الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
Université Akli Mohand Oulhadj - Bouira -
Tasdawit Akli Muḥend Ulḥağ - Tubirett -

وزارة التعليم العالي والبحث العلمي
جامعة أكلي محند أولحاج
-البويرة-

جامعة البويرة

Faculté des Sciences et des Sciences Appliquées

كلية العلوم والعلوم التطبيقية

Référence : ...../MM/2021

المرجع:...../م م/ 2021

Mémoire de Master

Présenté au

Département: Génie Électrique
Domaine: Sciences et Technologies
Filière: Télécommunications
Spécialité: Systèmes des Télécommunications

Réalisé par :

Silem Amira Et **Daoui Hibatellah**

Thème

Reconnaissance automatique des émotions par la voix

Soutenu le**:06/07/2022**

Devant la commission composée de :

Dr : CHELBI SALIM	M.C.B	Univ. Bouira	Président
Dr : ABDENNOUR ALIMOHAD	M.C.B	Univ. Bouira	Rapporteur
Mr : SAIDI MOHAMED	M.A.A	Univ. Bouira	Examinateur

Année Universitaire: 2020-2021

Dédicaces

Je dédie

Ce modeste travail comme un témoignage d'affectation, de respect et d'admiration

A ma chère mère ***Aicha***

La femme qui m'a toujours encouragé à Poursuivre Mes études, source de tendresse. Tous les mots ne Pourront exprimer L'amour que je te porte. Tu es mon bonheur Et ma raison de vivre. Qu'dieu te garde pour moi pour toujours.

A mon cher père ***Djelloul***

Pour son soutien, son amour, son affection et la confiance qu'il ma accordé.

A ma chère sœur ***Nadia*** *et son époux*

Pour tous les sacrifices qu'ils n'ont cessé de m'apporter tout au long de mes études .que dieu leur apporte le bonheur, les aide à réaliser tous leurs vœux et leur offre et un avenir pleine de succès

A mes chers frères ***Amar*** *et* ***Ouassim***

Pour leur aide tout en long de mon chemin, grâce à leur amour, leur compréhension et leur patience.

A mon cher ami ***Mourad***

Nulle dédicace ne pourrait exprimer ma profonde affection et mon immense gratitude pour tous les encouragements et soutiens qu'il a consentis à mon égard

A ma nièce ***Rahil*** *Mon plus beau cadeau de cette année*

A ma chère binômes ***Amira***

Pour son amour, son compréhension et son aide pour terminer ce travail.

Hiba

Dédicaces

Je tiens c'est avec grande plaisir que je dédie ce travail

A mon cher père ***Madjid***

Qui n'a pas cessée de me conseiller, encourager et soutenir tout au long de mes études et qui a été derrière moi pour me guider dans la bonne direction pour que je puisse atteindre mes objectifs.

A ma chère mère ***Faroudja***

Qui m'a soutenu, encouragé et donner l'amour et a vivacité durant toute ma vie. je te remercier forcement pour ton affection me couvre, ta bienveillance me guide et ta présence à mes côtés a toujours été ma source de force pour affronter les différents obstacles.

Que dieu garde mes parents et je les souhaite plus de bonheur et de joie dans leurs vie.

A ma seule sœur ***Kahina***

Qui a partagé avec moi tous les moments d'émotion lors de la réalisation de ce travail

A mon fiancé ***Hamza***

Qui était présent à mes côtés durant tout au long de ce travail et ma guide et encourager a tout moment.

A ma chère famille ceux qui me donnent de l'amour, Que Dieu leur donne une longue et joyeuse vie.

A tous mes amis qui m'ont toujours encouragé et à qui je souhaite plus de succès.

Sans oublier ma binôme ***Hiba***

Pour son soutien moral, sa patience et sa sympathie tout au long de ce projet.

Amira

Remerciements

Ce travail a été effectué au sein du Département de génie electrique de l'Université de Bouira.

On remercie dieu le tout puisant de nous avoir donné la santé et la volonté d'entamer et de terminer ce mémoire.

