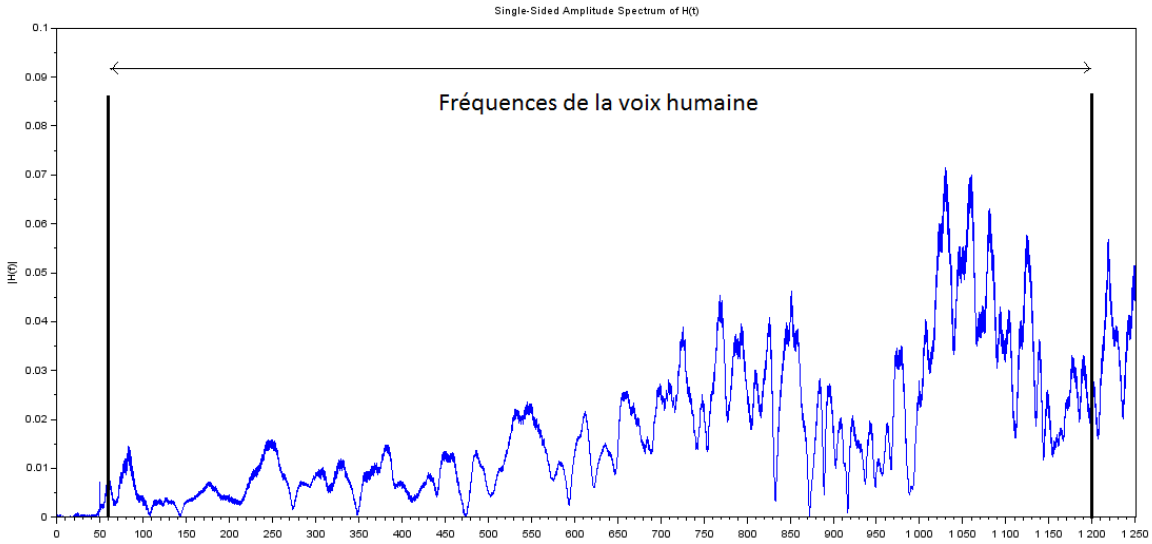


FIGURE A.3: Zoom sur la plage 60 Hz-1200 Hz.

voix humaine.

A.3.2 Choix du matériel

— **Deux filtres anti-réflexion LD Systems RF1** : permettent d'atténuer les réflexions sonores, bruits et échos indésirables dans des locaux sans traitement acoustique, et permet d'obtenir des enregistrements de voix ou d'instruments 'secs', sans réverbération ambiante.

— **Deux Rode NT1-A Complete Vocal Bundle** : pack contenant 1 microphone, une housse de protection noire, un câble XLR, un ensemble support micro "araignée " et un filtre anti-pop. Nous avons choisi ce microphone en particulier car il est à large membrane (moins de bruit de fond) et cardioïde, c'est à dire qu'il capte le son à l'avant et sur les côtés et le rejette à l'arrière ce qui permet de limiter les bruits de retour du microphone tandis que les microphones omnidirectionnels, eux, sont sensibles de manière uniforme à tous les angles.

— **Deux adaptateurs Cable XLR - Mini Jack pro snake** : permettent de brancher directement les microphones à l'ordinateur pour le traitement des enregistrements.

— **Une alimentation fantôme ART PRO AUDIO PHANTOM II PRO** : sert à alimenter les deux microphones. Ce sont des microphones à condensateur et il faut donc les alimenter avec une tension de +24V ou +48V.

— **Une prise secteur 9V EAGLETONE AD9** : prise secteur de l'alimentation fantôme.

A.3.3 Procédure d'enregistrement

Nous avons appliqué exactement la même procédure à chaque personne pour éviter d'avoir des différences inhérentes au locuteur. Le locuteur est donc assis en face de la structure composée des deux microphones et d'un tour en mousse. Ce dernier permet l'absorption des sons et donc supprime tout écho ce qui est important pour obtenir un enregistrement de qualité. La procédure est comme suit :

— Il y a sept fiches ,

— Chaque fiche correspond à une émotion,

— Pour chaque émotion il y a dix à douze phrases avec leur contexte,

— Le locuteur "lit" deux phrases de chaque émotion,

— Lors de la lecture, il est enregistré via notre installation (voir Figure A.4) à l'aide du logiciel libre Audacity,

— Les enregistrements sont stockés dans des fichiers audio (extension des fichiers .wav), nommés de cette façon : ⟨*id_locuteur*⟩ ⟨*id_phrase*⟩ ⟨*id_emotion*⟩.

Exemple : 10a01Pa.wav

- 10 représente l'identifiant du locuteur ;
- a01 représente l'identifiant du phrase ;
- Pa signifie que l'émotion jouée dans cet enregistrement est la peur, Ca pour la colère, Da pour le dégout, Ja pour la joie, Ta pour la tristesse, Sa pour la surprise et Na pour le neutre ;
- .wav est l'extension de l'enregistrement.

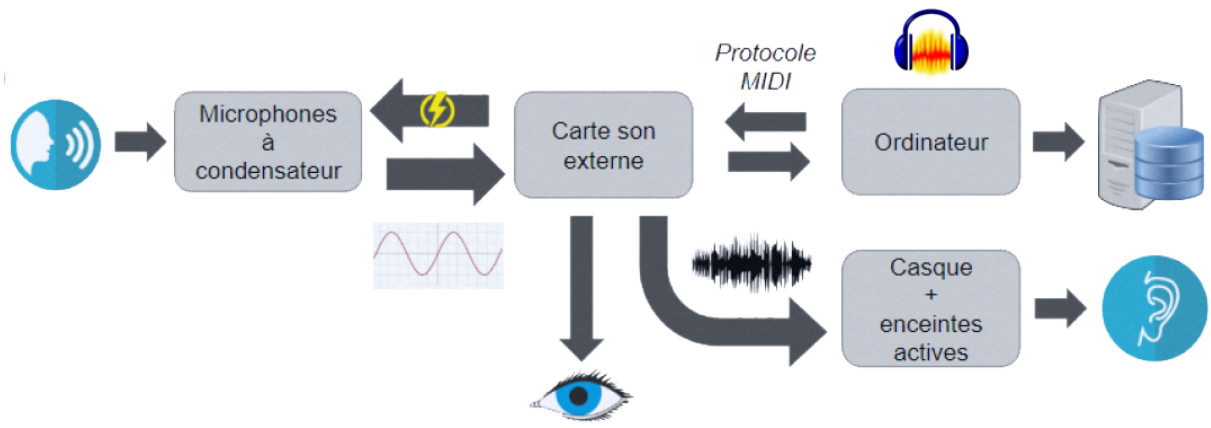


FIGURE A.4: Schéma de la chaine d'enregistrement

A.3.4 Résultat

Le tableau détaillé sur les locuteurs est le suivant :

