République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Ecole Nationale Supérieure Polytechnique

المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Polytechnique

مخبر الإشارة و الإتصالات
Signal & Communications Lab.

Département d'Electronique

Laboratoire Signal et communications

Thèse de doctorat en Electronique

Présentée par :
Mr Houari HORKOUS

Master/Ingénieur d'Etat en Électronique, ENP Alger

Pour l'obtention du titre de
Docteur Troisième Cycle (D/LMD) en Électronique

La Reconnaissance des Emotions dans le Dialecte Algérien

Soutenue publiquement le 11/12/2021 devant le jury composé de :

Mme Latifa HAMAMI	Professeur	ENP	Président
Mme Mhania GUERTI	Professeur	ENP	Directeur de thèse
Mme Nadjia BENBLIDIA	Professeur	USDB1	Examinateur
Mme Malika KEDIR-TALHA	Professeur	USTHB	Examinateur
Mr Hicham BOUSBIA-SALAH	MCA	ENP	Examinateur

ENP 2021

République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Ecole Nationale Supérieure Polytechnique

Département d'Electronique

Laboratoire Signal et communications

Thèse de doctorat en Electronique

Présentée par :
Mr Houari HORKOUS

Master/Ingénieur d'Etat en Électronique, ENP Alger

Pour l'obtention du titre de
Docteur Troisième Cycle (D/LMD) en Électronique

La Reconnaissance des Emotions dans le Dialecte Algérien

Soutenue publiquement le 11/12/2021 devant le jury composé de :

Mme Latifa HAMAMI	Professeur	ENP	Président
Mme Mhania GUERTI	Professeur	ENP	Directeur de thèse
Mme Nadjia BENBLIDIA	Professeur	USDB1	Examinateur
Mme Malika KEDIR-TALHA	Professeur	USTHB	Examinateur
Mr Hicham BOUSBIA-SALAH	MCA	ENP	Examinateur

ENP 2021

ملخص:

هناك أبحاث قليلة في مجال التعرف على العاطفة في الكلام في اللهجة الجزائرية بسبب ندرة قواعد البيانات المتاحة في الكلام. في هذا السياق هذا االعمل يقدم قاعدة بيانات عاطفية جديدة في اللهجة الجزائرية .تم إنشاء هذه الأخيرة من أفلام جزائرية معروفة تصف الحرب الأهلية الجزائرية (2000-1992) والفترة التي تلتها. هدفنا في هذا العمل هو التعرف على العواطف في اللهجة الجزائرية. إستخراج المعلمات التي تميز العاطفة بشكل فعال يعد خطوة مهمة في تطوير أنظمة التعرف على العواطف في الكلام. لذلك إخترنا معلمات صوتية شائعة جدا في مجال التعرف على العواطف في الكلام: معلمات العروض, معلمات جودة الصوت و معلمات طيفية. تم إجراء تحليل للكشف عن تأثير مختلف العواطف المدروسة على المعلمات الصوتية المستخرجة. العديد من التجارب أجريت لدراسة تأثير عوامل مختلفة: المعلمات الصوتية, عدد العواطف, الجنس و اللغة على نظام التعرف على العواطف في اللهجة الجزائرية. إستنتجنا من التجارب أن أداء نظام التعرف على العواطف تحسن باستعمال مزيج من المعلمات الصوتية المختلفة. إستنتجنا من النتائج أيضا أن نظام التعرف على العواطف في اللهجة الجزائرية يتأثر بعدد العواطف في النظام و يتأثر بفئات الجنس (الذكور والإناث).

كلمات مفتاحية: نظام التعرف, اللهجة الجزائرية, معلمات العروض, معلمات جودة الصوت, معلمات طيفية.

Abstract:

Little researches have implemented the systems of Speech Emotion Recognition (SER) in the Algerian dialect due to the scarcity of the databases available on emotional speech in the Algerian dialect. In this context, this document presents a new Algerian Dialect Emotional Database (ADED). This database constructed from six famous movies in the Algerian dialect. These movies describe the civil war of years (1992-2000) in addition to the period that followed. The aim of this work is the SER in the Algerian Dialect. Extraction of parameters that can effectively characterize the emotions is an important step to develop the SER systems. We have chosen very popular acoustic descriptors in the SER : prosodic, voice quality and spectral parameters. An analyze is done to detect the influence of different emotional states on the parameters that chosen. Several experiments are performed to study the influence of different factors such as acoustic parameters, number of emotions, classes of gender and language on the SER system in the Algerian dialect. We have concluded from the results obtained that the performance of recognition system is improved by using combination of different acoustic parameters. We have also noted that the performance of SER system in the Algerian dialect is influenced by the number of emotions included in the recognition systems and influenced by the gender classes.

Keywords: ADED, Algerian Dialect, SER, Prosodic, Voice quality, Spectral.

Résumé :

Peu de recherches ont mis en oeuvre des systèmes de la Reconnaissance des Emotions dans la Parole (REP) dans le dialecte algérien en raison de la rareté des Bases de Données(BD) disponibles sur la parole émotionnelle du Dialecte Algérien. Dans ce contexte, ce document présente une nouvelle base de données émotionnelle du dialecte algérien (ADED). Cette BD est créée à partir des films algériens. Ces films décrivent la crise de la guerre civile (1992 - 2000) ainsi que la période qui la suit. Notre but dans ce travail est basé sur la REP dans le dialecte algérien. L'extraction des paramètres qui peuvent caractériser efficacement des émotions est une étape importante pour développer les systèmes de REP. Nous avons choisi de considérer des descripteurs acoustiques très utilisés dans la REP : des paramètres prosodiques, des paramètres de la qualité de la voix et des paramètres spectraux. Plusieurs expériences sont effectuées pour étudier l'influence des différents facteurs tels que les paramètres acoustiques, le nombre d'émotions, le sexe et la langue sur les systèmes de REP dans le dialecte algérien. Nous avons conclu d'après les résultats obtenus que la performance de système de reconnaissance est améliorée par l'utilisation de combinaison des différents paramètres acoustiques. Nous avons constaté également que la performance de système de REP dans le dialecte algérien est influencée par le nombre des émotions inclus dans les systèmes de reconnaissance et influencée aussi par les classes de sexe (Hommes et Femmes).

Mots clés : ADED, Dialecte Algérien, REP, Paramètres Prosodiques, Qualité de la voix, Spectraux

Remerciements

Tout d'abord, je remercie Allah le tout puissant de m'avoir donné le courage et la patience nécessaires pour mener ce travail à son terme.

Je tiens à remercier tout particulièrement ma directrice de thèse Mme **Guerti Mhania**, Professeur à l'Ecole Nationale Polytechnique, pour l'aide compétente qu'elle m'a apportée, pour sa patience et son encouragement. Son œil critique m'a été très précieux pour structurer le travail et pour améliorer la qualité des différentes sections J'exprime ma gratitude et mes vifs remerciements à Mme **Latifa Hamami**, Professeur à l'Ecole Nationale Polytechnique, d'avoir bien voulu présider le jury.

Je tiens à remercier Mme **Nadjia Benblidia**, Professeur à l'Université Saâd Dahleb de Blida, Mme **Malika Kedir-Talha**, Professeur à Université des Sciences et de la Technologie Houari Boumediene et Monsieur **Hicham Bousbia Salah**, Maître de conférence à l'Ecole Nationale Polytechnique, qui m'ont fait l'honneur d'accepter de juger ce travail en tant qu'examinateurs de cette thèse.

Je souhaite aussi remercier l'équipe pédagogique et administrative de l'ENP pour leurs efforts dans le but de nous offrir une excellente formation. Pour finir, je souhaite remercier toute personne ayant contribué de près ou de loin à la réalisation de ce travail.

Dédicaces

Je dédie ce travail aux gens qui m'ont soutenu à préparer ce travail :

Mes parents,

Mes frères,

Mes sœurs,

Mes grands-parents,

Et mes amis.

Houari

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

ADED	Algerian Dialect Emotional Database
ACA	Agents Conversationnels Animés
AMCASC	Algerian Modern Colloquial Arabic Speech Corpus
API	Alphabet Phonétique International
AR	Auto Régressif
ANN	Artificial Neural Network
ASDF	Average Square Difference Function
BD	Base de Données
BHUES	Beihang University Mandarin Emotion Speech
BVP	Blood Volume Pulse
CLDC	Chinese Linguistic Data Consortium
CNN	Convolutional Neural Network
CREAMA-D	Crowd-sourced Emotional Multimodal Actors Dataset
DNN	Deep Neural Network
DCT	Discrete Cosine Transformation
DFT	Discrete Fourier Transform
DT	Decision Tree
ECG	ElectroCardioGram
EEG	ElectroEncephaloGraphy
EMASPEL	Emotional Multi-Agents System for Peer-to-peer E-learning
EMG	ElectroMyoGram
EMO-DB	Berlin Database of Emotional Speech
EYASE	Egyptian Arabic speech emotion
ESMBS	Emotional Speech of Mandarin and Burmese Speakers
F1	Formant1
F2	Formant2
F3	Formant3
F4	Formant4

Fc	Fréquence cardiaque
FFT	Fast Fourier Transform
FLDA	Fisher Linear Discriminant Analysis
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
HNR	Harmonic to Noise Ratio
IFFT	Inverse Fast Fourier Transform
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LDC	Linguistic Data Consortium
LEM	Line Edge Map
LFPC	Log Frequency Power Coefficients
LPC	Linear Predictive Coding coefficients
LPCC	Linear Prediction Cepstral Coefficients
MAA	Modèles d'Apparence Active
MEDC	Mel Energy spectrum Dynamic Coefficients
MFCC	Mel Frequency Cepstral Coefficients
MPEG	Moving Picture Experts Group
MLP	Multi Layer Perceptron
NAQ	Normalized Amplitude Quotient
NV	Non Voiseés
PAP	Périodique APériodique
PLP	Perceptual Linear Prediction
RDA	Regularized Discriminate Analysis
RED	Réponse Electrodermale
REP	Reconnaissance des émotions dans la Parole
RNN	Recurrent Neural Network
SAL	Sensitive Artificial Listener
SAVEE	Surrey Audio Visual Expressed Emotion
ShEMO	Sharif Emotional Speech Database
SKT	Skin Temperature
SMS	Short Message Sendig
STEVE	SOAR Training Expert for Virtual Environments

SVM	Support Vector Machines
SVNN	Support Vector Neural Network
TPZ	Taux de Passage par Zéro
TURES	Turkish Emotional Speech Database
UFR	Unvoiced Frames Rate
V	Voisées
VR	Volume Respiratoire

Notations

$H(z)$	Fonction de transfert d'un filtre linéaire auto régressif
a_i	Coefficients de prédiction du filtre
F_0	Fréquence fondamentale
Hz	Hertz
T	Période
s	Seconde
I	Intensité
dB	Décibel
$S(n)$	Signal de parole
w	Fenêtre d'analyse gaussienne
A_n	Amplitude du signal de parole
HNR	Rapport harmonique sur bruit
E_p	Énergie de la partie périodique
E_n	Énergie du bruit
$m(f)$	Expression de l'échelle des fréquences Mel
f	Fréquence
Mean	Moyenne
Max	Maximum
Min	Minimum
Range	Différence entre le Maximum et le Minimum
$w[k]$	Fenêtre de Hamming
$X[k]$	Formule qui calcule la transformée de Fourier Discrète
$d(x, y)$	Distance
$H(x)$	Hyperplan séparateur
W	Vecteur de *m* dimensions

Introduction générale

La parole est une moyenne de communication extrêmement riche et complexe. Elle véhicule non seulement l'information linguistique mais aussi des informations sur nos traits personnels tels que notre identité, âge, genre, état de santé, personnalité et surtout notre état émotionnel. En tant qu'humains, nous trouvons que la parole est la manière la plus naturelle de nous exprimer. Nous reconnaissons son importance lorsque nous devons utiliser d'autres moyens de communication, tels que les emails ou les SMS. Il n'est pas surprenant que les émotions soient devenues courantes dans les messages texte, car ces messages texte pourraient être mal compris, et nous aimerions transmettre l'émotion avec le texte comme nous les faisons dans la parole.

L'émotion est une réponse motivationnelle et adaptative d'un organisme à l'environnement social. Elle fait partie de la vie quotidienne chez l'homme et sa forme primitive se trouve aussi chez l'animal. La manifestation des émotions est un domaine particulier complexe de la communication humaine. Elle concerne des sciences pluridisciplinaires telles que la psychologie, la cognition, la sociologie et la physiologie. Cette pluridisciplinarité qui due a la haute variabilité du comportement humain, influence à la fois la production et la perception de ces émotions. Généralement, la communication vocale possède deux formes. La première forme est interpersonnelle, c'est l'interaction entre deux personnes servant à échanger l'information et le message. La deuxième forme est intrapersonnelle, c'est la relation avec l'aspect interne de la personne qui touche particulièrement aux émotions. Cette dernière est la partie non verbale de la communication vocale. La compréhension des émotions est essentielle dans les interactions sociales humaines.

L'émotion est devenue un sujet d'intérêt depuis 1872 dans le cadre des études de Darwin sur l'expression et la transmission de l'émotion entre des êtres humains et entre des animaux, et que les études sur la composante audio de l'émotion aient aussi été démarrées depuis les années 1970, la reconnaissance des émotions dans la parole ainsi que dans d'autres modalités n'attirent l'attention des chercheurs à une échelle importante que depuis une vingtaine années. L'interface entre l'Homme et la Machine deviendront plus significatifs si les machines peuvent reconnaître le contenu émotionnel. Dans cette zone, trois facettes différentes peuvent être envisagées : la reconnaissance vocale en présence d'une parole émotionnelle, la synthèse de discours émotionnel et la reconnaissance des émotions. Ce travail se concentre sur le 3 ème aspect, à savoir la Reconnaissance des Émotions dans la Parole (REP).

La reconnaissance des émotions peut être faite par plusieurs modalités : la parole, les expressions faciales, ou tout autre canal biologique [1-6]. Les recherches à propos de la reconnaissance des émotions dans les expressions faciales sont très riches [7-10]. Cependant, la reconnaissance de émotions à l'aide des expressions faciales est complexe en terme de calcul, puisqu'il nécessite des caméras de haute qualité pour capturer des images de visage, coût de la mise en ouvre est également élevé. Outre l'expression faciale, la parole s'est avéré une modalité plus prometteuse de reconnaissance des émotions. La REP existe depuis deux décennies [11] et a des applications dans l'interaction homme machine [12], ainsi que dans les robots [13], les services mobiles [14], les centres d'appels [15], les jeux informatiques [16] et l'évaluation psychologique [17].

La REP aide à identifier l'état émotionnel d'un être humain à partir de sa voix. Donc la REP a été formulée comme un problème de reconnaissance qui implique l'extraction de caractéristiques et la classification des émotions. Un système de REP est défini comme un ensemble de méthodologies qui traitent et classifient les signaux vocaux pour détecter les émotions qui y sont intégrées. En général, le processus de REP peut être fait en extrayant des caractéristiques qui contiennent des informations émotionnelles provenant de la voix de locuteur et utilisant des méthodes de reconnaissance de formes appropriées pour identifier les états émotionnels. Plusieurs caractéristiques et modèles de classifications sont déjà proposés dans la littérature pour la tache de REP. En outre, il est également important de recommander un classificateur car les performances des systèmes de reconnaissance des émotions dépendent fortement des modèles de classification [18].

Il existe plusieurs travaux qui rassemblent des études existant dans la REP. On peut citer le travail de Ververdis et Kotopoulos [19], qui se sont concentrés sur des données vocales collectées, des caractéristiques acoustiques et des classificateurs utilisés dans la REP. Koolagudi et Rao se sont également appuyés sur la classification des bases de données, des caractéristiques et des classificateurs exploités dans REP [20]. Anagnostopoulos et Giannoukos ont fourni une étude des publications concernant des travaux de la REP entre l'année 2000 et 2011 [21]. Un travail récent de Basu et ses collègues met en évidence des publications qui impliquent des bases de données, des caractéristiques et des classificateurs pour la reconnaissance de l'émotion, y compris des avancées récentes telles que les réseaux de neurones convolutifs et récurrents [22].

Nous vivons une époque où la REP a connu une grande progression. Il a eu beaucoup de recherches dans le domaine de la REP dans différentes langues comme l'Anglais, l'Espagnol, le Slovène, le Français, l'Allemand, etc. Mais très peu d'ouvrages ont été rapportés dans le dialecte algérien. Dans cette thèse, notre objectif est d'étudier la REP dans le dialecte algérien. L'Algérie a subi une guerre civile qui a duré plus de dix ans. La guerre civile a provoqué la propagation de la peur, le stress, la tristesse et d'autres émotions. Ces émotions ont conduit à l'apparition de nombreuses maladies psychologiques qui subsistent encore à l'heure actuelle. Le but dans ce travail est basé sur la REP dans le dialecte algérien pour aider les psychologiques à exploiter la parole dans les domaines de la psychologie. Une base de données émotionnelle du dialecte algérien (ADED) est construite. Cette base de données est créée à partir des films algériens. Ces films décrivent la crise de la guerre civile ainsi que la période qui la suit. ADED comprend quatre différentes émotions : la peur, la colère, la tristesse et le neutre. Des caractéristiques prosodiques, spectraux et qualité vocales sont extraites à partir de la base de données ADED et deux méthodes de classification (K-plus-Proches-Voisins (KNN) et Machines à Vecteurs Supports (SVM)) sont appliquées.

Organisation de la thèse

Ce document se compose de quatre chapitres. Le premier introduit en premier lieu quelques notions concernant les émotions telles que leurs définitions, leurs différentes théories, leurs types et leurs descriptions. Puis, il présente des notions sur la parole. Et enfin nous présentons des corrélations entre l'aspect acoustique et les émotions.

Dans le deuxième chapitre, nous allons présenter une revue de littérature sur la REP, compte tenu des différents types de corpus utilisés pour développer les systèmes de REP, des paramètres spécifiques aux émotions extraites de différents aspects de la parole et des méthodes de classification utilisées pour reconnaître les émotions. Ainsi des principales applications de REP seront présentées dans ce chapitre.

Dans le troisième chapitre, la base de données émotionnelle du dialecte algérien (ADED) qui est utilisée dans ce travail sera décrite. Ainsi, les paramètres acoustiques utilisés pour modéliser les différents états émotionnels seront présentés. Une analyse pour détecter l'influence de différents états émotionnels (peur, colère, tristesse et neutre) sur les paramètres acoustiques choisis sera faite.

Dans le quatrième chapitre, plusieurs expériences sont effectuées pour étudier l'influence des différents facteurs tels que les paramètres acoustiques, le nombre d'émotions, le sexe et la langue sur les systèmes de REP. Les techniques de classification utilisées dans notre système de reconnaissance seront exposées. Ainsi, les résultats des expériences effectuées seront présentés dans ce chapitre.

Ce rapport s'achève sur des conclusions générales, des perspectives ainsi que des références bibliographiques.

Chapitre 1 :

Emotions et Parole

1.1 Introduction

La parole est l'une des principales modalités naturelles de l'Interaction Homme-Machine. L'interface entre l'homme et la machine deviendra plus significative si la machine peut reconnaître le contenu émotionnel. Par conséquent il est nécessaire, pour la conception d'une interaction fluide entre un homme et la machine, que cette interaction soit fondée sur la maîtrise des états émotionnels par la machine. Les humains manifestent diverses émotions tout au long de leur vie quotidienne, tel que la joie, la colère, le dégoût, la tristesse, la peur, etc., en réponse aux différentes situations qu'ils rencontrent. Nous débutons ce chapitre par présenter quelques notions concernant les émotions telles que leurs définitions, leurs différentes théories, leurs types et leurs descriptions. Ensuite, nous présentons des notions sur la parole. Enfin, nous exposons des corrélations entre l'aspect acoustique et les émotions.

1.2 Emotions

Le problème de la définition des émotions, de les distinguer des autres états affectifs et les mesurer de manière significative a été un défi constant pour les chercheurs dans différentes disciplines des sciences sociales et comportement sur une longue période.

1.2.1 Définition du terme émotion

Il n'existe pas réellement de définition d'une émotion ou d'un état émotionnel. Cependant, il n'y a pas de consensus sur la définition de l'émotion et c'est encore un problème ouvert en psychologie.

Plusieurs définitions ont été données à l'émotion. Ces définitions diffèrent en fonction des différentes approches proposées. En 1879, Charles Darwin définit l'émotion comme une qualité innée, universelle et communicative, liée au passé de l'évolution de notre espèce [23]. Il a présenté l'idée que les émotions sont inséparables des schémas d'actions sélectionnés par l'évolution en raison de leur valeur de survie [24]. Cowie utilise le terme états émotionnels auquel il donne un sens large en incluant les états reliés à des émotions telles que les humeurs. Il distingue ensuite les différents états émotionnels en fonction de leur description temporelle, de leur focus et du contrôle de la personne. En 1986, Frijda décrit l'émotion comme le changement dans un état de promptitude pour maintenir ou modifier des rapports avec l'environnement. L'émotion a été décrite comme interface de l'organisme vers le monde extérieur. Murray et Arnott ont défini l'émotion comme un changement brusque en réponse à des stimuli particuliers qui dure pendant une période courte [24].

Ekman, Izard, Plutchik, Tomkins et Mac Lean ont développé la théorie des émotions de bases ou fondamentales, mais seules 6 émotions de base sont communes aux divers auteurs. Ces émotions sont : la tristesse, la colère, la joie, le dégoût, la peur et la surprise, connues sous le nom "Big Six". De leur côté, Kleinginna A.M et Kleinginna P.R ont recensé plus de 140 définitions, reflétant chacune des différents aspects du processus émotionnel [23].

Pour tenter de clarifier les divergences entre les différents états affectifs, Scherer a proposé de les classifier selon différents critères (intensité, durée, etc.). Le tableau 1.1 ci-

dessous récapitule cette classification en considérant les modalités d'intensité, de durée ou de vitesse d'adaptation pour différents états affectifs [25].

Tableau 1.1 - Classification des différents états affectifs [25].

Phénomènes Affectifs	Intensité	Durée	Synchronisation	Causalité	Evaluation	Vitesse d'adaptation	Impact sur le comportement
Emotions	+ + + + +	+	+ + +	+ + +	+ + +	+ + +	+ + +
Humeurs	+ + +	+ +	+	+	+	+ +	+
Position Relationnelles	+ + +	+ + +	+	+ +	+	+ + +	+ +
Attitudes (Préférences)	+ +	+ + + + +	0	0	+	+	+
Disposition Affectives (traits de personnalité)	+	+ + +	0	0	0	0	+

Ainsi, pour Scherer, les différents états affectifs ont les caractéristiques suivantes [26] :

- **Émotion** : épisode relativement bref d'une réponse synchronisée de tous ou de la plupart des sous-systèmes de l'organisme en réponse à l'évaluation d'un événement interne ou externe évalué comme étant très important (ex. : *colère, tristesse, joie, honte, fierté, euphorie, désespoir).*
- **Humeur** : état affectif diffus, ressenti comme un changement subjectif d'état, de faible intensité, mais de durée relativement importante, souvent sans cause évidente (ex. : gaîté, mélancolie, irritabilité, indifférence, déprime).
- **Attitude** : position prise par rapport à une autre personne dans une interaction qui colore l'échange (ex. : distant, froid, chaleureux, supportant, insolant, arrogant).
- **Sentiment** : croyances, préférences et prédispositions affectives relativement durables par rapport à des objets ou des personnes (ex. : aimer, détester, désirer, respecter).
- **Trait de personnalité** : dispositions et tendances comportementales affectives stables et typiques chez une personne (ex.: nerveux, impatient, insouciant, mélancolique, hostile, envieux, jaloux).

Comparativement aux autres types d'états affectifs, les émotions sont donc intenses, de courte durée, à haut degré de synchronisation, fortement liées à la situation, produit de l'évaluation cognitive et elles ont un impact fort sur le comportement qui se modifiera rapidement. De plus, les émotions s'accompagnent de modifications physiologiques qui ont un impact sur la voix, ce qui n'est pas le cas des attitudes par exemple.

1.2.2 Théories des émotions

Plusieurs théories ont été proposées dans la littérature :

1.2.2.1 Théorie de James-Lange

Parmi les premières théories de l'émotion celle qu'a été proposée en 1884 par le psychologue et le philosophe américain, William James, très proche de celle du psychologue danois, Cari Lange. C'est pour cette raison, on parle souvent de la théorie périphérique de James-Lange. La théorie de James-Lange, énonce que l'émotion traduit la réponse aux modifications physiologiques intervenant dans le corps. Pour James William « L'émotion est notre perception des modifications qui surviennent [dans notre Corps] ». Par exemple, on est triste parce que l'on pleure, plutôt que l'on pleure parce que l'on est triste.

Un exemple de la théorie de James- Lang est monté dans la figure 1.1 : lorsqu'une personne se retrouve en face d'un danger (un animal effrayant par exemple), l'amène à courir pour fuir, et le fait de courir entraîne la peur. C'est-à-dire, un stimulus conduit à des réactions viscérales (battements de cœur, forte respiration) et comportementales, qui à leur tour, sont interprétées comme une émotion [23].

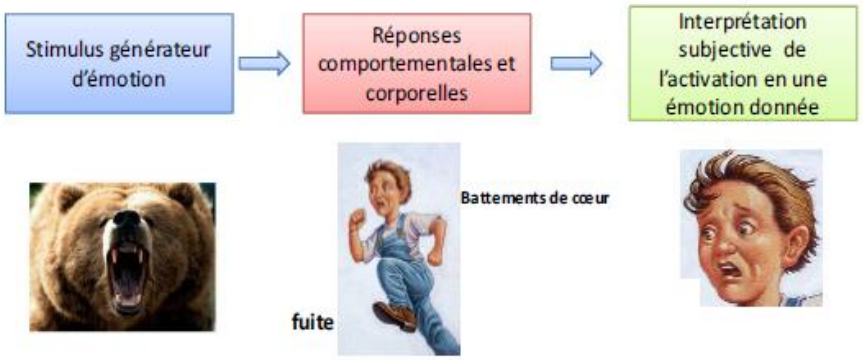


Figure 1.1 - Théorie des émotions de James-Lange [23]

1.2.2.2 Théorie de Cannon-Bard

La théorie de Cannon a été développée après par Philip Bard en 1934. Cannon et Bard ont noté que les réponses viscérales sont habituellement lentes et elles surviennent environ une à deux secondes après l'apparition du stimulus. A l'inverse, les réponses émotionnelles sont immédiates et précèdent souvent aussi bien les réactions viscérales que les comportements. Cannon et Bard ont donc proposé une théorie, selon laquelle les stimuli gérèrent des émotions produisant simultanément une expérience émotionnelle et des réponses corporelles (figure 1.2) [23].

1.2.2.3 Théorie de l'évaluation cognitive

Les théories de James-Lange et de Cannon-Bard mettent en évidence le fait que le phénomène émotionnel est accompagné de manifestations physiologiques, mais les aspects sociaux ne sont pas abordés. L'environnement est ici envisagé uniquement comme stimulus.

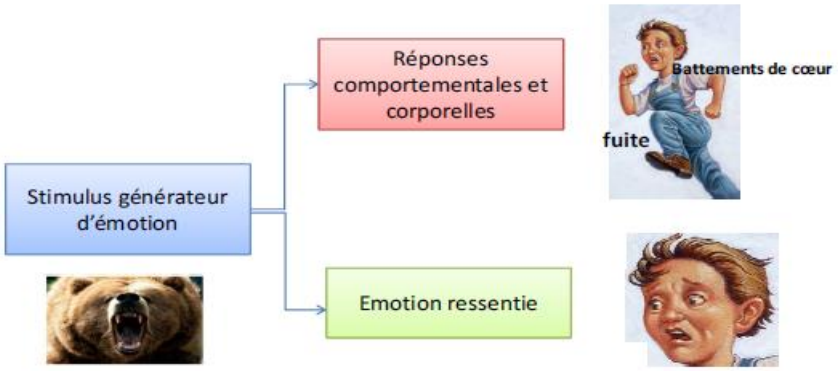


Figure 1.2 - Théorie des émotions de Cannon-Bard [23]

Or la particularité de l'environnement est qu'il peut affecter l'ampleur et l'intensité des manifestations physiologiques ressenties. Il est donc nécessaire de prendre en considération le facteur environnemental déterminé par l'apprentissage social pour l'évaluation de l'événement stimulus. C'est de fait, ce que les théories cognitives ont développé dans leurs modèles. Les modèles d'évaluation cognitive ont été introduits pour la première fois par Arnold en 1960. Ces modèles supposent qu'il est nécessaire de comprendre les évaluations que l'individu fait au sujet des événements de son environnement pour comprendre les émotions. Donc, une évaluation cognitive est définie comme un processus cognitif, rapide, automatique et inconscient dont la fonction est d'évaluer les stimuli perçus sur la base de critères particuliers. Selon cette théorie, une émotion est déclenchée par l'évaluation subjective d'un événement. La perception et l'évaluation cognitive d'un événement déterminent le type et l'intensité de l'émotion ressentie. Plus précisément, le déclenchement d'une émotion est issu de l'évaluation d'un ensemble de variables (appelés variables d'évaluation). Les valeurs des variables d'évaluation dépendent autant des facteurs culturels que de l'état mental de l'individu (buts, croyances, etc.) et de son profil (personnalité, préférences, etc.). Ainsi, une même situation peut déclencher deux émotions distinctes chez deux individus différents. La plupart des modèles sont fondés sur les théories de l'évaluation cognitive de Ortony, Clore et Collins "le modèle OCC" dans lequel 22 émotions types sont définies suivant leurs conditions de déclenchement [23].

1.2.3 Types d'émotion

Il existe trois types d'émotions : primaires, secondaires et sociales [27]:

1.2.3.1 Émotions primaires

Les émotions primaires sont déclenchées par des événements particuliers (tableau 1.2). Elles sont à la base de nos réactions qui ne sont pas seulement déterminées par notre jugement rationnel ou notre passe individuel mais aussi par notre passe ancestral. Les émotions primaires sont considérées comme une matière première, à partir de laquelle on peut fabriquer toutes les autres émotions.

1.2.3.2 Émotions secondaires

Les émotions secondaires ont pour base, un processus de pensée et sont l'aboutissement de l'apprentissage des émotions primaires. Les émotions secondaires sont celles qui sont engendrées à l'évocation de souvenirs et arrivent à maturation à l'âge adulte.

1.2.3.3 Émotions sociales

Les émotions sociales sont inhérentes à la relation par rapport aux autres, comme la culpabilité, la honte, la jalousie, la timidité, l'humiliation, etc. Toutes ces émotions sont apprises et sont constituées à partir des émotions primaires.

Tableau 1.2 - Emotions de base [27]

Emotion	Déclencheurs et circonstances d'apparition	Comportement
Joie	Désir Réussite Bien-être Accomplissement	Approche
Tristesse	Perte Deuil	Repli sur soi
Colère	Obstacle Injustice Dommage Atteindre a son intégrité physique ou psychique Limites de la personne Atteinte au système de valeurs	Attaque
Peur	Menace Danger Inconnu	Fuite Sidération Evitement Parfois attaque
Dégout	Substance ou personne nuisible Aversion physique ou psychique Contre quelqu'un Rejet	
Surprise	Danger immédiat Inconnu Imprévu	Retrait Sursaut

1.2.4 Description des émotions

Plusieurs approches ont été présentées pour décrire l'ensemble des émotions :

1.2.4.1 Approches discrètes

Les approches discrètes reposent sur l'existence d'un petit nombre d'émotions primaires discrètes, les émotions primaires sont supposées être discriminantes entre elles. Donc avec ces

approches, les autres émotions sont considérées comme des mélanges des émotions primaires. Les émotions les plus communément considérées comme émotions primaires sont : la joie, la tristesse, la peur, le dégoût, la colère et la surprise. Cependant, le nombre d'émotions primaires varie de deux à dix-huit selon des théoriciens de l'émotion. Le tableau 1.3 présente différentes listes d'émotions primaires proposées par les théoriciens, cette synthèse est proposée par Ortony et Turner [28]. On peut constater que certaines listes d'émotions primaires contiennent des éléments qui n'existent pas dans d'autres listes. Par exemple l'émotion de surprise : elle est incluse dans les listes d'émotions primaires (Ekman, Izard , Plutchik et Tomkins). Le désir est également un cas discutable bien que Descartes et quelques autres théoriciens (par exemple, Arnold et Frijda) aient inclus le désir dans leurs listes d'émotions primaires. Un autre exemple incertain dans les listes d'émotions primaires est l'intérêt. Quelques théoriciens (Exemple Frijda, Izard et Tomkins) le considèrent comme une émotion primaire contrairement à d'autres (Oatleyet Johnson-Laid, Ortony et Turner). La raison de ceci est que l'intérêt relève plus d'une attitude que d'un état émotionnel [24].

Comme décrit ci-dessus, les approches discrètes essaient de conceptualiser les émotions à partir de quelques émotions primaires et de traiter chacune d'elles comme une émotion discrète. Cependant, il n'y a pas de consensus sur la définition des émotions primaires et ces approches ne fournissent pas de description claire pour des émotions non basiques. Néanmoins, la notion d'émotion primaire forme une base pour la conceptualisation de l'émotion [24].

La plupart des systèmes de REP se concentrent sur ces catégories émotionnelles de base. Dans la vie quotidienne, les gens utilisent ce modèle pour définir leurs émotions observées. Néanmoins, ces catégories discrètes ne sont pas en mesure de définir certains des états émotionnels complexes observés dans la communication quotidienne [29].

1.2.4.2 Approches dimensionnelles

Les approches dimensionnelles considèrent les émotions comme un phénomène continu ou graduel. Les théoriciens essaient d'identifier les émotions en les plaçant dans une espace à plusieurs dimensions. Les approches dimensionnelles les plus rencontrées sont bidimensionnelles, tridimensionnelles et multidimensionnelles avec un nombre de dimensions supérieures à trois [30]. La plupart des théoriciens de l'approche dimensionnelle incluent une dimension « valence » (agréable/désagréable ou positif/négatif), une dimension d'activité (haute/basse ou actif/passif « arousal » en Anglais) et peu avec une autre dimension d'intensité (forte/faible) [24].

Tableau 1.3 - Listes d'émotions primaires avec les principes justificateurs, synthétisées par Ortony et Turner [28].

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

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

émotionnel ». La figure 1.3 présente le diagramme de ce modèle bidimensionnel. Schlosberg a proposé un modèle tridimensionnel en ajoutant une dimension de l'activité (haute/basse) [24].

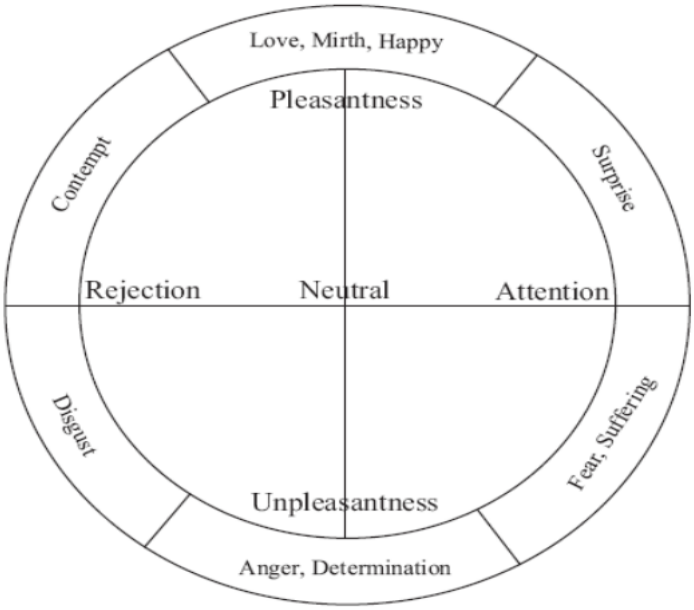


Figure 1.3 - Modèle bidimensionnel de Schlosberg [24]

1.2.4.2.2 Modèle circumplex de Russel : le modèle circumplex de Russell est un modèle bidimensionnel polaire. Il est illustré par la figure 1.4 avec 28 émotions déterminées expérimentalement. Russel a proposé ce modèle en analysant les résultats de la catégorisation des sujets et en les ordonnant sur un schéma polaire selon la graduation multidimensionnelle des états émotifs rapportés par ses expérimentateurs. L'axe horizontal du modèle est interprété comme la valence (agréable/désagréable) et l'axe vertical comme l'activité (haute/basse). Les étiquettes sont tracées par Russell sur ce circumplex selon le degré d'agrément et d'activité. Il explique que le nombre de catégories est extensible et que n'importe quelle catégorie concernant l'émotion pourrait être ajoutée à ce modèle. De même que Schlosberg, Russell a considéré que le centre du cercle est un point neutre ou un niveau d'adaptation. La distance entre le point neutre et la position d'une émotion particulière représente l'intensité de cette émotion. Tandis que l'activité est mesurée comme la déviation de l'état physiologique normal d'une personne, l'intensité est considérée comme le degré auquel l'expérience émotionnelle produit un changement de l'état neutre. Il a aussi divisé l'espace circulaire en espaces plus étroits pour des observations plus fines. Une division des hémisphères nous donne des dimensions de satisfaction. Une division en quatre parties nous donne quatre quarts de cercle [24] :

1) agréable/haute activité (exultation).
2) désagréable/haute activité (détresse).
3) désagréable/base activité (dépression).
4) agréable/base activité (calme).