Tous d'abord, ce travail ne serait pas aussi riche et n'aurait pas pu avoir le jour sans l'aide et l'encadrement de Mr Abdennour ALIMOHAD, on le remercie pour la qualité de son encadrement exceptionnel, pour sa patience, sa rigueur et sa disponibilité durant notre préparation de ce mémoire.

Nous remercions ensuite l'ensemble des membres de jury Mr Salim CHELBI et Mr Mohamed SAIDI qui nous ont fait l'honneur d'accepter d'évaluer et bien vouloir étudier avec attention notre travail.

Notre remerciement s'adresse également à tous nos amis pour leur soutien moral et leurs encouragements.

Résumé

La parole est l'un des moyens linguistiques les plus couramment utilisés par les humains pour transmettre les états émotionnels internes. Par conséquent, un système capable de la reconnaissance automatiquement des émotions humaines serait intéressant. Notre mémoire vise, donc, à concevoir un système de reconnaissance émotionnelle par la voix. Le système consiste à utiliser les trois paramètres spectraux MFCC, LPCC, PLP ainsi que les paramètres prosodiques ; énergie, formants et pitch. Chaque type d'émotion sera modélisé par deux techniques : le modèle GMM-UBM et DTW. Plusieurs tests ont été effectués afin de trouver le meilleur taux de reconnaissance correct. En termes de paramètre, la technique PLP a réalisé le meilleur résultat avec 87.50%. Une amélioration additionnelle des performances du système a été obtenue par la fusion entre les paramètres prosodiques et spectraux qui correspond à un taux de 89,58%.

Mots clés : Reconnaissance, émotions, paramètres prosodiques, DTW, paramètres spectraux, GMM-UBM.

Table des Matières

Chapitre 2 : Système de reconnaissance automatique des émotions

Liste des Figures

Liste des Tableaux

Listes des Abréviations

RAE	Reconnaissance Automatique des Emotions
ATM	Asynchronous Transfer Mode
AMDF	Average Magnitude Difference Function
TFD	Transformeé de Fourier Discrete
LPC	Linear Prediction Cepstral
LPCC	Linear Prediction Cepstral Coefficients
MFCC	Mel-Frequency Cepstral Coefficients
PLP	Perceptual Linear Prediction
FFT	Fast Fourier Transfer
DCT	Discret Cosinus Transform
GMM	Gaussian Mixture Model
HMM	Hidden Markov Models
VQ	Vector Quantisation
DTW	Dynamic Time Warping
EM	Expectation Maximisation
MAP	Maximum à posteriori
UBM	Universal Background Model

Introduction Générale

Les humains peuvent ressentir l'état émotionnel de leur partenaire de communication à travers leurs sens. Cette sensation émotionnelle est naturelle pour les humains, mais c'est une tâche très difficile pour les ordinateurs.

Avec le développement de la technologie actuel où le monde est réduit à un petit village semi-connecté, de nombreux nouveaux besoins sont apparus et sont même devenus vitaux. Nous citons, par exemple, dans les véhicules il est necessaire d'utiliser un système qui detecte si le conducteur est en état qui le permet pas de conduir et l'alerter sur les conséquences de cette état, la nécessite d'accéder à distance et de détecter certaines maladies surtout dans la médecine psychologiques, aussi il existe des entreprises qui utilise des applications pour mesurer la satisfaction des clients a partir de leur émotion. Ces types de tâches seraient probablement plus accessibles si elles pouvaient être faites oralement avec une utilisation minimale d'outils comme le clavier et la souris.

Dans cette optique, la mise en œuvre d'un système automatique est nécessaire. Nous nous intéressons, dans ce travail à la reconnaissance automatique des émotions par la voix. C'est un système qui permet de classer différents fichiers audio en différentes émotions telles que la neutralité, le calme, le bonheur, la tristesse, le colère, la surprise par les ordinateurs [1].