Locuteur	Genre	Nationalité	Nombre d'enregistrements							Nombre total
			Colère	Joie	Peur	Surprise	Dégoût	Tristesse	Neutre	
1	H	Française	2	2	2	2	2	2	2	14
2	H	Espagnole	2	2	2	2	2	2	2	14
3	H	Française	2	2	2	2	2	2	2	14
4	H	Française	2	2	2	2	2	2	1	13
5	F	Française	2	2	2	2	2	2	2	14
6	H	Française	2	2	2	2	2	2	2	14
7	H	Française	2	2	2	2	2	2	2	14
8	F	Française	2	2	2	2	2	2	2	14
9	F	Française	6	6	6	6	6	6	6	42
10	H	Française	2	2	2	2	2	2	2	14
11	F	Française	2	2	2	2	2	2	1	13
12	H	Française	2	2	2	2	2	2	2	14
13	H	Française	2	2	2	2	2	2	2	14
14	H	Française	2	2	2	2	2	2	2	14
15	H	Française	2	2	2	2	2	2	2	14
16	H	Française	2	2	2	2	2	2	2	14
17	F	Espagnole	6	6	6	6	5	6	6	41
18	H	Française	2	2	2	2	2	2	2	14
19	H	Française	2	2	2	2	2	2	2	14
20	H	Française	2	2	2	2	2	2	2	14
21	H	Française	2	2	2	2	2	2	2	14
22	H	Française	3	2	2	2	2	2	2	15
23	H	Française	2	2	2	2	2	2	2	14
24	H	Française	2	2	2	2	2	2	2	14
25	H	Française	2	2	2	2	2	2	2	14
26	H	Française	2	2	2	2	2	2	2	14
27	F	Française	2	2	2	2	2	2	2	14
28	H	Française	2	2	2	2	2	2	2	14
29	H	Française	2	2	2	2	2	2	2	14
30	H	Française	2	2	2	2	2	2	2	14
31	H	Française	2	2	2	2	2	2	2	14
32	H	Française	2	2	2	2	2	2	2	14

Table A.2: Tableau détaillé contenant les informations sur les locuteurs.

A.4 Validation du corpus

A.4.1 Étude statistique des tests de reconnaissance des émotions

A.4.1.1 Étude 1 : Impact du biais sémantique sur la reconnaissance de l'émotion par l'humain

Dans ce qui suit nous décrivons la formulation du problème avec les hypothèses de travail. L'ensemble de tests de reconnaissance est réparti en deux échantillons indépendants :

— un échantillon e_1, de taille n_1, constitué de tests de reconnaissance effectués sur le corpus avec présence du biais sémantique,

— et un échantillon e_2, de taille n_2, constitué de tests de reconnaissance effectués sur le corpus modifié. Cette modification consiste à tronquer les mots susceptibles d'indiquer l'état émotionnel.

La question de l'influence du biais sémantique consiste, donc, à savoir si les taux de reconnaissance d'une émotion sont comparables ou significativement différents pour les deux échantillons. La formulation pour répondre à cette question est identique pour chaque émotion. Nous détaillerons, donc, la formulation pour une émotion actée par l'ensemble des locuteurs. Pour bien poser le problème, la réponse à la $i^{ème}$ test de reconnaissance d'une émotion, peut être modélisée par une variable aléatoire de Bernoulli R_i tel que :

$$R_{i=}\begin{cases}1 & si\ lors\ de\ la\ i^{ème}\ test\ l'émotion\ a\ été\ reconnue\\ 0 & sinon\end{cases}$$

A partir de ces variables aléatoires nous pouvons définir deux variables aléatoire $\overline{T_1}$ et $\overline{T}_2$ qui correspondent aux taux de reconnaissance dans les deux échantillons e_1 et e_2.

$$\overline{T}_1 = \frac{\sum_{i\in e_1} R_i}{n_1}$$

et

$$\overline{T}_2 = \frac{\sum_{i\in e_2} R_i}{n_2}$$

$\overline{T}_1$ et $\overline{T}_2$ sont définies comme sommes de n variables aléatoires supposées indépendantes et identiquement distribuées (i.e. de même loi), vu la taille de nos échantillons le théorème centrale

limite nous permet d'approcher leurs lois par deux lois gaussiennes $LG\left(p_1, \sigma_1 = \sqrt{\frac{p_1(1-p_1)}{n_1}}\right)$ et $LG\left(p_2, \sigma_2 = \sqrt{\frac{p_2(1-p_2)}{n_1}}\right)$ respectivement.
Les hypothèses, que nous souhaitons tester leurs validités sont alors $H_0 : p_1 = p_2 = p$ et $H_1 : p_1 \neq p_2$. Dans ces conditions, si l'hypothèse H_0 est vraie alors on a la variable $D = \overline{T}_1 - \overline{T}_2$ suit une loi gaussienne $LG\left(p, \sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$.
On dispose de $\bar{t}_1$ et $\bar{t}_1$ qui sont les taux de reconnaissances empiriques de deux échantillons. Pour une probabilité de se tromper α on rejette H_0 si

$$|\bar{t}_1 - \bar{t}_2| > \underbrace{t_{\alpha/2}\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}_{Seuil_\alpha}$$

Pour $\alpha = 0.05$ on a $t_{\alpha/2} = 1.96$, et pour $\alpha = 0.01$ on a $t_{\alpha/2} = 2.58$. Pour estimer p on utilise l'estimateur $\hat{p} = \frac{n_1\bar{t}_1 + n_2\bar{t}_2}{n_1+n_2}$. Résultats par testeur Le tableau A.3 récapitule les hypothèses adoptées pour l'étude.