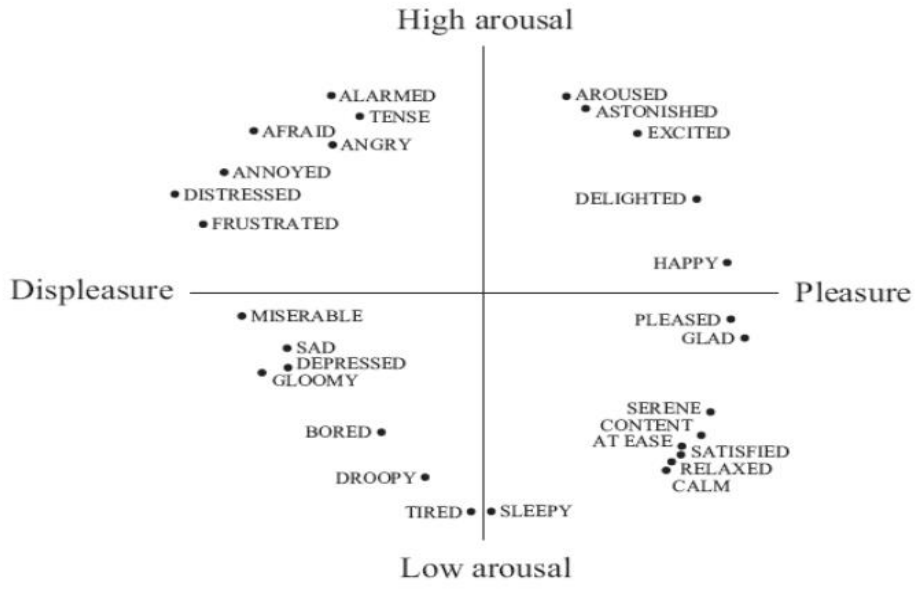


Figure 1.4 - Modèle circumplex de Russel [24]

Comme nous l'avons vu, le modèle de circumplex de Russell est simple et permet facilement de comprendre comment les émotions sont liées entre elles. Cependant, ce modèle a besoin d'être amélioré pour une analyse plus fine. Certains travaux ajoutent toutefois une troisième dimension, trouvant les deux premières insuffisantes. Cette troisième dimension est nommée contrôle ou dominance [24]. Elle correspond à l'effort du locuteur pour contrôler son émotion et permet de distinguer les émotions provoquées par le sujet lui-même ou par l'environnement. Elle permet de distinguer les émotions donnant lieu à des réactions d'approche et de combat comme la colère, de celles engendrant des comportements de fuite comme la peur [23].

1.2.4.2.3 Modèle du cône multidimensionnel de Plutchik : Plutchik [30] a développé sa théorie psychoévolutionnaire sur l'approche discrète de l'émotion. Il a choisi la colère, l'attente, la joie, l'acceptation, la peur, la surprise, la tristesse et le dégoût comme émotions primaires pour son modèle. Il a présumé un arrangement circulaire des émotions primaires comme présente la figure 1.5 et ces émotions sont arrangées avec les autres émotions « relatives » dans un modèle de cône tridimensionnel avec l'intensité, la polarité, et les dimensions de similitude. La figure 1.6 a présenté son diagramme de ce modèle sous une forme géométrique proche d'un épi de maïs. La figure 1.6 montre le modèle multidimensionnel de Plutchik où les 8 émotions primaires sont présentées sur une section dans le plan horizontal. Sur une section verticale, sont reportées les différentes intensités d'une même émotion primaire (par exemple : appréhension, peur et terreur). Dans les deux figures : figure 1.6 et la figure 1.7, la polarité est montrée par des émotions opposées autour du point neutre, par exemple : la joie face à la tristesse [24].

Plutchik a défini des règles d'association des émotions fondamentales pour former des émotions mixtes. En effet, les émotions ne s'associent pas n'importe comment et leur

combinatoire répond à des règles fondées sur la méthode des dyades et des triades. Plutchik a défini ainsi les dyades :

- Les dyades primaires comme étant la combinaison de deux émotions adjacentes.
- Les dyades secondaires comme étant la combinaison d'émotions proches à une émotion près.
- Les dyades tertiaires comme étant la combinaison d'émotions voisines à deux émotions près.

La figure 1.7 présente le cône des émotions de Plutchik tiré de Plutchik [30]. Les émotions élémentaires sont placées dans une roue. Les émotions secondaires correspondent à des mélanges d'émotions primaires. La roue peut être transformée en cône afin de représenter les différents degrés d'intensité des émotions primaires et secondaires. Ainsi pour prendre quelques exemples, les deux émotions adjacentes joie et admiration, forment un composé primaire : l'amour. Un composé secondaire : le désespoir, sera formé de la peur et de la tristesse, si l'on saute un élément. Enfin, un composé tertiaire sera formé de l'association de la peur au dégoût (si l'on saute deux éléments adjacents), composé qui aboutira à la honte [24].

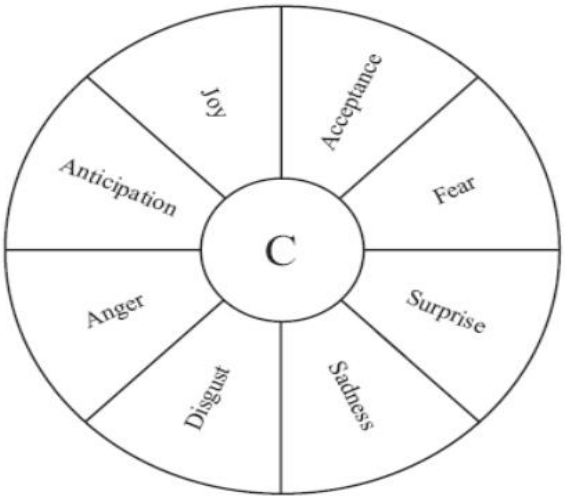


Figure 1.5 - Émotions primaires de Plutchik [30]

Il existe plusieurs inconvénients pour la représentation dimensionnelle. Ce n'est pas assez intuitif et une formation spéciale peut être nécessaire pour étiqueter chaque émotion. De plus, certaines émotions deviennent identiques, comme la peur et la colère, et certaines émotions comme la surprise ne peuvent pas être catégorisées et se situer en dehors de l'espace dimensionnel, car l'émotion surprise peut avoir une valence positive ou négative selon le contexte [24].

1.2.4.3 Approche hybride

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

de l'émotion. La couche la plus abstraite comporte seulement les deux catégories : « valence positive » et « valence négative ». Les catégories de la couche du milieu sont les catégories d'émotions primaires : la joie, l'amour, la colère, la tristesse et la peur. Ce sont des équivalents aux émotions primaires définies dans l'approche de l'émotion discrète. La plus basse couche se compose d'émotions non basiques et concrètes (par exemple : l'adoration, la tendresse pour l'amour, l'enthousiasme, le zèle pour la joie, l'agitation, la gêne pour la colère, etc.). On a conduit plus loin une analyse multidimensionnelle et tracé un diagramme avec deux dimensions orthogonales qui ressemble au modèle de circumplex de Russell. L'axe vertical du diagramme peut être considéré comme la dimension de « valence » (positive – négative) et l'axe horizontal peut être considéré comme la dimension d'activité [24].

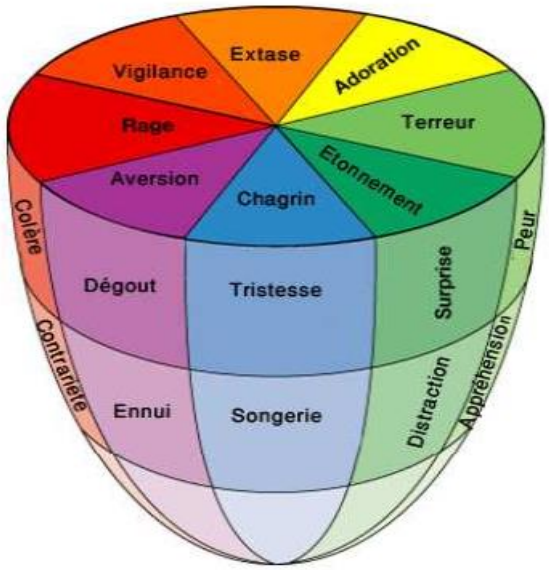


Figure 1.6 - Modèle du cône multidimensionnel [30]

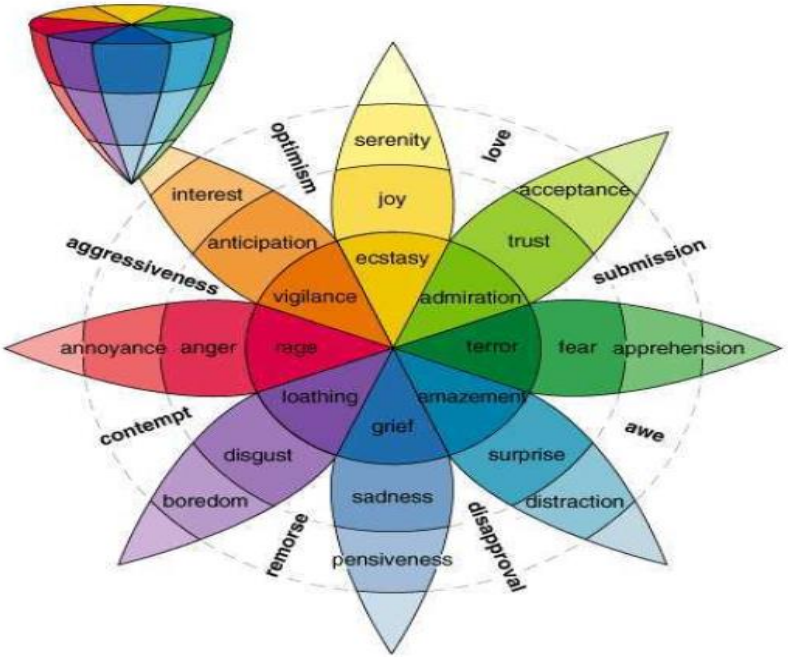


Figure 1.7 - Cône des émotions de Plutchik [30]

1.3 Notions sur la parole

La parole est le principal moyen de communication dans toute société humaine. Son apparition peut être considérée comme concomitante à l'apparition des outils, l'homme ayant alors besoin de raisonner et de communiquer pour les façonner [31]. La parole est très importante dans l'Interaction Homme-Machine. Elle permet de se dégager de toute obligation de contact physique avec la machine, libérant ainsi l'utilisateur qui peut alors effectuer d'autres tâches.

La parole est une faculté, propre à l'homme, de communication par des sons articulés. Elle met en jeu des phénomènes de natures très différentes et peut être analysés de bien des façons. On distingue généralement plusieurs niveaux de description non exclusifs : physiologique, phonologique, phonétique, acoustique, morphologique, syntaxique, sémantique et pragmatique [32].

1.3.1 Niveau physiologique

La production de la parole est assurée par plusieurs organes successifs. Les poumons sont indispensables dans ce processus puisqu'ils assurent la génération de l'air sous pression. Cet air traverse alors les cordes vocales qui entrent ou non en action pour produire un voisement. Ce voisement correspond à la fréquence fondamentale qui est le timbre de la voix [33].

Cette fréquence fondamentale étant produite, elle est propagée dans l'ensemble du conduit vocal. Ce conduit est de forme et de volume variable. Plusieurs organes concourent à ces possibles modifications qui permettent de produire des sons différents. Parmi ces organes se trouve la langue, qui peut agir par constriction ou occlusion du conduit vocal. Les dents et les lèvres agissent également par occlusion ou constriction, à des degrés cependant moindres. Le conduit vocal est, la plupart du temps, constitué du seul conduit buccal. La luette et son prolongement vers le palais, le vélum, assurent normalement la fermeture du conduit nasal pendant la production de parole. Le conduit nasal peut être connecté au conduit vocal dans certains cas. Cette connexion permet de générer des sons supplémentaires en modifiant le volume de la caisse de résonance normalement constituée par le seul conduit buccal [33]. Une coupe de l'appareil phonatoire humain est illustré en figure 1.8.

Les différents organes de la parole et leur agencement peuvent servir de base à des modélisations du conduit vocal.

Les différents organes de la parole et leur agencement peuvent servir de base à des modélisations du conduit vocal.

1.3.2 Niveaux phonétique et phonologique

La phonétique et la phonologie sont deux branches de la linguistique qui interprètent la parole. La phonétique étudie les sons des langues du monde en tant que réalité physique (production, transmission et perception de ces sons), tandis que la phonologie recherche les principes qui régissent leur apparition et leur fonction de codage d'une langue particulière. Autrement dit, la phonétique est l'étude scientifique des sons du langage humain.

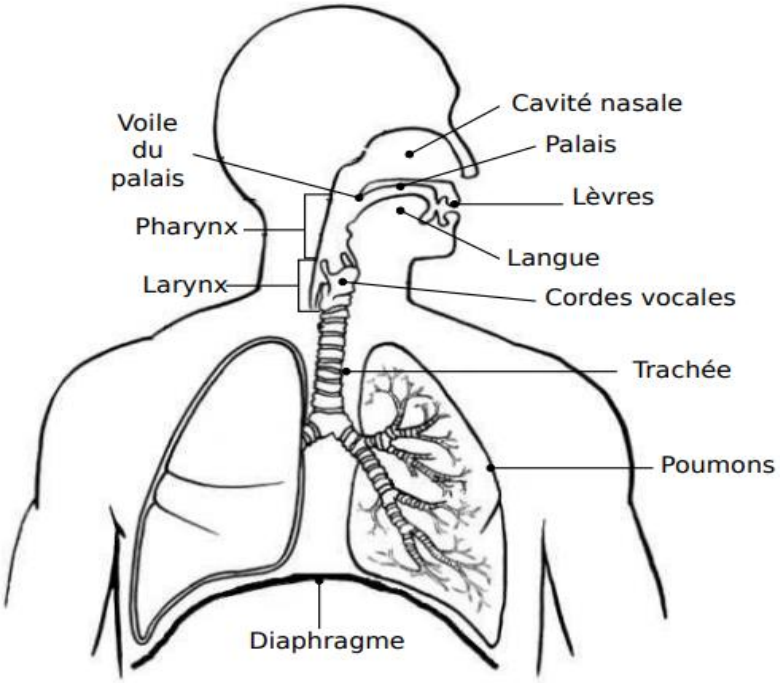


Figure 1.8 - Coupe de l'appareil phonatoire humain [34]

Elle exclut les autres sons produits par les êtres humains, même s'ils servent parfois à communiquer (les toux, les raclements de gorge). Elle exclut aussi les sons non humains. Elle se divise en trois domaines [32]:

• la phonétique articulatoire s'occupe de l'activité des cordes vocales, de la bouche, etc. qui rendent possible la parole.

• la phonétique acoustique examine les caractéristiques sonores des sons du langage.

• la phonétique auditive examine les phénomènes de perception des sons du langage par les êtres humains.

Chaque langue retient pour son fonctionnement un ensemble de sons, parmi ceux que pourrait produire l'appareil vocal. Les plus petites unités sonores distinctives utilisées dans une langue donnée sont appelées phonèmes.

L'ensemble des phonèmes généralement adopté pour une langue donnée sont regroupés par un système de transcription phonétique utilisé par les linguistes, représenté par l'Alphabet Phonétique International (API). Les phonéticiens regroupent les sons de parole en deux grandes classes phonétiques en fonction de leur mode articulatoire : les voyelles et les consonnes.

Les voyelles : cette classe correspond, à quelques nuances supplémentaires près, aux voyelles de l'écrit. Elles se caractérisent principalement par le voisement qui crée des formants. Ces formants, qui sont des zones fréquentielles de forte énergie, correspondent à une résonance dans le conduit vocal de la fréquence fondamentale produite par les cordes

vocales. Ces formants peuvent s'élever jusqu'à des fréquences de 5 kHz mais ce sont principalement les formants en basses fréquences qui caractérisent les voyelles. Cette caractéristique permet d'ailleurs de distinguer grossièrement les voyelles en fonction de leur premier et deuxième formant [33].

Les consonnes : contrairement aux voyelles, les consonnes sont produites lorsque le passage de l'air venant des poumons est partiellement ou totalement obstrué. Autrement dit, les consonnes correspondent à des mouvements rapides de constriction des organes articulateurs, donc souvent à des sons peu stables, qui évoluent dans le temps [32].

Les fricatives : dans cette classe sont regroupées les sons produits par la friction de l'air dans le conduit vocal lorsque celui-ci est rétréci au niveau des lèvres, des dents ou de la langue. Cette friction produit un bruit de hautes fréquences et peut être voisée ou sourde [33].

Les occlusives : les phonèmes de cette classe se caractérisent oralement par la fermeture du conduit vocal, fermeture précédant un brusque relâchement. Les occlusives sont donc constituées de deux parties successives : une première partie de silence, correspondant à l'occlusion effective, et une deuxième partie d'explosion, au moment du relâchement. Les occlusives peuvent être voisées, à la manière des voyelles, ou sourdes, c'est à dire non voisées. Les occlusives voisées peuvent également être appelées occlusives sonores [33].

Les nasales : les phonèmes sont formés par passage de l'air dans le conduit vocal depuis les cordes vocales. Ce passage exclut normalement toute connexion du conduit normal, le conduit buccal, avec le conduit nasal. Ce dernier peut cependant être employé, dans un nombre limité de cas puisque sa physionomie ne permet pas de créer des sons autrement qu'en modifiant le volume de la caisse de résonance qu'il constitue par l'intermédiaire de la langue, faisant occlusion dans le conduit buccal. Les nasales sont donc produites de la même manière que les occlusives nasales mais l'air n'est pas, cette fois, comprimé dans le conduit vocal. Le vélum est en effet abaissé pour permettre à l'air d'être expiré. Les nasales sont voisées. Il est à noter que certaines voyelles possèdent également un caractère de nasalité [33].

Les semi-voyelles : sont des consonnes voisées, mouvements rapides qui passent par la position articulatoire d'une voyelle brève. Enfin, les liquides résultent d'une excitation voisée et de rapides mouvements articulatoires principalement de la langue.

1.3.3 Niveau acoustique

La phonétique acoustique étudie le signal de parole en le transformant dans un premier temps en signal électrique grâce au transducteur approprié : le microphone. Le signal électrique résultant est souvent numérisé. Il peut alors être soumis à un ensemble de traitements statistiques qui visent à en mettre en évidence les traits acoustiques qui sont liés à sa production [32] :

• **La fréquence fondamentale (F_0)** : la hauteur de la voix, au cours d'une conversation variée selon les personnes, elle est essentiellement dépendue de la dimension et de la tension des cordes vocales, ainsi que des dimensions des résonateurs. Elle peut être volontairement

modifiée dans certaines limites, par l'intermédiaire des muscles respiratoires, en faisant varier la pression de l'air. L'association de ces éléments détermine la fréquence de vibration des cordes vocales, appelée fréquence fondamentale ou pitch, elle est variée selon l'âge et le sexe. Alors que la fréquence fondamentale de la voix parlée est [35] :

- De 60 à 250 Hz pour une voix masculine.
- De 150 à 500 Hz pour une voix féminine.
- De 200 à 600 Hz pour une voix d'enfant.

• **L'énergie ou l'intensité (I)** : l'intensité correspond à l'amplitude des vibrations sonores, dépend principalement de la pression produite par le souffle thoracique et de la résistance que peuvent lui opposer les cordes vocales (tension, affrontement); mais aussi des résonateurs. Elle se mesure généralement à l'aide d'un décibelmètre ou sonomètre qui doit être placé à environ 30 cm de la bouche. L'intensité moyenne de la voix parlée est de 60 dB [35].

• **Le spectre** : l'enveloppe spectrale ou spectre représente l'intensité de la voix selon la fréquence, elle est généralement obtenue par une analyse de Fourier à court terme. La quasi stationnarité du signal de parole permet de mettre en œuvre des méthodes efficaces d'analyse et de modélisation utilisées pour le traitement à court terme du signal vocal sur des fenêtres de durée généralement comprise entre *20 ms* et *30 ms* appelées trames, avec un recouvrement entre ces fenêtres qui assure la continuité temporelle des caractéristiques de l'analyse [36].

Chaque trait acoustique est lui-même intimement lié à une grandeur perceptuelle : pitch, intensité, et timbre [33].

Les modèles les plus classiques de représentation du signal de parole s'inspirent du mode de production de type source – filtre (figure 1.9). Le modèle est divisé en trois parties, la source (le voisement, la friction), le filtre (simulation des effets filtrants des conduits oraux et nasaux), et la radiation aux lèvres [37].

Le signal de source résulte de la production d'une onde acoustique au niveau de la glotte. Cette onde passe ensuite dans le conduit vocal (oral, nasal) et subit l'effet de radiation des lèvres. Les transformations du signal de source par ces différents organes peuvent être modélisées par un simple filtrage linéaire (figure 1.10).

Les caractéristiques acoustiques des cavités supra -glottiques peuvent être modélisées à l'aide d'un filtre linéaire AR (Autorégressif) dont la fonction de transfert s'exprime comme suit :

$$H(z) = \frac{1}{1+\sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \qquad (1.1)$$

Où les a_i sont les coefficients de prédiction du filtre.

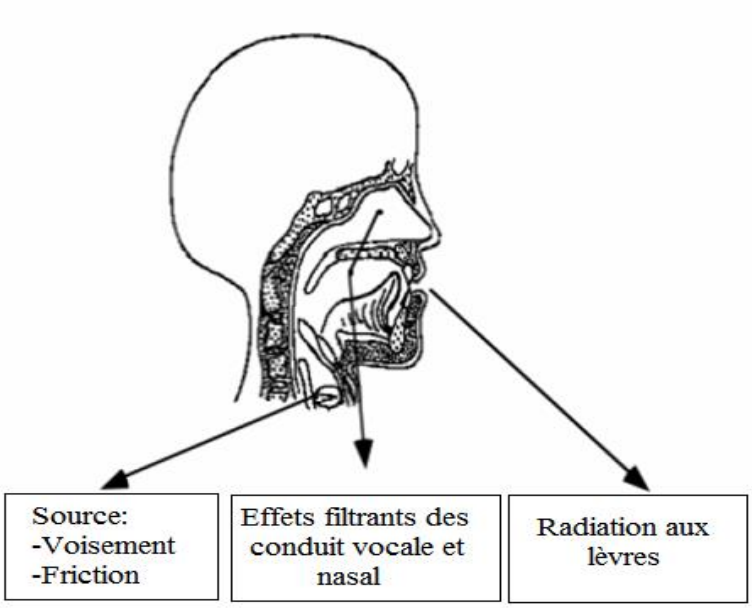


Figure 1.9 - Conceptualisation fondamentale du modèle source - filtre [36]

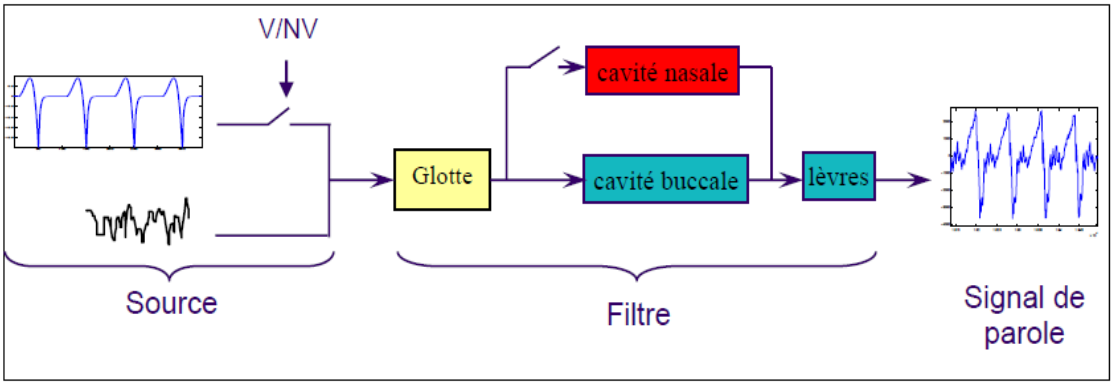


Figure 1.10 - Modèle source – filtre [35]
(V = Voisement et NV = Bruit (aspiration, frication, explosion))

Pour une parole intelligible, le nombre de coefficients a_i est fixé de telle façon que la fonction de transfert du filtre présente un nombre suffisant de résonances pour modéliser correctement les 3 à 5 premiers formants des segments voisés.

1.4 Corrélation entre l'aspect acoustique et les émotions

La reconnaissance des émotions contenues dans la parole est possible si et seulement s'il existe des corrélations fiables entre les émotions et les caractéristiques acoustiques du signal de la parole [37]. Plusieurs travaux ont étudié la corrélation entre les émotions et les caractéristiques acoustiques et leurs résultats s'accordent avec des corrélations venant des contraintes physiologiques pour des classes d'émotions primaires [38-39].

Dans les travaux de Picard [40], quelques états émotifs sont souvent corrélés avec des états physiologiques particuliers qui ont à leur tour des effets assez « mécaniques » et prévisibles sur la parole, particulièrement sur la fréquence fondamentale, le débit et la qualité

de la voix. Lorsqu'on est dans un état de colère, de peur ou de joie, la voix est donc forte, rapide et parlée avec une forte énergie de haute fréquence, la moyenne et la variation de la fréquence fondamentale sont également plus importantes. Au contraire, lorsque on est fatigué, ennuyé ou triste, la voix est donc lente, à basse intonation et avec peu d'énergie à haute fréquence [41].

Plusieurs études comme celles de [42] et [43] ont réalisé des expérimentations sur le signal acoustique de la parole. Dans ces expérimentations les auteurs ont demandé aux évaluateurs (américains et japonais) d'identifier les émotions de la joie, la tristesse, la colère, la peur ou le calme en se basant uniquement sur l'information acoustique. A partir de ces expérimentations, des résultats ont été conclus : il y a peu de différence de performance entre la détection des émotions exprimées dans la même langue ou dans une autre langue [43].

Plusieurs revues de la littérature font état de résultats convergents quant aux patrons acoustiques qui caractérisent l'expression vocale des principales émotions. Voici quelques principaux corrélats présentés dans la littérature [26] :

- ***Colère*** *:* les revues qui ont étudié l'émotion de colère relèvent une augmentation de la moyenne de la fréquence fondamentale (F_0) et de l'intensité moyenne. On peut citer les travaux de [38][46], qui suggèrent que les études aient porté sur la colère chaude sont celles ayant trouvé une augmentation du débit, de la variabilité et de l'étendue de $F0$, alors que celles n'ayant pas trouvé d'augmentation portaient sur la colère froide. Johnstone et Scherer [44] rapportent un contour de F_0 descendant alors que Juslin et Laukka [45] rapportent un contour de F_0 montant, un F_0 irrégulier, une intensité variable, un F_1 élevé et ayant une largeur de bande étroite.

- ***Dégoût*** *:* Murray et Amott [46] ont conclu une diminution du débit, de la moyenne de F_0 et de l'intensité, une augmentation de l'étendue de F_0 et un contour de F_0 descendant dans l'émotion de dégoût. Johnstone et Scherer [44] notent toute fois que les résultats des différentes études sont contradictoires. En fait, les études ayant procédé par induction (visionnement de films) auraient enregistré une diminution de la moyenne de F_0, contrairement à celles utilisant des acteurs.

- ***Joie*** *:* les revues qui ont étudié l'émotion de la joie relèvent une augmentation de la moyenne de F_0, de l'étendue de F_0, de l'intensité et du débit. Murray et Amott [46] rapportent une qualité de voix soufflée et stridente. Les auteurs ont noté aussi que toutes les fréquences formantiques sont augmentées pendant la joie étant donné l'étirement des lèvres (sourire). Juslin et Laukka [45] rapportent aussi un F_0 régulier, une F_1 élevé ayant une largeur de bande étroite et un contour de F_0 montant.

- ***Peur*** *:* dans la littérature, les revues qui ont étudié l'émotion de peur relèvent une augmentation de la moyenne de F_0, de l'étendue de F_0 et du débit. Juslin et Laukka [45], Murray et Arnott [46] rapportent une F_0 irrégulier. Juslin et Laukka rapportent également une intensité variable, un $F1$ bas ayant une largeur de bande importante et un contour de F_0 montant. Johnstone et Scherer [44] rapportent une augmentation de l'intensité alors que Murray et Amott [46] rapportent une intensité normale.

- ***Surprise*** *:* dans la littérature, peu d'études ont porté sur la surprise. Oster et Risberg [47] ont observé une augmentation de l'étendue de *F*0 et dans certains cas de la moyenne de *F*0. Fonagy et Magdics [48] ont trouvé une qualité de voix soufflée. Les résultats concernant le débit sont contradictoires : Fonagy et Magdics [48] ont trouvé une réduction du débit alors qu'Oster et Risberg [49] n'ont pas trouvé de modification du débit.

- ***Tristesse*** *:* dans la littérature, les revues consultées qui ont étudié l'émotion de tristesse relèvent une diminution de la moyenne de F_0, de l'étendue de F_0, de l'intensité et du débit. Juslin et Laukka [44] rapportent également une intensité peu variable, un F_0 irrégulier, un F_1 bas et ayant une largeur de bande importante. Murray et Arnott [46] rapportent un contour de F_0 descendant.

Selon Scherer [49], les différences acoustiques liées aux émotions sont principalement liées au degré d'excitation. Scherer note que les émotions à activation élevée sont caractérisées par une augmentation de la moyenne, de l'étendue de F_0 et de l'intensité moyenne alors que l'inverse est vrai pour les émotions à faible activation due à la réponse du système sympathique.

1.5 Conclusion

La communication est une faculté fondamentale, basée non seulement sur des énoncés linguistiques mais aussi sur la partie émotionnelle. Dans ce chapitre, nous avons présenté dans un premier temps quelques notions concernant les émotions telles que leurs définitions, leurs différentes théories, leurs types et leurs descriptions. Nous avons pu constater qu'il n'existe pas de consensus pour la définition de l'émotion. Puis, dans un second temps, nous avons présenté des notions sur la parole. Enfin, nous avons présenté la corrélation entre l'aspect acoustique et les émotions.

.

Chapitre 2 :

Reconnaissance des Emotions dans la Parole

2.1 Introduction

Les systèmes de Reconnaissance des Émotions dans la Parole (REP) varient entre eux dans les caractéristiques de la Base de Données (la langue, le nombre des émotions, comment les émotions ont été induites, etc.) ainsi dans les paramètres extraits du signal de la parole et dans les classificateurs implémentés pour la reconnaissance des émotions. Dans ce chapitre, nous présentons une revue de la littérature sur la REP, compte tenu des différents types de corpus utilisées pour développer les systèmes de REP, les paramètres spécifiques aux émotions extraites de différents aspects de la parole et les méthodes de classification utilisées pour reconnaître les émotions. D'abord, nous présentons brièvement dans la première section la détection des émotions à partir des différentes modalités : l'expression faciale, les textes, le signal acoustique et les signaux physiologiques. Ensuite, dans la deuxième section, nous présentons un résumé de l'état de l'art sur la REP et des principales applications de ce domaine.

2.2 Reconnaissance des émotions

L'émotion joue un rôle majeur dans les processus cognitifs humains. Pour concevoir une interaction fluide entre un être humain et la machine, il est nécessaire que cette interaction soit basée sur la maîtrise des états émotionnels par la machine. La reconnaissance des émotions est un domaine de recherche très important pour améliorer les interfaces homme machine. L'émotion contrôle et modifie presque tous les modes de communication humaine tels que l'expression faciale, les gestes, la posture, la tonalité de la voix, le choix des mots, la respiration, le rythme du cœur, la pression du sang, la température, l'humidité de la peau, etc. La détection des émotions à partir des différentes modalités est présentée dans les paragraphes qui suivent.

2.2.1 Détection des émotions dans les images et les vidéos

Les expressions faciales sont des éléments plus importants dans le processus de communication et elles présentent une source d'informations importante concernant l'état émotionnel des personnes. Les expressions faciales humaines peuvent être utilisées pour détecter les émotions. Elles sont plus fréquemment capturées à partir des images ou des séquences d'images. Il y a eu de nombreuses études sur la reconnaissance des émotions basées uniquement sur des enregistrements vidéo des expressions faciales. Le visage fournit des signaux conversationnels qui clarifient notre d'attention et règlent nos interactions avec l'environnement. Les informations utilisées sur le visage peuvent être le mouvement de la bouche, les lèvres et les yeux [23].

De nombreuses études ont été effectuées sur la reconnaissance des émotions à partir des expressions faciales. Un modèle automatique a été proposé dans le but de classifier des expressions faciales en se basant sur les Modèles d'Apparence Active (MAA). Chaque image de visage a été représentée par un vecteur MAA correspondant. En 2007, Hammal a proposé une méthode de classification basée sur les distances entre points d'intérêt du visage. Une méthode de classification de trois émotions : neutre, sourire et colère a été proposée par Gao.

Son approche utilise la ligne du contour (LEM : Line Edge Map) comme une expression descriptive. Deux méthodes de classification des expressions faciales : statique et dynamique ont été décrites par Cohen. Ces méthodes sont appliquées sur un corpus de six émotions actées : la joie, la surprise, la colère, le dégoût, la tristesse et la peur [23].

2.2.2 Détection des émotions à partir du texte

Les études sur la détection des émotions à partir du texte se sont concentrées sur l'utilisation de « mots-clés émotifs », c'est-à-dire sur l'identification de mots spécifiques qui indiquent l'état émotionnel. Yanaru a suivi des locuteurs tandis qu'ils parlaient dans un contexte naturel en utilisant des mots-clés émotifs. Un groupe de mots émotifs a été construit en marquant manuellement le degré d'émotion pour chacun d'entre eux. Devillers et d'autres ont recherché les états émotionnels en calculant la probabilité conditionnelle entre les mots-clés émotifs et les émotions. Tao et ses collègues supposent également que le contenu émotionnel dans le texte d'une phrase est porté par le type des mots. Ils ont donc classé les mots en deux groupes : les mots de contenu et les mots fonctionnels d'émotion. Lee a montré que la détection des états émotionnels à partir du texte peut améliorer en combinant les trois types d'informations : linguistique, pragmatique et mots-clés [24].

2.2.3 Détection des émotions dans le signal acoustique

Les émotions sont également communiquées à travers la voix. De nombreuses caractéristiques de la voix traduisent une émotion. Par exemple, l'émotion de la tristesse correspond à un débit de parole lent, une intensité faible et une tonalité basse. Plusieurs travaux se sont intéressés à la reconnaissance des émotions dans le signal acoustique. Les premières études sur la parole émotionnelle se sont basées sur les paramètres prosodiques. La prosodie qui rassemble divers phénomènes tels que le ton, l'accent, la mélodie, le rythme ou encore le débit. La prosodie décrite par trois paramètres acoustiques : la fréquence fondamentale, l'intensité et la durée [23]. Claval a développé un système automatique de reconnaissance d'émotions de type peur en situation anormale en se basant sur les manifestations vocales des émotions [50]. Un system de reconnaissance de sept émotions : ennui, dégoût, colère, panique, joie, tristesse et neutre a été construit. Les statistiques de la prosodie (la fréquence fondamentale, l'énergie, la durée) ont été utilisé dans le système de la reconnaissance [51].