Dans ce travail, la reconnaissance automatique des émotions (RAE) utilise la voix comme source principale et unique d'information. Le système RAE comporte deux phases essentielles ; La phase d'apprentissage pour créer les modèles des émotions et la phase de test permettent de reconnaitre une émotion par rapport aux modèles existants. Toutes ces phases font appel à une multitude d'étapes comme l'extraction de paramètres dans le but de choisir les caractéristiques efficaces permettant d'avoir un système RAE à haute performance, la modélisation pour créer des modèles représentant chaque émotion enregistrés dans une base de données, et enfin les tests vont comparer des voix d'entrée avec les modèles pour identifier l'émotion. Dans notre travail la performance du système fait par un critère bien déterminé qui est le taux de reconnaissance correct.

Ce mémoire est divisé en trois chapitres, nous consacrons le premier à des généralités sur l'émotion, la production de la parole et la reconnaissance émotionnelle, ainsi que ses applications. Par la suite, le deuxième chapitre est consacré à l'étude du système de reconnaissance automatique d'émotion, où nous décrivons les paramètres prosodiques (Pitch, Energie, formant) et les paramètres spectraux (MFCC, LPC, LPCC, PLP). Ensuite nous allons introduire les méthodes de modélisation les plus utilisées. Dans le troisième chapitre, nous mettrons en pratique le système RAE par le canal vocal sous MATLAB et nous présenterons les résultats des différentes expériences effectuées.

Ce travail sera finalisé par une conclusion générale.

1.Introduction

Dans notre vie quotidienne, il existe de nombreuses formes de communication, par exemple : le langage corporel, le langage textuel, le langage pictural et la parole. Cependant, parmi ces formes, la parole est toujours considérée comme la forme la plus puissante en raison de sa dimensionnalité : le texte du discours, la référence au sexe, l'attitude, l'état de santé, l'identité d'un locuteur et l'émotion. Ces informations sont très importantes pour une communication efficace [2].

Dans le développement du domaine homme-machine, la parole joue un rôle très important dans les systèmes de reconnaissances automatiques telle que la reconnaissance de locuteur, reconnaissance de la parole, reconnaissance de la langue, et la reconnaissance des émotions. Cette dernière sera traitée dans notre travail.

Les émotions sont définies comme des sentiments intenses qui se produisent sous l'influence d'évènements externes, tels que des réponses à des situations heureuse, surprenants, stressantes, d'agression verbale ou de violence, qui peuvent produire des états émotionnels d'intensité variable [3].

Dans ce chapitre nous présentons les notions de base de l'émotion, de la parole et de la reconnaissance automatique des émotions.

2. Reconnaissance émotionnelle

2.1. Définition de l’émotion

De nombreux scientifiques s'accordent à dire que l’émotion est une tâche difficile en termes de définition, car elle est caractérisée par des impacts physiques et mentaux. L'émotion est une réaction physiologique courte et transitoire à une situation donnée. Elle a des répercussions physiques différentes. Par exemple, la peur peut déclencher des pleurs et un rythme cardiaque rapide, la tristesse peut déclencher des larmes ou la joie déclenche un sourire. L'émotion est supposée comme un système de cinq composantes dont : La première est les réaction physiologiques telles que l'accélération cardiaque, l'augmentation de la température corporelle ou par une modification du rythme et de l'intensité de la respiration. La deuxième composante est basée sur les réponses expressives, c'est-à-dire des changements volontaires ou involontaires du visage et de la voix. La troisième comprend des réponses cognitives et expérientielles. La quatrième est la motivation. Enfin, le sentiment est la dernière composante [4].

2.2. Classification des émotions

Les émotions sont classée en deux classes: les émotions primaires et les émotions secondaires.

2.2.1. Emotions primaires (émotions de base)

Ce sont des émotions innées qui sont présentes chez le bébé, il existe six émotions primaires avec une neutralité (aucune émotion) dont chacune correspond à une expression du visage qui sont les mêmes entre plusieurs personnes d'âge, de culture ou de sexe différents. Ces émotions sont : la tristesse, la colère, la joie, le dégoût, la peur et la surprise. Ces sept émotions sont aujourd'hui mondialement connues [4].