A.4.1.2 Étude 2 : Intervalle de confiance la reconnaissance de l'émotion par l'humain

Dans cette partie, nous établissons l'intervalle de confiance des différents taux de reconnaissance des émotions (voir tableau A.4). Le but de cet intervalle est d'indiquer qu'il est possible à x% de chance que le résultat obtenu soit statistiquement valable. Pour chaque pourcentage, nous calculons l'intervalle de confiance à 95%. La formule suivante permet de calculer le rayon de l'intervalle de confiance :

$$r = t_{\alpha/2}\sqrt{\frac{p * (1-p)}{n}} \tag{A.1}$$

Avec p représente le taux de reconnaissance correcte et n est le nombre d'instances. Pour $\alpha = 0.05$ on a $t_{\alpha/2} = 1.96$, et pour $\alpha = 0.01$ on a $t_{\alpha/2} = 2.58$.
L'intervalle de confiance du pourcentage est alors : $IC = [p - r; p + r]$

	Colère	Joie	Peur	Surprise	Dégoût	Tristesse	Neutre
$\bar{t}_1$	0,6581	0,8293	0,4957	0,7119	0,7304	0,5088	0,8136
$\bar{t}_2$	0,8095	0,7805	0,5181	0,7073	0,8902	0,5125	0,9157
$\lvert t_1 - t_2 \rvert$	0,1514	0,0488	0,0223	0,0045	0,1598	0,0037	0,1021
n_1	117,0000	123,0000	117,0000	118,0000	115,0000	114,0000	118,0000
n_2	84,0000	82,0000	83,0000	82,0000	82,0000	80,0000	83,0000
$\hat{p}$	0,7214	0,8098	0,5050	0,7100	0,7970	0,5103	0,8557
$Seuil_{0.05}$	0,1257	0,1097	0,1406	0,1279	0,1140	0,1429	0,0987
$Seuil_{0.01}$	0,1654	0,1444	0,1851	0,1683	0,1500	0,1881	0,1299
Hypothèse pour un risque 5%	H1	H0	H0	H0	H1	H0	H1
Hypothèse pour un risque de 1%	H0	H0	H0	H0	H1	H0	H0

TABLE A.3: Les hypothèses pour un risque de 1% et de 5% des différents taux de reconnaissance des émotions pour le corpus avec et sans biais sémantique. $\bar{t}_1$ et $\bar{t}_1$ sont respectivement les taux de reconnaissances des échantillons avec et sans biais sémantique.

	Colère	Joie	Peur	Surprise	Dégout	Tristesse	Neutre
$\bar{t}_1$ $[IC_{95}]$	65,81% [57,22% ; 74,41%]	82,93 % [76,28% ; 89,58 %]	49,57% [40,51% ; 58,63%]	71,19% [63,01% ; 79,36%]	73,04% [64,93% ; 81,15%]	50,88% [41,70% ; 60,05%]	81,36% [74,33% ; 88,38%]
$\bar{t}_2$ $[IC_{95}]$	80,95% [72,55% ; 89,35%]	78,05% [70,37% ; 85,73%]	51,81% [43,38% ; 60,24%]	70,73% [62,63% 78,83%]	89,02% [82,22% ; 95,83%]	51,25% [42,61% ; 59,89%]	91,57% [85,21% ; 97,92%]
$\bar{t}_1$ $[IC_{99}]$	65,81% [54,50% ; 77,13%]	82,93 % [69,89% ; 95,96%]	49,57% [31,82% ; 67,33%]	71,19% [55,17% ; 87,20%]	73,04% [57,15% ; 88,94%]	50,88% [32,89% ; 68,86%]	81,36% [67,58% ; 95,13%]
$\bar{t}_2$ $[IC_{99}]$	80,95% [69,90% ; 92,01%]	78,05% [66,26% ; 89,84%]	51,81% [37,66% ; 65,96%]	70,73% [57,77% 83,70%]	89,02% [80,12% ; 97,93%]	51,25% [36,83% ; 65,67%]	91,57% [83,70% ; 99,44%]

Table A.4: L'intervalle de confiance à 95% et à 99% des différents taux de reconnaissance des émotions pour le corpus avec et sans biais sémantique. $\bar{t}_1$ et $\bar{t}_1$ sont respectivement les taux de reconnaissances des échantillons avec et sans biais sémantique.

A.4.2 Impact du testeur

Les tableaux A.5 et A.6 présentent les scores par testeur selon le genre et la culture pour la base avec et sans biais sémantique.

TABLE A.5: Taux de reconnaissance par testeur pour la base avec biais sémantique.

Testeur	Genre	Culture	Taux de reconnaissance
1	H	Berbère	100%
2	H	Française	100%
3	H	Française	92,85%
4	F	Arabe	92,00%
5	F	Française	85,71%
6	H	Française	85,71%
7	H	Française	85,71%
8	F	Française	85,71%
9	H	Française	85,71%
10	H	Non renseignée	83,33%
11	F	Non renseignée	78,58%
12	F	Non renseignée	78,57%
13	H	Arabe	78,57%
14	F	Arabe	78,57%
15	H	Française	78,57%
16	H	Tunisienne	78,57%
17	H	Tunisienne	78,57%
18	F	Tunisienne	78,57%
19	H	Tunisienne	78,57%
20	H	Tunisienne	78,57%
21	H	Française	78,57%
22	H	Française	78,57%
23	H	Française	75,00%
24	H	Camerounais	71,42%
25	H	Française	71,42%
26	H	Tunisienne	71,42%
27	H	Tunisienne	71,42%

Table A.5: Taux de reconnaissance par testeur pour la base avec biais sémantique(suite).

Testeur	Genre	Culture	Taux de reconnaissance
28	F	Française	71,42%
29	H	Française	71,42%
30	H	Française	71,42%
31	F	Tunisienne	71,42%
32	H	Française	71,42%
33	F	Française	71,42%
34	F	Orientale	71,00%
35	F	Tunisienne	71,00%
36	H	Arabe	70,00%
37	F	Tunisienne	66,66%
38	H	Non renseignée	64,28%
39	F	Française	64,28%
40	F	Tunisienne	64,28%
41	H	Arabe	64,28%
42	H	Française	64,28%
43	F	Non renseignée	57,14%
44	F	Tunisienne	57,14%
45	F	Tunisienne	57,14%
46	H	Française	57,14%
47	F	Française	57,14%
48	F	Française	57,14%
49	F	Arabe	57,00%
50	F	Tunisienne	57,00%
51	F	Française	50,00%
52	F	Non renseignée	50,00%
53	F	Tunisienne	50,00%
54	F	Non renseignée	50,00%
55	F	Tunisienne	50,00%
56	H	Française	50,00%
57	H	Française	42,85%
58	H	Tunisienne	42,85%

TABLE A.5: Taux de reconnaissance par testeur pour la base avec biais sémantique(suite).

Testeur	Genre	Culture	Taux de reconnaissance
59	F	Tunisienne	42,85%
60	H	Arabe	42,85%
61	H	Française	42,85%
62	H	Française	42,85%
63	F	Tunisienne	0%
64	F	Arabe	0%

TABLE A.6: Taux de reconnaissance par testeur pour la base sans biais sémantique.