2.2.4 Multimodalité

Pour améliorer la performance de reconnaissance des émotions, les chercheurs ont utilisé l'information multimodale comme la combinaison entre le signal audio avec l'information linguistique, la combinaison entre l'audio et la vidéo, ou la combinaison des trois aspects. Schüller et d'autres chercheurs ont combiné des informations acoustiques et des informations linguistiques pour extraire l'état émotionnel parmi différentes émotions : la colère, la joie, le dégoût, la tristesse, la surprise et le neutre. Trois aspects de la parole ont été combinés pour détecter les états émotionnels. Ces aspects sont : l'information acoustique, l'information lexicale et l'information du discours. Le corpus utilisé contient des conversations collectées à

partir d'un centre d'appels. Chen et d'autres ont utilisé la combinaison des deux modalités en exploitant les paramètres prosodiques du canal audio et les paramètres des mouvements des yeux, des joues, et de la bouche du canal vidéo. De Silva et ses collègues ont proposé une méthode de classification des émotions dans des données audiovisuelles. Les mouvements et la vitesse de certains signes faciaux : des lèvres, de la bouche, des sourcils sont utilisées dans le canal vidéo. Dans le canal audio, la fréquence fondamentale a été utilisée. D'après les résultats obtenus, la classification sur le canal vidéo est meilleure que celle sur le canal acoustique, l'approche bimodale donne les meilleurs résultats [24].

2.2.5 Détection des émotions par les signaux physiologiques

Différentes activations physiologiques sont associées aux états émotionnels. Elles concernent des modifications au niveau du système nerveux autonome par exemple : la modification du rythme cardiaque, la pression artérielle, la température corporelle, etc. Les indices physiologiques plus utilisés pour déterminer les états émotionnels sont : la Réponse ElectroDermale (RED), le volume sanguin impulsionnel (Blood Volume Pulse : BVP), le signal du Volume Respiratoire (VR), l'activité ElectroMyoGraphique (EMG), la température cutanée (Skin Temperature : SKT), la fréquence cardiaque (Fc) et le rythme ElectroEncéphaloGramme (EEG) [23].

Des mesures plus précises du rythme cardiaque ont exploité par ABDAT. Ces mesures ont permis de mettre en évidence des variations subtiles du patron de la rythmicité cardiaque entre la peur et la colère. Des chercheurs ont montré que la peur, la tristesse et la colère sont associées à une augmentation du rythme cardiaque. Le dégoût diminue le rythme cardiaque. La conductance de la peau augmente après un état d'amusement et diminue après l'état neutre [23].

Plusieurs travaux ont été réalisés dans la reconnaissance des émotions à partir les signaux physiologiques. Le modèle cartes psycho-physiologiques émotionnelles a été proposé par Lisetti et Villon [52]. Ce modèle se constitue une représentation paramétrique permettant l'interprétation émotionnelle à partir des signaux physiologiques. Les paramètres physiologiques utilisés sont la conductance cutanée et les battements de cœur. Ces paramètres ont été interprétés pour détecter l'état affectif de l'utilisateur durant une interaction homme-machine. Nasoz et d'autres chercheurs [53] ont développé un système qui reconnaît l'émotion à partir de ses signaux physiologiques. Ils ont utilisé des films pour induire l'émotion d'un participant avant de la mesurer. Un système pour la reconnaissance des trois états émotionnels à partir des signaux physiologiques a été proposé [54]. Les paramètres physiologiques utilisés dans le système de reconnaissance sont : la pression sanguine, la fréquence cardiaque, la respiration, la réponse galvanique de la peau et les signaux EEG. Haag et ses collègues [55] ont utilisé des photos pour provoquer les émotions des participants. Ils se sont intéressés à mesurer les manifestations physiologiques (la température, la respiration, l'EMG, la conductance cutanée, le BVP, l'activité électrique du cœur (ElectroCardioGram : ECG).

2.3 Reconnaissance des Emotions dans la Parole

Nous définissons un système de reconnaissance des émotions comme un ensemble de méthodologies qui traitent et classifient les signaux vocaux pour détecter les émotions qui y sont intégrées. Ces dernières années les systèmes de REP ont connu une grande progression. Les phases principales qui constituent les systèmes de reconnaissance se sont beaucoup développées.

Un système de REP se base sur quatre phases principales [23] :

- l'extraction de paramètres acoustiques : le signal de parole est analysé afin de le transformer en une séquence de vecteurs contenant les valeurs des différents paramètres retenus. Cette phase permet d'obtenir une représentation compacte des différentes caractéristiques acoustiques du signal de parole ;
- l'apprentissage : cette phase permet de rassembler les vecteurs acoustiques correspondant aux segments de parole d'une même classe dans un modèle caractéristique de cette classe. Ce modèle est obtenu à partir d'une base de données (dite d'apprentissage) ;
- la classification : cette phase s'agit de comparer les vecteurs acoustiques du signal vocal à analyser aux modèles de chaque classe ;
- la décision : cette phase permet d'associer une classe à un segment de parole.

Les sections suivantes présentent un résumé de l'état de l'art dans le domaine de REP et les principales applications de ce domaine. L'objectif de présenter un état de l'art dans la REP est de connaître les différents corpus, caractéristiques et les techniques de classification utilisés dans ce domaine.

2.3.1 Corpus de parole émotionnelle

Une Base de Données(BD) de parole émotionnelle appropriée est nécessaire pour un système de REP. Un point important à prendre en compte dans l'évaluation des systèmes vocaux est la qualité des bases de données utilisées pour développer et évaluer la performance des systèmes. La première étape pour construire un système de reconnaissance est de créer un corpus de données. Les corpus doivent être à la fois importants et naturels. Dans cette section on cite les différentes catégories de corpus émotionnel, ainsi un bref aperçu de certaines des bases de données de parole émotionnelle est donné.

2.3.1.1 Types de corpus de parole émotionnelle

Nous distinguons essentiellement différentes catégories de corpus émotionnels utilisés dans le domaine de REP : naturels, actés, élicités et extraits.

- **Corpus naturels :** contiennent des émotions associées à la parole spontanée ou quasispontanée obtenue lors d'interactions dans la vie quotidienne. Obtenir des émotions spontanées est une tâche complexe. Les bases de données de la parole naturelle sont principalement obtenues à partir des émissions de discussion, des enregistrements de centres d'appels, des conférences radiophoniques et des sources

similaires. Les corpus naturels collectés sont généralement basés sur une interaction humain-humain ou humain-machine. Les corpus vocaux en interaction humain humain les plus courants sont ceux collectés en centre d'appels par exemple : Corpus EmoVox [56] et corpus CEMO [57]. Cette méthode permet d'avoir un grand nombre de données et de plusieurs locuteurs. Parmi les corpus collectés en interaction homme-machine on peut citer le cas du corpus AIBO [58], ce corpus enregistré au cours d'une interaction entre AIBO (le robot de Sony) et un enfant.

Les inconvénients principaux des corpus naturels sont que les données sont très limitées en nombre de locuteurs, de courtes durées, et d'être très difficiles à collecter et à étiqueter en classes d'émotions [59].

- **Corpus actés :** la majorité des BD utilisées dans le domaine de REP font intervenir des émotions simulées (actées). Les émotions simulées sont des émotions exprimées par des acteurs professionnels ou semi-professionnels. Cette méthode représente le moyen préféré pour construire les bases des données dans le domaine de REP [60]. Elles sont relativement plus faciles de créer une telle base des données par rapport aux autres méthodes. Les BD actées présentent l'avantage de la facilité de la réalisation mais elles sont aussi moins naturelles. Plusieurs techniques sont utilisées permettant d'améliorer le naturel des émotions produites. L'utilisation des acteurs ou des actrices est une très bonne solution pour améliorer le naturel des émotions. Le principal avantage des bases de données actées est la possibilité de regrouper des données avec distribution uniforme vis à vis de chaque locuteur et vis-à-vis de chaque émotion. Ça nous permettons de trouver les mêmes phrases (mots, passages …) interprétées avec des émotions différentes par des locuteurs différents. Cet avantage permet d'étudier la différence des caractéristiques de la parole selon les émotions et selon les locuteurs [24]. Parmi les BD actées fameuses, on peut citer : Danish Emotional Speech Database [61] et Berlin Dataset [62].

- **Corpus élicitées :** sont créés en plaçant les locuteurs dans une situation émotionnelle simulée qui peut stimuler diverses émotions. Les émotions de cette catégorie sont induites expérimentalement dans des laboratoires en utilisant des techniques d'induction. Nombreuses techniques d'induction ont été créées pour recueillir des manifestations émotionnelles naturelles. Parmi ces techniques on peut citer, le type de scénario très répandu dans le domaine de l'Interaction Homme- Machine repose sur le paradigme du magicien d'Oz [63]. Le paradigme nommé « données spaghetti » développé par Cowie et son équipe, a permis de produire des enregistrements d'émotions spontanées de divers types. Pour améliorer les méthodes d'induction dans un contexte d'interactions sociales, Cowie et son équipe [64] ont également développé le paradigme SAL (Sensitive Artificial Listener). SAL est un agent artificiel doté de compétence affective et qui peut fonctionner en mode automatique [65]. D'autres techniques ont été proposées pour induire des émotions, on peut citer : présentation de films, images ou jeux induisant une réponse émotionnelle.

- **Corpus extraits :** sont collectés par extraction de segments à partir des films, des interviews ou des journaux télévisés [24]. Nous pouvons citer quelques travaux qui utilisent ce type des corpus. Cowie et ses collègues ont construit un corpus en extrayant des segments de programmes télévisés [66]. Schüller a également construit une partie de leur corpus en utilisant des segments émotionnellement colletés à partir des films américains [67]. Clavel, dans ses travaux sur les manifestations de type peur [50], a utilisé des séquences de films en anglais pour construire son corpus. Elle a décrit les avantages et les inconvénients de ce type de corpus. Les avantages de ce type de corpus présentent dans la spontanéité et les films contiennent souvent des passages émotionnellement chargés donc l'obtention d'un grand corpus. Les inconvénients de l'utilisation de ce type de données sont qu'elles sont souvent accompagnées de bruitages et restent susceptibles de ne pas refléter des comportements réels.

2.3.1.2 Aperçu de certaines des BD de parole émotionnelle

Le tableau 2.1 résume les caractéristiques de certaines BD couramment utilisées dans la REP. De ce tableau, nous remarquons qu'il y a des BD de parole émotionnelles développées pour un usage public et d'autres privées. Nous remarquons aussi que les émotions sont généralement stimulées par des acteurs professionnels ou des acteurs non-professionnels. Nous observons qu'il existe des BD en différentes langues. De plus, nous observons dans le tableau 2.1 que la plupart des BD partagent les émotions suivantes : colère, joie, tristesse, surprise, ennui, dégoût et neutre.

2.3.2 Descripteurs utilisés dans la REP

Une étape très importante dans la conception d'un système de REP est l'extraction des caractéristiques appropriées qui efficacement caractériser les différentes émotions. On distingue en général deux types d'informations qui sont utilisées : les informations paralinguistiques et les informations linguistiques. Ces types d'informations vont servir à la compréhension et à l'interprétation d'un message et la perception d'une émotion découlera de leur interprétation. D'après le modèle de Fónagy [81], le message oral est transmis par deux actes successifs d'encodage. Un encodage paralinguistique, qui correspond à la manière dont les sons vont être exprimés. Un encodage linguistique, qui transforme un message global en une séquence de phonèmes.

2.3.2.1 Informations paralinguistiques

Un grand nombre des descripteurs acoustiques sont utilisés pour caractériser les différents états émotionnels : des descripteurs prosodiques, des descripteurs de qualité de la voix et des descripteurs spectraux.

2.3.2.1.1 Descripteurs prosodiques

La prosodie est un canal parallèle au contenu sémantique du message parlé dans les conversations à travers lequel l'auditeur peut percevoir les intentions et l'état émotionnel de locuteur, ou encore distinguer une déclaration d'une question ou d'une commande [82].

Tableau 2.1 - Caractéristiques de certaines BD utilisées dans REP

Corpus	Langue	Source	Emotions	Accès
Berlin emotionnel database [62]	Allemande	Acteurs professionnels	Colère, joie, tristesse, peur, dégoût, ennui, neutre	Oui
Danish emotional database [61]	Danois	Acteurs non professionnels	Colère, joie, tristesse, surprise, neutre.	Oui
LDC Emotional Prosody Speech and Transcripts [68]	Anglais	Acteurs professionnels	Neutre, panique, anxiété, colère, désespoir, tristesse, exaltation, joie, intérêt, ennui, honte, fierté, contempt.	Oui
ESMBS (Emotional Speech of Mandarin and Burmese Speakers [69]	Mandarin	Acteurs non professionnels	Colère, joie, tristesse, dégoût, peur, surprise.	Non
INTERFACE [70]	Anglais, Français, Espagnole Slovène	Acteurs professionnels	Colère, dégout, peur, joie, tristesse, neutre.	Oui
MPEG-4 [71]	Anglais	Films américains	Joie, colère, dégout, peur, tristesse, surprise, neutre.	Non
KISMET [72]	Anglais	Acteurs non professionnels	Approbation, attention, interdiction, apaisante, neutre.	Non
CLDC [73]	Chinoise	Acteurs non professionnels	Joie, colère, surprise, peur, neutre, tristesse	Non
Turkish Emotional Speech Database (TURES) [74]	Turque	Acteurs non professionnels	Bonheur, surprise, tristesse, colère, peur, neutre, valence, activation et Dominance	Oui
Natural [75]	Mandarin	Centre d'appels	Colère , neutre	Non
Sharif Emotional Speech Database (ShEMO) [76]	Persane	Acteurs non professionnels	Surprise, joie, tristesse, peur, colère, neutre	Oui
Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [77]	Anglais	Acteurs non professionnels	Joie, tristesse, peur, dégout, colère, neutre	Oui
Egyptian Arabic speech emotion (EYASE) [78]	Arabe	Acteurs non professionnels	Anger, happiness, neutral and sadness	Non
Tunisian dialect database [79]	Arabe	Acteurs professionnels	Joie, colère, peur, tristesse, neutre	Non
Polish Emotional Speech Database [80]	Polonaise	Acteurs non professionnels	Colère, tristesse, joie, peur, surprise, neutre.	Oui

Les caractéristiques prosodiques et leur corrélation avec les états émotionnels sont étudiées dans [83] [84]. On distingue en général trois types des caractéristiques prosodiques : le pitch, l'intensité et la durée :

- **La fréquence fondamentale ou pitch :** correspond à la fréquence de vibrations des cordes vocales. Donc la fréquence fondamentale ou pitch caractérise les parties voisées du signal de parole et est liée à la sensation de hauteur de la voix (aigüe ou grave). Elle dépend de facteurs spécifiques aux locuteurs comme le sexe, l'âge, la langue maternelle ou l'accent. Les parties voisées ont une structure pseudopériodique et sur ces portions, le signal est généralement modélisé comme la somme d'un signal périodique *T* et d'un bruit blanc. La fréquence fondamentale est l'inverse de la période *T*, $F_0 = 1/T$. Il existe plusieurs méthodes pour l'extraction de la fréquence fondamentale : les méthodes temporelles (autocorrélation, fonctions de différences moyennées (ASDF : Average Square Différence Function)), les méthodes d'estimation par maximum de vraisemblance, les méthodes reposant sur une analyse du cepstre, etc [50].

- **L'intensité (l'énergie) :** correspond à la variation de l'amplitude de signal de la parole causée par une énergie plus ou moins forte provenant du diaphragme et provoquant une variation de la pression de l'air sous la glotte : si la pression sous glottale augmente, l'intensité de la voix augmente également et vice versa [50]. Ce descripteur apporte une mesure globale de la force sonore de la voix (faible ou forte). Elle se mesure généralement en décibel (dB). L'intensité est calculée sur une portion de signal de longueur N à l'aide du logiciel PRAAT de la manière suivante :

$$I = 10\log\left(\sum_{n=1}^{N} s^2(n)w(n)\right) \quad (2.1)$$

  Où *w* est une fenêtre d'analyse gaussienne.

- **Le rythme et la durée :** la notion de rythme comprend le débit de la parole, la longueur et la répartition des pauses, les allongements syllabiques, la durée de divers événements sonores (syllabes, phonèmes). Pour calculer cette mesure du rythme on segmente le signal de la parole en syllabes [50].

  La durée représente l'aspect temporel du signal de la parole. La durée est un paramètre difficile à préciser, par ce qu'elle n'est pas directement associable à un corrélat biologique du système phonatoire. Elle comprend : le débit de la parole, les pauses et la durée des phonèmes qui forment le message. En général, la durée est en corrélation avec les informations linguistiques de la parole telles que : les phonèmes, les syllabes, les mots et les phrases [85].

Dans la littérature, il existe de nombreuses études qui se concentrent sur différents aspects des caractéristiques prosodiques. Les caractéristiques prosodiques telles que l'énergie, la durée, le pitch et ses dérivés sont traitées en corrélation avec les émotions [86-88]. Caractéristiques statistiques telles que : le minimum, le maximum, la moyenne, la variance et le standard déviation d'énergie, et des caractéristiques similaires de pitch sont utilisées

comme des informations prosodiques pour discriminer les émotions [89-90]. La relation entre les paramètres de pitch, de durée et d'énergie est exploitée pour détecter les émotions dans la parole [91]. Les performances des caractéristiques prosodiques locales et globales, et leurs combinaisons sont comparées dans le but de classifier les émotions dans la parole [92]. Les caractéristiques globales sont calculées à partir des paramètres statistiques des caractéristiques prosodiques. Les caractéristiques locales sont rassemblées à partir de la durée des syllabes, les trames de pitch et les valeurs d'énergie. Les résultats ont indiqué que la performance du système de reconnaissance utilisant les caractéristiques locales est meilleure que la performance des caractéristiques globales. Lorsque les caractéristiques prosodiques locales et globales sont combinées, la performance de classification est légèrement augmentée.

2.3.2.1.2 Descripteurs de qualité de la voix

Les paramètres de qualité de la voix décrivent les propriétés de l'excitation glottale. En paralinguistique, le rôle de la qualité de la voix est de véhiculer des informations de type affectif ou attitude, ainsi que sur l'état émotionnel du locuteur [85]. Il existe une forte corrélation entre la qualité de la voix et le contenu émotionnel dans la parole [93]. Parmi les paramètres de la qualité de la voix les plus utilisés dans la reconnaissance des émotions on trouve : le jitter, le shimmer et le rapport harmonique sur bruit (Harmonic to Noise Ratio :HNR), le Taux de de Trames Non Voisées (Unvoiced frames rate :UFR)), et le Taux de Passage par Zéro(TPZ) :

- **Jitter ou la modulation fréquentielle :** représente la variation de la fréquence fondamentale, et correspond à la mesure d'un bruit sur la fréquence. Le jitter permet de modéliser ces oscillations autour de la fréquence fondamentale de la voix [50]. Il est principalement affecté par une insuffisance de contrôle de la vibration du champ vocal. Jitter est utilisé pour identifier l'âge des locuteurs et déterminer le degré de pathologie à partir de la voix. Les jitters brutes et normalisés sont définis respectivement comme [85]:

$$jitter = {\sum_{i=1}^{N-1} |T_i - T_{i+1}|}/{N-1} \tag{2.2}$$

  Où T_i est la période et N représente le nombre de périodes.

- **Shimmer ou la modulation en amplitude :** indique la perturbation ou la variabilité de l'amplitude sonore. Il est relié aux variations d'intensité de l'émission vocale et partiellement affecté par la réduction de la résistance glottique. Shimmer est souvent utilisé de pair avec le jitter. Shimmer est estimé de façon similaire à celle du jitter, sauf qu'il utilise l'amplitude comme paramètre. Shimmer est généralement mesuré sur une échelle logarithmique en décibel. On peut calculer le paramètre shimmer par la formule suivante [85]:

$$shimmer(dB) = \frac{1}{N-1} = \sum_{i=1}^{N-1} |20\log(A_{i+1} - A_i)| \tag{2.3}$$

Où A_n est l'amplitude de la période T_i.

- **Taux de Trames Non Voisées (UFR) :** le signal sonore peut être divisé en plusieurs trames voisées ou non voisées. Le taux de trames non voisées est correspond à la proportion de fenêtres considérées comme non voisées sur une portion du signal de la parole [50]. Le pourcentage des trames non voisées dans la phrase est révélateur de la quantité de pauses sur cette phrase. Par exemple une phrase prononcée à un rythme normal contiendra beaucoup plus de pauses donc plus de trames non voisées. Au contraire une phrase prononcée avec un rythme élevé comme le cas de la colère ou de la peur contiendra moins de pauses donc moins de trames non voisées.

- **Rapport Harmonique sur Bruit (HNR) :** l'idée de ce paramètre est de trouver un indicateur du niveau de souffle dans la voix en calculant le rapport harmonique sur bruit du signal de la parole connu sous le nom de PAP (Périodique APériodique) ou HNR (Harmonic to Noise Ratio) [50]. L'harmonicité est mesurée en dB, et calculée comme le Ratio de l'Energie de la partie Périodique avec l'Energie du Bruit :

$$HNR = 10\log(\frac{E_p}{E_n}) \tag{2.4}$$

Où E_p est l'énergie de la partie périodique et E_n est l'énergie du bruit.

- **Taux de Passage par Zéro (TPZ) :** comme son nom l'indique, il est défini par le nombre de passages par zéro dans une région définie de signal, divisé par le nombre d'échantillons de cette région. Le TPZ de courte durée représente la fréquence contenue du signal vocale [94-96]. Un TPZ élevé implique une haute fréquence et un faible TPZ implique une basse fréquence.

- **Normalized Amplitude Quotient (NAQ) :** il est décrit comme une mesure de souffle dans la voix. Il dépend de la fréquence fondamentale et de la configuration glottale au travers du quotient d'amplitude [97]. Ce coefficient a été utilisé pour la reconnaissance des émotions à partir de la parole.

Les paramètres de qualité de la voix sont utilisés dans plusieurs travaux de reconnaissance et de classification des émotions. On peut citer les travaux de Zhang qui a utilisé les paramètres prosodiques, paramètres de qualité vocale (jitter, shimmer et HNR) et les formants pour reconnaître des émotions dans la parole. Lorsqu'une combinaison de paramètres prosodiques et de paramètres qualité vocale a été utilisée le taux de reconnaissance est amélioré de 10% par rapport à l'utilisation des paramètres prosodiques seuls [98]. Li et ses collègues ont utilisé les paramètres jitter et shimmer avec des paramètres spectraux pour classifier les émotions dans la parole [99]. Un taux de reconnaissance élevé est obtenu par l'addition des paramètres jitter et shimmer au système de reconnaissance. Les paramètres de qualité vocale (jitter, shimmer, HNR et TPZ) et les paramètres prosodiques (pitch et intensité) sont utilisés pour la classification de sept émotions (colère, joie, tristesse, peur, dégoût, ennui et neutre) dans la parole [100]. Les auteurs dans ce travail ont exploité le corpus Berlin dataset comme base de données du système de reconnaissance. Le taux de reconnaissance obtenu est

égale 81.13%. Les trames voisées et non voisées sont utilisées par l'auteur Vasquez-Correa pour la REP [101].

2.3.2.1.3 Descripteurs spectraux et cepstraux

Les caractéristiques spectrales sont obtenues en transformant le signal du domaine temporel en signal du domaine fréquentiel en utilisant la Transformée de Fourier. Plusieurs paramètres spectraux et cepstraux sont utilisés dans le domaine de REP. Parmi les paramètres les plus utilisées on peut citer : les formants, les coefficients MFCC (Mel-Frequency Cepstral Coefficients), les coefficients du Codage par Prédiction Linéaire (LPC : Linear Predictive Coding coefficients), les coefficients cepstraux de prédiction linéaire (LPCC : Linear Prediction Cepstral Coefficients), et les coefficients PLP (Perceptual Linear Prediction) :

- **Les formants et leur largeur de bande :** le conduit vocal présent des fréquences de résonance, ce qui se manifeste dans le spectre par l'apparition de pics formantiques. Un formant est un pic d'amplitude dans le spectre d'un son composé de fréquences harmoniques, inharmoniques et/ou de bruit. On ajoute en général la largeur de bande du formant. Ce dernier est défini comme la largeur de la bande du spectre entre les points à -6 dB par rapport à la crête du formant [50].

- **Mel-Frequency Cepstral Coefficients (MFCC):** les paramètres de MFCC appartiennent à la famille des descripteurs cepstraux qui se basent sur une représentation cepstrale du signal. Le cepstre présente l'avantage de permettre une séparation des contributions respectives de la source et du conduit vocal. Pour obtenir les coefficients MFCC, en utilisant pour le calcul du cepstre, l'échelle des fréquences Mel. L'échelle des fréquences Mel est une échelle fréquentielle non linéaire tenant compte des particularités de l'oreille humaine. L'échelle Mel correspond à une approximation de la sensation psychologique de la hauteur d'un son qui prend notamment en compte que la sélectivité en fréquence est plus grande dans les graves que dans les aigus [50]. On peut obtenir l'échelle des fréquences Mel par l'expression suivante :

$$m(f) = 2595 \log(1 + \frac{f}{700}) \quad (2.5)$$

  Où f est la fréquence en Hertz.

  Pour calculer les paramètres de MFCC, les énoncés de la parole sont divisés en segments, puis chaque segment est converti dans le domaine fréquentiel en utilisant une transformée de Fourier discrète de courte durée. Un filtrage des énergies par un filtre banc en échelle de fréquence Mel est effectué. Enfin, une transformée de Fourier inverse est appliquée pour obtenir les paramètres MFCC [102].

- **Linear Predictive Coding Coefficients (LPC):** les coefficients LPC sont une autre représentation de l'enveloppe spectrale du signal. Les coefficients LPC sont basés sur le modèle de production de la parole qui considère en première approximation que l'appareil de production de la parole (cordes vocales et conduit vocal complet) est

constitué d'une source (source pseudopériodique ou source de bruit) et d'un filtre se comportant comme un résonateur (conduit vocal).

- **Linear Prediction Cepstral Coefficients (LPCC) :** les coefficients LPCC sont des coefficients dérivés des LPC. Ces paramètres montrent des différences avec des émotions particulières.
- **Perceptual Linear Prediction (PLP) :** les coefficients PLP correspondent à une amélioration des LPC puisqu'ils sont fondés sur le filtre perceptif de Bark.

Les paramètres spectraux sont fortement utilisés dans le domaine de REP. Les paramètres MFCC donnent une performance élevée par rapport au paramètre de pitch dans un système de REP [103]. Les paramètres MFCC avec le pitch et l'intensité sont utilisés pour reconnaître les émotions de colère, joie, tristesse et neutre, le corpus SAVEE est utilisé comme BD dans le système de la reconnaissance [104]. Combinaison de différents paramètres spectraux tels que MFCC, LPCC, PLP et LFPC (Log Frequency Power Coefficients) est utilisée pour classer les émotions de colère, ennui, bonheur, neutre et tristesse en Mandarin [105]. La performance de paramètres LFPC est comparée avec les performances des paramètres LPCC and MFCC afin de classifier les émotions, les LFPC fonctionnent légèrement mieux [106]. Les paramètres LPCC et les paramètres prosodiques (pitch et énergie) sont utilisés pour discriminer les émotions de colère, dégoût, peur, joie, neutre, tristesse et surprise [107]. Les formants et leur largeur de bande ainsi les paramètres de pitch et l'intensité sont exploités pour reconnaître les émotions dans la langue Telugu [108].

Dans beaucoup de travaux, l'utilisation des combinaisons entre les différents paramètres : prosodiques, paramètres de qualité de la voix et les paramètres spectraux améliore la performance des systèmes de reconnaissance. Pour analyser la performance des différents corpus dans le domaine de REP, différentes combinaisons de différents paramètres sont formés pour trouver la combinaison de paramètres qui donne la meilleure performance. Les paramètres utilisés sont : le pitch, l'intensité, le jitter, le shimmer, les paramètres MFCC et les formants [109]. Dans [110], les paramètres d'énergie, la fréquence fondamentale, les MFCC, les formants, les PLP et les LPC sont utilisés pour classifier les émotions dans la parole. Pour classifier trois émotions : la colère, la joie et le neutre, différents paramètres de différents types (intensité, TPZ, formants, MFCC, spectral centroid, short term energy) sont exploités dans le système de la reconnaissance [111]. Pour reconnaître les émotions dans les BD de parole émotionnelle de Berlin et d'Espagne, le système qui a utilisé une combinaison des caractéristiques prosodiques (énergie et pitch) et spectrales (MFCC) a donné des meilleures performances par rapport aux performances des systèmes qui ont utilisé des caractéristiques individuelles (caractéristiques prosodiques ou spectrales) [102]. Une combinaison de différents paramètres acoustiques (pitch, intensité, durée, UFR, jitter, shimmer et MFCC) est utilisée dans une système de reconnaissance des émotions de peur et de neutre [112]. Les résultats obtenus montrent que l'utilisation des paramètres MFCC donne un taux de classification important 86, 14%. Dans [113], des paramètres prosodiques (pitch, intensité et durée) et les paramètres MFCC sont utilisés pour reconnaitre les émotions de joie et de tristesse. Emo-dB est exploitée comme BD dans les systèmes de reconnaissance. Les résultats

obtenus indiquent que la combinaison entre les paramètres prosodiques et les paramètres MFCC donne le meilleur taux de reconnaissance.

2.3.2.2 Informations linguistiques

Le contenu linguistique de l'énoncé parlé est une partie importante de l'émotion véhiculée [114]. Les informations linguistiques ont été introduites dans le domaine de REP par plusieurs auteurs. Elles se présentent au niveau lexical et au niveau du discours. Au niveau lexical l'information linguistique est représentée par la saillance émotionnelle des mots. La saillance émotionnelle est plus ou moins liée à l'information mutuelle entre un mot particulier et une certaine catégorie émotionnelle [115]. Au niveau du discours l'information linguistique est représentée par : les actes de dialogue tels que le rejet, la répétition, la reformulation ou la demande de répétition. Dans le contexte de REP, les informations de discours peuvent faire référence à la manière dont un utilisateur interagit avec la machine [87].

Dans le domaine de la parole émotionnelle, les informations linguistiques ont été utilisées séparément [116] ou en combinaison avec les informations paralinguistiques par exemple dans [117-120].

Le niveau linguistique peut apporter des informations pour la détection des émotions pour des données enregistrées au téléphone par exemple dans la BD CEMO, les émotions négatives peuvent être liées à certains termes comme « problème ». Et la détection du soulagement peut être attribuée à certains termes spécifiques comme « merci » [121]. Boufaden et ses collègues [122] ont montré que l'utilisation d'un vocabulaire composé uniquement des deux mots Yes et No, donnés comme réponse d'un centre d'appels, permettait d'améliorer les performances de reconnaissances des émotions.

Les actes dialogiques ont été exploités dans les travaux de Devillers [123]. Des mesures de corrélation ont montré que les émotions négatives : peur et colère sont susceptibles de générer des assertions, réassertions, requêtes et répétitions, alors que les émotions positives comme la neutre et satisfaction sont corrélées avec l'acceptation. Les auteurs dans [87] ont utilisé cinq actes de dialogue du type réjection, répétition, en plus d'indices lexicaux et paramètres prosodiques pour améliorer la performance de détection des émotions. Les informations sur le discours ont été combinées avec des corrélats acoustiques afin d'améliorer la performance des systèmes de reconnaissance des émotions [124].

2.3.3 Techniques de classifications

Un système de REP comprend une phase de classification qui consiste à attribuer une catégorie d'informations, Exemple l'émotion, à un ensemble de caractéristiques mesurables, Exemple les paramètres acoustiques. En général, un système de classification automatique comporte deux étapes : la phase d'apprentissage et la phase de test. La phase d'apprentissage s'agit de créer un modèle caractéristique de chaque classe. La phase de test permet d'évaluer les performances du système de classification. Elle consiste à associer une classe d'émotion au vecteur de caractéristiques à l'aide d'une fonction de décision [125].

On distingue deux types de classification : supervisée et non supervisée. Lors d'une classification supervisée la classe de chaque objet (représentée par son étiquette) est fournie au programme d'apprentissage en même temps que les données. Lors d'une classification non supervisée, les classes sont déterminées automatiquement en fonction de la structure des données. Les systèmes de classification automatique d'émotions utilisent essentiellement des méthodes supervisées où les classes considérées sont des classes d'émotions souvent déterminées en fonction de l'application visée [51]. Il existe de familles de classifieurs : génératifs et discriminants.

Les classifieurs génératifs exploitent des méthodes statistiques pour estimer l'appartenance aux classes des données testées parmi ces classifieurs on peut citer : Analyse Linéaire Discriminante (Linear Discriminant Analysis : LDA), Analyse Discriminante de Fisher (Fisher Linear Discriminant Analysis : FLDA), Modèles de Mélanges Gaussiens (Gaussian Mixture Models: GMM). Les classifieurs discriminants cherchent à définir le meilleur moyen de séparer les données à travers leur espace de représentation et selon leur classe associée parmi ces classifieurs on peut citer : Perceptron et ses variantes Multicouches (Multi Layer Perceptron : MLP), Machines à Vecteur Support (Support Vector Machines : SVM).

Dans la littérature, les techniques de classifications les plus utilisés dans le domaine de la REP sont : SVM, GMM, K plus proches voisins (K-Nearest Neighbors : KNN), Modèles de Markov Cachés (Hidden Markov Models : HMM)), Réseaux de Neurones Artificiels (Artificial Neural Network : ANN).

La technique de SVM est largement utilisée dans de nombreuses applications de reconnaissance de formes et elle a montrée des performances supérieures par rapport des autres classificateurs bien connus [126]. SVM est également largement utilisée dans les systèmes de REP dans de nombreuses études. Shen et ses collègues ont utilisé la méthode de SVM pour classer les émotions dans la base de données de Berlin. Les auteurs dans ce travail ont utilisée une combinaison de caractéristiques prosodiques et spectrales dans le système de reconnaissance [127]. SVM est utilisé avec les paramètres de pitch, énergie, MFCC, LPCC et MEDC pour classer les émotions dans les bases de données émotionnelles de Berlin et de Chinoise [128].

La technique de classification GMM a donné des résultats acceptables et souvent comparables à d'autres méthodes de classification [129]. Neiberg et Coll ont utilisé le classificateur GMM pour classer les émotions dans la parole en utilisant les paramètres de pitch et MFCC [130]. GMM a atteint une précision de 92% lorsqu'il est utilisé pour reconnaître les émotions dans le corpus de Basque [131]. Les caractéristiques spectrales ont été utilisées pour le paramétrage des signaux vocaux pour la reconnaissance des émotions en utilisant le classificateur GMM [132].

De nombreux travaux se sont concentrés sur les HMM en tant que classificateur pour reconnaître les émotions de la parole. Les HMM ont été utilisés comme classificateur dans le système de reconnaissance pour classer six états émotionnels de la parole : la colère, le bonheur, la joie, la peur, la tristesse, le dégoût [133]. Les classificateurs HMM et SVM avec

les paramètres spectraux sont utilisés pour la classification de sept émotions discrètes : la colère, la surprise, la joie, la peur, la tristesse, le dégoût et le neutre [134].

KNN est une technique de classification simple utilisée dans les systèmes de REP. Le classificateur KNN est utilisé pour reconnaître les émotions dans deux corpus émotionnels, naturel et acté en Mandarin. Les auteurs dans ce travail ont utilisé les paramètres de pitch, énergie, TPZ et les paramètres spectraux [135]. KNN est utilisé pour différencier l'émotion de l'anxiété entre sept émotions en utilisant des combinaisons des paramètres spectraux et en exploitant la base de données de Berlin [136].

ANN est une méthode couramment utilisée pour plusieurs types de problèmes de classification. La technique de classification d'ANN est utilisée pour la reconnaissance des émotions dans beaucoup de travaux. Quatre types d'émotion : le neutre, la joie, la tristesse et la colère sont classés en utilisant le classificateur ANN [137]. Dans [138], ANN est utilisée comme classificateur dans un système de REP en utilisant les paramètres de pitch et LPC et en utilisant une base de données de la langue japonaise. Réseaux de Neurones Convolutifs (Convolutional Neural Network : CNN) sont des types particuliers de réseaux neuronaux qui sont exploités pour la classification des émotions dans la parole [139].