Figure 1.1 : Les six émotions primaires [04].

2.2.2. Emotions secondaires (sociales et acquises)

Ces émotions ne sont pas innées mais acquises au cours de la vie par l'influence de la famille, la religion ou de la société elle-même. On peut citer : tendresse, addoration, la culpabilité, la jalousie, la honte, l'orgueil et la vanité [5].

On l'appelle parfois d'émotions mixtes car ce sont des émotions complexes qui résultent d'un mélange d'émotions primaires [4].

2.3. Catégories d'émotions

Il existe deux types de catégories d'émotions : les émotins negative et les émotion positives.

2.3.1. Les émotions négatives

C'est la catégorie d'émotions désagréables, liées au mal-être comme : le colère, la tristesse et la peur [3].

2.3.2. Les émotions positives

C'est la catégorie d'émotions agréables, liées au bien-être comme : l'amour, le bonheur et la joie [3].

2.4. Représentation des émotions

En psychologie, plusieurs modèles ont été conclus pour décrire l'ensemble des émotions, nous pouvons présenter : les modèles catégoriels, les modèles dimensionnels et les modèles hybride.

2.4.1. Approche catégorielle (discrète)

Cette approche basée sur l'existence de nombreuses émotions primaires discrètes, de sorte que ces émotions devraient être capables de les distinguer. Cette méthode peut présenter d'autres émotions, qui sont considérées comme des mélanges d'émotions principales. Elle tente de conceptualiser l'émotion à partir de plusieurs émotions majeures et de traiter chacune d'elles comme une émotion discrète [6]. La majorité des études d'effet vocal sur l'émotion ont utilisé ce modèle et ont choisi d'examiner les effets du bonheur, de la tristesse, de la peur, de la colère et de la surprise [7]. Un exemple de cet approche est le modele de Plutchik qui est illustré dans la Figure 1.2, ce modèle se compose de 8 émotions de base, composées de 4 paires opposées de deux (heureux-triste, anticipation-surprise, colère-peur et dégoût-confiance) et de multiples variations [4].