Testeur	Genre	Culture	Taux de reconnaissance
1	H	Française	92,85%
2	F	Française	92,85%
3	H	Française	92,85%
4	H	Française	92,85%
5	H	Bresilienne	85,71%
6	H	Française	85,71%
7	H	Française	85,71%
8	H	Française	85,71%
9	H	Française	85,71%
10	H	Française	85,71%
11	H	Française	85,71%
12	F	Française	78,57%
13	F	Française	78,57%
14	F	Française	78,57%
15	H	Française	78,57%
16	F	Française	78,57%
17	H	Française	78,57%
18	F	Française	78,57%
19	F	Française	78,57%
20	H	Française	78,57%
21	F	Française	78,57%

22	H	Française	78,57%
23	H	Française	71,42%
24	H	Française	71,42%
25	H	Africaine	71,42%
26	H	Française	71,42%
27	F	Française	71,42%
28	F	Française	71,42%
29	H	Française	64,28%
30	H	Française	64,28%
31	H	Américaine	64,28%
32	H	Française	64,28%
33	F	Asiatique	64,28%
34	H	Française	57,14%
35	H	Française	57,14%
36	H	Française	57,14%
37	H	Française	57,14%
38	F	Française	57,14%
39	H	Française	50,00%
40	H	Française	50,00%
41	H	Française	50,00%
42	F	Française	42,85%

Table A.6: Taux de reconnaissance par testeur pour la base sans biais sémantique

Annexe B

Résultats détaillés des expériences du système de RAE

Résultats avec fusion de caractéristiques

Dans ce qui suit, nous présentons les résultats détaillés des expériences sur la fusion de caractéristiques.

Caractéristique	Taux de reconnaissance (%)		
	RL	SVM	RNN
MFCC+MSF	73.20	76.60	63.67
SMFCC+MAF	68.11	76.98	72.59
SMFCC+ECC	65.47	76.60	72.95
SMFCC+MFF	62.45	76.03	68.88
MFF+MAF	60.18	68.11	64.52
ECC+EFCC	50.18	52.45	52.00
ECC+MAF	60.56	70.00	67.56
ECC+MFF	61.13	67.54	61.90
EFCC+MAF	61.50	65.84	67.10
EFCC+MFF	61.69	64.33	60.24
ECC+MFF+MAF	58.49	70.18	69.57
EFCC+MFF+MAF	58.67	69.81	69.16
ECC+EFCC+MAF	57.54	67.16	67.35
ECC+EFCC+MFF	54.90	65.28	63.24
SMFCC+MFF+MAF	59.43	76.60	73.81
EFCC+MFF+MAF+SMFCC	54.52	79.24	76.50
SMFCC+ECC+EFCC+MAF+MFF	35.47	**80.56**	77.09

TABLE B.1: Récapitulatif des résultats des trois classifieurs avec les différentes combinaisons sur la base allemande.

Caractéristique	Taux de reconnaissance (%)		
	RL	SVM	RNN
MFCC+MSF	84.25	91.83	91.36
SMFCC+MAF	91.25	95.69	89.97
SMFCC+ECC	90.49	96.20	89.54
SMFCC+MFF	89.83	95.76	90.35
MFF+MAF	70.03	78.47	72.88
ECC+EFCC	58.47	70.43	69.01
ECC+MAF	71.55	78.82	77.93
ECC+MFF	61.50	72.54	71.38
EFCC+MAF	70.76	78.72	77.06
EFCC+MFF	60.91	71.49	68.61
ECC+MFF+MAF	74.95	81.40	79.87
EFCC+MFF+MAF	74.48	80.77	79.72
ECC+EFCC+MAF	73.50	78.55	78.40
ECC+EFCC+MFF	65.29	73.84	73.05
SMFCC+MFF+MAF	92.38	96.00	90.52
EFCC+MFF+MAF+SMFCC	92.99	**96.24**	90.72
SMFCC+ECC+EFCC+MAF+MFF	93.04	95.81	90.65

Table B.2: Récapitulatif des résultats des trois classifieurs avec les différentes combinaisons sur la base espagnole.

Résultats avec et sans sélection de caractéristiques

Nous présentons ici la sélection des caractéristiques pour les 2 corpus (allemand et espagnol). La sélection de caractéristiques donne de meilleurs résultats avec la base allemande allant jusqu'à +37 points pour la combinaison SMFCC, ECC, EFCC, MAF et MFF en utilisant le classifieur RL (voir tableau B.3).

Caractéristiques	Résultats sans sélection	Nombre de caractéristiques sélectionnées	Résultats avec sélection
MFCC	71.69%	28 sur 60	76.22%
MSF	67.16%	75 sur 95	66.03%
SMFCC	75.28%	55 sur 60	77.54%
ECC	54.71%	40 sur 60	54.33%
EFCC	57.92%	44 sur 60	57.92%
MAF	59.24%	120 sur 132	59.24%
MFF	63.96%	69 sur 96	62.45%
MFCC+MSF	75.66%	103 sur 155	**80.56**%
SMFCC+ECC	72.07%	95 sur 120	75.47%
SMFCC+MAF	70.00%	175 sur 192	72.26%
SMFCC+MFF	70.75%	124 sur 156	74.15%
SMFCC+MFF+MAF	63.58%	244 sur 288	72.64%
EFCC+MFF+MAF+SMFCC	53.58%	288 sur 348	70.37%
SMFCC+ECC+EFCC+MAF+MFF	29.05%	328 sur 408	66.79%

TABLE B.3: Comparaison des résultats sans et avec sélection de caractéristiques en utilisant le classifieur RL sur la base allemande (avec SN).

Caractéristiques	Résultats sans sélection	Nombre de caractéristiques sélectionnées	Résultats avec sélection
MFCC	73.77%	28 sur 60	76.79%
MSF	72.45%	75 sur 95	71.32%
SMFCC	75.66%	55 sur 60	75.84%
ECC	57.16%	40 sur 60	57.35%
EFCC	56.41%	48 sur 60	58.49%
MAF	65.47%	120 sur 132	66.41%
MFF	65.66%	69 sur 96	68.11%
MFCC+MSF	80.00%	103 sur 155	80.00%
SMFCC+ECC	78.67%	95 sur 120	81.69%
SMFCC+MAF	79.43%	175 sur 192	81.13%
SMFCC+MFF	80.18%	124 sur 156	81.13%
SMFCC+MFF+MAF	83.96%	244 sur 288	83.96%
EFCC+MFF+MAF+SMFCC	85.84%	288 sur 348	84.15%
SMFCC+ECC+EFCC+MAF+MFF	84.52%	328 sur 408	**85.47**%

TABLE B.4: Comparaison des résultats sans et avec sélection de caractéristiques en utilisant le classifieur SVM sur la base allemande (avec SN).