Réseaux de Neurones Récurrents (Recurrent Neural Network : RNN) sont une famille de réseaux de neuronaux spécialisés dans le traitement de données séquentielles. RNN sont utilisés dans plusieurs travaux de parole émotionnelle [140] [141].

Dans la littérature, il y a des autres méthodes de classification qui sont utilisées dans le domaine de REP. On peut citer : LDA, Analyse Discriminante Régularisée (Regularized Discriminate Analysis : RDA) [142], FLDA [143], MLP [144] et Arbre de décision (Decision Tree : DT) [145].

Différentes techniques de classification ont été comparées pour développer les systèmes de REP. Dans [146], les performances des classificateurs ANN et SVM ont été comparées. Les résultats ont indiqué que le classificateur ANN a donné des performances élevées autour de 88, 4 % et 78, 2% pour le classificateur SVM. La performance du classificateur KNN a été comparée à celle du classificateur SVM. Le classificateur SVM est mieux performant que KNN pour classer les émotions de la parole en utilisant la base de données EYASE [78]. Deux méthodes de classifications GMM et SVM sont exploitées pour la classification des émotions. Les deux méthodes obtiennent des performances similaires, 76% pour les SVM et 75% pour les GMM [147]. Pour classer les émotions dans la parole, différents classificateurs (KNN, SVM, LDA et RDA) ont été comparées. Les résultats obtenus indiquent que la meilleure performance a été donnée par RDA [142]. Les méthodes KNN et LDA sont utilisées dans [148]. Les meilleures performances sont obtenues avec la méthode KNN. Dans [149], SVM et DT sont utilisées pour classifier les émotions sur des données réelles (centres d'appel). Les résultats obtenus n'ont pas montré de différences significatives entre les performances par les deux techniques. Le système de reconnaissance dans [112] est basé sur les méthodes de classification KNN, SVM et ANN. Les résultats obtenus indiquent que la méthode d'ANN donne les meilleures performances par rapport aux autres méthodes utilisées.

Dans plusieurs travaux, différents classificateurs ont été combinés pour améliorer la performance de reconnaissance. Par exemple on peut citer : un classificateur hybride (GMM-DNN) composé de deux classificateurs GMM et DNN (Deep Neural Network). Ce classificateur hybride a été comparé avec les classificateurs de SVM et MLP. Les résultats obtenus indiquent que la performance de GMM-DNN est supérieure par rapport aux celles de SVM et MLP [150]. Multi-SVNN (Multiple Support Vector Neural Network) est un classificateur composé de SVM et ANN [151]. Ce classificateur est exploité pour identifier les émotions dans le signal de parole. Un système de classification parallèle composé de trois techniques : SVM, KNN et LDA est utilisé comme classificateur pour étudier l'impact de l'âge et du sexe sur la reconnaissance des émotions dans le dialecte algérien [152].

2.3.4 Discussion sur certaines recherches importantes dans la REP

Dans cette section certaines recherches importantes liées à la REP sont discutées en bref [153]:

- la majorité des systèmes de REP a utilisé des bases de données avec un nombre limité des locuteurs ;
- la plupart des recherches sur la parole émotionnelle sont focalisées principalement sur la caractérisation des émotions. Par conséquent, la tâche principale effectuée était de dériver des informations spécifiques aux émotions tirées de la parole et les utiliser pour classer les émotions;
- l'expression des émotions est un phénomène universel, qui peut être indépendant du locuteur, du sexe et de la langue ;
- la majorité des travaux de REP réalisés dans la littérature sont effectués à l'aide de bases de données simulées. Le défi principal est de reconnaître les émotions naturelles. Les caractéristiques et les techniques discutées dans la littérature peuvent être appliquées aux corpus naturels pour analyser la reconnaissance des émotions, dont la réalisation nécessite la collection des corpus de parole émotionnel, couvrant un large éventail d'émotions ;
- plus souvent dans la littérature, la tâche de classification des émotions est effectuée à laide d'une seule modèle (GMM, ANN, KNN ou SVM, etc.). Des modèles hybrides de classificateurs sont utilisés pour améliorer la performance dans le cas de la reconnaissance des émotions ;
- la tendance de la reconnaissance des émotions est connue dans de nombreuses langues et n'est pas connue en d'autres langues ;
- l'étude sur la discrimination des émotions a étendu aux dimensions de l'émotion (excitation et valence), qui sont issus de la psychologie de la production et la perception des émotions. Les caractéristiques liées aux dimensions de l'émotion ont été explorées pour améliorer la performance de REP ;

- l'expression des émotions est une activité multimodale. Par conséquent, d'autres modalités comme l'expression faciale, les signaux biologiques ont été utilisée avec le signal de la parole pour développer la reconnaissance des émotions ;

- dans des applications en temps réel telles que l'analyse des appels en cas de services d'urgence, vérification d'émotions pour analyser l'authenticité des demandes est importante. Dans ce contexte, dans le cadre de vérification d'émotions, des caractéristiques et des modèles peuvent être explorés ;

- l'effet de l'expression d'émotions dépend également du contenu linguistique de la parole. L'identification de la saillance émotionnelle des mots dans la parole émotionnelle et les caractéristiques extraits de ces mots a amélioré la performance de la reconnaissance des émotions.

2.3.5 Applications de la REP

La REP a plusieurs applications dans la vie de tous les jours. L'intégration des émotions dans des applications de l'intelligence artificielle connaît un grand développement dans les dernières années. Et plusieurs domaines s'intéressent de plus en plus à l'état émotionnel des utilisateurs. Nous exposerons dans ce qui suit quelques applications de la REP :

- **l'amélioration du service à la clientèle :** la détection de l'état émotionnel de l'utilisateur permet d'adapter la stratégie dialogique pour fournir des réponses plus adaptées [154]. L'identification automatique de l'émotion à travers la voix permettra de faire le suivi de la qualité de la relation avec les clients. La détection d'un appel problématique à travers la détection d'un état émotionnel négatif d'un client permettra à la machine d'entreprendre plusieurs stratégies de gestion de l'échec de l'appel (Exemple, restreindre et guider le dialogue, traiter l'appel en priorité par un opérateur humain, etc.). Intégration un système de reconnaissance des émotions aux serveurs vocaux interactifs des centres d'appels commerciaux permet d'améliorer les services à la clientèle [155]. Les systèmes de dialogues développés pour des banques ou des services téléphoniques commerciaux s'intéressent à des réactions de frustration ou d'irritation. Et les systèmes développés pour les centres d'appels d'urgence s'intéressent à l'inquiétude ou à l'anxiété ;

- **les systèmes tutoriels :** un système tutoriel est un système capable de savoir si l'utilisateur est ennuyé, découragé ou irrité par la matière enseignée et pourra par conséquent changer le style et le niveau de la matière dispensée, fournir une compensation et un encouragement émotionnel ou accorder une pause à l'utilisateur [156][157]. L'apprentissage de piano assisté par ordinateur est un exemple d'un système tutoriel. Dans ce cas le professeur peut analyser l'état émotionnel de son élève avec les notes jouées, la position du pianiste, etc. Ce professeur serait bien plus efficace en étant capable de déterminer si son élève prend du plaisir à jouer, ou si au contraire, il présente des signes de stress. Pour l'élève, il est bien plus agréable d'avoir un professeur non seulement parfait techniquement, mais également compréhensif et

patient. Un système d'apprentissage a été développé nommé STEVE (SOAR Training Expert for Virtual Environments). Ce système intègre un agent virtuel qui permet d'aider les étudiants à apprendre à exécuter des tâches physiques basées sur des procédures. Le rôle principal de l'agent STEVE est le contrôle sur l'état émotionnel de l'apprenant. Un système de tuteurs intelligents nommé DARWAR a été développé destiné à l'apprentissage d'une langue étrangère. A coté de l'apprentissage dans ce système, un agent virtuel permet l'aide le conseil et le soutient émotionnellement. Un système d'apprentissage appelé EMASPEL a été développé intégrant plusieurs agents conçus pour la gestion des émotions [23];

- **l'indicateur d'aptitude :** la reconnaissance de l'état émotionnel peut être utilisée comme indicateur d'aptitude pour exploiter dans la sécurité des personnes, telle que la conduite ou le pilotage afin d'activer les routines de sécurité. L'exemple qui est utilisé la reconnaissance d'émotions comme un indicateur d'aptitude est le système de bord de voiture (car board system). La reconnaissance des émotions dans le système de bord de voiture aide l'humain à effectuer diverses tâches telles que l'accélération, le freinage, la vitesse, la distance par rapport a un autre véhicule, etc. Le système de reconnaissance des émotions peut être utilisé dans un système de conduite automobile, où les informations sur l'état d'un conducteur peuvent être utilisées pour le garder alerte pendant la conduite. Cela permet d'éviter certains accidents dus à état mental stressé du conducteur [155];

- **la surveillance et monitoring :** la REP soutient de nombreuses situations comme la gestion de crises liées à la sécurité et la surveillance. Par exemple la détection de la présence d'émotions principalement la peur, dans le cadre de la surveillance dans les lieux publics [158]. Les systèmes de lutte contre le terrorisme peuvent être plus efficaces en détectant les émotions telles que l'agressivité d'agresseurs ou la peur des victimes potentielles [159][160]. La reconnaissance des émotions utilisée pour l'évaluation de l'urgence d'un appel pour prendre une décision dans le cadre d'un centre d'appels médical offrant un service de conseils médicaux aux patients [161];

- **l'indexation des films :** indexation automatique d'émotion d'acteurs ou d'évènements émotionnels utiles pour la récapitulation automatique des films [162]. L'analyse du contenu affectif des films peut fournir des informations sur le genre du film. La détection des événements émotionnels par exemple les rires ou les manifestations d'horreur à travers la bande audio dans des comédies et des films d'horreur [163];

- **les systèmes de dialogues :** les émotions font partie intégrante de la communication dans le sens où elles ont la faculté d'agir sur la sémantique du message transmis. Les émotions interviennent de deux manières dans les systèmes de dialogue, en tant que phénomène susceptible d'altérer la reconnaissance des mots prononcés par l'utilisateur, et en tant que phénomène permettant de mieux comprendre son comportement [23]. Les traits spéciaux véhiculés par les émotions sont utilisés pour le développement de systèmes de vérification automatique de locuteurs ou de systèmes de reconnaissance automatique de la parole [164] [165]. Parmi les systèmes de

dialogues, les travaux de Chavel [50] qui a développé une application de surveillance. Chavel s'est intéressé à la reconnaissance des états émotionnels dans la parole pour identifier des situations de menace pour la vie humaine. On peut citer également les travaux de Devillers et Vidrascu [166] qui ont étudié les réactions de frustration ou d'irritation dans des interactions orales enregistrées dans des centres d'appels ;

- **les robots :** peuvent interagir avec les gens et peuvent les aider dans leurs activités quotidiennes, dans des lieux communs tels que les maisons, les supermarchés, les hôpitaux, les bureaux, etc. Pour accomplir ces tâches, les robots doivent reconnaître les émotions des humains afin de créer un environnement convivial. Sans reconnaître l'émotion, le robot ne peut pas interagir avec l'humain de manière naturelle [155]. Les robots thérapeutiques ont un fort potentiel pour améliorer l'état de l'activité cérébrale chez les patients souffrant de démence. Ils ont montré que l'interaction avec un robot rend les personnes âgées plus actives et plus communicatives [23]. Le robot peut servir comme assistants thérapeutiques pour personnes avec handicap cognitif ou social, les autistes en l'occurrence, pour améliorer leurs habilités de communication sociale et leurs apprendre à exprimer leurs émotions [167]. Les agents artificiels cherchent à analyser et reproduire les comportements humains pour interagir socialement avec l'homme. Par exemple, le robot AIBO et le robot Kismet [168] intègrent les différentes émotions dans leur modèle d'interaction. AIBO est capable d'exprimer les émotions de la joie, la tristesse, la colère, la surprise, la peur et le mécontentement. Kimset est aussi capable d'exprimer les émotions de calme, colère, dégoût, peur, joie, tristesse et intérêt ;

- **les jeux :** peuvent être contrôlés par les émotions de la parole humaine. L'ordinateur peut reconnaître l'émotion humaine à partir de leurs discours et détecter le niveau du jeu (facile, moyen, difficile). Par exemple, si le discours humain est de nature agressive, le niveau devient difficile. Supposons que si l'humain est trop détendu, le niveau devient facile [155]. Des agents conversationnels animés (ACA) ont été développés dans le but d'améliorer les interfaces homme-machine graphiques traditionnelles ou en tant que personnage intelligent d'un jeu vidéo [169]. Les chercheurs Jones et Sutherland ont porté sur la détection explicite des émotions du joueur dans le but de modifier dynamiquement l'environnement. Par exemple, un jeu sur ordinateur peut s'adapter et réagir en ajoutant de nouveaux adversaires pour stimuler le joueur lorsqu'il commence à se lasser (émotion détectée : l'ennui), où au contraire, diminuer la difficulté du jeu si le joueur est trop excité, stressé ou fatigué [155];

- **le domaine de la santé :** l'utilisation de la communication vocale dans le domaine médical permet au patient de décrire son état de santé. Dans l'analyse clinique, les émotions humaines sont analysées en fonction des caractéristiques liées aux paramètres prosodiques et aux paramètres de qualité vocale. Les applications dans le domaine de la santé peuvent servir d'une surveillance acoustique détectant les douleurs et classant automatiquement les appels de détresse par la reconnaissance automatique de l'émotion [170]. Les médecins peuvent utiliser le contenu émotionnel

de la parole d'un patient comme un outil de diagnostic de divers troubles (la maladiede Parkinson, l'ablation du larynx et les effets pathologiques) [171];

- **le détecteur de mensonges :** utilisant la REP aide à décider si quelqu'un ment ou non. Ce mécanisme est notamment utilisé dans des domaines tels que le bureau central d'enquête pour trouver les criminels. X13-VSA PRO Voice Lie Detector 3.0.1 PRO est un système avancé et sophistiqué, ce système analyse le stress dans la voix qui nous permet de détecter la vérité instantanément [155];
- **la messagerie vocale :** est un système électronique d'enregistrement et de stockage de messages vocaux pour une récupération ultérieure par le destinataire prévu. La reconnaissance des émotions appliquée pour trier les messages vocaux exprimés par l'appelant [155].

2.4 Conclusion

Ce chapitre a présenté une revue de travaux sur la REP du point de vue des bases de données émotionnelles, des paramètres spécifiques aux émotions extraites de différents aspects de la parole et des modèles de classification. Certaines recherches importantes dans le domaine de la REP sont également discutées dans ce chapitre. Ainsi des principales applications de ce domaine de reconnaissance de l'émotion sont présentées. Dans ce chapitre nous avons identifié les paramètres qui composent un système de REP. Ces systèmes nécessitent des données d'apprentissage fournies par des bases de données émotionnelles de différents types : naturels, actés, élicités et extraits. Les systèmes de reconnaissance utilisent les plus souvent des paramètres prosodiques, spectraux et de qualités vocales. Une fois que tous les paramètres sont extraits, les systèmes disposent d'un large éventail d'algorithmes de classification.

Chapitre 3 :

Base de données de parole émotionnelle et extraction des paramètres acoustique

3.1 Introduction

La Reconnaissance des Émotions dans la Parole (REP) vise à identifier l'état émotionnel d'un être humain à partir de sa voix. L'architecture générale du système de REP comporte quatre étapes principales illustrées à la figure 3.1: une entrée vocale émotionnelle, l'extraction des caractéristiques appropriées du signal de la parole, la classification et un état émotionnel dans la sortie.

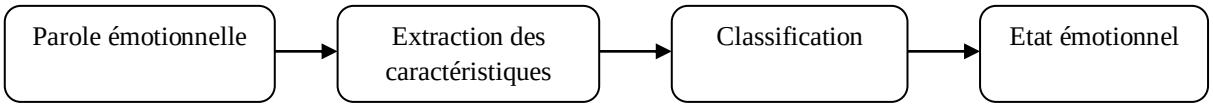


Figure 3.1- Architecture générale du système de REP

Dans le domaine de REP, les caractéristiques acoustiques se sont avérées être très efficaces pour reconnaître les émotions. Comme mentionné dans le deuxième chapitre, l'étude des paramètres est une partie indispensable dans les systèmes de classification et de la reconnaissance des émotions. L'extraction des caractéristiques acoustiques est effectuée à partir d'une BD de parole émotionnelle. Les caractéristiques de la BD dépendent de l'objectif de la recherche. Dans ce chapitre, la base de données émotionnelle du dialecte algérien (ADED) qui est utilisée dans ce travail est décrite dans la deuxième section. Dans la troisième section, les paramètres acoustiques utilisés pour modéliser les différents états émotionnels sont présentés. Ainsi nous avons fait une analyse pour détecter l'influence de différents états émotionnels (peur, colère, tristesse et neutre) sur les paramètres acoustiques choisis.

3.2 Base de données de parole émotionnelle

Les corpus sont essentiels pour l'entraînement et l'évaluation des systèmes de REP. Dans ce domaine, des corpus sont déjà disponibles pour la plupart des langues occidentales et orientales comme l'Anglais, le Français, le Chinois, etc. Cependant, le dialecte algérien ne fait pas partie des langues étudiées dans le contexte de la reconnaissance des émotions. Dans ce travail, nous avons construit une BD émotionnelle du dialecte algérien. Une idée sur le dialecte algérien est donnée dans cette section. Ainsi la procédure de création de notre BD de parole émotionnelle est décrite.

3.2.1 Dialecte algérien

L'Algérie est un grand pays, administrativement divisé en 58 départements. Sa première langue officielle est l'Arabe Standard Moderne. Cependant, les dialectes algériens sont largement le moyen de communication prédominant.

Le dialecte arabe algérien est un dialecte maghrébin. Il présente de nombreuses variantes qui se développent principalement à la suite des deux phases d'arabisation et de l'histoire de la colonisation profonde. Selon les phases d'arabisation, nous pouvons diviser les dialectes en trois grands groupes : pré-Hilal, Hilal et les dialectes mixtes. Les dialectes pré-Hilal sont appelés dialectes sédentaires, qui sont parlés dans la région qui est affectée par l'expansion de

l'Islam au 7ème siècle. Les dialectes Hilal prennent le nom de Banu Hilal, ils sont appelés dialectes Bédouins, parlés dans la région qui est influencée par l'immigration arabe au 11ème siècle. Les dialectes mixtes ou Hilal urbain sont parlés dans une région affectée par les deux phases d'arabisation [172][173].

Les mots du dialecte algérien sont majoritairement dérivés de la langue arabe, bien qu'ils aient également emprunté de nombreux mots en Français, dans une moindre mesure de mots berbères [174]. Les dialectes algériens sont également influencés par la longue période de colonisation profonde. En fait, le dialecte algérien est affecté par d'autres langues telles que le Turc, l'Italien et l'Espagnol [152]. Ce dialecte algérien, bien qu'assez proche des dialectes marocain et tunisien, se distingue assez facilement de chacun d'eux par son vocabulaire et certaines des constructions syntaxiques, il est assez différent de l'égyptien et des autres dialectes du Moyen-Orient. Pour être plus précis, il n'y a pas de dialecte algérien unique : différentes régions d'Algérie parlent de légères variations du même dialecte avec des accents différents sur la prononciation. Par exemple le dialecte de la ville de Constantine qui est située dans l'est d'Algérie est un peu différent de celui parlé à Alger, la capitale qui est située au centre-nord du pays, et les deux sont différentes du dialecte parlé dans les villes d'Oran ou de Tlemcen dans la région ouest, ou d'Adrar et de Béchar dans le sud-ouest du pays.

Les dialectes algériens ont été étudiés dans plusieurs travaux. On peut citer le travail de Meftouh et ses collègues [174], ce travail présente une analyse linguistique d'une classe particulière de dialecte algérien qui est le dialecte d'Annaba (Algérie EST). Une étude concernant les sous-dialectes algériens a été menée par Bougrine [175] dans lequel les auteurs ont créé un corpus de discours parallèles en Arabe. Plus précisément, le corpus a été créé via un enregistrement direct, et les auteurs ont préparé une série de questions à poser à 109 participants de 17 villes différentes. Djellab et ses collègues ont proposé un nouveau corpus de parole pour la reconnaissance de l'accent régional algérien (Algerian Modern Colloquial Arabic Speech Corpus : AMCASC) [176]. Le corpus est réel (environ 88 heures) enregistré sur différents supports. Dans le travail de Harrat [177], le but est de faire une traduction entre le dialecte algérien et l'arabe classique. Dans ce travail, les auteurs ont étudié plusieurs dialectes pour les adapter à la traduction automatique. Ils ont concentré sur deux types de dialecte algérien, ALG qui est le dialecte d'Alger et ANB qui est le dialecte d'Annaba. Dans [178], une approche qui identifie le dialecte arabe algérien dans la parole a été conçue. L'approche est basée sur l'information prosodique de la parole, à savoir l'intonation et le rythme. Pour le traitement de la parole émotionnel, un échantillon de données du dialecte algérien a été présenté [179]. Il contient deux heures d'enregistrements audiovisuels extraits à partir d'une émission algérienne « Ligne rouge ». Les données se composent de 14 locuteurs avec 1443 énoncés qui sont des phrases complètes. 15 émotions étudiées dont cinq sont dominants : l'enthousiasme, l'admiration, la désapprobation, la neutralité et la joie.

3.2.2 Base de données émotionnelle du dialecte algérien

Dans notre travail, une base de données émotionnelle du dialecte algérien (Algerian Dialect Emotional Database : ADED) est créée. Nous avons décidé d'utiliser des films algériens pour collecter les données, car les émotions extraites à partir des films sont plus

réalistes que les émotions enregistrées en studio exprimées par des locuteurs lisant des phrases prédéfinies. Quatre états émotionnels de base sont considérés dans la base de données émotionnelle ADED : la colère, la peur, le neutre et la tristesse.

3.2.2.1 Acquisition de données

La procédure de création de notre base de données de parole émotionnelle est illustrée à la figure 3.2. Donc la procédure de création s'est déroulée en plusieurs étapes. Dans la première étape, six films en dialecte algérien sont sélectionnés, ces films décrivent la crise de la guerre civile ainsi que la période qui la suit. Les titres des films sélectionnés sont : Le repentant (التائب), Rachida (رشيدة), Bab El Oued cité (سيتي الواد باب), Les portes du soleil (الشمس أبواب), Fugitif (الهارب), Le prix du rêve (الحلم ثمن). Ces films contiennent un vocabulaire émotionnel riche. Et les acteurs participant aux films sont des acteurs célèbres. Dans la deuxième étape nous avons extrait des séquences (la séquence est une portion de film référant à un même contexte situationnel) des vidéos qui contiennent les émotions concernées (la colère, la peur, le neutre et la tristesse). Ensuite, nous avons extrait les informations audio (séquences d'audio) à partir les séquences des vidéos. Après, nous avons segmenté les séquences d'audio en parties plus petites de telle sorte que chaque segment couvre l'échantillon de parole d'un seul locuteur avec un contenu émotionnel homogène et sans aucun bruit de fond. Grâce à la procédure ci-dessus, un total de 260 segments de parole émotionnelle ont été collectées. Les nombres de segments pour chaque émotion sont présentés dans le tableau 3.1. Ces segments sont stockés sous forme des fichiers audio (48 kHz, wav audio data).

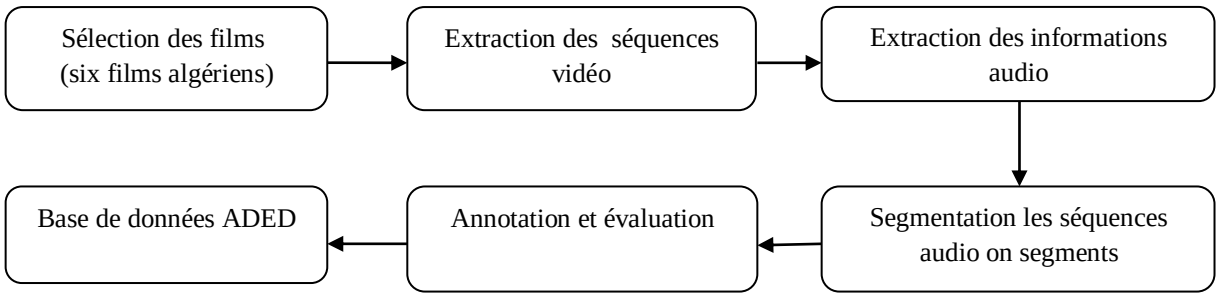


Figure 3.2- Schéma de la procédure de création de la base de données émotionnelle du dialecte algérien.

Tableau 3.1- Nombre de segments pour chaque émotion

Emotions	**Nombre des segments**
Peur	62
Neutre	68
colère	64
Tristesse	66
Total	260

3.2.2.2 Annotation et évaluation

L'annotation est nécessaire pour déterminer la véritable émotion exprimée dans les données vocales. La stratégie d'annotation adoptée se focalise sur la description des manifestations orales de l'émotion. L'unité d'annotation de l'émotion, c'est-à-dire l'intervalle temporel sur lequel l'émotion est décrite. Dans ce travail nous utilisons le segment comme unité d'annotation. Pour s'assurer la qualité des données vocales, un test de perception a été effectué. L'émotion de chaque segment a été évaluée dans un test d'auditeur par 3 annotateurs (auditeurs humains). Les annotateurs sont tous des locuteurs natifs de dialecte algérien sans déficience auditive ni problèmes psychologiques. Les annotateurs ont été invités à écouter les fichiers audio et à attribuer une étiquette d'émotion à partir d'une liste de termes (colère, peur, neutre, tristesse et inconnu) pour chaque segment. Le terme « inconnu » a été attribué lorsque plus d'une émotion est transmise à partir d'un segment ou que l'émotion sous-jacente ne faisait pas partie des états émotionnels spécifiés. Les annotateurs n'ont pris en compte que les informations audio. Les segments sont présentés dans un ordre aléatoire aux annotateurs dans un environnement calme et la dénomination de l'émotion par les annotateurs est réalisée sans restriction particulière. Les annotateurs peuvent écouter les segments autant de fois.

Les résultats de test de perception sont montrés dans le tableau 3.2. Un vote majoritaire dans notre travail représente au moins 2 sur 3 des annotateurs. Par exemple, si deux annotateurs ou plus ont attribué la même étiquette à un segment. Ce dernier dispose donc d'un vote majoritaire. Nous avons conservé les segments qui ont obtenu un vote majoritaire des annotateurs correspondant avec la base de données. Et nous avons supprimé les segments qui n'ont pas eu un vote majoritaire ainsi que ceux qui sont annotés par le terme "inconnu". Au total, environ 200 segments sur 260 ont été conservés dans notre base de données.

Tableau 3.2 - Résultats des tests de perception

Emotions	**Nombre des segments ayant un vote majoritaire**	**Nombre des segments n'ayant pas un vote majoritaire**	**Nombre des segments annoté par « Inconnu »**	**Total**
Peur	52	7	3	62
Colère	52	10	6	68
Neutre	48	11	5	64
Tristesse	48	10	8	66
Total	200	38	22	260

Les confusions entre les émotions sont montrées dans le tableau 3.3, dans lequel les émotions en verticale sont les émotions que les acteurs prononcés et les émotions en horizontale sont celles perçues par les annotateurs. Les taux moyens de perception par les annotateurs sont montrés sur le tableau 3.3. Les émotions ont été correctement perçues dans 82.14% des cas. La peur et la colère sont légèrement confondues aussi bien que le neutre et la tristesse.

Tableau 3.3 - Matrice de confusion entre les émotions pour tous les locuteurs et annotateurs

	Taux de perception %				
Emotions	Peur	Colère	Neutre	Tristesse	Inconnu
Peur	**87.63%**	3.23%	2.15%	2.15%	4.84%
Colère	5.88%	**81.37%**	1.96%	1.96%	8.82%
Neutre	2.08%	0%	**81.77%**	6.25%	8.85%
Tristesse	3.53%	1.51%	6.06%	**77.77%**	11.61%

Finalement, la base de données qui a été créée composée de 200 fichiers audio d'une durée allant de 0,5 s à 3 s. Les nombres de segments pour chaque émotion sont indiqués dans le tableau 3.4. Les fichiers audio inclus dans cette base de données sont exprimés par 32 acteurs (16 hommes et 16 femmes) d'âges différents entre 18 et 60 ans. La répartition des segments sur les locuteurs montre la diversité de la base de données comme illustre dans le tableau 3.5. Les enregistrements ont été pris avec une fréquence d'échantillonnage de 48 kHz.

Tableau 3.4 - Nombres de segments pour chaque émotion

	Nombre de segments		
Emotions	Homme	Femme	**Total**
Peur	19	33	52
Neutre	26	22	48
Colère	24	28	52
Tristesse	14	34	48
Total	**87**	**113**	**200**

Tableau 3.5 - Répartition des segments sur les locuteurs

Nombre de segments	**Nombre de locuteurs**
>10	7
6 à 10	6
2 à 5	14
1	5

Pour connaître le contenu de la base de données émotionnelle du dialecte algérien (ADED), quelques phrases en dialecte algérien appartiennent à la base de données ADED sont présentées dans le tableau 3.6. La prononciation des phrases représentées et leurs équivalents en arabe standard et en français sont également montrés. D'après le tableau 3.6, on remarque qu'il existe des mots en arabe standard (سمعتني, ركبنا, تمنيت), des mots en dialecte natif (علاش, مانيش, أغصب) et quelques mots en langue française (طوموبيل, ليلاتخ).

Tableau 3.6 - Quelques phrases du dialecte algérien appartiennent à ADED

Emotions	Phrases en langue française	Phrases en arabe standard	Phrases en dialecte algérien	Prononciation des phrases
Peur	N'aie pas peur!	لا تخافي!	ما تخافيش!	Ma tkhafich!
	Plus vite! Plus vite!	أسرع أسرع !	أغصب أغصب!	Aghssab! aghssab!
	Non non!	لا لا!	لا لا!	Lala!
Colère	Ça suffit Djamila!	يكفي يا جميلة!	خلاص جميلة!	Khlass Djamila!
	Maintenant tu m'as entendu!	الآن سمعتني!	دركا سمعتني!	Dorka smaattni!
	Je ne suis pas ton frère!	لست أخاك!	مانيش خوك!	Manich khouk.
Tristesse	J'ai souhaité la mort.	تمنيت الموت.	تمنيت الموت.	Tmaniit el mout.
	Pourquoi? Pourquoi?	لماذا ؟ لماذا ؟	علاش ؟ علاش؟	Alach? alach?
	Quand nous sommes montés dans la voiture.	عندما ركبنا في السيارة.	كي ركبنا في طوموبيل.	Ki rkabna fi tomobil.
Neutre	Vous ne pouvez pas venir avec nous.	لا تستطيع المجيء معنا.	ماتطيقش تجي معانا.	Mattikch tji maana.
	Ne me laisse pas dormir.	لا يتركني أنام.	ما يخليني نرقد.	Ma ykhalinich nargod
	Messages de menaces	رسائل التهديد.	ليلاتخ دو مونا ص.	.Lilatkh de menace .

3.3 Analyse acoustique des émotions

Une étape importante dans la conception d'un système de REP est l'extraction des caractéristiques appropriées qui caractérisent efficacement les différentes émotions [180]. Nous présentons dans cette section les paramètres acoustiques que nous avons choisis afin de caractériser les émotions cibles de notre application, c'est à dire. les émotions de type peur, colère, neutre et tristesse. Une liste de paramètres pertinents pour la caractérisation du contenu émotionnel est sélectionnée et extraite à partir de la base de données émotionnelle du dialecte algérien. Une analyse est faite pour étudier l'influence des émotions traitées sur les paramètres que nous avons choisis.

3.3.1 Extraction des paramètres

L'extraction des paramètres vocaux est une étape nécessaire pour toutes les applications de traitement de la parole. Nous avons choisi pour caractériser les états émotionnels des descripteurs acoustiques parmi les trois familles suivantes : les descripteurs prosodiques, les descripteurs de qualité de la voix et les descripteurs spectraux. Ces paramètres sont les valeurs statistiques de pitch (Mean, Max, Min et Range) et les paramètres similaires d'intensité, le taux des trames non voisées(UFR), le jitter, le shimmer, le HNR, les formants (formant 1 et formant 2) et les paramètres MFCC. Mean, Max et le Min correspondant à la valeur moyenne, la valeur maximale et la valeur minimale respectivement du paramètre dans un segment. Range correspondant à la gamme du paramètre pour un segment (Max-Min). Les paramètres

sont extraits par le logiciel PRAAT [181], sauf les paramètres des MFCC sont extraits par MATLAB.

PRAAT est un logiciel scientifique open source pour analyser le signal de la parole. Il est très connu dans le domaine du traitement de la parole par sa simplicité dans l'utilisation, sa rapidité dans les calculs et sa flexibilité dans les modes de fonction. PRAAT permet d'afficher la forme d'onde de la parole et un spectrogramme à large bande montrant l'énergie spectrale du son en fonction du temps. De plus, les contours de pitch, les formants, et l'intensité peuvent également être visualisés.

PRAAT nous permet d'enregistrer un son avec un microphone ou tout autre périphérique d'entrée audio, ou lire un son à partir d'un fichier. La Figure 3.3 montre une fenêtre de son de PRAAT. La moitié supérieure de la fenêtre montre une représentation visible du son (la forme d'onde). La moitié inférieure montre plusieurs analyses: les spectrogramme (une représentation de la quantité de hautes et basses fréquences dans le signal) est peint dans les tons en gris. Le contour de pitch (fréquence fondamentale) est dessiné en courbe bleue. Le contour d'intensité est montré en courbe verte et les contours des formants sont représentés par des points rouges.

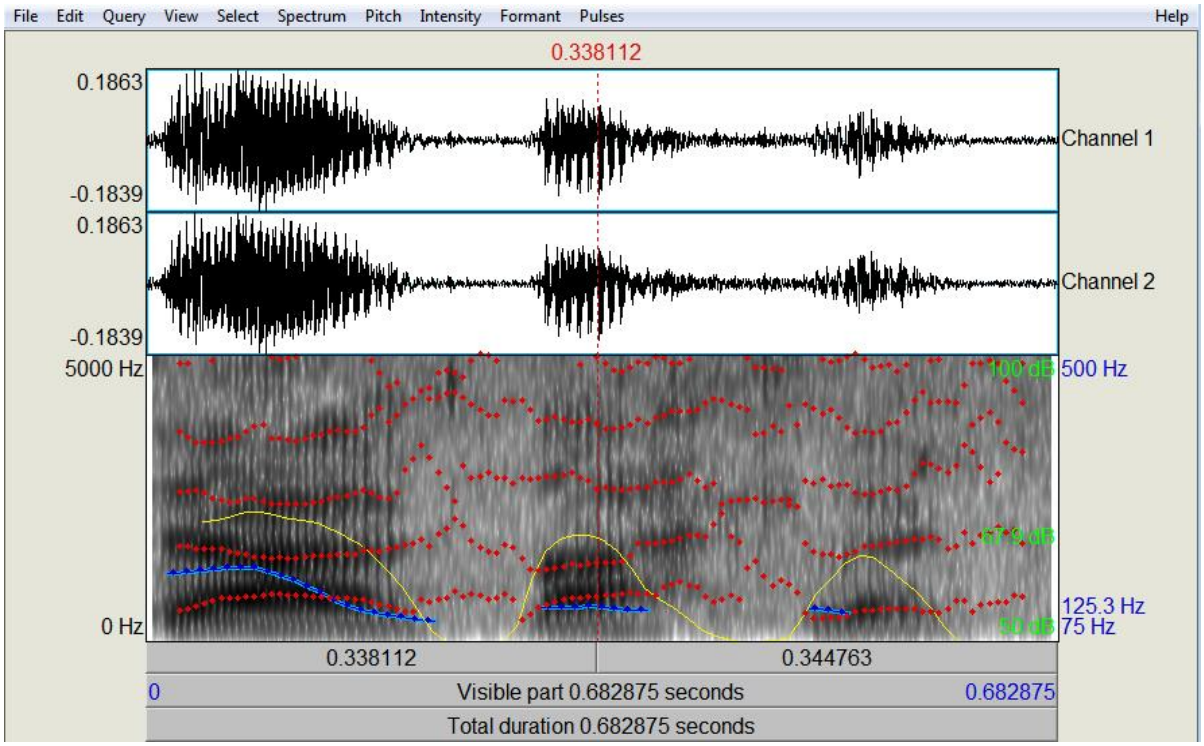


Figure 3.3- Fenêtre de son de PRAAT

Les moyennes des valeurs statistiques des paramètres extraits par le logiciel PRAAT pour les quatre émotions : peur, colère, neutre et tristesse (sans distinction de sexe) sont présentées dans le tableau 3.7. Pour analyser l'influence des émotions précédentes sur les caractéristiques extraites en fonction du sexe (masculin et féminin), les valeurs moyennes des caractéristiques extraites de chaque émotion en fonction du sexe sont montrées dans le tableau 3.8.