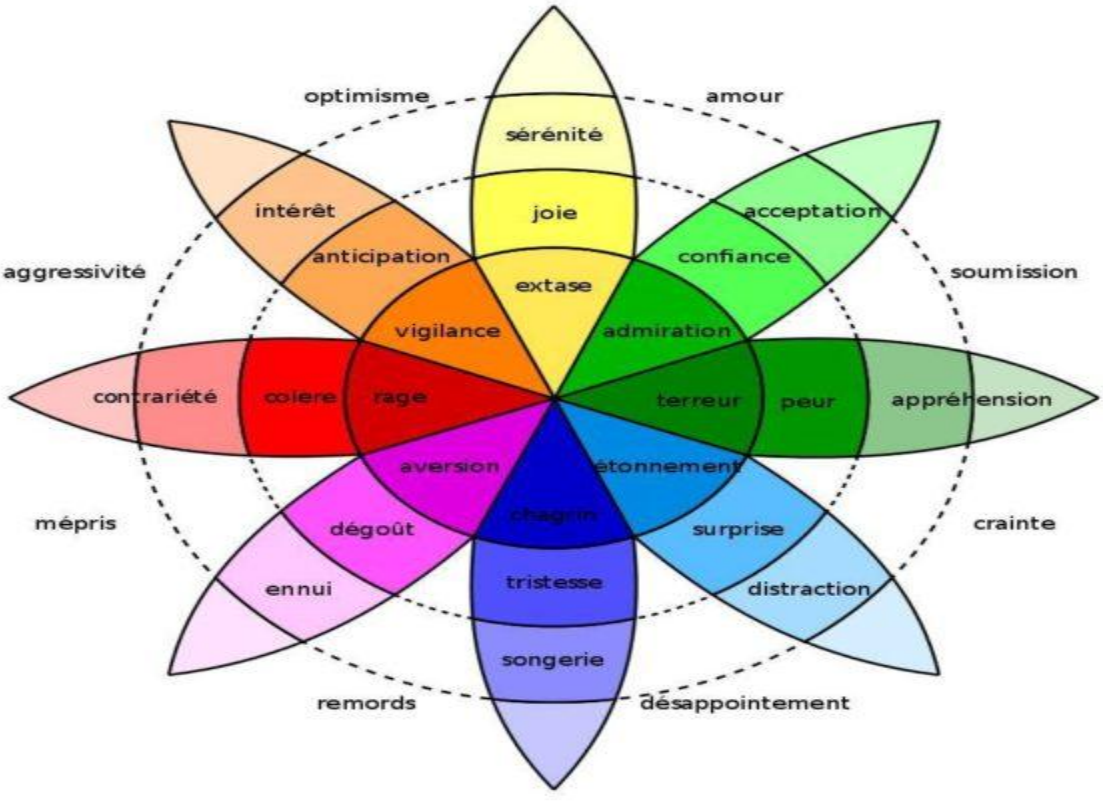


Figure 1.2: La roue des émotions de Plutchik [4].

2.4.2. Approche dimensionnelle (continue)

Ils considèrent l'émotion comme un phénomène continu ou progressif [6] et s'intéressent principalement aux descriptions verbales des sentiments subjectifs. Les émotions sont situées dans un espace à deux ou trois dimensions. Les deux dimensions principales incluent la dimension de valence (agréable-désagréable) et la dimension d'activité (actif/passif), la troisième dimension représentant le contrôle ou l'intelligence [7].

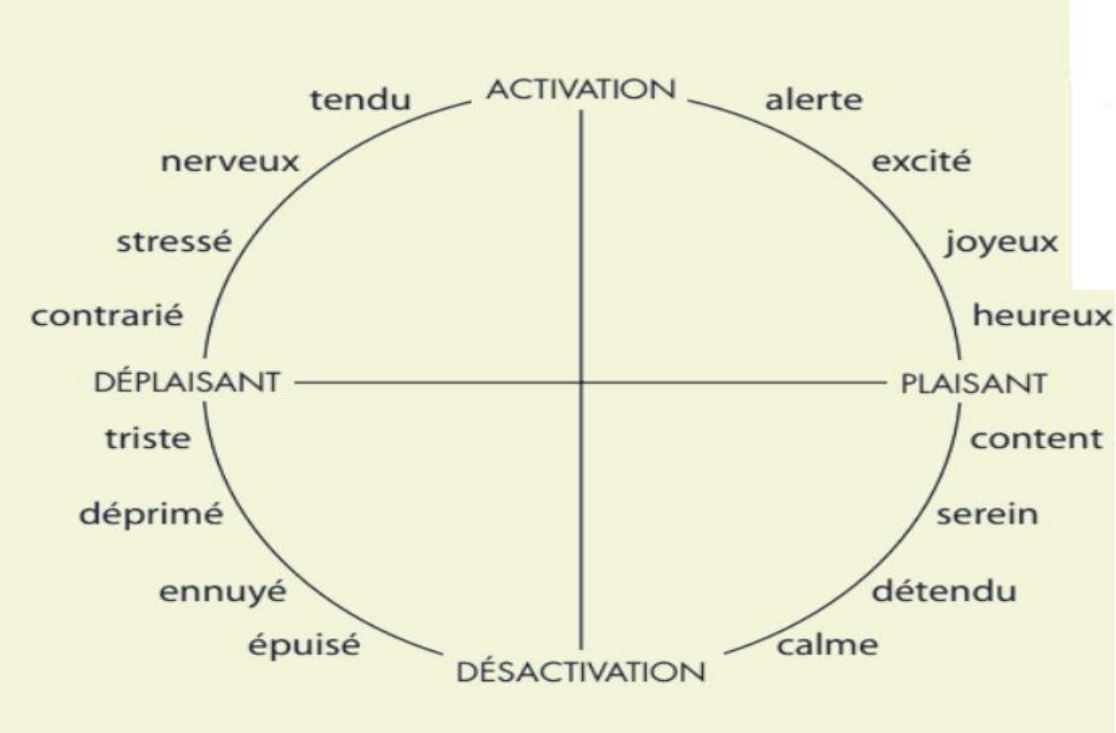


Figure 1.3: La boussole de Russell [4].