Caractéristiques	Résultats sans sélection	Nombre de caractéristiques sélectionnées	Résultats avec sélection
MFCC	46.97%	51 sur 60	76.05%
MSF	45.29%	85 sur 95	67.59%
SMFCC	83.06%	29 sur 60	83.90%
ECC	51.67%	33 sur 60	51.68%
EFCC	51.37%	48 sur 60	51.62%
MAF	65.36%	58 sur 108	54.12%
MFF	49.08%	49 sur 90	30.84%
MFCC+MSF	55.33%	136 sur 155	82.15%
SMFCC+ECC	86.19%	105 sur 120	86.63%
SMFCC+MAF	88.17%	150 sur 168	86.92%
SMFCC+MFF	86.30%	141 sur 150	84.03%
SMFCC+MFF+MAF	89.40%	199 sur 258	87.11%
EFCC+MFF+MAF+SMFCC	89.65%	247 sur 318	88.52%
SMFCC+ECC+EFCC+MAF+MFF	89.95%	280 sur 378	**89.30**%

Table B.5: Comparaison des résultats sans et avec sélection de caractéristiques en utilisant le classifieur RL sur la base espagnole (avec SN).

Caractéristiques	Résultats sans sélection	Nombre de caractéristiques sélectionnées	Résultats avec sélection
MFCC	85.38%	51 sur 60	85.84%
MSF	79.93%	85 sur 95	81.92%
SMFCC	85.09%	29 sur 60	77.77%
ECC	64.67%	33 sur 60	65.02%
EFCC	62.73%	48 sur 60	61.96%
MAF	68.25%	58 sur 108	60.32%
MFF	60.39%	49 sur 90	33.12%
MFCC+MSF	88.20%	136 sur 155	88.20%
SMFCC+ECC	86.95%	105 sur 120	72.35%
SMFCC+MAF	87.40%	150 sur 168	81.21%
SMFCC+MFF	86.15%	141 sur 150	87.06%
SMFCC+MFF+MAF	88.18%	199 sur 258	88.63%
EFCC+MFF+MAF+SMFCC	87.25%	247 sur 318	**88.93**%
SMFCC+ECC+EFCC+MAF+MFF	87.25%	280 sur 378	87.67%

Table B.6: Comparaison des résultats sans et avec sélection de caractéristiques en utilisant le classifieur RNN sur la base espagnole (avec SN).

Bibliographie

[1] P. Gayathri et Sarprasatham Matilda : Emotion recognition : A survey. *International journal of advanced research in computer science and applications*, 2015.

[2] Paul Ekman et Richard J. Davidson : *The nature of emotion : Fundamental questions.* New York : Oxford University Press, 1994.

[3] Jacques Lecomte : *Les 30 notions de la psychologie.* Dunod, 2013.

[4] Vidrascu Laurence : *Analyse et détection des émotions verbales dans les interactions orales.* Thèse de doctorat, Université Paris Sud - Paris XI, 2007.

[5] Paul Ekman : An argument for basic emotions. *Cognition & emotion*, (3-4):169–200, 1992.

[6] Andrew Ortony et Terence J Turner : What's basic about basic emotions ? *Psychological review*, (3):315, 1990.

[7] Paul Ekman, Wallace V Friesen et Phoebe Ellsworth : *Emotion in the Human Face : Guide-lines for Research and an Integration of Findings : Guidelines for Research and an Integration of Findings.* Pergamon, 1972.

[8] Paul Ekman : *Basic emotions*, chapitre 3, pages 45–60. Wiley Online Library, 1999.

[9] Charles Darwin : *L'expression des émotions chez l'homme et les animaux.* 1877.

[10] Wallace V Friesen et Paul Ekman : *Pictures of facial affect.* Consulting psychologists press, 1976.

[11] Nathalie Golouboff : *La reconnaissance des émotions faciales : développement chez l'enfant sain et épileptique.* Thèse de doctorat, Paris 5, 2007.

[12] Dimitrios Ververidis et Constantine Kotropoulos : Emotional speech recognition : Resources, features, and methods. *Speech communication*, (9):1162–1181, 2006.

[13] Marie Tahon : *Analyse acoustique de la voix émotionnelle de locuteurs lors d'une interaction humain-robot.* Thèse de doctorat, Paris 11, 2012.

[14] James A Russell : A circumplex model of affect. *Journal of personality and social psychology*, (6):1161, 1980.

[15] Lisa Feldman Barrett et James A Russell : Independence and bipolarity in the structure of current affect. *Journal of personality and social psychology*, (4):967, 1998.

[16] Carroll E Izard : The face of emotion. *New York : Appleton-Century-Crofts*, 1971.

[17] Paul Ekman : What emotion categories or dimensions can observers judge from facial behavior ? *Emotions in the human face*, pages 39–55, 1982.

[18] Judith A Hall et David Matsumoto : Gender differences in judgments of multiple emotions from facial expressions. *Emotion*, (2):201, 2004.

[19] James A Russell : Is there universal recognition of emotion from facial expression ? a review of the cross-cultural studies. *Psychological bulletin*, (1):102, 1994.

[20] Hamza Hamdi : *Plate-forme multimodale pour la reconnaissance d'émotions via l'analyse de signaux physiologiques : application à la simulation d'entretiens d'embauche.* Thèse de doctorat, Université d'Angers, 2012.

[21] Swati Johar : Psychology of voice. *In Emotion, Affect and Personality in Speech*, pages 9–15. Springer, 2016.

[22] Klaus R Scherer : Vocal communication of emotion : A review of research paradigms. *Speech communication*, (1-2):227–256, 2003.

[23] Dimitrios Ververidis et Constantine Kotropoulos : A state of the art review on emotional speech databases. *In Proceedings of 1st Richmedia Conference*, pages 109–119. Citeseer, 2003.

[24] Monorama Swain, Aurobinda Routray et P Kabisatpathy : Databases, features and classifiers for speech emotion recognition : a review. *International Journal of Speech Technology*, pages 1–28, 2018.

[25] Thurid Vogt et Elisabeth André : 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*, pages 474–477. IEEE, 2005.

[26] Robert Ruiz : Analyse acoustique de la voix pour la détection de perturbations psychophysiologiques : Application au contexte aéronautique. *Habilitation à Diriger des Recherches, Laboratoire de Recherche en Audiovisuel (LA. RA)*, 2012.

[27] Chloé Clavel : *Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales.* Thèse de doctorat, Télécom ParisTech, 2007.

[28] Ruth Menahem : La voix et la communication des affecta. *L'année psychologique*, 83(2): 537–560, 1983.

[29] Erwan Pépiot : *Voix de femmes, voix d'hommes : différences acoustiques, identification du genre par la voix et implications psycholinguistiques chez les locuteurs anglophones et francophones*. Thèse de doctorat, Université Paris VIII Vincennes-Saint Denis, 2013.