Tableau 3.7 - Moyennes des valeurs statistiques des paramètres extraits avec le logiciel PRAAT

Paramètres	**Peur**	**Colère**	**Neutre**	**Tristesse**
Mean de pitch (Hz)	**279.79**	269.09	190.60	196.89
Max de pitch (Hz)	**398.84**	387.36	273.25	299.61
Min de pitch (Hz)	163.18	131.87	**120.62**	125.09
Range de pitch (Hz)	235.65	**255.50**	152.62	174.52
Mean d'intensité (dB)	69.97	**71.29**	70.30	62.69
Max d'intensité (dB)	75.16	**77.35**	76.98	70.09
Min d'intensité (dB)	57.20	55.45	51.11	**42.90**
Range d'intensité (dB)	17.96	21.91	25.86	**27.18**
UFR (%)	24.12	23.92	24.36	**32.82**
Jitter (%)	3.29	3.19	2.89	**3.61**
Shimmer (%)	**16.03**	15.40	13.40	15.28
HNR (dB)	7.10	7.23	**8.55**	7.59
Formant1(Hz)	**715.89**	665.40	626.35	655.81
Formant2(Hz)	**1814.45**	1763.86	1755.50	1807.67

Tableau 3.8 - Moyennes des valeurs des paramètres extraits avec le logiciel PRAAT en fonction du sexe

Sexe	**Masculin**				**Féminin**			
Paramètres	**Peur**	**Colère**	**Neutre**	**Tristesse**	**Peur**	**Colère**	**Neutre**	**Tristesse**
Mean de pitch (Hz)	219.49	**226.36**	157.18	113.20	**314.51**	305.66	230.10	231.35
Max de pitch (Hz)	313.44	**334.56**	230.17	207.71	**448.02**	438.89	324.16	337.45
Min de pitch (Hz)	139.33	114.37	99.22	**81.36**	176.92	**140.24**	145.91	143.09
Range de pitch (Hz)	174.11	**219.79**	130.95	126.35	271.10	**298.65**	178.24	194.36
Mean d'intensité (dB)	69.24	**71.88**	69.64	62.37	70.38	70.77	**71.07**	63.21
Max d'intensité (dB)	74.45	**77.93**	76.00	69.40	75.57	76.87	**78.13**	70.37
Min d'intensité (dB)	56.40	57.24	52.87	**45.95**	57.65	53.91	49.03	**41.65**
Range d'intensité (dB)	18.05	20.69	23.12	**23.45**	17.92	22.95	**29.11**	28.72
UFR (%)	30.67	22.86	25.38	**33.35**	20.35	24.84	23.15	**32.53**
Jitter (%)	**3.93**	3.50	2.92	3.82	2.92	2.92	2.86	**3.53**
Shimmer (%)	**18.09**	16.83	14.24	17.55	**14.85**	14.17	12.40	14.35
HNR (dB)	6.19	6.12	**7.41**	5.78	7.63	8.19	**9.91**	8.34
Formant1(Hz)	742.4	664.7	**606.1**	642.1	700.7	666.0	**650.25**	661.5
Formant2(Hz)	1805.6	1682	**1645.1**	1698.7	1814.5	1833.6	**1885.9**	1852.5

3.3.2 Analyse des paramètres acoustiques

Dans cette section, nous analysons les paramètres acoustiques choisis pour détecter l'influence des émotions ciblées sur ces paramètres. Les états émotionnels sont comparés en fonction des différents paramètres extraits.

3.3.2.1 Analyse du pitch (fréquence fondamentale)

Le pitch est un paramètre acoustique fondamental lors du processus d'analyse de la parole. Il a des informations sur l'émotion, car il dépend de la tension des cordes vocales. Ce paramètre est fortement exploité dans le domaine de REP. D'après le tableau 3.7 et la figure

3.4, nous constatons que les valeurs statistiques de pitch (Mean, Max, Min et Range) sont différentes entre les quatre émotions traitées. Nous observons que les émotions de peur et de colère ont des valeurs plus élevées de Mean et de Max de pitch, tandis que l'état neutre a une valeur basse de pitch. Les états de tristesse et de neutre sont associés à des valeurs basses de Range de pitch. Nous pouvons regrouper les quatre émotions en deux groupes : les émotions « hautes » et les émotions « basses » en se basant sur la valeur de pitch. La peur et la colère appartiennent aux émotions hautes car elles donnent des grandes valeurs de pitch. Au contraire, le neutre et la tristesse se trouvent dans la zone des émotions basses dont les valeurs de pitch sont assez faibles. À partir du tableau 3.8, nous remarquons que la moyenne des valeurs statistiques de pitch (Mean, Max, Min et Range) chez le sexe féminin est élevée en comparant avec les valeurs statistiques de pitch chez le sexe masculin dans toutes les quatre émotions étudiées.

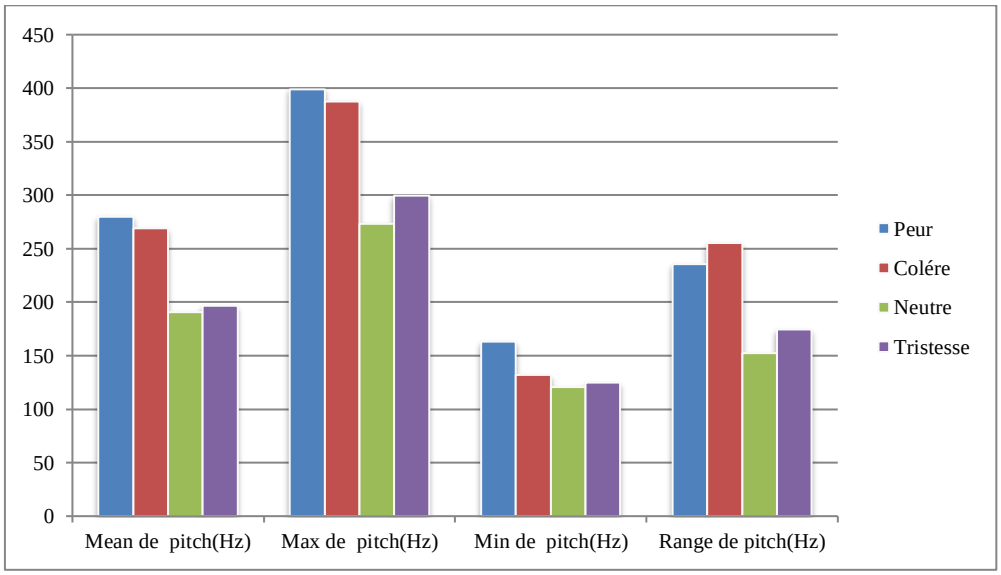
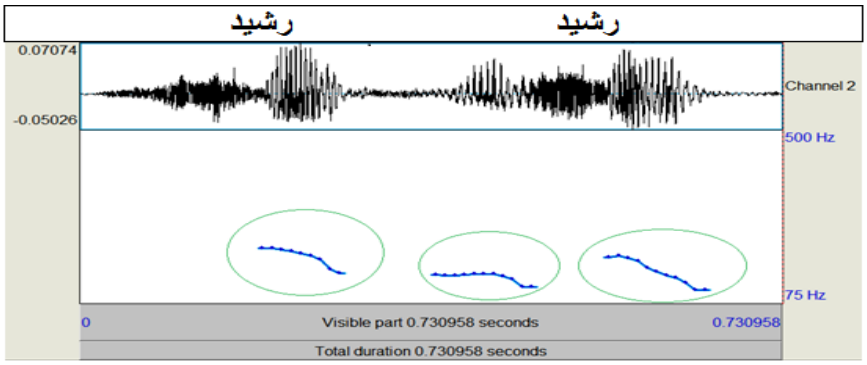
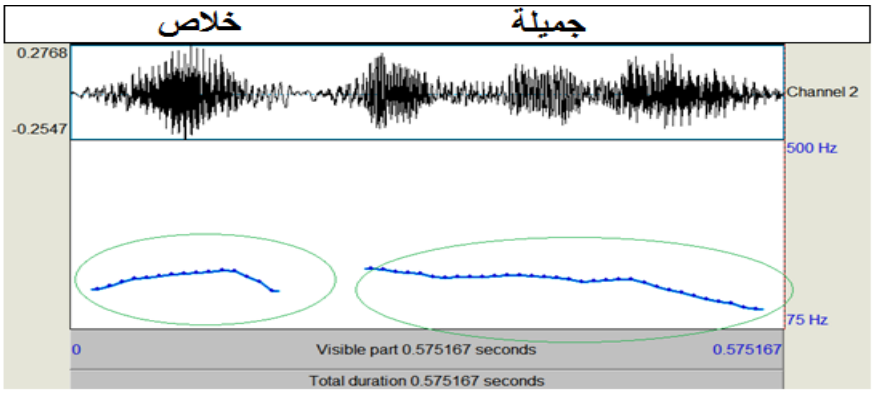


Figure 3.4 - Comparaison entre les valeurs statistiques de pitch dans chaque émotion

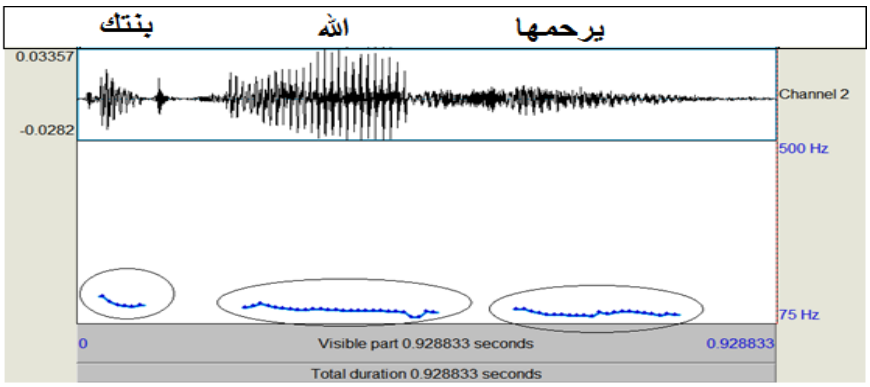
Les figures 3.5, 3.6, 3.7 et 3.8 illustrent des segments de parole avec les contours de pitch correspondant la peur, la colère, la tristesse et le neutre respectivement. Les figures sont obtenues à l'aide du logiciel PRAAT. Les contours de pitch sont montrés par la ligne bleue. On remarque que le discours dans les états peur et colère montre plus de montée et de baisse de contour de pitch par rapport à un discours neutre. Nous remarquons aussi que le contour de pitch varie de manière visible lors du dernier mot dans les états émotionnels de peur et de colère.



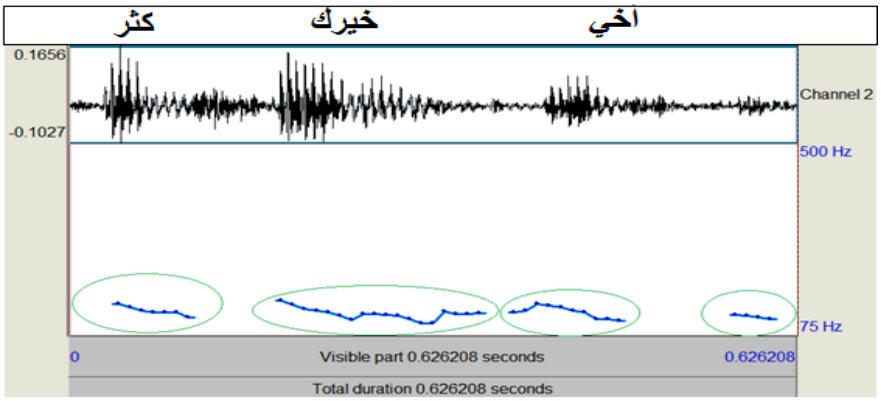
Figures 3.5- Contour de pitch concernant l'émotion de peur



Figures 3.6- Contour de pitch concernant l'émotion de colère



Figures 3.7 - Segment de parole concernant l'émotion de tristesse



Figures 3.8 - Segment de parole concernant l'état neutre

3.3.2.2 Analyse d'intensité

L'intensité est un paramètre acoustique permet de fournir une mesure de la force sonore de la voix. Elle est généralement représentée en décibels. L'intensité est considérée comme l'une des caractéristiques distinctives des émotions dans les systèmes de REP. Il est remarqué dans le tableau 3.7 et la figure 3.9 que l'émotion de tristesse a une faible valeur d'intensité et une large gamme (Range d'intensité) par rapport aux autres états émotionnels. L'émotion de colère a des valeurs élevées d'intensité. Les moyennes des valeurs statistiques d'intensité en fonction du sexe sont montrées dans le tableau 3.8. Nous observons la même remarque précédente c'est-à-dire l'état de tristesse a des valeurs faibles d'intensité (Min d'intensité et Mean d'intensité) dans les deux sexes. On remarque aussi que l'émotion de colère a des valeurs élevées d'intensité (Mean d'intensité et Max d'intensité) chez le sexe masculin. Et chez le sexe féminin l'état neutre a des valeurs élevées d'intensité (Mean d'intensité et Max d'intensité).

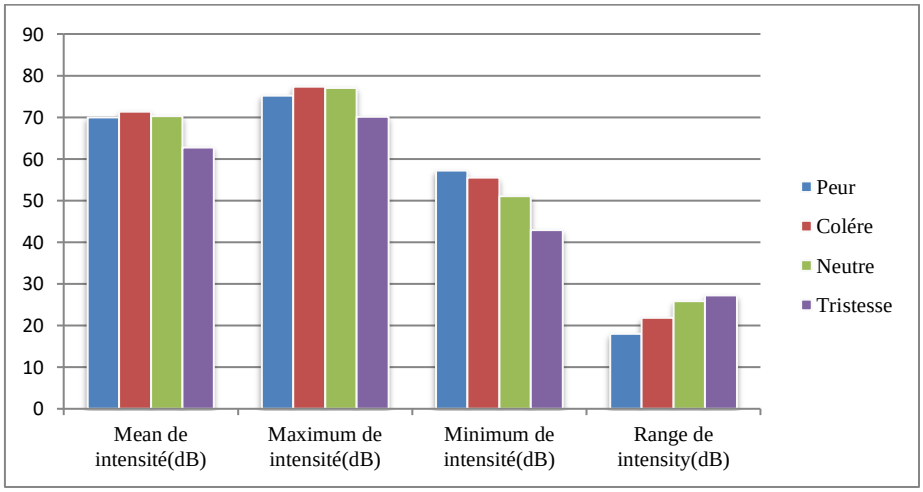
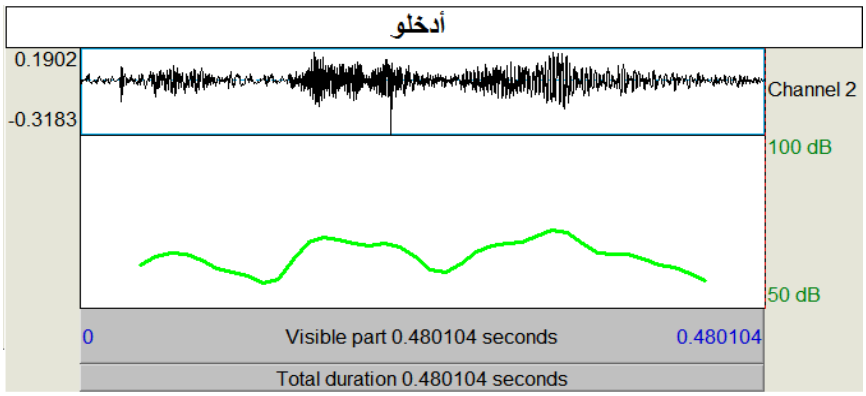
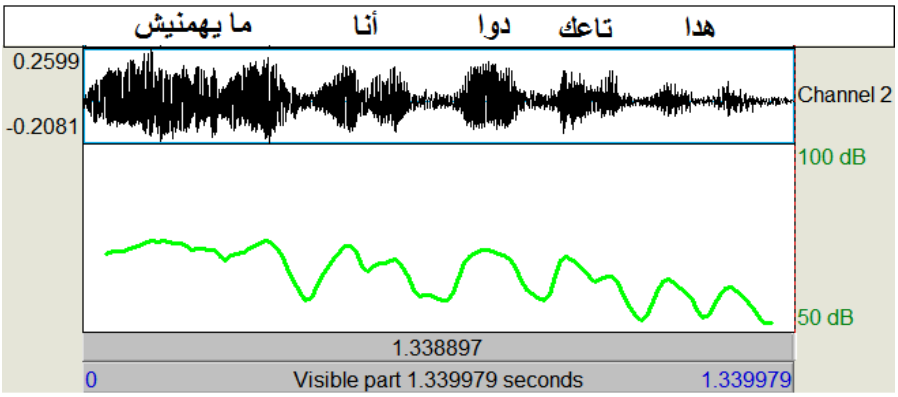


Figure 3.9 - Comparaison entre les valeurs statistiques d'intensité dans chaque émotion

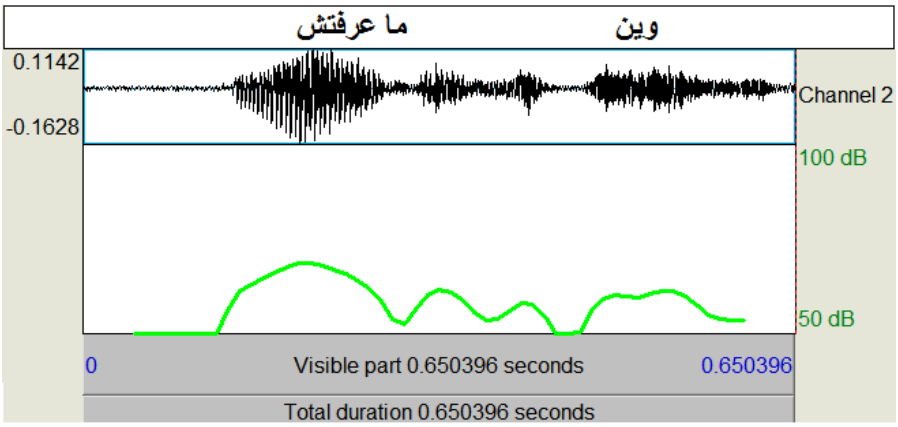
Les figures 3.10, 3.11, 3.12 et 3.13 illustrent des segments de parole avec les contours d'intensité correspondant concernant la peur, la colère, la tristesse et le neutre respectivement. Les figures sont obtenues à l'aide du logiciel de PRAAT. Les contours d'intensité sont montrés par la ligne verte. Nous remarquons que le contour d'intensité de colère contient plus de variation par rapport aux contours des autres états émotionnels.



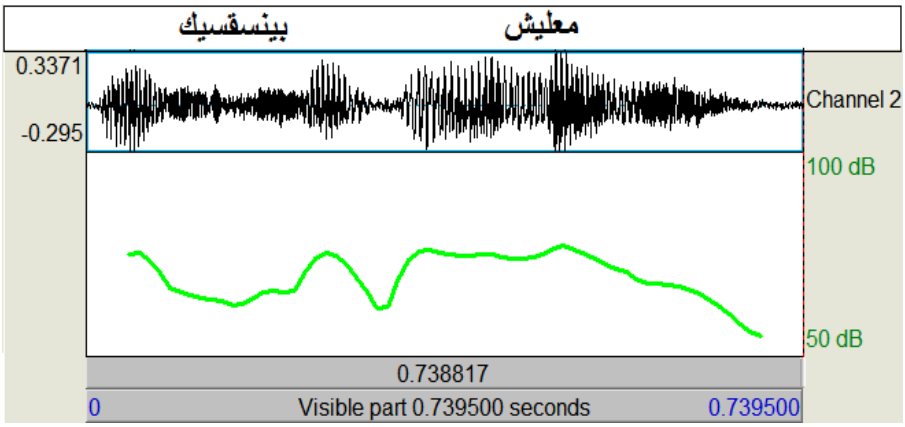
Figures 3.10 - Contour d'intensité concernant l'émotion de peur



Figures 3.11 - Contour d'intensité concernant l'émotion de colère



Figures 3.12 - Contour d'intensité concernant l'émotion de tristesse



Figures 3.13 - Contour d'intensité concernant l'émotion de neutre

3.3.2.3 Analyse des trames non voisées (UFR)

Le signal de la parole peut être divisé en trames voisées et non voisées. Le UFR dans un segment de la parole est révélateur de la quantité des pauses sur ce segment. Ainsi, une phrase prononcée à un rythme normal contiendra beaucoup plus de pauses (donc plus de trames non voisées) qu'une phrase prononcée avec un rythme élevé. Le UFR a été utilisé dans de nombreux systèmes de REP. D'après le tableau 3.7 et la figure 3.14, nous constatons que le nombre de trames non voisées diffère entre les émotions traitées. L'émotion de tristesse a le grand nombre de trames non voisées bien que l'émotion de colère a la valeur la plus faible. Et

le nombre de trames non voisées est modéré dans les autres états émotionnels. Les mêmes remarques sont observées dans le tableau 3.8 concernant le sexe masculin mais pour le sexe féminin, l'état de peur a le nombre le plus faible des de trames non voisées.

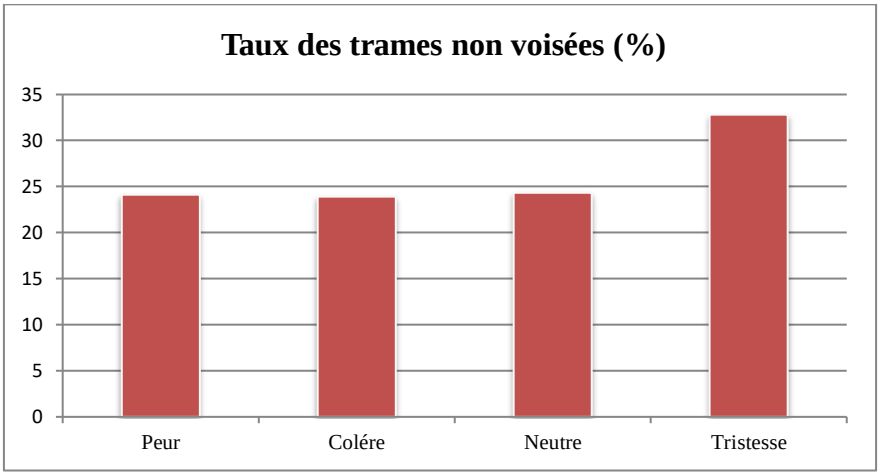


Figure 3.14 - Comparaison entre les moyennes des valeurs des taux de trames non voisées dans chaque émotion

3.3.2.4 Analyse des paramètres de jitter, shimmer et HNR

Jitter, shimmer et HNR ont été largement utilisés pour améliorer et développer les systèmes de REP . Nous remarquons dans le tableau 3.7 et la figure 3.15 que les valeurs de jitter, shimmer et HNR diffèrent entre les différents types d'émotions. La valeur de jitter est légèrement plus élevée dans l'état de tristesse, l'émotion de peur a la valeur la plus élevée du shimmer et la valeur HNR est faible dans les émotions de peur, de colère et de tristesse par rapport à l'état neutre. Concernant l'influence de sexe, nous observons dans le tableau 8 que la valeur de jitter dans l'état de peur est légèrement plus élevée par rapport aux autres états chez le sexe masculin. Mais chez le sexe féminin, l'émotion de tristesse a la valeur la plus grande de jitter. L'émotion de peur a la plus grande valeur de shimmer dans les deux sexes. La valeur de HNR est élevée dans l'état neutre chez les deux sexes.

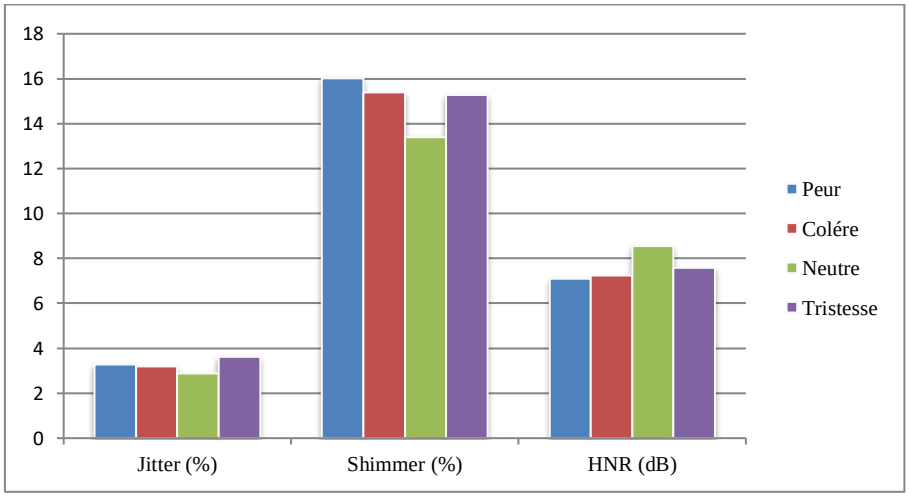


Figure 3.15 - Comparaison entre les moyennes des valeurs des jitter, shimmer et HNR dans chaque émotion

3.3.2.5 Analyse des paramètres formantiques

Les formants ont été utilisés comme descripteurs des émotions dans la parole et ils ont été fortement exploités dans les systèmes de REP.

En théorie, il existe une infinité de formants mais les quatre premiers formants sont largement exploités dans la pratique. Les formants sont numérotés à partir de la fréquence la plus basse vers le haut (F1, F2, F3, etc.). L'émotion a une influence considérable sur le positionnement des formants, en particulier sur le placement des deux premiers formants [182]. Les moyennes des valeurs des deux premiers formants (formant 1 et formant 2) de chaque émotion sont illustrées dans le tableau 3.7. Il est noté dans le tableau 3.7 et la figure 3.16 que les valeurs de formant 1 et de formant 2 sont plus élevées dans l'émotion de peur que les autres émotions, et l'état neutre a les valeurs les plus basses des formants. La même remarque est constatée dans le tableau 3.8 concernant le sexe masculin mais pour le sexe féminin, l'état de neutre a une moyenne des valeurs de formant 2 plus élevée par rapport aux autres états émotionnels.

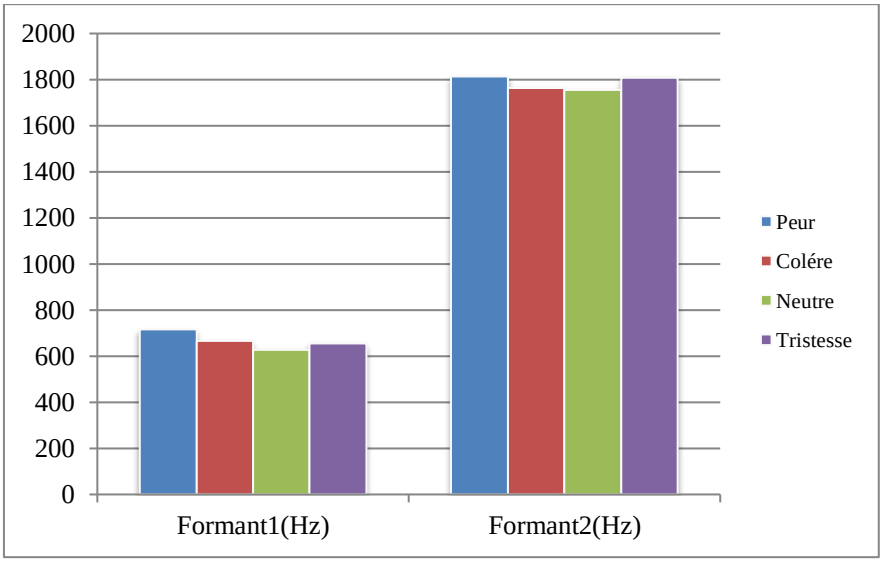
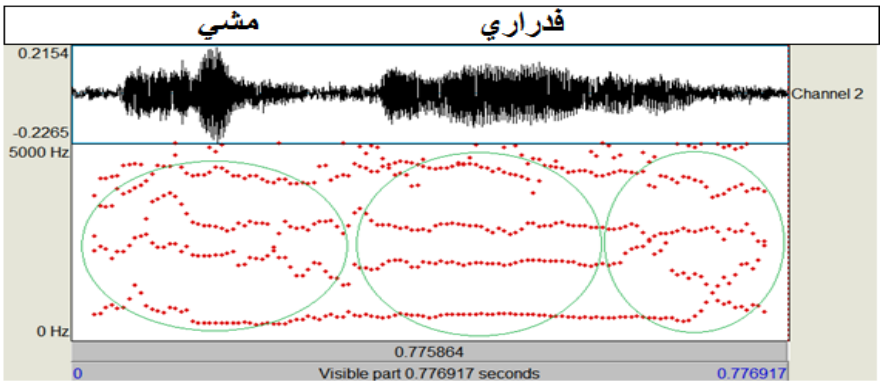
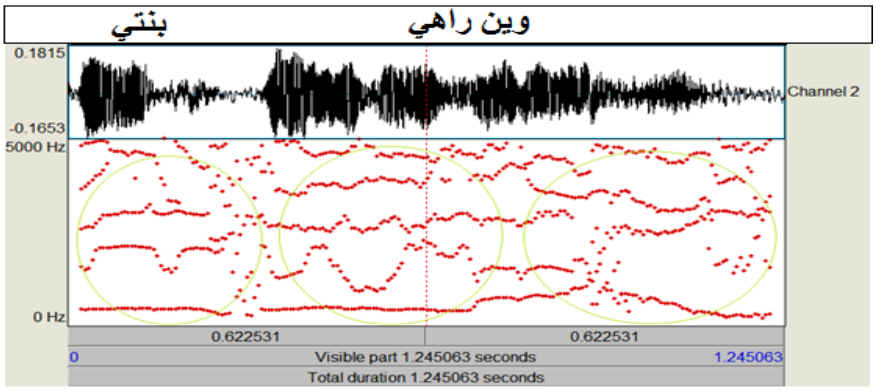


Figure 3.16 - Comparaison entre les moyennes des valeurs des formants dans chaque émotion

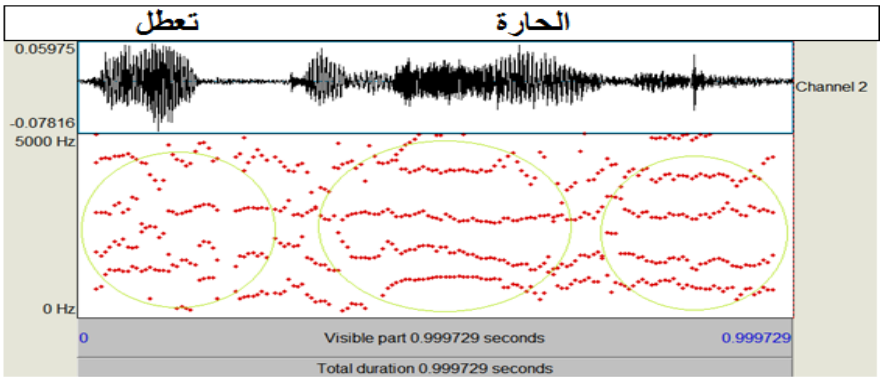
Les figures 3.17, 3.18, 3.19 et 3.20 illustrent des segments de parole avec les contours des formants correspondants concernant la peur, la colère, la tristesse et le neutre respectivement. Nous obvervons que le contour des formants varie selon les émotions, comme le montrent les figures 3.17, 3.18, 3.19 et 3.20, le contour des formants dans l'état de colère a la puissance la plus élevée, tandis que nous avons des faibles puissances spectrales dans les états de peur et neutre.



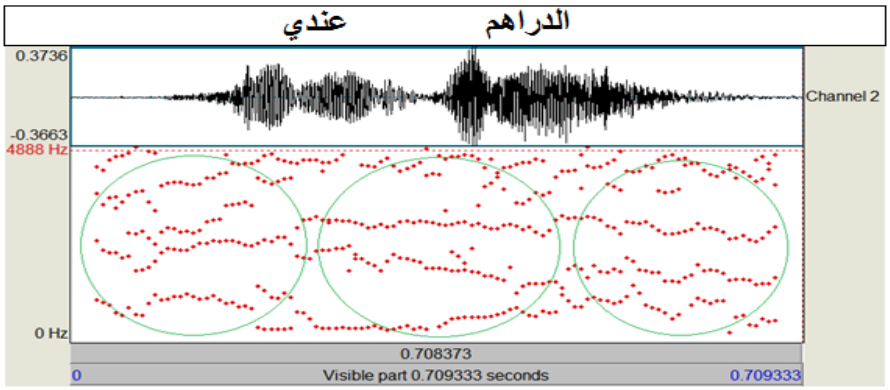
Figures 3.17 - Contour des formants concernant l'émotion de peur



Figures 3.18 - Contour des formants concernant l'émotion de colère



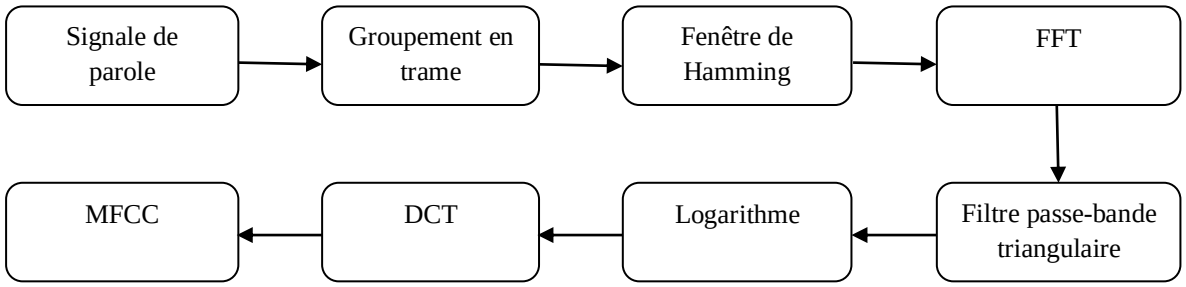
Figures 3.19 - Contour des formants concernant l'émotion de tristesse



Figures 3.20 - Contour des formants concernant l'émotion de neutre

3.3.2.6 Analyse des paramètres MFCC

Les paramètres des MFCC ont été fortement exploités pour développer et améliorer les performances des systèmes de REP. Ces coefficients sont dérivés du spectre de puissance en appliquant un banc de filtres uniformément espacés sur une échelle fréquentielle modifiée, appelée échelle de Mel. La figure 3.21 montre la procédure de calcul des paramètres MFCC, cette procédure comprend les blocs suivants : groupement en trame (*Frame blocking*), fenêtre de Hamming (*Hamming window*), transformée de Fourier rapide (*Fast Fourier Transform (FFT)*), filtre passe-bande triangulaire (*Triangular band-pass filter*), logarithme, transformation cosinus discrète (*Discrete Cosine Transformation (DCT)*) [183]:



Figures 3.21 - Schéma de procédure de calcul des paramètres MFCC

- Groupement en trame (*Frame blocking*) : le signal de la parole est segmenté en trames. La longueur de chaque trame correspondant environ de 20 à 40 ms de parole car la variation du signal de parole dans cette gamme est négligeable.