2.4.3. Approche hybride

C'est un compromis entre l'approche discrète et l'approche dimensionnelle. Le modèle a trois couches pour la perception des sentiments.

-La couche la plus abstraite contien deux classes : « valence » positive et « valence» négative.

-La couche moyenne contient les catégories d'émotions primaires : joie, colère, dégout, surprise, tristesse et peur.

-La couche inférieure est constituée des émotion secondaire: adoration et tendresse pour l'amour ; l'enthousiasme et le zele pour la joie ; l'insomnie et le gêne pour la colère [6].

2.5. Type de corpus des émotions

Il y a trois catégories principales de corpus émotionnels utilisées dans le domaine de la reconnaissance automatique des émotions : naturelles, induites, et simulées.

2.5.1. Corpus naturel (réaliste)

Ce corpus est obtenu en enregistrant des états émotionnels naturels et spontanés. Il se caractérise par une validité écologique très élevée. Son inconvénient est que ces données sont très limitées en nombre de locuteurs, de courte durée, souvent de mauvaise qualité, et difficiles à étiqueter en catégories d'émotion [7]. Le contexte dans lequel ces données sont collectées est varié (émissions de télévision, centres d'appels, entretiens avec des consommateurs) [4].

2.5.2. Corpus induit

Dans ce type les émotions sont induites expérimentalement, par exemple, en exposant des sujets à des tâches difficiles pendant une courte période de temps pour induire un stress, ou en présentant des images animées ou des films [7]. Ce corpus est utilisé dans le domaine de la reconnaissance automatique des émotions, car les émotions induites sont généralement de faible intensité et n'induisent pas le même état émotionnel chez un individu si le même protocole d'induction est utilisé [4].

2.5.3. Corpus acté (simulés)

Les émotions dans ce corpus sont générées par des acteurs professionnels ou semi-professionnels à partir de noms des classes émotionnelles et/ou de scènes typiques. Cette méthode représente la méthode préférée pour composer les données utilisées dans l'étude de la reconnaissance automatique des émotions (RAE) [7]. Ces corpus présentent plusieurs avantages, ils permettent d'obtenir de grandes quantités de données très typiques et faciles à collecter [4]. Cependant, certaines critiques ont été dirigées contre ce type car les émotions simulées sont caractérisées par des émotions plus fortes que naturelles [7].

2.6. Canaux de communication émotionnelle

Nous distingue trois canaux de communication émotionnelle qui sont : l’expression faciale, les signaux physiologiques et la voix.

2.6.1. Expressions faciales

Les expressions faciales sont des changements et des mouvements des traits de visage qui transmettent les états émotionnels d'une personne. Par exemple, le sentiment de joie peuvent être reconnus par les sourcils détendus, une bouche large avec les coins tirés vers les oreilles. Il est la source d’information la plus utilisée en reconnaissance automatique d’émotion [8].

2.6.2. Signaux physiologiques

Les changements physiologiques sont des marqueurs d'état physique qui reflète par les émotions. Les manifestations utilisés pour caractériser l'émotion sont :
l'activité électromyographie, l'activité électrodermale, le rythme cardiaque, le signal du volume respiratoire, la température cutanée [8].

2.6.3. Voix

La voix est parmi les moyens ou les sources les plus fiables d'interaction humaine. Elle est considéré comme une mesure de l'émotion et un reflet de la personnalité dans la reconnaissance des émotions. Par exemple, le sentiment de tristesse peuvent être traduits par une voix faible et cassée accompagnée d'un discours non motivé [8].

La reconnaissance des émotions par la voix est le cœur de notre travail.