[30] Josiane Clarenc : Les caractéristiques articulatoires et acoustiques, 2006. http://asl.univ-montp3.fr/e58fle/caracteristiquesarticulatoiresetacoustiques.pdf.

[31] Elena Sender : La voix : véhicule des émotions, Novembre 2013. https://www.sciencesetavenir.fr/sante/la-voix-vehicule-des-emotions_26363.

[32] Geneviève Caelen-Haumont : *Les états émotionnels et la Prosodie : paradigmes, modèles, paramètres*, pages 397–424. Nguyen, Noël. Phonologie et phonétique : Forme et substance, Hermès, 2005.

[33] Juliette Sasserath *et al.* : Voix et émotions : impact de l'intensité émotionnelle sur la hauteur tonale en fonction de l'étendue fréquentielle des participants. Mémoire de D.E.A., Université de Liège, Liège, Belgique, 2018.

[34] Yazid Attabi : *Reconnaissance automatique des émotions à partir du signal acoustique*. Thèse de doctorat, École de technologie supérieure, 2008.

[35] Alison Tickle : Englishand japanese speaker's emotion vocalizations and recognition : a comparison highlighting vowel quality. *In ISCA Workshop on Speech and Emotion*. Citeseer, 2000.

[36] Siqing Wu : *Recognition of Human Emotion in Speech Using Modulation Spectral Features and Support Vector Machines*. Thèse de doctorat, 2009.

[37] Safa Chebbi et Sofia Ben Jebara : On the use of pitch-based features for fear emotion detection from speech. *In Advanced Technologies for Signal and Image Processing (ATSIP), 2018 4th International Conference on*, pages 1–6. IEEE, 2018.

[38] Chloé Clavel et Gaël Richard : Système d'interaction émotionnelle, chapitre reconnaissance acoustique des émotions. c. pelachaud. *Hermès*, page 21, 2010.

[39] Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman et M Wilkes : Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, 47(7):829–837, 2000.

[40] Sophia Favreau : *Psychoacoustique : étude comparative des seuils différentiels d'intensité et de hauteur en fonction du niveau de perte auditive*. Thèse de doctorat, Université de Lorraine, 2015.

[41] Miriam Kienast et Walter F Sendlmeier : Acoustical analysis of spectral and temporal changes in emotional speech. *In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[42] Rui Cai, Lie Lu, Hong-Jiang Zhang et Lian-Hong Cai : Highlight sound effects detection in audio stream. *In Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 3, pages III–37. IEEE, 2003.

[43] Clément Chastagnol : *Reconnaissance automatique des dimensions affectives dans l'interaction orale homme-machine pour des personnes dépendantes.* Thèse de doctorat, Université Paris Sud-Paris XI, 2013.

[44] Ismail Shahin, Ali Bou Nassif et Shibani Hamsa : Emotion recognition using hybrid gaussian mixture model and deep neural network. *IEEE Access*, 7:26777–26787, 2019.

[45] G Adda, G Chollet, S Essid, T Fillon, M Garnier-Rizet, C Hory et L Zouari : *Traitement des modalités ≪audio≫ et ≪parole≫*, chapitre 4. Sémantique et multimodalité en analyse de l'information, 2011.

[46] M Campedel et Pierre Hoogstoël : Sémantique et multimodalité en analyse de l'information. *Lavoisier-Hermes Science*, 2011.

[47] David W Aha et Richard L Bankert : Feature selection for case-based classification of cloud types : An empirical comparison. *In Proceedings of the AAAI-94 workshop on Case-Based Reasoning*, page 112, 1994.

[48] Huan Liu et Lei Yu : Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4):491–502, 2005.

[49] Pierre-yves Oudeyer : Novel useful features and algorithms for the recognition of emotions in human speech. *In Speech Prosody 2002, International Conference*, 2002.

[50] Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh et Wen-Yuan Liao : Detecting emotions in mandarin speech. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 10, Number 3, September 2005 : Special Issue on Selected Papers from ROCLING XVI*, 10(3):347–362, 2005.

[51] Dimitrios Ververidis, Constantine Kotropoulos et Ioannis Pitas : Automatic emotional speech classification. *In Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–593. IEEE, 2004.

[52] Bo Xie, Ling Chen, Gen-Cai Chen et Chun Chen : Feature selection for emotion recognition of mandarin speech. *Journal-Zhejiang University Engineering Science*, 41(11):1816, 2007.

[53] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao et Te-Won Lee : Emotion recognition by speech signals. *In Eighth European Conference on Speech Communication and Technology*, 2003.

[54] Kwee-Bo Sim, In-Hun Jang et Chang-Hyun Park : The development of interactive feature selection and ga feature selection method for emotion recognition. *In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 73–81. Springer, 2007.

[55] Tetsuya Noda, Yoshikazu Yano, Shinji Doki et Shigeru Okuma : Adaptive emotion recognition in speech by feature selection based on kl-divergence. *In Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, volume 3, pages 1921–1926. IEEE, 2006.

[56] Imran Naseem, Roberto Togneri et Mohammed Bennamoun : Linear regression for face recognition. *IEEE transactions on pattern analysis and machine intelligence*, (11):2106–2112, 2010.

[57] A Milton, S Sharmy Roy et S Tamil Selvi : Svm scheme for speech emotion recognition using mfcc feature. *International Journal of Computer Applications*, (9), 2013.

[58] Weishan Zhang, Dehai Zhao, Zhi Chai, Laurence T Yang, Xin Liu, Faming Gong et Su Yang : Deep learning and svm-based emotion recognition from chinese speech for smart affective services. *Software : Practice and Experience*, (8):1127–1138, 2017.

[59] GS Divya Sree, P Chandrasekhar et B Venkateshulu : Svm based speech emotion recognition compared with gmm-ubm and nn. *International Journal of Engineering Science*, 2016.

[60] Rahul B Lanjewar, Swarup Mathurkar et Nilesh Patel : Implementation and comparison of speech emotion recognition system using gaussian mixture model (gmm) and k-nearest neighbor (k-nn) techniques. *Procedia computer science*, 49:50–57, 2015.

[61] B. Ingale Ashish et D S Chaudhari : Speech emotion recognition using hidden markov model and support vector machine. *International Journal of Advanced Engineering research Study*, pages 316–318, 2012.

[62] Martin Vondra et Robert Vích : Recognition of emotions in german speech using gaussian mixture models. *In Multimodal Signals : Cognitive and Algorithmic Issues*, pages 256–263. Springer, 2009.

[63] Ruinian Chen, Ying Zhou et Yanmin Qian : Emotion recognition using support vector machine and deep neural network. *In National Conference on Man-Machine Speech Communication*, pages 122–131. Springer, 2017.