- Fenêtre de Hamming (*Hamming window*) : le découpage du signal en trames produit des discontinuités aux frontières des trames. Pour réduire ces problèmes, des fenêtres de pondération sont appliquées. Parmi les fenêtres les plus courantes, nous pouvons citer la fenêtre de Hamming [184] :

$$w[k] = \begin{cases} 0.54 + 0.46\cos\left(\frac{2\pi k}{N-1}\right) & si \quad 0 \le k \le N-1 \\ 0 & sinon \end{cases} \qquad (3.1)$$

- Transformée de Fourier Rapide (*Fast Fourier Transform (FFT)*) : FFT transforme le signal du domaine temporel au domaine fréquentiel. La FFT est un algorithme rapide pour le calcul de la transformée de Fourier discret (*Discrete Fourier Transform (DFT)*) et est définie par la formule (3.2). Les valeurs obtenues sont appelées le spectre.

$$X[k] = \sum_{n=0}^{N-1} x_a[n] e^{-j2\pi nk/N} \quad , \quad 0 \le k \le N \qquad (3.2)$$

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

- Filtre passe-bande triangulaire (*Triangular band-pass filter*) : l'énergie du spectre est calculée à travers un banc de filtres numériques couvrant la bande passante, ce qui permet de ne conserver qu'un sous ensemble de ces paramètres. Les filtres triangulaires sont les plus utilisés. Ils sont préférés pour leur simplicité et leur effet de lissage sur le spectre. Ces filtres sont les plus souvent répartis sur l'échelle Mel qui est non linéaire. Le but d'utilisé banc Mel est de simuler les filtres des bandes critiques du mécanisme d'audition. La relation entre la fréquence en échelle Hertz et sa correspondance en Mels est la suivante :

$$Mel(f) = x.\log\,(1 + \frac{f_{Hz}}{y}) \tag{3.3}$$

  Où f est la fréquence, x = 2595 et y = 700.

- Logarithme : dans cette étape, le logarithme est appliqué sur la sortie du filtre Mel.
- Transformation cosinus discrète (*Discrete Cosine Transformation (DCT)*) : dans cette dernière étape, le signal est reconverti dans le domaine temporel du domaine fréquentiel en utilisant la transformation cosinus discrète. L'inverse de la transformée de Fourier rapide (*Inverse Fast Fourier Transform (IFFT)*) peut également être utilisé pour cette conversion mais DCT est préférée en raison de son efficacité.

Dans ce travail la longueur de chaque trame utilisée est d'environ 20 ms. Et la fréquence d'échantillonnage utilisée égale 8 kHz. La figure 3.22 illustre des formes des MFCC de quatre émotions : peur, colère, tristesse et neutre. Cette figure est obtenue par MATLAB. Dans ce travail, 20 MFCC sont utilisés dans le système de reconnaissance. On remarque sur la figure 3.22 que les formes de MFCC de chaque émotion varient selon l'émotion. Cette différence qui existe entre les formes des MFCC nous permet d'utiliser les paramètres des MFCC dans le système de REP.

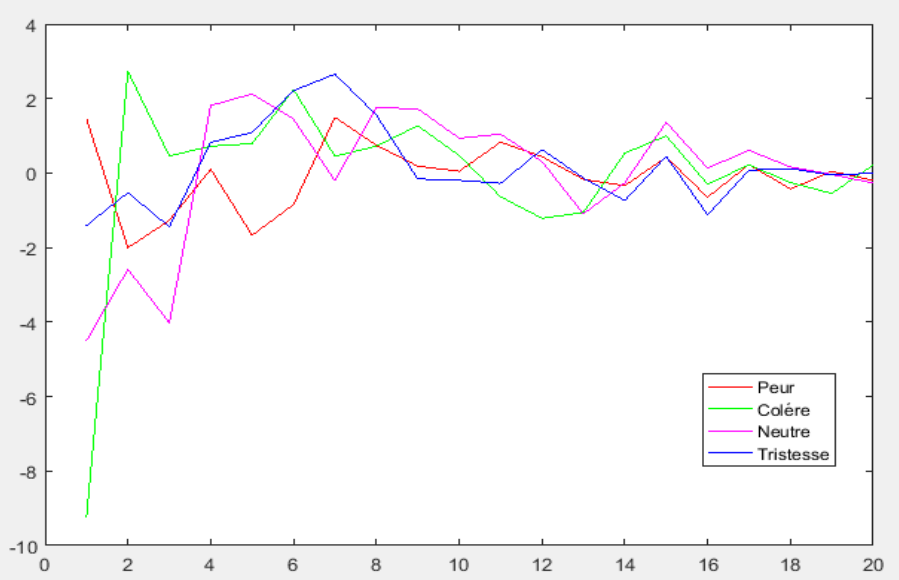


Figure 3.22 - Formes des MFCC de quatre émotions

3.4 Conclusion

Dans ce chapitre, nous avons décrit la procédure de création de notre base de données ADED qui comprend l'acquisition des données, l'annotation et la validation des émotions. La base de données qui a été créée est composée de 200 fichiers audio de quatre états émotionnels. Ensuite, nous avons décrit aussi les paramètres acoustiques exploités dans notre travail tel que les paramètres prosodiques, les paramètres de qualité de la voix et les paramètres spectraux. Après nous avons analysées les paramètres acoustiques choisis pour détecter l'influence des émotions ciblées sur ces paramètres.

Chapitre 4 :

Expériences et résultats obtenus à partir des différentes classifications

4.1 Introduction

Dans le chapitre précédent, nous avons étudié la phase d'extraction des paramètres à partir les segments de la base de données émotionnelle du dialecte algérien (ADED). Dans le présent chapitre, nous nous intéressons au développement des systèmes de REP dans le dialecte algérien, expérimenté sur la base de données ADED. Ce travail est basé sur la reconnaissance de quatre émotions de base : la peur, la colère, la tristesse et le neutre. Plusieurs expériences sont effectuées pour étudier l'influence des différents facteurs tels que les paramètres acoustiques, le nombre des émotions, le sexe et la langue sur les systèmes de reconnaissance. Ces systèmes sont basés sur des techniques de classification. Les techniques de classifications exploitées dans nos expériences sont présentées brièvement dans la deuxième section. Ensuite, la performance des paramètres acoustiques sur le système de REP dans le dialecte algérien est étudiée dans la troisième section. Dans la quatrième section, l'influence du nombre d'émotions incluse dans le système de reconnaissance est étudiée. Après, l'influence du sexe sur le système de reconnaissance est étudiée dans la cinquième section. Finalement, la performance des paramètres acoustiques sur le système de reconnaissance des émotions de colère et de neutre dans différentes langues est étudiée dans la cinquième section.

4.2 Classification dans le système de REP

Les systèmes de REP reposent sur des méthodes de classification, car elles sont basées sur une procédure d'apprentissage capable à partir d'une quantité de données suffisantes, de caractériser les propriétés acoustiques de chaque classe d'émotion. La phase de classification est la dernière étape de REP. Les paramètres extraits de l'étape d'extraction sont données en entrée du classificateur pour détecter la classe d'émotion. Donc le but de classificateur est de classer de manière optimale l'état émotionnel d'un échantillon de parole. Dans la littérature plusieurs techniques de classification sont utilisées dans les systèmes de REP. Les systèmes de reconnaissance effectuée dans ce travail sont basée sur les méthodes de classification KNN et SVM. La description de fonctionnement de ces deux méthodes est décrite dans cette section.

4.2.1 K-plus-Proches-Voisins (KNN)

KNN (*K Nearest Neighbors*) est le plus simple algorithme de tous les algorithmes d'apprentissage automatique. La méthode des k plus proches voisins est utilisée pour la classification et la régression. Le principe de la méthode KNN est de trouver les k plus proches voisins, à partir de l'échantillon d'apprentissage, à une nouvelle instance qu'on cherche à classer. La classe de la nouvelle instance est la classe majoritaire (la plus représentée) parmi ces k voisins. Dans le cas d'une régression, la valeur de sortie est une valeur continue qui peut être, par exemple, la moyenne des valeurs des k voisins. KNN est une méthode supervisée. Elle a été utilisée dans l'estimation statistique et la reconnaissance des modèles comme une technique non paramétrique, cela signifie qu'elle ne fait aucune hypothèse sur la distribution des données [185].

La méthode de KNN utilise principalement deux paramètres : une fonction de similarité pour comparer les individus dans l'espace de caractéristiques et le nombre k qui décide combien de voisins influencent la classification [186]. La méthode de KNN [187] se base sur une comparaison directe entre le vecteur caractéristique de l'instance à classer et les vecteurs des instances de la base d'apprentissage. La comparaison consiste en un calcul de distances entre ces instances. Puis à l'instance à classer est assignée la classe majoritaire parmi les classes des k instances les plus proches.

Pour trouver la classe d'un exemple donné *x*, l'algorithme cherche les k plus proches voisins de ce nouveau cas et prédit la réponse la plus fréquente de ces k plus proches voisins. Le principe de décision consiste tout simplement donc à calculer la distance de l'exemple inconnu *x* à tous les échantillons fournis. L'exemple est alors affecté à la classe majoritairement représentée parmi ces k échantillons. La méthode utilise deux paramètres : le nombre k et les fonctions de similarité pour comparer le nouvel exemple aux exemples déjà classés. La fonction de similarité elle permet de mesurer le degré de différence entre deux vecteurs. Il existe plusieurs fonctions pour calculer la distance entre deux voisins, notamment, la distance euclidienne, la distance de Manhattan, la distance de Minkowski [188]:

- La distance Euclidienne :

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (4.1)$$

Où : *x*, *y* sont des vecteurs.

- La distance de Minkowsky:

$$d(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p} \quad (4.2)$$

Où : *x*, *y* sont des vecteurs.
p : paramètre.

- La distance de Manhattan :

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (4.3)$$

Où : *x*, *y* sont des vecteurs.

La distance utilisée dans notre travail est la distance euclidienne.

La méthode de KNN est très utilisée dans le domaine de REP. Cette méthode de classification a donné des meilleurs résultats par rapport autres méthodes de classification dans différents systèmes de la REP [148][189].

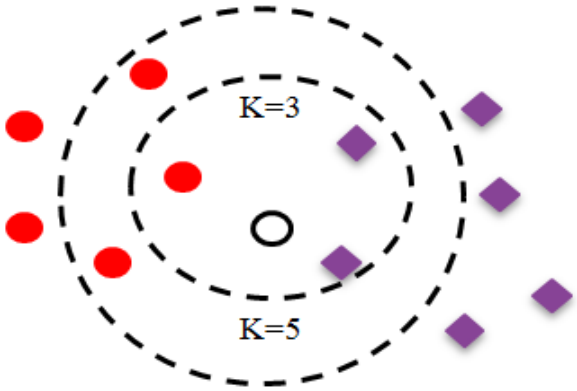


Figure 4.1 - Exemple de K-plus proche voisin pour k=3 et k=5

4.2.2 Machines à Vecteurs Supports (SVM)

La méthode de SVM (*Support Vector Machines*) est une technique simple et efficace utilisée dans les domaines de classification et de reconnaissance. Elle a été introduite par Vapnik en 1995 [190]. Initialement prévue pour résoudre des problèmes de classification à deux classes, après elle est généralisé pour les problèmes multi-classe [191]. SVM reposent sur deux notions principales : la notion de marge maximale et la notion de fonction noyau [186].

Le principe des SVM consiste à projeter les données de l'espace d'entrée (appartenant à deux classes différentes) non-linéairement séparables dans un espace de plus grande dimension appelé espace de caractéristiques de façon à ce que les données deviennent linéairement séparables. Le but du SVM binaire est de trouver un hyperplan optimal qui sépare les deux classes en maximisant la distance. Cette distance est appelée marge. Dans le cas d'une classification binaire, figure 4.2, l'hyperplan est une droite. Les points les plus proches, qui seuls sont utilisés pour la détermination de la marge, sont appelés vecteurs de support [192].

L'hyperplan séparateur est représenté par l'équation (4.4) :

$$H(x) = W^T x + b \tag{4.4}$$

W est un vecteur de *m* dimensions et *b* est un terme. La fonction de décision, pour un exemple *x*, peut être exprimée comme suit :

$$\begin{cases} Classe = 1 & si\ H(x) > 1 \\ Classe = -1 & si\ H(x) < -1 \end{cases} \tag{4.5}$$

Maximiser la marge revient maximiser $\frac{2}{\|W\|}$ et qui vaut à minimiser $\|W\|$.

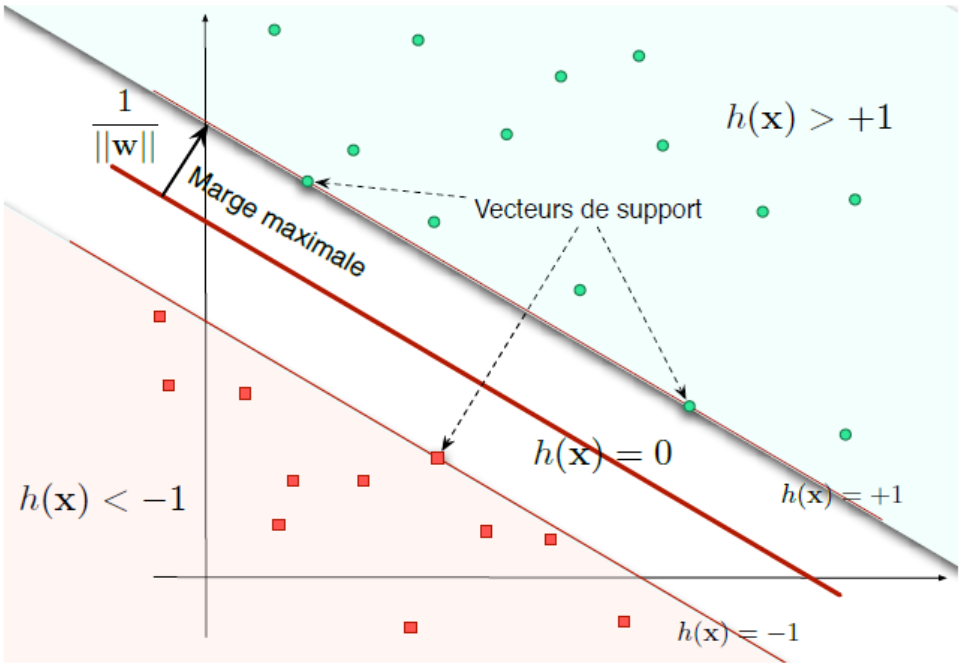


Figure 4.2 - SVM classification binaire

La classification d'un nouvel exemple de test est donnée par sa position dans l'espace de redescription par rapport à l'hyperplan optimal défini lors de l'apprentissage. Les SVM fournissent une distance à l'hyperplan dont le signe détermine la classe de l'exemple testé. SVM réduit le problème multi classe à une composition de plusieurs hyperplans bi-classe permettant de tracer les frontières de décision entre les différentes classes. Il décomposel'ensemble d'exemples en plusieurs sous-ensembles représentant chacun un problème de classification binaire.

A chaque fois un hyperplan de séparation est déterminé par la méthode SVM binaire. On construit lors de la classification une hiérarchie des hyperplans binaires qui est parcourue de la racine jusqu'à une feuille pour décider de la classe d'un nouvel exemple.

Parmi les modèles SVM nous distinguons le cas linéairement séparable et le cas non linéairement séparable. Pour surmonter les inconvénients du cas non linéairement séparable, l'idée des SVM est de transformer l'espace des données, afin de passer d'un problème de séparation non linéaire à un problème de séparation linéaire dans un espace de re-description de plus grande dimension. Cette transformation non linéaire est effectuée via une fonction, dite fonction noyau. L'un des paramètres à fixer lors de l'implémentation des SVM va donc être le choix du noyau. On peut citer les exemples de noyaux suivants : linéaire, polynomiale, gaussien, sigmoïde et laplacien [50]. La fonction noyau utilisé dans notre travail est le noyau linéaire.

La méthode de SVM est couramment utilisée dans le domaine de REP. SVM a obtenu des performances élevées par rapport aux autres classificateurs dans nombreux de travaux de la REP [78][189][193].

4.3 Performance des paramètres acoustiques sur le système de REP

Nous avons choisi de focaliser dans cette section sur la performance des paramètres acoustiques dans le système de REP dans le dialecte algérien en utilisant la base de données ADED.

4.3.1 Méthodologie

La méthodologie de cette section est illustrée dans la figure 4.3. La base de données émotionnelle du dialecte algérien (ADED) est utilisée pour extraire les paramètres acoustiques. Ces paramètres sont les valeurs statistiques de pitch (Mean, Max, Min et Range) et les paramètres similaires d'intensité, le UFR, le jitter, le shimmer, le HNR et les paramètres MFCC. Les paramètres extraits sont utilisés comme vecteurs de paramètres dans un classificateur pour identifier les différents états émotionnels. Cette étape est basée sur le classificateur KNN. La performance est évaluée en fonction du taux de reconnaissance qui définit par le nombre des échantillons classés en émotion sur le nombre totale des échantillons en émotion.

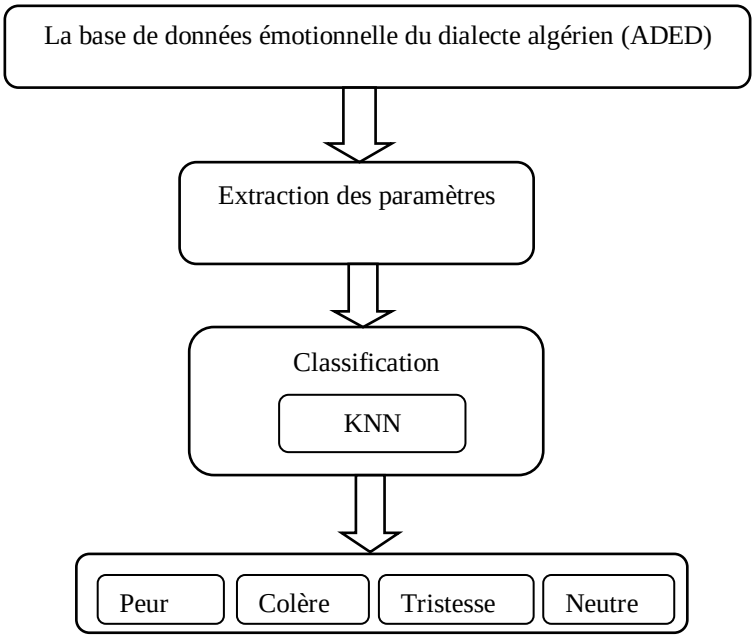


Figure 4.3 - Schéma du système de reconnaissance pour étudier la performance des paramètres acoustiques

4.3.2 Expériences et résultats

Différents ensembles des paramètres sont formés pour détecter l'ensemble des paramètres qui donne de meilleures performances. Les différents ensembles sont illustrés dans le tableau 4.1.

Afin d'obtenir des meilleurs résultats de classification, la performance de nombre de voisins k dans le classificateur KNN est ainsi évaluée. Quatre valeurs de k (k=3, 5,7 et 9) sont

utilisées dans les expériences. Les résultats sont présentés dans le tableau 4.2, ces résultats sont obtenus par MATLAB.

Tableau 4.1- Différents ensembles des paramètres acoustiques utilisés

Ensembles des paramètres	Paramètres acoustiques utilisés
1	Mean de pitch, Max de pitch, Min de pitch, Range de pitch Mean d'intensité, Max d'intensité, Min d'intensité, Range d'intensité.
2	Taux des trames non voisées, jitter, shimmer, HNR
3	MFCC
4	Mean de pitch, Max de pitch, Min de pitch, Range de pitch, Mean d'intensité, Max d'intensité, Min d'intensité, Range d'intensité, taux des trames non voisées, jitter, shimmer, HNR
5	Mean de pitch, Max de pich, Min de pitch, Range de pitch, Mean d'intensité, Max d'intensité, Min d'intensité, Range d'intensité, taux des trames non voisées, jitter, shimmer, HNR, MFCCs.

Tableau 4.2 - Résultats de différentes expériences

Ensembles des paramètres	Taux de reconnaissance (KNN)			
	K=3	**K=5**	**K=7**	**K=9**
1	**81.25%**	72.92%	66.67%	64.06%
2	**77.60%**	61.46%	57.81%	55.73%
3	**80.21%**	65.10%	59.37%	51.56%
4	**81.77%**	73.44%	68.23%	65.62%
5	**82.29%**	72.17%	66.15%	66.15%

D'après le tableau 4.2, les meilleurs résultats ont été obtenus avec une valeur de k est égale 3 dans le classificateur KNN. La figure 4.4 présente une comparaison entre les taux de reconnaissance de système de chaque ensemble de paramètres pour k= 3.

Dans les trois premiers ensembles (1, 2 et 3), chaque ensemble contient des paramètres de même type, l'ensemble 1 contient des paramètres prosodiques (les valeurs statistiques du pitch et intensité), l'ensemble 2 contient des paramètres de qualités vocales (UFR, jitter, shimmer et HNR) et l'ensemble 3 contient des paramètres spectraux (MFCC). Il est observé que la performance de paramètres prosodiques est élevée par rapport aux performances des autres paramètres. Dans les ensembles (4 et 5), des combinaisons des différents paramètres sont utilisées. Il est remarqué que la performance de reconnaissance est améliorée. Le meilleur taux de reconnaissance (82.29%) est obtenu par l'ensemble 5 qui contient une combinaison de paramètres de différents types (Mean de pitch, Max de pitch, Min de pitch,

Range de pitch, Mean d'intensité, Max d'intensité, Min d'intensité, Range d'intensité, UFR, jitter, shimmer, HNR et MFCC).

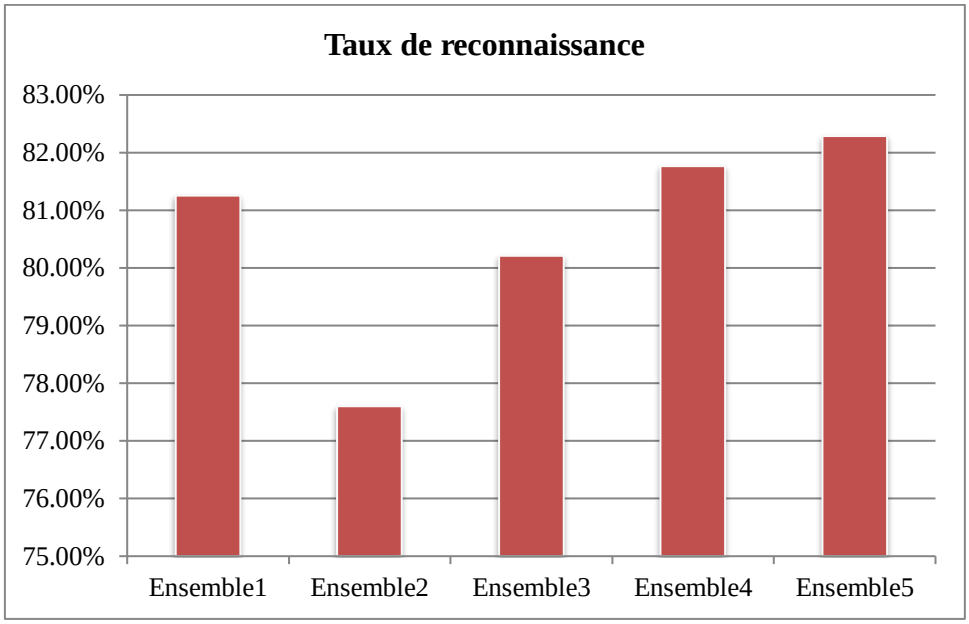


Figure 4.4 - Comparaison entre les taux de reconnaissance de système de chaque ensemble pour k=3

Il est conclu selon les résultats que la performance du système de REP dans le dialecte algérien est améliorée par l'utilisation de combinaison des paramètres dans le système de reconnaissance. Les mêmes observations sont notées dans autres travaux c'est à- dire la performance de reconnaissance des émotions est augmentée par l'utilisation de combinaison des paramètres acoustiques. Pour reconnaître les émotions dans la langue arabe égyptienne [78], les systèmes qui utilisaient des combinaisons des paramètres ont atteint des taux de reconnaissance plus élevés que les systèmes qui utilisaient les mêmes types de paramètres. Différents paramètres ont été utilisés dans ce travail tels que : le pitch, l'intensité, les formants, les MFCC, etc.

Pour reconnaître les émotions dans les bases de données de parole émotionnelle de Berlin et d'Espagne, le système qui a utilisé une combinaison des paramètres prosodiques (pitch et énergie) et des paramètres spectraux (MFCC) a donné de meilleures performances par rapport aux performances du système qui utilise des paramètres de même type (prosodique ou spectral). Dans la base de données de parole émotionnelle de Berlin, le taux de reconnaissance de système qui utilise uniquement les paramètres prosodiques égale 55%, et 68.75% pour le système qui utilise uniquement les paramètres spectraux. Le taux de reconnaissance est amélioré (75.5%) dans le système qui utilise une combinaison des paramètres prosodique et spectraux. Dans la base de données d'Espagne, la performance de reconnaissance est améliorée est atteint 73% en comparant avec le système qui utilise uniquement les paramètres prosodiques (60.25%) et le système qui utilise uniquement les paramètres spectraux (64.5%) [194].

4.4 Performance du nombre d'émotions sur le système de REP

Dans cette section, l'influence du nombre d'émotions incluses dans le système de REP dans le dialecte algérien est étudiée en utilisant la base de données ADED.

.

4.4.1 Méthodologie

La méthodologie de cette section est présentée dans la figure 4.5. La procédure est divisée en trois étapes. Dans la première étape, la reconnaissance d'émotion de peur par rapport à l'émotion neutre est effectuée. Dans cette partie uniquement les segments de parole d'états de peur et neutre sont utilisés. Dans la deuxième étape, l'état de colère est inséré dans le système de reconnaissance avec les deux premières émotions. Par conséquent les segments de parole concernant les émotions de peur, colère et neutre sont exploités. Enfin, l'émotion de tristesse est ajoutée au système de reconnaissance dans la troisième étape. Une combinaison de paramètres contient : Mean de pitch, Max de pitch, Min de pitch, Range de pitch, Mean d'intensité, Max d'intensité, Min d'intensité, Range d'intensité, UFR, jitter, shimmer, HNR et MFCC est utilisée dans le système de reconnaissance. Ainsi la méthode de KNN avec k =3 est exploitée comme classificateur dans les trois étapes.

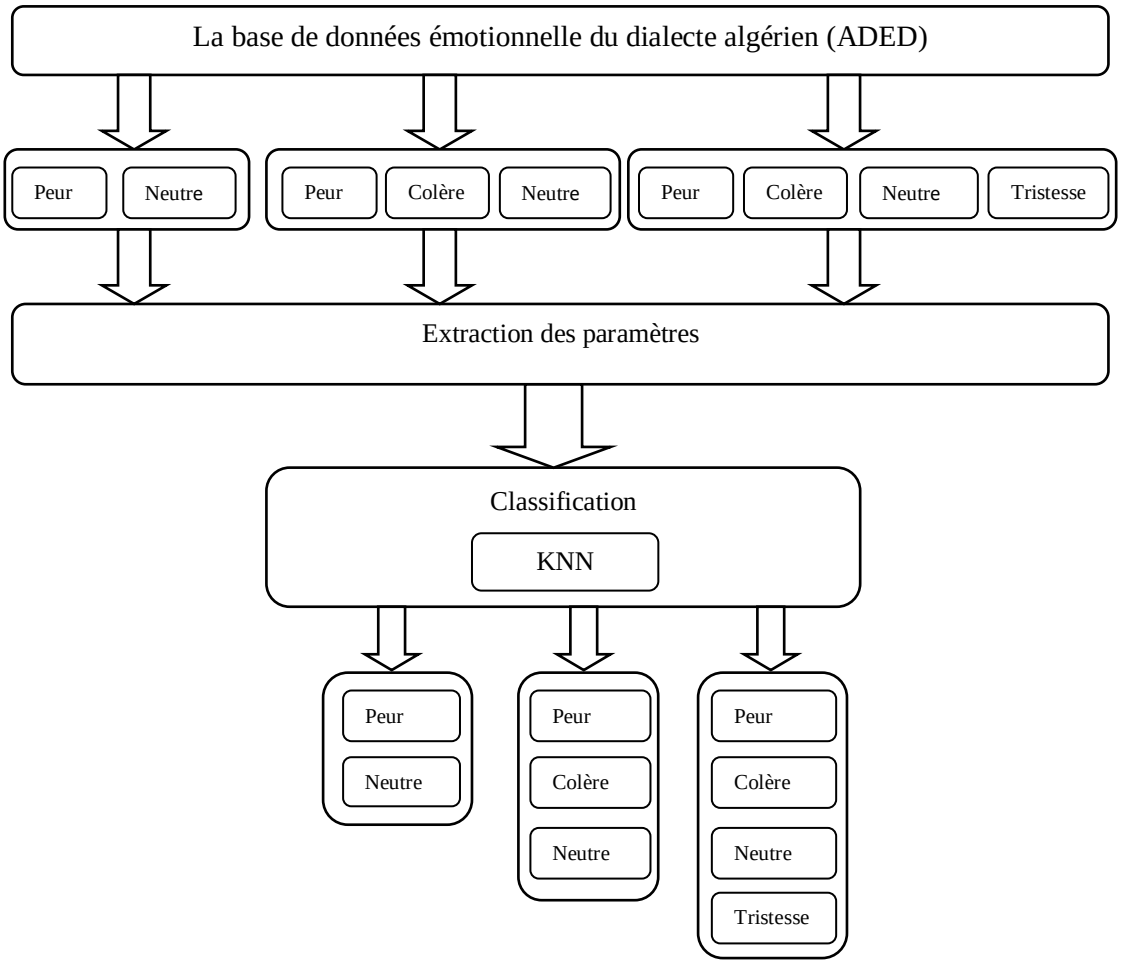


Figure 4.5 - Schéma des systèmes de reconnaissance pour étudier l'influence du nombre d'émotions

4.4.2 Expériences et résultats

Dans cette section, trois expériences sont menées pour étudier l'influence de nombre des émotions sur le système de reconnaissance. Les trois expériences sont effectuées par MATLAB. Les taux de reconnaissance des trois expériences sont montrés dans le tableau 4.3. Une comparaison entre les taux de reconnaissance obtenus de chaque expérience est illustrée sur la figure 4.6. À partir de tableau 4.3 et la figure 4.6, on observe que la performance de reconnaissance est démunie si le nombre des émotions augmente dans le système de reconnaissance. Le taux de reconnaissance est égale 87.50% s'il y a deux émotions dans le système de reconnaissance. Lorsque l'état de colère est ajouté au système, le taux de reconnaissance démunie d'environ 84.03%. Lorsque l'émotion de tristesse est insérée avec les autres émotions dans le système de reconnaissance, le taux atteint 82.29%. Cette diminution est causée par les émotions qui partagent presque les mêmes valeurs de certains paramètres : l'émotion de peur et de colère ont une valeur élevée de pitch, les émotions de tristesse et de neutre sont associées à une faible Range de pitch, les UFR sont modérés dans les états de neutre et de peur, la valeur de HNR est faible dans les émotions de peur et de colère.

Tableau 4.3 - Taux de reconnaissance des trois expériences

Emotions	**Taux de reconnaissance (KNN, k=3)**
Peur et neutre	**87.50%**
Peur, colère et neutre	84.03%
Peur, colère, tristesse et neutre	82.29%

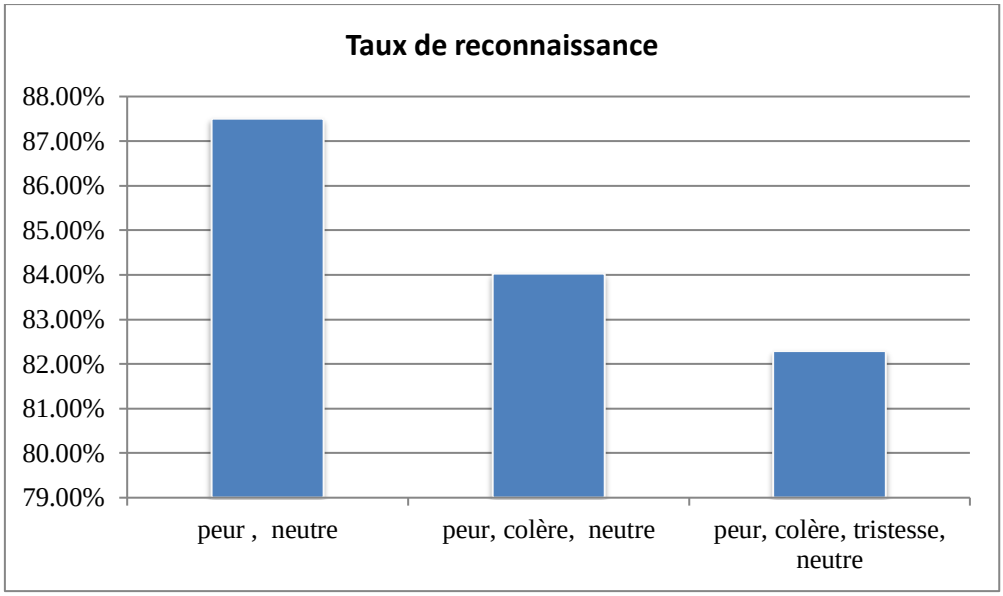


Figure 4.6 - Comparaison entre les taux de reconnaissance de chaque expérience

Les tableaux 4.4, 4.5 et 4.6 illustrent les matrices de confusion de chaque expérience. Le tableau 4.4 présente la matrice de confusion entre les émotions de peur et neutre. La matrice de confusion entre les émotions de peur, colère et neutre est illustrée dans le tableau 4.5. Le tableau 4.6 montre la matrice de confusion entre les émotions de peur, colère, tristesse et neutre. Il est noté que le taux de reconnaissance de l'état neutre est le plus élevé dans les trois expériences mais pour l'émotion de tristesse le taux de reconnaissance est faible. Il est aussi observé dans les tableaux 4.5 et 4.6 qu'il y a une confusion entre les états de peur et colère et entre les états de tristesse et neutre.

Tableau 4.4 - Matrice de confusion entre les émotions de peur et neutre

Emotion	Peur	Neutre
Peur	81.25 %	18.75%
Neutre	6.25%	93.75%

Tableau 4.5 - Matrice de confusion entre les émotions de peur, colère et neutre

Emotion	Peur	Colère	Neutre
Peur	81.25 %	10.42%	8.33%
Colère	6.25%	79.17%	14.58%
Neutre	4.17%	4.17%	91.66%

Tableau 4.6 - Matrice de confusion entre les émotions de peur, colère, tristesse et neutre

Emotion	Peur	Colère	Neutre	Tristesse
Peur	79.17 %	10.42%	6.25%	4.17%
Colère	4.17%	83.33%	10.42%	2.08%
Neutre	2.08%	2.08%	91.67%	4.17%
Tristesse	8.33%	2.08%	14.58%	75.00%

Il est conclu selon les résultats que la performance de système de reconnaissance des émotions dans le dialecte algérien est influencée par le nombre des émotions dans le système de reconnaissance. La même remarque est observée dans d'autres travaux, on peut citer le travail qui reconnaît les émotions dans la langue arabe égyptienne [78]. La performance était élevée (81%) lors de la classification de trois émotions : colère, tristesse et neutre que lors de la classification de quatre émotions : colère, tristesse, neutre et le bonheur, le taux de reconnaissance est égale 66.8%.

4.5 Influence de sexe sur le système de REP

Notre objectif dans cette section est d'étudier l'impact de classes de sexe sur la reconnaissance des émotions dans le dialecte algérien en utilisant la base de données ADED.

4.5.1 Méthodologie

Le schéma de notre méthodologie dans cette section est illustré dans la figure 4.7. Notre méthodologie consiste à diviser la base de données ADED en fonction du sexe, c'est-à-dire des segments de parole masculins et des segments de parole féminins ainsi on utilise des segments sans distinction entre les classes du sexe. Dans l'étape d'extraction de paramètres, différents paramètres tels que Mean de pitch, Max de pitch, Min de pitch, Range de pitch, Mean d'intensité, Max d'intensité, Min d'intensité, Range d'intensité, UFR, jitter, shimmer, HNR et MFCC sont extraits à partir les segments de parole utilisés. Les paramètres sont entrés dans le classificateur en tant que vecteur de paramètres. Dans l'étape de classification deux classificateurs KNN (pour k =3) et SVM sont exploités.