3. Parole et reconnaissance automatique

3.1. Parole

Physiquement la parole définie comme une variation de la pression de l'air provoqué et diffusée par le système articulatoire, elle comporte plusieurs informations qui servent a la communication entre l'individu ainsi qui permettent d'identifier l'état émotionnel de la personne. Aujourd'hui le traitement de la parole est considérer comme une tâche fondamentale dans le développement des technique de télécommunication car ils situés au croissement du signal numérique et du traitement de langage [9].

3.2. Production de la parole

La parole traduit comme le résultat de l'action volontaire de plusieurs muscles c-à-d sa production se déroule par l'interaction entre le système nerveux qui joue le rôle de contrôleur des actions et le système physiologique « appellé l'appareil phonatoire » qui considére l'air comme une source d'énergie appliquer sur les poumons, au sommet de celle-ci se trouve le larynx qui comporte les cordes vocales comme principal organe vibratoire, la langue et les lèvres comme organes vibratoires accessoires, ainsi, les cavités buccale et nasale comme des résonateurs [9]. Pour faire l'étude sur la parole il est nécessaire d'abord de connaitre comment il se produit par le système phonatoire humain [10] qui se décrit par trois phases : la phase respiratoire, phonatoire, articulatoire :

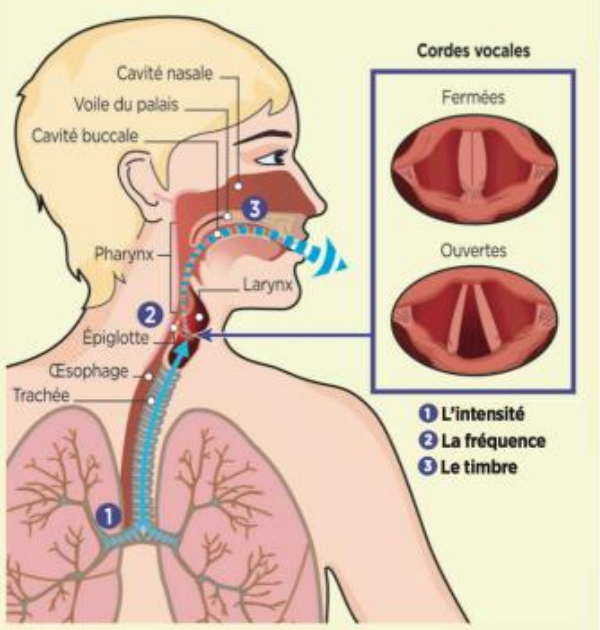


Figure 1.4 : Description détaillée de l'appareil vocal [4].

3.2.1. La phase respiratoire (Les poumons et la trachée)

La fonction primordiale des poumons est évidemment de permettre au corps de s'oxygèner, ainsi, les poumon sont la pièce maitresse de la production du son, car elles produisissent un flux d'air pour les deux phases de la respiration. Lors de la phase d'inspiration, le diaphragme se contracte et s'abaisse et les muscles intercostaux permettent de créer un vide dans les poumons qui est rempli par la pénétration d'air. Lors de l'expiration, le diaphragme se relâche et laisse, ainsi s'échapper l'air des poumons qui est conduit vers les cordes vocales a travers la trachée pour permettre leur vibration: plus le flux est faible, moins les cordes vocales vibrent et moins la voix s'entend [11].

3.2.2. La phase phonatoire (cordes vocale et glotte)

Le larynx est l'organe responsable de la vocalisation et se compose de plusieurs morceaux de cartilage et de muscle. Les cartilages les plus importants et les plus connus sont les cordes vocales et la glotte, cette dernière est un petit espace triangulaire entre les plis vocaux. Les cordes vocales sont des cartilages qui vibrent sous la pression de l'air expulsé des poumons. Et ils s'ouvre et se ferme très rapidement à travers la glotte, produisant un son voisé [11].

3.2.3. La phase d'articulatoire (les cavités supra glottiques)

Les résonateurs sont des organes qui permettent d'amplifier et de modifier le son conduit par la glotte. Lorsque le son sort de la glotte, il circule et filtre à travers les résonateurs (cavités) de la gorge, de la bouche et des fosses nasales et prend sa couleur, son timbre spécifique et ses harmoniques qui permettront de différencier le son entre eux [4] par les quarte cavités : pharyngale, buccale, labiale et nasale.