[64] Seyedmahdad Mirsamadi, Emad Barsoum et Cha Zhang : Automatic speech emotion recognition using recurrent neural networks with local attention. *In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 2227–2231. IEEE, 2017.

[65] Yixiong Pan, Peipei Shen et Liping Shen : Speech emotion recognition using support vector machine. *International Journal of Smart Home*, (2):101–108, 2012.

[66] Lei Zhang, Shuai Wang et Bing Liu : Deep learning for sentiment analysis : A survey. *arXiv preprint arXiv :1801.07883 (2018)*, 2018.

[67] Ling He, Margaret Lech, Namunu C Maddage et Nicholas B Allen : Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. *Biomedical Signal Processing and Control*, 6(2):139–146, 2011.

[68] Mansour Sheikhan, Mahdi Bejani et Davood Gharavian : Modular neural-svm scheme for speech emotion recognition using anova feature selection method. *Neural Computing and Applications*, 23(1):215–227, 2013.

[69] Mina Hamidi et Muharram Mansoorizade : Emotion recognition from persian speech with neural network. *International Journal of Artificial Intelligence & Applications*, 3(5): 107, 2012.

[70] Monorama Swain, Subhasmita Sahoo, Aurobinda Routray, Prithviraj Kabisatpathy et Jogendra N Kundu : Study of feature combination using hmm and svm for multilingual odiya speech emotion recognition. *International Journal of Speech Technology*, 18(3):387–393, 2015.

[71] Sih-Huei Chen, Jia-Ching Wang, Wen-Chi Hsieh, Yu-Hao Chin, Chin-Wen Ho et Chung-Hsien Wu : Speech emotion classification using multiple kernel gaussian process. *In 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–4. IEEE, 2016.

[72] D Pravena et D Govind : Development of simulated emotion speech database for excitation source analysis. *International Journal of Speech Technology*, 20(2):327–338, 2017.

[73] Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu et Guan-Zheng Tan : Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, 273:271–280, 2018.

[74] Dimitrios Ververidis et Constantine Kotropoulos : A review of emotional speech databases. *In Proc. Panhellenic Conference on Informatics (PCI)*, volume 2003, pages 560–574, 2003.

[75] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier et B. Weiss : A Database of German Emotional Speech. *INTERSPEECH (2005)*, 2005.

[76] Emotional speech synthesis database. http://catalog.elra.info/en-us/repository/browse/emotional-speech-synthesis-database/629db920a9e811e7a093ac9e1701ca021bdb22603cbc4702a3b6f592d250e427/.

[77] Motamed Sara, Setayeshi Saeed et Azam Rabiee : Speech emotion Recognition Based on a Modified Brain Emotional Learning Model. *Elsevier*, pages 32–38, 2017.

[78] Reinhard Pekrun : Emotions as drivers of learning and cognitive development. *In New perspectives on affect and learning technologies*, pages 23–39. Springer, 2011.

[79] Isabelle Puozzo : Pédagogie de la créativité : de l'émotion à l'apprentissage. *Éducation et socialisation. Les Cahiers du CERFEE*, (33), 2013.

[80] Soumaya Chaffar et Claude Frasson : Apprentissage machine pour la prédiction de la réaction émotionnelle de l'apprenant. *Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation*, 14:13–pages, 2007.

[81] Le ministère de l'Enseignement supérieur de la Recherche et de l'Innovation : Chiffres-clés de l'égalité femmes-hommes. Journée internationale des droits des femmes, 2018. http://cache.media.enseignementsup-recherche.gouv.fr/file/Brochures/32/8/parite2018_stats_A5_11d_908328.pdf.

[82] Chul Min Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee et Shrikanth Narayanan : Emotion recognition based on phoneme classes. *In Eighth International Conference on Spoken Language Processing*, 2004.

[83] Adi Lausen, Kurt Hammerschmidt et Annekathrin Schacht : Emotion recognition and confidence ratings predicted by vocal stimulus type and acoustic parameters. *PsyArXiv*, 2019.

[84] Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh et Wen-Yuan Liao : Combining acoustic features for improved emotion recognition in mandarin speech. *In International conference on affective computing and intelligent interaction*, pages 279–285. Springer, 2005.

[85] Md Jahangir Alam, Yazid Attabi, Pierre Dumouchel, Patrick Kenny et Douglas D O'Shaughnessy : Amplitude modulation features for emotion recognition from speech. *In INTERSPEECH*, pages 2420–2424, 2013.

[86] Siqing Wu, Tiago H Falk et Wai-Yip Chan : Automatic speech emotion recognition using modulation spectral features. *Speech communication*, 53(5):768–785, 2011.

[87] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung et Henry H. Liu : The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London A : Mathematical, Physical and Engineering Sciences*, pages 903–995, 1998.

[88] Jean-Christophe Cexus, Abdel-ouahab Boudraa et Abdelkhalek Bouchikhi : Tht et transformation de hough pour la détection de modulations linéaires de fréquence. *In XXIIe colloque GRETSI (traitement du signal et des images), Dijon (FRA), 8-11 septembre 2009*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images, 2009.

[89] Rajib Sharma, Leandro Vignolo, Gastón Schlotthauer, Marcelo A Colominas, H Leonardo Rufiner et SRM Prasanna : Empirical mode decomposition for adaptive am-fm analysis of speech : a review. *Speech Communication*, pages 39–64, 2017.

[90] Steven B Davis et Paul Mermelstein : Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. pages 357–366. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980.

[91] Rachid Harba : *Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole*. Thèse de doctorat, Université d'Orléans (France, 2012.

[92] Hacine Gharbi Abdenour : *Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole*. Thèse de doctorat, Université Ferhat Abbas de Sétif 1, 2012.

[93] Oumar Niang, Mouhamed Ould Guerra, Abdoulaye Thioune, Eric Deléchelle, Mary Teuw Niane et Jacques Lemoine : A propos de l'orthogonalité dans la décomposition modale empirique. *In Conference GretSI*, 2011.

[94] Petros Maragos, James F Kaiser et Thomas F Quatieri : Energy separation in signal modulations with application to speech analysis. *IEEE transactions on signal processing*, 41(10):3024–3051, 1993.

[95] Xiang Li et Xin Li : Speech emotion recognition using novel hht-teo based features. *JCP*, 6(5):989–998, 2011.

[96] Kais Khaldi, Abdel-Ouahab Boudraa et Ali Komaty : Speech enhancement using empirical mode decomposition and the teager–kaiser energy operator. *The Journal of the Acoustical Society of America*, (1):451–459, 2014.