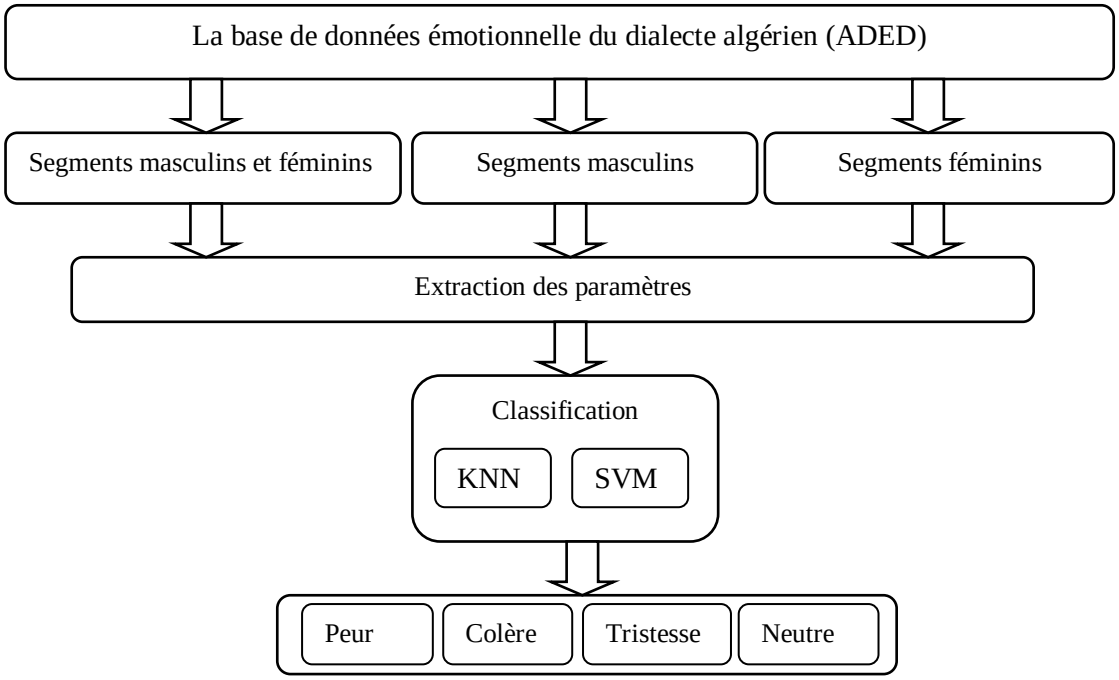


Figure 4.7 - Schéma des systèmes de reconnaissance pour étudier l'influence de classes de sexe

4.5.2 Expériences et résultats

Plusieurs expériences sont effectuées pour atteindre l'objectif de cette section. Les expériences sont effectuées par MATLAB. Les résultats obtenus sont montrés dans le tableau 4.7. Ce tableau présent les matrices de confusion pour la reconnaissance des émotions en fonction des classes de sexe. Ainsi une comparaison entre les deux méthodes de classification est montrée dans le tableau 4.7. La figure 4.8 montre aussi une comparaison entre les résultats obtenus. D'après le tableau 4.7 et la figure 4.8, il est noté que les taux de la reconnaissance dans les systèmes qui n'utilisaient que les segments de parole masculins ou les segments

féminins se sont améliorés par rapport aux taux des systèmes qui utilisaient des segments sans distinction du sexe. Il est aussi noté que les taux de reconnaissance des émotions chez les masculines sont élevés que les féminines. Il est observé également dans la figure 4.8 que la performance du classificateur SVM est élevée par rapport au classificateur KNN.

Tableau 4.7 - Matrices de confusion pour la reconnaissance des émotions en fonction du sexe

Segments de parole		**KNN**				**SVM**			
masculine et féminine	Emotion	Peur	Colère	Neutre	Tristesse	Peur	Colère	Neutre	Tristesse
	Peur	78.57%	7.14%	14.29%	0%	100%	0%	0%	0%
	Colère	7.14%	85.72%	7.14%	0%	7.14%	92.86%	0%	0%
	Neutre	0%	0%	100%	0%	7.14%	0%	92.86%	0%
	Tristesse	7.14%	0%	28.57%	64.29%	7.14%	0%	28.57%	64.29%
	Moyen	**82.14%**				**87.50%**			
masculine	Emotion	Peur	Colère	Neutre	Tristesse	Peur	Colère	Neutre	Tristesse
	Peur	71.44%	0%	14.28%	14.28%	100%	0%	0%	0%
	Colère	0%	92.86%	7.14%	0%	0%	100%	0%	0%
	Neutre	0%	0%	100%	0%	0%	7.14%	92.86%	0%
	Tristesse	0%	0%	14.29%	85.71%	0%	0%	0%	100%
	Moyen	**87.50%**				**98.21%**			
féminine	Emotion	Peur	Colère	Neutre	Tristesse	Peur	Colère	Neutre	Tristesse
	Peur	78.57%	21.43%	0%	0%	100%	0%	0%	0%
	Colère	0%	92.86%	7.14%	0%	0%	71.43%	0%	28.57%
	Neutre	0%	0%	92.86%	7.14%	0%	0%	100%	0%
	Tristesse	0%	0%	21.43%	78.57%	0%	0%	7.14%	92.86%
	Moyen	**85.71%**				**91.07%**			

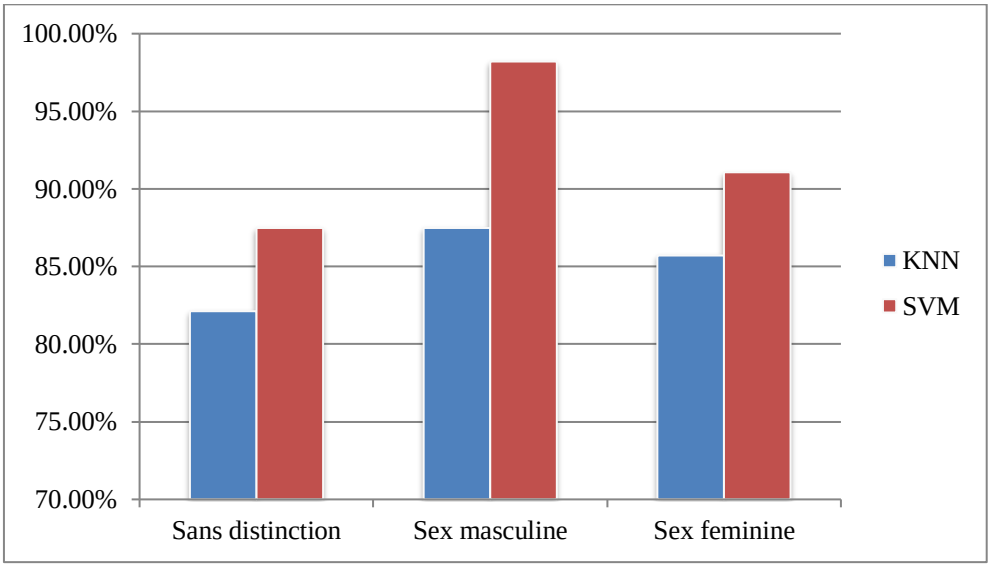


Figure 4.8 - Comparaison entre les résultats obtenus pour la reconnaissance des émotions en fonction du sexe

Il est conclu des résultats que la reconnaissance des émotions dans le dialecte algérien est influencée par les classes de sexe. En comparant les résultats obtenus avec ceux des autres travaux qui ont utilisé différentes bases de données : la base de données BHUES (BeiHang University Mandarin Emotion Speech database) [195] et les bases de données Berlin and SmartKom [196]. La même remarque était observée, c'est-à-dire que la performance des systèmes qui n'utilisent que les énoncés de parole masculine ou féminine est élevée que la performance du systèmes qui utilisent les énoncés de parole sans distinction entre les classes de sexe.

4.6 Performance des descripteurs acoustiques sur le système de la REP dans différentes langues

Le but de cette section est d'étudier la performance des paramètres acoustiques sur le système de reconnaissance des émotions de colère et de neutre dans différentes langues.

4.6.1 Méthodologie

Le schéma qui décrit la méthodologie de cette section est illustré sur la figure 4.9. Pour atteindre l'objectif, quatre bases de données de différentes langues : ADED, EMO-DB, ShEMO et CREAMA-D sont exploitées dans le système de reconnaissance. Dans cette section, nous utilisons uniquement les segments concernant les émotions de colère et de neutre de chaque base de données. Dans l'étape d'extraction de paramètres, différents paramètres acoustiques sont extraits à partir des segments de parole choisis. Ces paramètres acoustiques sont montrés dans le tableau 4.8. Dans l'étape de classification, la méthode de SVM est exploitée comme classificateur dans les systèmes de reconnaissance.

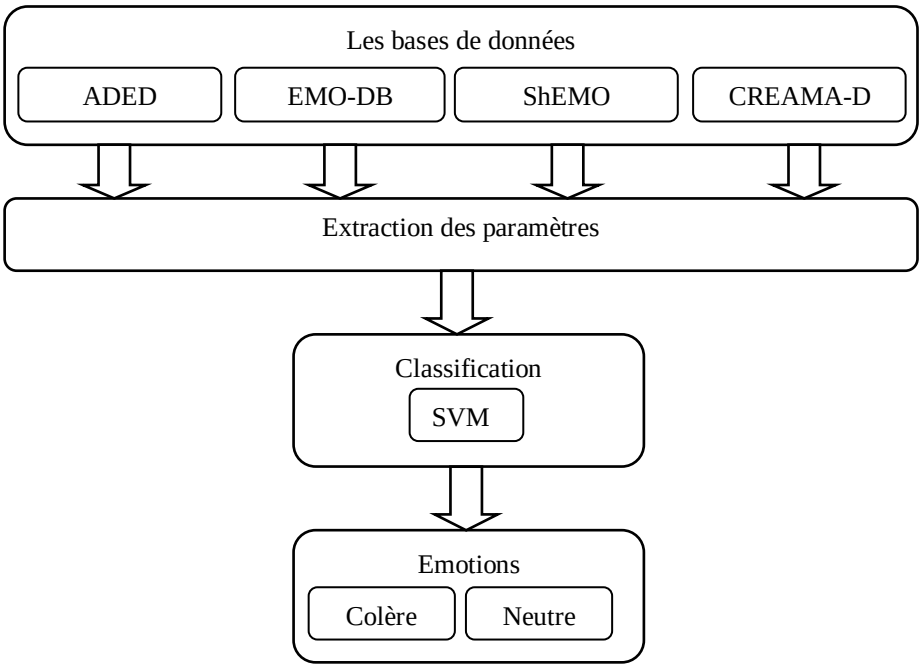


Figure 4.9 - Schéma des systèmes de reconnaissance pour étudier la performance des paramètres acoustiques dans différentes langues

Tableau 4.8 - Paramètres acoustiques utilisés dans les systèmes de la REP

	Paramètres
Prosodiques	Mean de pitch
	Max de pitch
	Min de pitch
	Déviation standard de pitch
	Mean d'intensité
	Max d'intensité
	Min d'intensité
Formants	Formant1
	Formant2
	Formant3
	Formant4
MFCC	13MFCCs

4.6.2 Bases de données

Quatre bases de données de différentes langues : ADED, EMO-DB, ShEMO et CREAMA-D sont utilisées dans cette section. ADED est décrit dans le troisième chapitre. Les autres bases de données sont décrites brièvement dans cette sous section [197]:

4.6.2.1 EMO-DB (Berlin database of emotional speech)

EMO-DB est une base de données allemande de parole émotionnelle enregistrée par l'université technologie de Berlin [62]. La base de données EMO-DB contienne sept types d'émotions : le bonheur, la peur, le dégoût, la colère, la tristesse, l'ennui et le neutre. Dans cette base de données, 535 fichiers audio ont été enregistrés par 5 hommes et 5 femmes âgés de 21 à 35 ans. Le nombre de fichiers audio de chaque émotion est indiqué dans le tableau 4.9.

Tableau 4.9 - Nombre de fichiers audio de chaque émotion dans EMO-DB

Emotions	Nombre des fichiers audio
Peur	69
Neutre	79
Colère	127
Tristesse	62
Bonheur	71
Dégout	46
Ennui	81
Total	535

4.6.2.2 ShEMO (Sharif Emotional Speech Database)

ShEMO est une base de données pour la langue persane. Cette base de données comprend 3000 énoncés de parole exprimés par 87 persans natifs (31 femmes, 56 hommes). La base de données ShEMO classée en six états émotionnels de base : la surprise, le bonheur, la tristesse,

la peur, la colère et l'état neutre [76]. Le nombre d'énoncés de chaque émotion est illustré dans le tableau 4.10.

Tableau 4.10 - Nombre d'énoncés de chaque émotion dans ShEMO

Emotions	Nombre des énoncés
Peur	38
Neutre	1028
Colère	1059
Tristesse	449
Bonheur	201
Surprise	46
Total	3000

4.6.2.3 CREAMA-D(Crowd-sourced Emotional Multimodal Actors Dataset)

CREAMA-D est une base de données composée d'expressions faciales et vocales en langue anglaise [77]. Cette base de données composée de 7442 fichiers audio exprimés en six émotions de base : le bonheur, la tristesse, la peur, le dégoût, la colère et le neutre. Les fichiers audio ont été enregistrés par 91 acteurs (48 hommes et 43 femmes) âgés de 20 à 74 ans. Le nombre de fichiers audio de chaque émotion est indiqué dans le tableau 4.11.

Tableau 4.11- Nombre de fichiers audio de chaque émotion dans CREAMA-D

Emotions	Nombre des fichiers audio
Peur	1271
Neutre	1087
Colère	1271
Tristesse	1271
Bonheur	1271
Dégout	1271
Total	7442

4.6.3 Expériences et résultats

Dans cette section, plusieurs expériences sont menées pour étudier l'influence des paramètres extraits sur la reconnaissance de colère et de neutre dans quatre bases de données différentes. Le tableau 4.12 présente le nombre de segments de chaque base de données utilisée dans les expériences. Différents ensembles de paramètres sont formés pour identifier les ensembles de paramètres qui donnent des meilleures performances. Les paramètres sont entrés dans le classificateur SVM en tant que vecteurs de paramètres. Les expériences sont effectuées par MATLAB. Les résultats sont présentés dans les tableaux 4.13, 4.14 et 4.15. Le tableau 4.13 présente les taux de reconnaissance obtenus avec les différents ensembles de paramètres dans chaque base de données. Des comparaisons entre les performances de chaque ensemble de paramètres sur chaque base de données sont illustrées sur les figures 4.10, 4.11, 4.12 et 4.13. Les tableaux 4.14 et 4.15 présentent les taux de reconnaissance pour le sexe masculin et le sexe féminin respectivement.

Tableau 4.12 - Nombre de segments de parole de chaque BD utilisée dans les expériences

Base de données	Colère	Neutre
ADED	52	48
EMO-DB	52	48
ShEMO	52	48
CREMA-D	52	48

Tableau 4.13 - Taux de reconnaissance obtenus avec les différents ensembles de paramètres dans chaque base de données sans distinction du sexe

Paramètres	ADED	EMO-DB	ShEMO	CREMA-D
Prosodiques	78.12%	93.75%	78.47%	93.05%
Formants	61.45%	79.17%	60.42%	69.44%
MFCC	63.54%	89.58%	83.33%	69.44%
Prosodiques + Formants	78.12%	96.87%	86.46%	93.75%
Prosodiques + MFCC	83.33%	95.83%	88.54%	95.14%
Formants + MFCC	80.21%	92.71%	83.33%	79.86%
Prosodiques + Formants + MFCC	**85.42%**	**97.92%**	**90.97%**	**95.14%**

Tableau 4.14 - Taux de reconnaissance obtenus avec les différents ensembles de paramètres dans chaque base de données pour les Hommes

Paramètres	ADED	EMO-DB	ShEMO	CREMA-D
Prosodiques	79.54%	100%	87.50%	91.67%
Formants	63.63%	73.91%	62.50%	68.05%
MFCC	79.54%	91.67%	84.72%	75.00
Prosodiques + Formants	88.63%	100%	87.50%	91.66%
Prosodiques + MFCC	93.18%	100%	94.44%	93.05%
Formants + MFCC	81.81%	91.67%	77.78%	86.11%
Prosodiques+ Formants + MFCC	**100%**	**100%**	**97.22%**	**94.44%**

Tableau 4.15 - Taux de reconnaissance obtenus avec les différents ensembles de paramètres dans chaque base de données pour les Femmes

Paramètres	ADED	EMO-DB	ShEMO	CREMA-D
Prosodiques	88.63%	98.61%	87.50%	94.94%
Formants	59.09%	84.72%	76.39%	70.83%
MFCCs	70.45%	97.22%	70.83%	70.83%
Prosodiques + Formants	90.91%	98.61%	94.44%	94.44%
Prosodiques + MFCC	97.73%	100%	93.05%	98.61%
Formants + MFCC	79.54%	97.22%	83.33%	76.39%
Prosodiques+Formants + MFCC	**100%**	**100%**	**95.83%**	**100%**

Nous avons observé dans le tableau 4.13 et les figures 4.10, 4.11, 4.12 et 4.13 que le taux de reconnaissance est élevé dans les bases de données EMO-DB et CREMA-D, et ce taux de reconnaissance est faible dans les bases de données ADED et ShEMO lorsque uniquement les paramètres prosodiques sont utilisés dans le système de reconnaissance. Des taux de reconnaissance très faibles sont obtenus lors l'utilisation uniquement les paramètres des formants dans toutes les bases de données. Des taux de reconnaissance acceptables sont notés

lorsque les paramètres MFCC sont utilisés dans les bases de données EMO-DB et ShEMO par rapport aux bases de données ADED et CREMAD. Les performances de reconnaissance sont améliorées lors de l'utilisation des différentes combinaisons des paramètres dans toutes les bases de données. Les meilleurs taux de reconnaissance sont obtenus par la combinaison des paramètres prosodiques, formants et MFCC dans toutes les bases de données. D'après les tableaux 4.14 et 4.15, les mêmes remarques sont observées par rapport aux expériences ayant utilisées les bases de données sans distinction de sexe (tableau 4.13). Les performances les plus élevées sont obtenues par l'utilisation d'une combinaison de paramètres (prosodiques, formants et MFCC) dans chaque base de données.

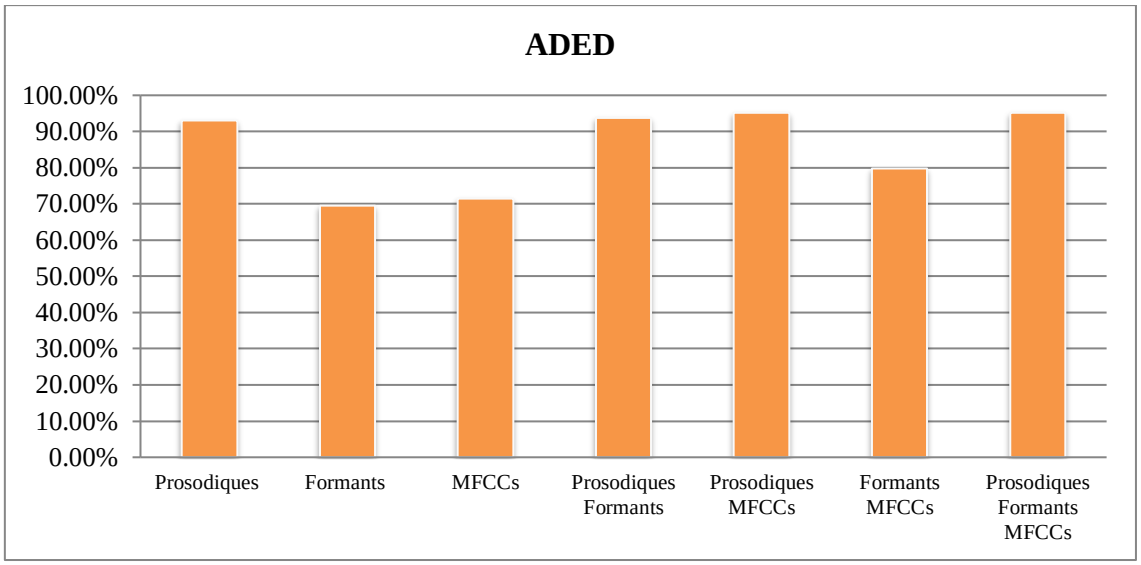


Figure 4.10 - Comparaison entre la performance de chaque ensemble de paramètres dans la base de données ADED

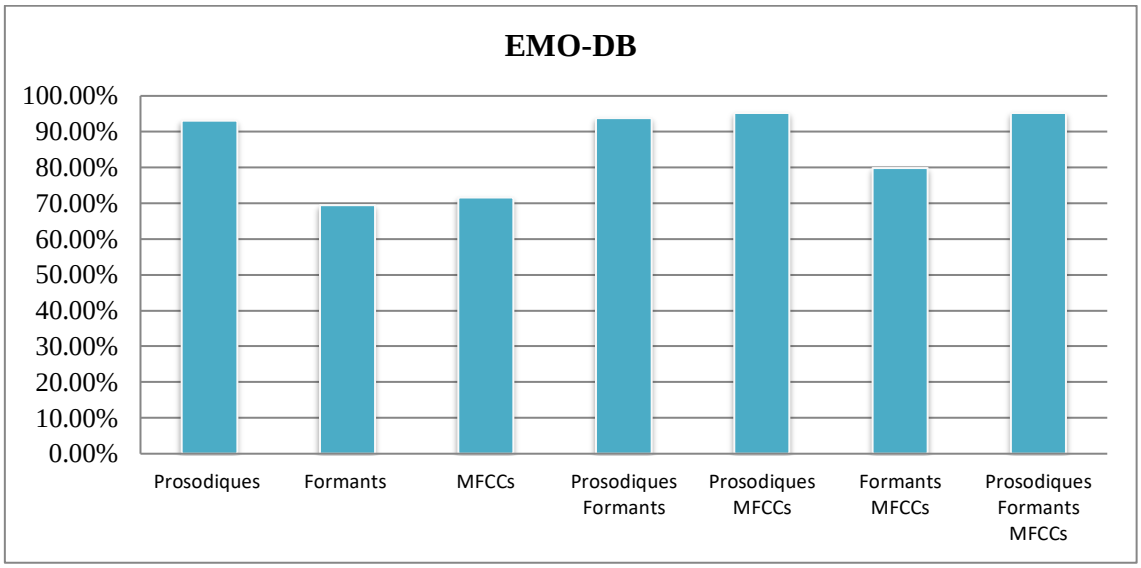


Figure 4.11- Comparaison entre la performance de chaque ensemble de paramètres dans la base de données EMO-DB

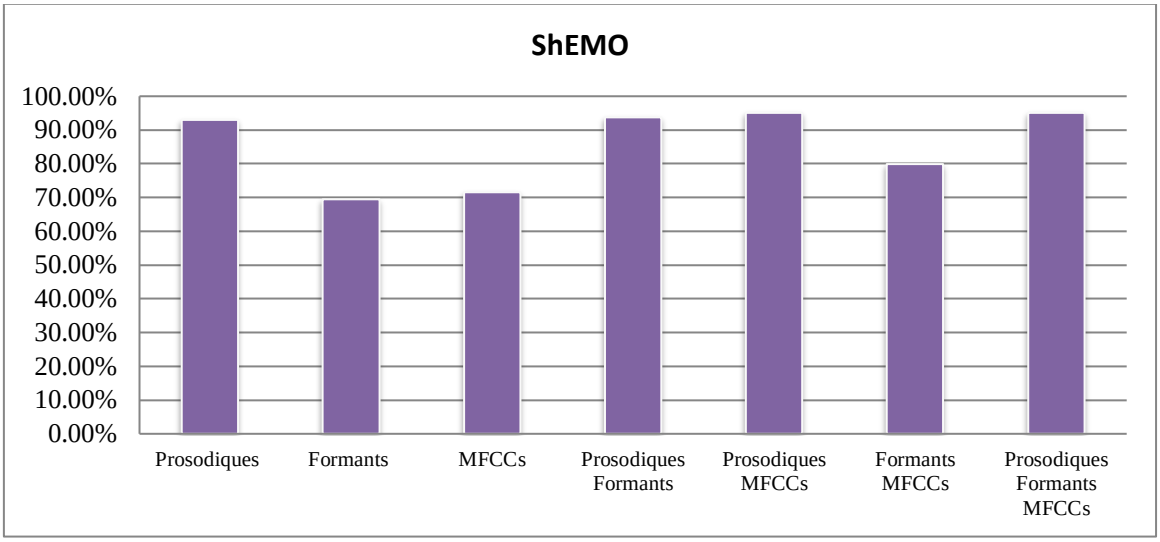


Figure 4.12 - Comparaison entre la performance de chaque ensemble de paramètres dans la base de données ShEMO

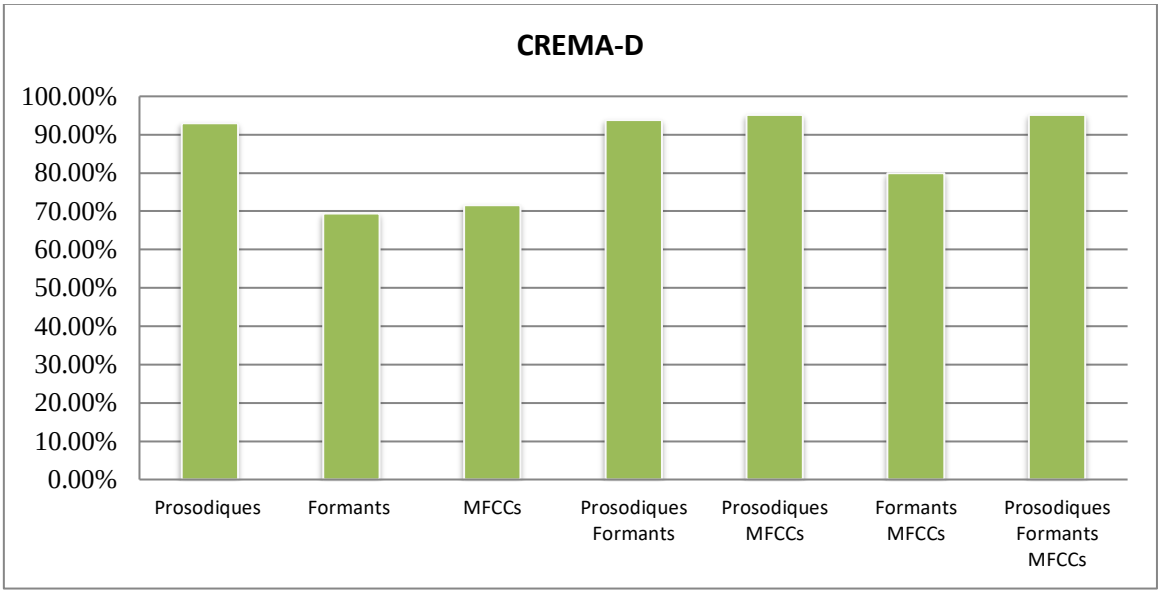


Figure 4.13 - Comparaison entre la performance de chaque ensemble de paramètres dans la base de données CREMA-D

Il est conclu selon les résultats que les performances des systèmes de reconnaissance sont améliorées lors de l'utilisation de combinaison entre les paramètres prosodiques, les formants et les paramètres MFCC dans les différentes bases de données. Dans les travaux précédents sur la reconnaissance des émotions, la combinaison de différents paramètres améliore la performance des systèmes de reconnaissance dans des bases de données avec différentes langues. La combinaison des paramètres prosodiques (pitch et énergie) et spectraux a obtenu de meilleurs résultats pour reconnaître les émotions dans les bases de données de parole émotionnelle de Berlin et d'Espagne [194]. Différents paramètres tels que le pitch, l'intensité, les formants, les MFCC, etc. ont été utilisés pour reconnaître les émotions dans la langue arabe égyptienne. Les résultats obtenus ont montré que les performances des systèmes qui utilisent une combinaison des paramètres ont atteint des performances supérieures à celles des

systèmes qui utilisent les mêmes types de paramètres [78]. Dans la base de données émotionnelle en Mandarin, la performance du système qui utilise une combinaison des paramètres spectraux et prosodiques est plus élevée que le système qui utilise uniquement des paramètres spectraux ou prosodiques [198].

4.7 Conclusion

Dans ce chapitre nous avons réalisé des systèmes de reconnaissance des émotions de peur, colère, tristesse et neutre en utilisant des paramètres prosodiques, paramètres de qualité de la voix et paramètres spectraux. Les systèmes de reconnaissance sont basés sur les classificateurs de KNN et SVM. Plusieurs expériences sont menées afin d'étudier des systèmes de reconnaissance des émotions dans le dialecte algérien (ADED). Nous avons notamment étudié la performance des paramètres acoustiques sur le système de reconnaissance. Les taux de reconnaissance obtenus ont montré que la performance de système est améliorée par l'utilisation de combinaison des paramètres. Ainsi, l'influence du nombre d'émotions incluses dans le système de reconnaissance est étudiée. Et nous avons conclu que la performance de système de reconnaissance des émotions dans ADED est influencée par le nombre des émotions incluses dans les systèmes. Nous avons également évalué de l'influence de classes de sexe sur les systèmes de reconnaissance. Les résultats obtenus ont montré que la reconnaissance des émotions dans ADED est influencée par les classes de sexe. Enfin, la performance des descripteurs acoustiques sur le système de reconnaissance des émotions de colère et de neutre dans différentes langues est aussi étudiée. Les taux de reconnaissance obtenus ont montré que l'utilisation de combinaison entre les paramètres prosodiques, les formants et les paramètres MFCC dans les différentes bases de données améliore les performances de système de reconnaissance.

Conclusions générales et Perspectives

Le traitement des émotions dans la parole aide à assurer le naturel dans la performance des systèmes vocaux existants. Dans cette thèse nous nous sommes focalisés sur la REP dans le dialecte algérien pour aider les psychologues à exploiter la parole dans le domaine de la psychologie. L'originalité de notre travail réside dans la construction d'une Base de Données Emotionnelle du Dialecte Algérien (ADED). Nous avons choisi des films algériens comme matériel de base pour la construction de notre Base de Données. Ces films décrivent la crise de la guerre civile ainsi que la période qui la suit en Algérie. Quatre états émotionnels de base sont considérés dans l'ADED : la colère, la peur, la tristesse et le neutre. Nous avons choisi de considérer des paramètres acoustiques modélisant des contenus très variés : paramètres prosodiques, paramètres de la qualité de la voix et paramètres spectraux. Ces paramètres sont les valeurs statistiques de pitch et les paramètres similaires d'intensité, UFR, le jitter, le shimmer, le HNR, les formants ainsi que les paramètres MFCC. Les systèmes de reconnaissance effectués dans ce travail sont basés sur les méthodes de classification KNN et SVM. Plusieurs expériences sont menées pour étudier l'influence des différents facteurs tels que les paramètres acoustiques, le nombre d'émotions, le sexe et la langue, sur les systèmes de REP dans le dialecte algérien.

Nous avons présenté dans un premier temps quelques notions concernant les émotions et la parole ainsi des corrélations entre les différents états émotionnels et les caractéristiques acoustiques du signal de la parole. Ensuite nous avons présenté une revue de travaux de REP du point de vue des bases de données émotionnelles, des paramètres spécifiques aux émotions extraites de différents aspects de la parole et des modèles de classification. Ainsi que les principales applications de ce domaine sont présentées. Puis, nous avons décrit la procédure de création de notre BD qui comprend l'acquisition des données, l'annotation et la validation des émotions. La base de données que nous avons créée est composée de 200 fichiers audio de quatre états émotionnels : colère, peur, tristesse et neutre. Les fichiers audio inclus dans cette BD sont exprimés par 32 acteurs (16 Hommes et 16 Femmes) d'âges différents entre 18 et 60 ans. Après, nous avons décrit aussi les paramètres acoustiques exploités dans notre travail tels que les paramètres prosodiques (pitch et intensité), les paramètres de qualité de la voix (UFR, jitter, shimmer et HNR) et les paramètres spectraux (formants et MFCC).

Plusieurs expériences sont effectuées pour étudier l'influence des différents facteurs tels que les paramètres acoustiques, le nombre des émotions, le sexe et la langue sur les systèmes de REP dans le dialecte algérien. Nous avons conclu d'après les résultats obtenus que la performance du système de reconnaissance est améliorée par l'utilisation de la combinaison des différents paramètres acoustiques (Mean de pitch, Max de pitch, Min de pitch, Range de pitch, Mean d'intensité, Max d'intensité, Min d'intensité, Range d'intensité, UFR, jitter, shimmer, HNR et MFCC). Nous avons constaté également que la performance du système de reconnaissance est influencée par le nombre des émotions incluses dans les systèmes de reconnaissance, autrement dit la performance de la reconnaissance est démunie si le nombre des émotions augmente dans ce système. Nous avons remarqué aussi que les classes de sexe ont influé sur la performance du système de reconnaissance des émotions. Les Taux de la Reconnaissance dans les systèmes qui n'utilisaient que les segments de parole prononcés par des Hommes ou bien des Femmes se sont améliorés par rapport aux taux des systèmes qui

utilisaient des segments mixtes sans distinction du sexe. Enfin, nous avons montré que l'utilisation de la combinaison entre les paramètres prosodiques, les formants et les paramètres MFCC dans les différentes BD (ADED, EMO-DB, ShEMO et CREAMA-D) en différentes langues, améliore les performances du système de reconnaissance.

Les attentes dans le domaine de la REP par la machine sont énormes. Les travaux réalisés dans le cadre de cette thèse peuvent être appliqués dans plusieurs directions. La BD émotionnelle du dialecte algérien (ADED) pourrait être élargie en augmentant le nombre des segments. Nous avons mis en place un système de reconnaissance de quatre états émotionnels, le nombre d'émotions pourrait être augmenté dans le système de reconnaissance en insérant d'autres états émotionnels tels que la surprise et la joie. D'autres paramètres acoustiques pourraient être extraits pour améliorer la performance de reconnaissance. Il existe de nombreuses méthodes de classification. Des modèles de classification pourraient être utilisés comme les Réseaux de Neurones Convolutifs (CNN) et les Réseaux de Neurones Récurrents (RNN). Ainsi la fusion des résultats obtenus par les différents classificateurs pourrait être une piste pour l'amélioration des performances. L'Algérie est un grand pays, administrativement divisé en 58 départements. Il n'y a pas de dialecte algérien unique : différentes régions d'Algérie parlent de légères variations du même dialecte avec des accents différents sur la prononciation. La reconnaissance des émotions dans chaque région d'Algérie pourrait être étudiée.

Références Bibliographiques

Références Bibliographiques

[1] Petrantonakis, P. C., & Hadjileontiadis, L. J. (2010). Emotion recognition from EEG using higher order crossings. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 186– 197.

[2] Frantzidis, C. A., Bratsas, C., et al. (2010). On the classification of emotional bio-signals evoked while viewing affective pictures : an integrated data-mining-based approach for healthcare applications. IEEE Transactions on Information Technology in Biomedicine, 14(2), 309–318.

[3] Murugappan, M., Rizon, M., Nagarajan, R., Yaacob, S., Zunaidi, I., Hazry, D. (2007). EEG feature extraction for classifying emotions using FCM and FKM. International Journal of Computers and Communications, 2(1), 21–25.

[4] Schaaff, K., Schultz, T. (2009). Towards an EEG-based emotion recognizer for humanoid robots. In The 18th IEEE international symposium on robot and human interactive communication, Toyama, Japan, Sept 27–Oct 2, 792–796.

[5] Lin, Y. P.,Wang, C. H., Jung, T. P.,Wu, T. L., Jeng, S. K., Duann, J. R., Chen, J. H. (2010). EEG-based emotion recognition in music listening. IEEE Transactions on Biomedical Engineering, 57(7), 1798–1806.

[6] Jing, C., Liu, G., Hao, M. (2009). The Research on Emotion recognition from ECG signal. International conference on information technology and computer science, Kiev, July 25–26.

[7] Richoz, A. R., Lao, J., Pascalis, O., Caldara, R. (2018). Tracking the recognition of static and dynamic facial expressions of emotion across the life span. Journal of Vision, 18(9), 1–27.

[8] Campanella, S., Belin, P. (2007). Integrating face and voice in person perception. Itrends in Cognitive Sciences, 11(12), 535–543.

[9] Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., Lepore, F. (2008). Audio-visual integration of emotion expression. Brain Res, 126–135.

[10] Grimm, M., Kroschel, K., Narayanan, S. (2008). The Vera Am Mittag German audio-visual emotional Speech Database. IEEE international conference on multimedia and expo, Hannover, Germany, June 23–26.

[11] Schuller, B. W. (2018). Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM, 61(5), 90–99.

[12] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. G. (2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18(1), 32–80.

[13] Huahu, X., Jue, G., Jian, Y. (2010). Application of speech emotion recognition in intelligent household robot. International Conference on Artificial Intelligence and Computational Intelligence, 1, 537–541.

[14] Yoon, W. J., Cho, Y. H., Park, K. S. (2007). A study of speech emotion recognition and its application to mobile services. Ubiquitous Intelligence and Computing, Springer, Berlin, Heidelberg, 758–766.

[15] Gupta, P., Rajput, N. (2007). Two-stream emotion recognition for call center monitoring. Interspeech 2007, Antwerp, Belgium, August 27-31, 2241–2244.

[16] Szwoch, M. , Szwoch, W. (2015). Emotion recognition for affect aware video games. Image Processing Communications Challenges 6, Springer International Publishing, 227–236.

[17] Low, L. S. A., Maddage, M. C., Lech, M., Sheeber, L. B., Allen, N. B. (2011). Detection of Clinical Depression in Adolescents' Speech During Family Interactions. IEEE Transactions on Biomedical Engineering, 58(3), 574–586.

[18] Reddy, S. Arundathy, Singh, Amarjot, Kumar, N. Sumanth, Sruthi, K. S. (2011). The decisive emotion identifier. 3rd international conference on electronics computer technology (ICECT), Kanyakumari, India, April 8–10, 28–32.
[19] Ververidis, D., Kotropoulos, C. (2006). Emotional speech recognition: resources, features, and methods. Speech Communication, 48(9), 1162–1181.
[20] Ayadi, M. E., Kamel, M. S., Karray, F. (2011). Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognition, 44(3), 572–587.
[21] Anagnostopoulos, C. N., Iliou, T., Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech : a survey from 2000 to 2011. Artificial intelligence Review, 43(2), 155–177.
[22] Basu, S., Chakraborty, J., Bag, A., Aftabuddin, M. (2017). A review on emotion recognition using speech. International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, Tamilnadu, India, March 10-11, 109–114.
[23] Tayari Meftah, I. Modélisation, détection et annotation des états émotionnels à l'aide d'un espace vectoriel multidimensionnel. PhD thesis, Université de Nice - Sophia Antipolis, France, Avril 2013.
[24] Hùng, L. X. Indexation des émotions dans les documents audiovisuels à partir de la modalité auditive. PhD thesis, Institut Polytechnique de Grenoble, France, Juillet 2009.
[25] Vaudable, C. Analyse et reconnaissance des émotions lors de conversations de centres d'appels. PhD thesis, Université Paris Sud-Paris XI, France, 2012.
[26] Thieeault, M. Les émotions : étude articulatoire, acoustique et perceptive. PhD thesis, Université du Québec à Montréal, Canada, Octobre 2011.
[27] ABDAT, F. Reconnaissance automatique des émotions par données multimodales : expressions faciales et signaux physiologiques. PhD thesis, Université Paul Verlaine de Metz, France, juin 2010.
[28] Ortony, A., Turner, T. J. What's basic about basic emotions ? Psychol Rev, 1990.
[29] Akçay, B. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, 56-76.
[30] Plutchik, R. Emotion, a psychoevolutionary synthesis. New York, 1980.
[31] Leroi-Gourhan, A. (1992). Le geste et la parole ; tome 1 : technique et langage. Collection sciences d'aujourd'hui, Éditions Albin Michel, 324.
[32] Chantir, A. Etude de la Micro prosodie en vue de la Synthèse de la parole en Arabe Standard. PhD thesis, Ecole Nationale Supérieure Polytechnique, Algérie, Octobre 2009.
[33] Buniet, L. Traitement automatique de la parole en milieu bruité : étude de modèles connexionnistes statiques et dynamiques. PhD thesis, Université Henri Poincaré - Nancy 1, France, février 1997.
[34] Delebecque, L. Étude, analyse et modélisation physique de la production de la parole avec applications aux troubles liés à une surdité profonde. PhD thesis, Université de Grenoble, France, aout 2006.
[35] Maalem, H. Les Statistiques d'ordre Supérieur : Application au Traitement du Signal Parole Pathologique (Patients à Audition Déficiente – Patients Trachéotomisés). PhD thesis, Université Mentouri - Constantine, Algérie, Avril 2007.
[36] Yassamine, A. Modélisation AR et ARMA de la Parole pour une Vérification Robuste du Locuteur dans un Milieu Bruité en Mode Dépendant du Texte. Magister, Université Ferhat Abbas – Setif1, Algérie, Octobre 2013.
[37] Makhlouf, A. Reconnaissance automatique de la parole en milieu réel bruité par fusion audiovisuelle. PhD thesis, Université d'Annaba, Algérie, 2016.
[38] Banse, R. Scherer, K. R.(1996). Acoustic Profiles in Vocal Emotion Expression. Journal of Personality and Social Psychology, 70(3), 614- 636.
[39] Burkhardt, F., Sendlmeier, W.(2000). Verification of acoustical correlates of emotional speech using formant-synthesis. Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, UK, September 5-7, 151-156.
[40] Picard, R. W, Affective Computing, MIT Press, Cambridge, 1997.
[41] Breazeal, C. Sociable Machines: Expressive Social Exchange Between Humans and Robots. PhD Thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, University of California at Santa Barbara, May 2000.

[42] Abelin, A., Allwood, J. (2000). Cross-linguistic interpretation of emotional prosody. Proceedings of the ISCA Workshop on Speech and Emotion, Belfast, Ireland, Textflow, Belfast, 1-18.

[43] Tickle, A. (2000). English and Japanese speaker's emotion vocalizations and recognition: a comparison highlighting vowel quality. ISCA Workshop on Speech and Emotion, Newcastle, UK, September 5-7.

[44] Johnstone., Klaus, R., Scherer. (2000). Vocal Communication of Emotion. In Handbook of Emotions, New-York, ÉtatsUnis : Guilford, 220-235.

[45] Juslin., Patrik, N., Laukka, P. (2003). Communication of Emotions in Vocal Expression and Music Performance : Different Channels, Same Code ? Psychological Bulletin, 129(5), 770-814.

[46] Murray., Iain, R., John, L., Arnott. (1993). Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. Journal of the Acoustical Society of America, 93(2), 1097-1108.

[47] Oster., Marie, A., Risberg, A. (1986). The Identification of the Mood of a Speaker by Hearing Impaired Listeners, Quarterly Progress and Status Report, 27(4), 79-90.

[48] Fonagy., Ivan., Magdics, K. (1963). Emotional Patterns in Intonation and Music. Zeitschrift für Phonetik, 16, 293-326.

[49] Scherer. (1989). Vocal Correlates of Emotional Arousal and Affective Disturbance. In Handbook of Psychophysiology : Emotion and Social Behavior, Londres, Royaume Uni, 165-197.

[50] Clavel, C. Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales. PhD thesis, Ecole Nationale Supérieure des Télécommunications, France, mars 2007.

[51] Lee, C. M. Recognizing emotions from spoken dialogs: A signal processing approach. University of Southern California, ProQuest Dissertations Publishing, 2004.

[52] Villon, O, Lisetti, C. L. (2006). A user-modeling approach to build user's psycho-physiological maps of emotions using bio-sensors. 15th IEEE International Symposium on Robot and Human Interactive Communication, Session Emotional Cues in Human-Robot Interaction, Hateld, United Kingdom, September 2006.

[53] Nasoz, F., Alvarez, K., Lisetti, C. L., Finkelstein, N. (2003). Emotion recognition from physiological signals for user modeling of affect. 9th International Conference on User Model, Pittsburg, USA, June 22-26.

[54] Chanel, G., Ansari-Asl, K., Pun, T. (2007). Valence-arousal evaluation using physiological signals in an emotion recall paradigm. 2007 IEEE International Conference on Systems, Man and Cybernetics, 2662-2667, Montreal, Canada, October 7-10.

[55] Haag, A., Goronzy. S., Schaich, P., Williams, J. (2004). Emotion Recognition Using Bio-Sensors: First Steps Towards an Automatic System. Affective Dialogue Systems, 36-48.

[56] Vaudable, C., Rollet, N., Devillers, L. (2010). Annotation of affective interaction in real-life dialogs collected in a call-center. International Conference of Language Resources and Evaluation, Workshop on Emotion and Affect, Valetta, Malta, May 23.

[57] Devillers, L., Vidrascu, L. and Lamel, L. (2005). Challenges in real-life emotion annotation and machine leraning based detection. Journal of Neural Networks, Special Issue on Emotion and Brain, 18(4), 407-422.

[58] Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M., Wong, M. (2004). "You stupid tin box" - children interacting with the AIBO robot: A cross linguistic emotional speech corpus. Proceedings of the 4th International Conference of Language Resources and Evaluation LREC 2004, Lisbon, Portugal, 171-174.

[59] Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech communication, 40(1), 227-256.

[60] Banziger, T, Scherer. K. R. (2010). Introducing the Geneva multimodal emotion portrayal (GEMEP) corpus. A Blueprint for Affective Computing: A sourcebook and manual, New York : Oxford University Press, 271-294.

[61] Engberg, I. S., Hansen, A. V., Andersen, O., Dalsgaard, P. (1997). Design, Recording and Verification of a Danish Emotional Speech Database, EUROSPEECH'97: 5th European Conference on Speech Communication and Technology, Rhodes, Greece, September 22-25, 1695-1698.

[62] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. and WeissI, B. A. (2005). Database of German emotional speech. Interspeech, Lisbon, Portugal, 1517-1520.

[63] Audibert, N., Aubergé, V., Rilliard, A. (2004). Ewiz : contrôle d'émotions authentiques. Dans Actes des Journées d'Étude sur la Parole, Fès, Maroc, 49–52.

[64] Cowie, Roddy, Douglas-Cowie, E., Sneddon, I., McRorie, M., Hanratty, J., McMahon, E., McKeown, G. (2010). Induction techniques developed to illuminate relationships between signs of emotion and their context, physical and social. A Blueprint for Affective Computing : A sourcebook and manual, New York : Oxford University Press, 295-307.

[65] Attabi, Y. Reconnaissance automatique des émotions spontanées à partir du signal de parole. PhD thesis, Université du Québec, Montréal, Canada, 2015.

[66] Douglas-Cowie, E., Cowie, R., Schröder, M. (2000). A new emotion database: Considerations, sources and scope. Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Belfast, Textflow, September 5-7, 39-44.

[67] Schuller, B., Müller, R., Lang, M., Rigoll, G. (2005). Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. Interspeech, Special Session : Emotional Speech Analysis and Synthesis : Towards a Multimodal Approach , Lisbon, Portugal, September 4-8, 805-809.

[68] University of Pennsylvania Linguistic Data Consortium, Emotional prosody speech and transcripts/http://www.ldc.upenn.edu/Catalog/CatalogEntry. Jsp ?catalogId=LDC2002S28S, July 2002.

[69] Nwe, T., Foo, S., De Silva, L., (2003). Speech emotion recognition using hidden Markov models. Speech Communications Journal, 41(4), 603-623.

[70] Hozjan, V., Moreno, Z., Bonafonte, A., Nogueiras, A. (2002). Interface databases: design and collection of a multilingual emotional speech database. Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02), Las Palmas de Gran Canaria, Spain, May, 2019–2023.

[71] Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., Rigoll, G. (2005). Speaker independent speech emotion recognition by ensemble classification. IEEE International Conference on Multimedia and Expo (ICME), Amsterdam, Netherlands, July 6-9, 864–867.

[72] Breazeal, C., Aryananda, L. (2002). Recognition of affective communicative intent in robot-directed speech. Autonomous Robots, 12(1), 83–104.

[73] Zhou, J., Wang, G., Yang, Y., Chen, P. (2006). Speech emotion recognition based on rough set and svm. 5th IEEE International Conference on Cognitive Informatics, ICCI 2006, Beijing, July 17-19, 1, 53–61.

[74] Oflazoglu, C. , Yildirim, S. (2013). Recognizing emotion from Turkish speech using acoustic features. EURASIP Journal on Audio, Speech, and Music Processing, 1, 1-11.

[75] Morrison, D., Wang, R., De Silva, L. (2007). Ensemble methods for spoken emotion recognition in call-centres. Speech Communications, 49(2), 98–112.

[76] Mohamad Nezami, O., Jamshid Lou, P., Karami, M. (2018). ShEMO: a large-scale validated database for Persian speech emotion detection. Language Resources and Evaluation, 53(1), 1–16.

[77] Singh, R., Puri, H., Aggarwal, N., Gupta, V. (2019). An Efficient Language-Independent Acoustic Emotion Classification System. Arabian Journal for Science and Engineering, 45(4), 3111–3121.

[78] Abdel-Hamid, L. (2020). Egyptian Arabic Speech Emotion Recognition using Prosodic. Spectral and Wavelet Features. Speech communication, 122, 19-30.

[79] Meddeb, M., Hichem, K., Alimi. A. (2015). Speech Emotion Recognition Based on Arabic Features. 15th international conference on Intelligent Systems design and Applications (ISDA15), Marrakesh, Morocco, IEEE conference, December 14-16.

[80] Staroniewicz, P., Majewski, W. (2009). Polish Emotional Speech Database – Recording and Preliminary Validation. Lecture Notes in Computer Science, 42-49.

[81] Fónagy, I. (1983). La vive voix. Langage et société, 26, 65-69.

[82] Huang, X., Acero, A., Hon, H., Foreword By-Reddy, R. (2001). Spoken language processing: A guide to theory, algorithm, and system development. Prentice Hall PTR, 1, 2005.

[83] Frick, R. W. (1985). Communicating emotion : the role of prosodic features. Psychol. Bull, 97(3), 412–429.

[84] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. G. (2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18(1), 32–80.

[85] Gharsallaoui, S. Détection et classification de traits paralinguistiques par des métriques prythmique de la parole. PhD thesis, Univercité du Québec à Trois- Rivière, Canada, juillet 2016.

[86] Dellaert, F., Polzin, T., Waibel, A. (1996). Recognizing emotions in speech. Proceeding of Fourth International Conference on Spoken Language Processing (ICSLP), Philadelphia, PA, USA, October 3-6, 1970–1973.

[87] Lee, C. M., Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. IEEE Transactions on Audio, Speech, and Language Processing, 13(2), 293–303.

[88] Nwe, T. L., Foo, S.W., Silva, L. C. D. (2003). Speech emotion recognition using hidden Markov models. Speech Communication, 41(4), 603–623.

[89] Schroder, M. (2001). Emotional speech synthesis: A review. Seventh European conference on speech communication and technology, Eurospeech Aalborg, Denmark, September 3-7.

[90] Murray, I. R., Arnott, J. L. (1995). Implementation and testing of a system for producing emotion by rule in synthetic speech. Speech Communication, 16(4), 369–390.

[91] Iida, A., Campbell, N., Higuchi, F., Yasumura, M. (2003). A corpus based speech synthesis system with emotion. Speech Communication, 40(1-2), 161–187.

[92] Rao, K. S., Koolagudi, S. G., Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. International Journal of Speech Technology, 16 (2), 143–160.

[93] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. G. (2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine, 18(1), 32–80.

[94] Gouyon, F., Pachet, F., Delerue, O. (2000). On the use of zero-crossing rate for an application of classification of percussive sounds. Proceedings of COST G6 Conference on Digital Audio Effects, Verona, Italy, December 7-9, 147-152.

[95] Giannakopoulos, T., Pikrakis, A., Theodoridis, S. (2009). A dimensional approach to emotion recognition of speech from movies. Proceedings of IEEE international conference on acoustics, speech and signal processing, Taipei, Taiwan, April 19-24, 65–68.

[96] Atassi, H., Esposito, A. (2008). A speaker independent approach to the classification of emotional vocal expressions. Proceedings of 20th IEEE international conference on tools with artificial intelligence, Dayton, Ohio, USA, November, 2, 147–152.

[97] Alku, P., Backstrom, T. and Vilkman, E. (2002). Normalized amplitude quotient for parameterization of the glottal flow. Journal of the Acoustical Society of America, 112(2), 701-710.

[98] Zhang, S. (2008). Emotion recognition in Chinese natural speech by combining prosody and voice quality features. International Symposium on Neural Networks, Springer, Berlin/Heidelberg, Germany, September 457–464.

[99] Li, X. , Tao, J. , Johnson, M. T. , Soltis, J. , Savage, A. , Leong, K. M. , Newman, J. D. (2007). Stress and emotion classification using jitter and shimmer features. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, Hawaii, April 15-20, 1081-1084.

[100] Lalitha, S., Madhavan, A., Bhushan, B., Saketh, S., (2015). Speech emotion recognition. International Conference on Advances in Electronics, Computers and Communications (ICAECC), October, 1–4.

[101] Vasquez-Correa, J. C., Garcia, N., Orozco-Arroyave, J. R., Arias-Londono, J. D.,Vargas-Bonilla, J. F., Noth, E. (2015). Emotion recognition from speech under environmental noise conditions using wavelet decomposition. International Carnahan Conference on Security Technology (ICCST), Taipei,Taiwan, September 21-24, 247-252.

[102] Kuchibhotla, S., Vankayalapati, H., Vaddi, R., Anne, K. R. (2014). A comparative analysis of classifiers in emotion recognition through acoustic features. International Journal of Speech Technology, 17(4), 401–408.

[103] Neiberg, D., Elenius, K., Laskowski, K. (2006). Emotion recognition in spontaneous speech using GMMs. International Conference on Spoken Language Processing (ICSLP), Pittsburgh, Pennsylvania, September 17–19, 809–812.

[104] Sinith, M. S., Aswathi, E., Deepa, T. M., Shameema, C. P., Rajan, S. (2016). Emotion recognition from audio signals using Support Vector Machine. 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), December 10-12, 139–144.

[105] Pao, T.L., Chen, Y.T., Yeh, J.-H., Liao, W.Y. (2005). Combining Acoustic Features for Improved Emotion Recognition in Mandarin Speech. Lecture Notes in Computer Science, 279–285.

[106] Kamaruddin, N., Wahab, A. (2009). Features extraction for speech emotion. Journal of Computational Methods in Science and Engineering, 9(9), 1–12.

[107] Nicholson, J., Takahashi, K., Nakatsu, R. (2000). Emotion Recognition in Speech Using Neural Networks. Neural Computing Applications, 9(4), 290–296.

[108] Mannepalli, K., Sastry, P. N., Suman, M. (2017). Analysis of Emotion Recognition System for Telugu Using Prosodic and Formant Features. Speech and Language Processing for Human-Machine Communications, 137-144.

[109] Koolagudi, S. G., Murthy, Y. V. S., Bhaskar, S. P. (2018). Choice of a classifier, based on properties of a dataset : case study-speech emotion recognition. International Journal of Speech Technology, 21(1), 167–183.

[110] Kamińska, D., Sapiński, T., Anbarjafari, G. (2017). Efficiency of chosen speech descriptors in relation to emotion recognition. EURASIP Journal on Audio, Speech, and Music Processing, 1, 1-9.

[111] Chatterjee, J., Mukesh, V., Hsu, H.H., Vyas, G., Liu, Z. (2018). Speech emotion recognition using cross-correlation and acoustic features. 16th International Conference on Dependable, Autonomic and Secure Computing, 16th International Conference on Pervasive Intelligence and Computing, 4th International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), 243–249.

[112] Horkous, H., Guerti, M. (2019). Use of differents Classifiers for Recognition of Fear Emotions in speech. Models Optimisation and Mathematical Analysis Journal, 7(1), 21-25.

[113] Horkous, H., Guerti, M. (2018). Speech Emotions Recognition of Joy and Sadness Based on Prosodic and MFCCs parameters. Models Optimisation and Mathematical Analysis Journal, 6(1), 15-18.

[114] Devillers, L., Lamel, L. (2003). Emotion detection in task-oriented dialogs. In Proceedings of the International Conference on Multimedia and Expo (ICME), USA, July 6-9, 3, 549–552.

[115] Gorin, A. (1995). On automated language acquisition. The Journal of the Acoustical Society of America, 97(6), 3441-3461.

[116] Devillers, L., Lamel, L. (2003). Emotion detection in task-oriented dialogs. In Proceedings of the International Conference on Multimedia and Expo (ICME), USA, July 6-9, 3, 549–552.

[117] Ang, J, Dhillon, R., Krupski, A., Shriberg, E.,Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human computer dialog. 7th International Conference on Spoken Language Processing (ICSLP), Denver, Colorado, USA, September 16-20, 2037-2040.

[118] Lee, C. M., Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing, 13(2), 293-303.

[119] Planet, S., Iriondo, I. (2012). Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition. Conference: Information Systems and Technologies (CISTI), 2012 7th Iberian Conference, Madrid, Spain, June 20-23, 1-6.

[120] Schuller, B., Müller, R., Manfred, K. L., Rigoll, G. (2005). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles, Interspeech, 805-808.

[121] Vidrascu, L. Analyse et détection des émotions verbales dans les interactions orales. PhD thesis, Université Paris-Sud 11, décembre 2007.

[122] Boufaden, N., Dumouchel, P. (2008). Leveraging emotion detection using emotions from yes-no answers. Interspeech, Brisbane, Australia, September 22-26, 241-244.

[123] Devillers, L., Vasilescu. I., Lamel, L. (2002). Annotation and detection of emotion in a task oriented human-human dialog corpus. International Standards for Language Engineering, Edinburgh, December 16-17.

[124] Batliner, A., Fischer, K., Huber, R., Spiker, J., Noth, E.(2000). Desperately seeking emotions: actors, wizards and human beings. Proceedings of the ISCA Workshop Speech Emotion, 195–200.

[125] Chen, L. S., Tao, H., Huang, T. S., Miyasato, T., Nakatsu. R.(1998). Emotion recognition from audiovisual information. IEEE Workshop on Multimedia Signal Processing, Los Angeles, CA, USA, Dec 7-9, 83–88.

[126] Lee, C., Narayanan, S., Pieraccini, R. (2002). Classifying emotions in human–machine spoken dialogs. Proceedings of the IEEE International Conference on Multimedia Expo (ICME), Lausanne, Switzerland, 1, 737–740.

[127] Shen, P., Changjun, Z., Chen, X. (2011). Automatic speech emotion recognition using support vector machine. International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT), 2, 621–625.

[128] Pan, Y. , Shen, P. , Shen, L.(2012). Speech emotion recognition using support vector machine. International Journal of Smart Home, 6(2), 101–108.

[129] Batliner, A., Fisher, K., Huber, R., Spilker, J., Nöth, E. (2000). Desperately seeking emotions or : Actors, wizards and human beings. Proceedings of ISCA Workshop on Speech and Emotion, Belfast, United Kingdom, 195–200.

[130] Neiberg, D., Elenius, K., Laskowski, K. (2006). Emotion recognition in spontaneous speech using gmms. Ninth International Conference on Spoken Language Processing (ICSLP), Pittsburgh, PA, USA September 17-21, 809-812.

[131] Luengo, I., Navas, E., Hernez, I., Snchez, I. (2005). Automatic emotion recognition using prosodic parameters. Interspeech, Lisbon, Portugal, September 4-8, 493–496.

[132] Kuchibhotla, S., Vankayalapati, H. D., Anne, K. R. (2016). An optimal two stage feature selection for speech emotion recognition using acoustic features. International Journal of Speech Technology, 19(4), 657–667.

[133] Nwe, T. L., Foo, S. W., De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. Speech Communication, 41(4), 603–623.

[134] Swain, M., Sahoo, S., Routray, A., Kabisatpathy, P., Kundu, J. N. (2015). Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition. International Journal of Speech Technology, 18(3), 387–393.

[135] Rong, J., Li, G., Chen, Y. P. P. (2009). Acoustic feature selection for automatic emotion recognition from speech. Information Processing Management, 45(3), 315–328.

[136] Dan Zbancioc, M., Feraru, S. M. (2015). A study about the automatic recognition of the anxiety emotional state using Emo-DB. E-Health and Bioengineering Conference (EHB), Iasi, Romania, November 19-21.

[137] Firoz, S. A., Raji, S. A., Babu, A. P. (2009). Automatic emotion recognition speech using artificial neural networks with gender dependent databases. International Conference on Advances in Computing, Control, and Telecommunication Technologies, Trivandrum, Kerala, India, December, 162–164,

[138] Nicholson, J., Takahashi, K., Nakatsu, R. (2000). Emotion recognition in speech using neural networks. Neural Computing Applications, 9(4), 290–296.

[139] Trigeorgis, G. , Ringeval, F. , Brueckner, R. , Marchi, E. , Nicolaou, M. A. , Schuller, B. , Zafeiriou, S. (2016). Adieu features ? end-to-end speech emotion recognition using a deep convolutional recurrent network. International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5200–5204.

[140] Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., Cowie, R. (2008).Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. Proceedings Interspeech, 2008, 9th Annual Conference of the International Speech Communication Association, incorporating 12th Australasian International Conference on Speech Science and Technology (SST), Brisbane, Australia, January, 597–600.

[141] Kaya, H., Fedotov, D., Yesilkanat, A., Verkholyak, O., Zhang, Y., Karpov, A.(2018). Lstm based cross-corpus and cross-task acoustic emotion recognition. Interspeech, Hyderabad, India, September 2-6, 13, 521–525.

[142] Kuchibhotla, S., Vankayalapati, H. D., Anne, K. R. (2016). An optimal two stage feature selection for speech emotion recognition using acoustic features. International Journal of Speech Technology, 19(4), 657–667.

[143] Wang, Y., Guan, L. (2004). An investigation of speech-based human emotion recognition. IEEE 6th workshop on multimedia signal processing, Siena, Italy September 29- October 1, 15–18.

[144] Deusi, J. S., Popa, E. I.(2019). An Investigation of the Accuracy of Real Time Speech Emotion Recognition. International Conference on Artificial Intelligence, Cambridge, UK, December, 336-349.

[145] Lee, C. C., Mower, E., Busso, C., Lee, S., Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. Speech Communication, 53 (9–10), 1162–1171.

[146] Rajisha, T. M., Sunija, A. P., Riyas, K. S. (2016). Performance Analysis of Malayalam Language Speech Emotion Recognition System Using ANN/SVM. Procedia Technology, 24, 1097–1104.

[147] Schuller, B., Rigoll, G., Lang, M.(2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal,1 , 577– 580.

[148] Dellaert, F., Polzin, T., Waibel, A. (1996). Recognizing emotion in speech. Proceeding of Fourth International Conference on Spoken Language Processing(ICSLP), Philadelphia, PA, USA, October 3-6, 1970-1973.

[149] Vidrascu, L., Devillers, L. (2005). Detection of real-life emotions in call centers. Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbonne, September 4-8, 97-110.

[150] Shahin, I., Nassif, A. B., Hamsa, S. (2019). Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network. IEEE Access, 7, 26777–26787.

[151] Mannepalli, K., Sastry, P. N., Suman, M.(2018). Emotion recognition in speech signals using optimization based Multi-SVNN classifier. Journal of King Saud University, Computer and Information Sciences. Computer and Information Sciences, 32(10), 1218

[152] Horkous, H., Guerti, M. (2020). Study the Influence of Gender and Age in Recognition of Emotions from Algerian Dialect Speech. Traitement du Signal, 37(3), 413-423.

[153] Shashidhar, G., Koolagudi, K., Sreenivasa, R. (2012). Emotion recognition from speech: a review. International Journal of Speech Technology, 15(2), 99–117.

[154] Lee, C., Narayanan, S., Pieraccini, R. (2002). Classifying emotions in human machine spoken dialogs. IEEE International Conference on Multimedia and Expo(ICME), Lausanne, Switzerland, Aug 26-29.

[155] Ramakrishnan,S., Emary, I.M.M.E.I. (2013). Speech emotion recognition approaches in human computer interaction. Telecommunication Systems, 52(3), 1467–1478.

[156] Li, W., Zhang, Y., Fu, Y. (2007). Speech emotion recognition in e-learning system based on affective computing. Third International Conference on Natural Computation (ICNC), Haikou, Hainan, China, Aug 24-27, 5, 809-813.

[157] Zhu, A., Luo, Q. (2007). Study on speech emotion recognition system in E-learning ». Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments, Springer, 4552, 544- 552.

[158] Clavel, C., Vasilescu, I., Devillers, L., Ehrette, T., Richard, G. (2006). Safe corpus : fear-type emotions detection for surveillance application. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, May, 1099-1104.

[159] Kwon, H., Berisha, V., Spanias, A. (2008). Real-time sensing and acoustic scene characterization for security applications. 3rd International Symposium on Wireless Pervasive Computing, Santorini, Greece, May 7-9, 755-758.

[160] Clavel, C., Vasilescu, L., Devillers, L., Richard, G., Ehrette, T. (2008). Fear-type emotion recognition for future audio-based surveillance systems. Speech Communication, 50(6), 487- 503.

[161] Devillers, L., Vidrascu, L., Layachi, O. (2010). Automatic detection of emotion from vocal expression. In A Blueprint for Affective Computing : A sourcebook and manual, Series in Affective Science, 132-144.

[162] Malandrakis, N., Potamianos, A., Evangelopoulos, G., Zlatintsi, A. (2011). A supervised approach to movie emotion tracking. IEEE International Conference on in Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, May 22-27, 2376-2379.

[163] Xu, M., Chia, L. T., Jin, J. (2005). Affective content analysis in comedy and horror videos by audio emotional event detection. IEEE International Conference on Multimedia and Expo (ICME), Amsterdam, Netherlands, July 6-8.

[164] Varadarajan, V., Hansen, J., Ayako, I. (2006). Ut-scope - a corpus for speech under cognitive/physical task stress and emotion. Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, May 22-28.

[165] Ten Bosch, L. (2003). Emotions, speech and the as framework. Speech Communication, 40(1-2), 213–225.

[166] Devillers, L., Vidrascu, L. (2006). Real-life emotions detection on Human- Human spoken dialogs. Proceedings of IPMU'08, Málaga, Spain, June 22–27, 1590–1596.

[167] Marchi, E., Schuller, B., Batliner, A., Fridenzon, S., Tal, S. Golan, O. (2012). Emotion in the speech of children with autism spectrum conditions : Prosody and everything else. In Proceedings 3rd Workshop on Child, Computer and Interaction (WOCCI), Portland, Oregon, U.S.A., September 14.

[168] Breazeal, C., Aryananda, L. (2002). Recognizing affective intent in robot directed speech. Autonomous Robots, 12(1), 83–104.

[169] Pelachaud, C. (2005). Multimodal expressive embodied conversational agent. Proceedings of the 13th annual ACM international conference on Multimedia, Singapour, November 6–11, 683–689.

[170] Jones, C., Sutherland, J. (2008). Affect and emotion in human computer interaction: From theory to applications. chapitre Acoustic Emotion Recognition for Affective Computer Gaming, 209-219.

[171] Schoentgen, J. (2006). Vocal cues of disordered voices: an overview. Acta Acustica United with Acustica, 92(5), 667-680.

[172] Palva, H. (2006). Dialects: classification. Encyclopedia of Arabic Language and Linguistics, 1, 604–613.

[173] Pereira, C. (2011). Arabic in the North African Region. Semitic Languages, An International Handbook, Berlin, 944–959.

[174] Meftouh, K., Bouchemal, N., Smali, K. (2012). A study of a non-resourced language: an Algerian dialect. Proceedings of the third international workshop on spoken language technologies for underresourced languages (SLTU), Cape Town, South Africa, May 2012, 1-7.

[175] Bougrine, S., Cherroun, H., Ziadi, D., Lakhdari, A. Chorana, A. (2016). Toward a Rich Arabic Speech Parallel Corpus for Algerian sub-Dialects. Proceedings of the 2nd Workshop on Arabic Corporaand Processing Tools, Portorož, Slovenia, May 27, 2-10.

[176] Djellab, M., Amrouche, A., Bouridane, A., Mehallegue, N. (2016). Algerian Modern Colloquial Arabic Speech Corpus (AMCASC) : regional accents recognition within complex socio-linguistic environments. International Journal of Language Resources and Evaluation, 51(3), 1-29.

[177] Harrat, S., Meftouh, K., Abbas, M., Smaili, K. (2014). Building Resources for Algerian Arabic Dialects. 15th Annual Conference of the International Communication Association Interspeech (ISCA), Singapour, Singapore, September 2014.

[178] Bougrine, S., Cherroun, H., Ziadi, D. (2015). Prosody-based Spoken Algerian Arabic Dialect Identification. International Conference on Natural Language and Speech Processing (ICNLSP), Procedia Computer Science, 9–17.

[179] Dahmani, H., Hussein, H., Meyer-Sickendiek, B., Jokisch, O. (2019). Natural Arabic Language Resources for Emotion Recognition in Algerian Dialect. Arabic Language Processing : From Theory to Practice, Springer International Publishing, 18-33.

[180] Ayadi, M. E., Kamel, M. S., Karray, F. (2011). Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognition, 44(3), 572–587.

[181] Boersma, P., Weenink, D. (2002). Praat, a system for doing phonetics by computer. Glot International, 5(9/10), 341-345.

[182] Goudbeek , M., Goldman, J. P., Klaus, R. Scherer. (2009). Emotion dimensions and formant position. INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10.

[183] Zhang, G., Yin, Q., Yang, C. (2011). The Fixed-Point Optimization of Mel Frequency Cepstrum Coefficients for Speech Recognition. IEEE International Forum on Strategic Technology, Heilongjiang, Harbin, China, Aug 22-24, 1172-1175.

[184] Jamoussi, S. Méthodes statistiques pour la compréhension automatique de la parole. Thèse de doctorat, l'université Henri Poincaré-Nancy 1, France, 2004.

[185] Thirumuruganathan, S. A detailed introduction to k-nearest neighbor (knn) algorithm, may 2010.

[186] Hilali, H. Application de la classification textuelle pour l'extraction des règles d'association maximales. Thèse de maitrise en informatique, Université du Québec à Trois-rivières, Trois-rivières, Canada, 2009.

[187] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor non parametric regression. The American Statistician, 46(3), 175–185.

[188] Descôteaux, S. Les règles d'association maximale au service de l'interprétation des résultats de la classification. Thèse de maitrise en informatique, université du Québec à Trois-Rivières, Trois-Rivières, 2014.

[189] Demircan, S. Kahramanli, H. (2016). Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech. Neural Computing and Applications, 29(8), 59–66.

[190] Vapnik, V. The Nature of Statistical Learning Theory. Information Science and Statistics, Springer, 1995.

[191] Hsu, C. W., Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks, 13(2), 415–425.

Références Bibliographiques

[192] Djeffal, A. Utilisation des méthodes Support Vector Machine (SVM) dans l'analyse des bases de données. Thèse de doctorat, Université Mohamed Khider-Biskra, Algérie, 2012.

[193] Kerkeni, L., Serrestou, Y., Raoof, K., Mbarki, M., Mahjoub, A., Cleder, C. (2019). Automatic Speech Emotion Recognition using an Optimal Combination of Features based on EMD-TKEO, Speech Communication, 114, 22-35.

[194] Kuchibhotla, S., Vankayalapati, H. D., Vaddi, R. S., Anne, K. R. (2014). A comparative analysis of classifiers in emotion recognition through acoustic features. International Journal of Speech Technology, 17(4), 401–408.

[195] Fu, L., Wang, C., Zhang, Y. (2010). A study on influence of gender on speech emotion classification. 2010 2nd International Conference on Signal Processing Systems, Dalian, China, July 5–7, 2, V1-534 - V1-537.

[196] Vogt, T., Andre, E. (2006). Improving automatic emotion recognition from speech via gender differentiation. Proceedings of Language Resources and Evaluation Conference (LREC 2006), Genoa, Italy, May, 1123-1126.

[197] Horkous, H., Guerti, M.(2021). Recognition of Anger and Neutral Emotions in Speech with Different Languages. International Journal of Computing and Digital Systems, 10(1).

[198] Zhou, Y., Sun, Y., Zhang, J., Yan, Y. (2009). Speech Emotion Recognition Using Both Spectral and Prosodic Features. 2009 International Conference on Information Engineering and Computer Science, Wuhan, China, December 19-20.