[97] Xiang Li, Xin Li, Xiaoming Zheng et Dexing Zhang : Emd-teo based speech emotion recognition. *Life System Modeling and Intelligent Computing*, pages 180–189, 2010.

[98] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung et Henry H Liu : The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *In Proceedings of the Royal Society of London A : mathematical, physical and engineering sciences*, volume 454, pages 903–995. The Royal Society, 1998.

[99] Jean-Christophe Cexus et Abdel-Ouahab Boudraa : Nonstationary signals analysis by teager-huang transform (tht). *In Signal Processing Conference, 2006 14th European*, pages 1–5. IEEE, 2006.

[100] Petros Maragos, Thomas F Quatieri et James F Kaiser : Speech nonlinearities, modulations, and energy operators. *In Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 421–424. IEEE, 1991.

[101] Alexandros Potamianos et Petros Maragos : A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation. *Signal processing*, 37(1):95–120, 1994.

[102] Les Atlas et Shihab A Shamma : Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing*, 2003:668–675, 2003.

[103] Aron Arnold : *La voix genrée, entre idéologies et pratiques–Une étude sociophonétique*. Thèse de doctorat, Sorbonne Paris Cité, 2015.

[104] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub et Catherine Cleder : Automatic speech emotion recognition using machine learning. *In Social Media and Machine Learning*. IntechOpen, 2019.

[105] Svm and kernel methods matlab toolbox. `http://asi.insa-rouen.fr/enseignants/~arakoto/toolbox/`.

[106] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Mohamed Ali Mahjoub, Kosai Raoof et Catherine Cléder : Speech emotion recognition : Recurrent neural networks compared to svm and linear regression, 2017.

[107] Divya Sree G.S., Chandrasekhar P. et Venkateshulu B. : SVM Based Speech Emotion Recognition Compared with GMM-UBM and NN. *IJESC*, 6, 2016.

[108] Wootaek Lim, Daeyoung Jang et Taejin Lee : Speech emotion recognition using convolutional and recurrent neural networks. *In 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–4. IEEE, 2016.

[109] Ian Goodfellow, Yoshua Bengio et Aaron Courville : *Deep learning*. MIT press, 2016.

[110] The python deep learning library. `https://keras.io/`.

[111] Faiza Abdat : Reconnaissance automatique des émotions par données multimodales : expressions faciales et signaux physiologiques. *Université de Metz, France*, 2010.

[112] Fengxi Song, Zhongwei Guo et Dayong Mei : Feature selection using principal component analysis. *In 2010 international conference on system science, engineering design and manufacturing informatization*, volume 1, pages 27–30. IEEE, 2010.

[113] Belkacem Fergani *et al.* : News schemes for activity recognition systems using pca-wsvm, ica-wsvm, and lda-wsvm. *Information*, 6(3):505–521, 2015.

[114] Yahya Slimani, Mohamed Amir Essegir, Mouhamadou Lamine Samb, Fodé Camara et Samba Ndiaye : Approche de sélection d'attributs pour la classification basée sur l'algorithme rfe-svm. *Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées*, 17:197–219, 2014.

[115] Theodoros Iliou et Christos-Nikolaos Anagnostopoulos : Statistical evaluation of speech features for emotion recognition. *In 2009 Fourth International Conference on Digital Telecommunications*, pages 121–126. IEEE, 2009.

[116] Ali Harimi et Zeynab Esmaileyan : A database for automatic persian speech emotion recognition : collection, processing and evaluation. *International Journal of Engineering*, 27(1):79–90, 2014.

[117] Jesús B Alonso, Josué Cabrera, Manuel Medina et Carlos M Travieso : New approach in quantification of emotional intensity from the speech signal : emotional temperature. *Expert Systems with Applications*, 42(24):9554–9564, 2015.

[118] Iker Luengo, Eva Navas et Inmaculada Hernáez : Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Transactions on Multimedia*, 12(6):490–501, 2010.

[119] Houwei Cao, Ragini Verma et Ani Nenkova : Speaker-sensitive emotion recognition via ranking : Studies on acted and spontaneous speech. *Computer speech & language*, 29(1):186–202, 2015.

[120] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang et Lian Li : Speech emotion recognition using fourier parameters. *IEEE Transactions on affective computing*, 6(1):69–75, 2015.

[121] Angelo Farina : Simultaneous measurement of impulse response and distortion with a swept-sine technique. *In Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.

Titre : **Analyse acoustique de la voix pour la détection des émotions du locuteur**

Mots clés : Reconnaissance automatique des émotions, décomposition en modes empiriques, extraction de caractéristiques, RNN, SVM.

Résumé : L'objectif de cette thèse est de proposer un système de reconnaissance automatique des émotions (RAE) par analyse de la voix pour une application dans un contexte pédagogique d'orchestration de classe. Ce système s'appuie sur l'extraction de nouvelles caractéristiques, par démodulation en amplitude et en fréquence, de la voix ; considérée comme un signal multi-composantes modulé en amplitude et en fréquence (AM-FM), non-stationnaire et issue d'un système non-linéaire. Cette démodulation est basée sur l'utilisation conjointe de la décomposition en modes empiriques (EMD) et de l'opérateur d'énergie de Teager-Kaiser (TKEO). Dans ce système, le modèle discret (ou catégoriel) a été retenu pour représenter les six émotions de base (la tristesse, la colère, la joie, le dégoût, la peur et la surprise) et l'émotion dite neutre. La reconnaissance automatique a été optimisée par la recherche de la meilleure combinaison de caractéristiques, la sélection des plus pertinentes et par comparaison de différentes approches de classification. Deux bases de données émotionnelles de référence, en allemand et en espagnol, ont servi à entrainer et évaluer ce système. Une nouvelle base de données en Français, plus appropriée pour le contexte pédagogique a été construite, testée et validée.

Title : **Detection and analysis of human emotions through voice**

Keywords : Automatic emotion recognition, empirical mode decomposition, feature extraction, classification, RNN, SVM.

Abstract : The aim of this thesis is to propose a speech emotion recognition (SER) system for application in classroom. This system has been built up using novel features based on the amplitude and frequency (AM-FM) modulation model of speech signal. This model is based on the joint use of empirical mode decomposition (EMD) and the Teager-Kaiser energy operator (TKEO). In this system, the discrete (or categorical) emotion theory was chosen to represent the six basic emotions (sadness, anger, joy, disgust, fear and surprise) and neutral emotion. Automatic recognition has been optimized by finding the best combination of features, selecting the most relevant ones and comparing different classification approaches. Two reference speech emotional databases, in German and Spanish, were used to train and evaluate this system. A new database in French, more appropriate for the educational context was built, tested and validated.