3.2.3.1. Cavité pharyngale

C'est le premier résonateur rencontré par des sons produits au niveau du larynx. Sa longueur est d'environ 8 cm. Il est vertical au-dessus de la gorge. Le pharynx est le canal musculo-tendino-membraneux qui relie le larynx à la cavité nasale et buccale, elle est constituée de trois sous-parties : laryngo-pharynx situé en arrière de l'épiglotte; oro-pharynx s'étend de l'épiglotte à la cavité buccale ; et le naso-pharynx s'étend du voile de palais jusqu'aux les fosses [12].

3.2.3.2. Cavité buccale

Sa longueur est d'environ 8 cm. Elle représente la partie du tractus vocal qui est délimitée par les lèvres en avant, la cavité pharyngale en arrière, la langue vers le bas, le palais et le voile du palais vers le haut et les dents en latérale. Cette cavité orale comprend les articulateurs supralaryngés fixes (palais, dents du haut et du bas) et mobiles (luette, langue, lèvres). La configuration de cette cavité est importante car elle contribue à l'articulation de presque tous les phonèmes. La cavité buccale joue un rôle important dans la résonance lorsque la mâchoire est ouverte, ainsi que dans la production de certaines consonnes comme les latérales (l, j). Elle est comprise dans la cavité orale, cette cavité est tres essentielle dans la production de parole du fait de ses nombreux changements de configuration induits par les mouvements de la langue et de la mâchoire [12].

3.2.3.3. Cavité labiale

Elle est distinguée et constitué à l'extrémité antérieure du canal vocal. Elle est formée par la protrusion (c'est la projection vers l'avant) des lèvres, dont l'arrondissement forme un dernier résonateur le long de la chaîne formée par les cavités pharyngales et buccales [12].

3.2.3.4. Cavité nasale (Les fosses nasales)

Sont deux cavités tapissés de muqueuses séparées par le septum nasal. Pendant la production des sons laryngés, le voile du palais s'abaisse alors les cavités pharyngales, orales et nasales permettent a l'air expiré par la cavité nasale produisant un son nasal (m, n, A, E, ç) [12].

3.3. Classification des sons du langage

Une classification des sons fait à la sortie du larynx selon la nature de la source d'excitation, il distingue deux types de son : voisés et non voisés.

3.3.1. Sons voisés

La production des sons voisés résulte du passage de l'air des poumons vers la trachée, provoquant la vibration des cordes vocales. En général les sons voisés telle que les voyelles et semi-voyelles sont caractérisé par une quasi-périodicité, une énergie élevée et il représente la majorité du temps de phonation [10].

Les voyelles : Elle sont toutes voisées car elle sont produites par les vibrations des cordes vocales qui provoquent un écoulement dans le larynx. Ce flux d'air n'a rencontré aucun obstacle dans la cavités supra glottique, les differences de timbre entre les voyelles sont le resultat de changements de forme, de volume de nombre de resonateurs, de mouvements de la langue, de la machoire, des levres, du passage de l'air a travers des vibrations a differents niveaux d'organes supra glottique [13].

3.3.2. Sons non voisés

Le son non voisée est produit lorsque la glotte est ouverte, l'air circule librement et les cordes vocales sont écartées et ne vibre pas. Ils ont une structure non périodique, une énergie concentrée dans les hautes fréquences et correspond avec le bruit [10].

Les consonnes : Ils sont produites par des occlusions ou des constrictions en un point quelconque du conduit vocal, d'où une impression de bruit ou de frottement [13].

4. Reconnaissance automatique des émotions

La reconnaissance des émotions s'avère très utile dans le développement de la communication homme-homme et la communication homme-machine. Car il permet à l'ordinateur de reconnaître les émotions de l'utilisateur à partir de sa voix. Le système de la reconnaissance automatique des émotions est un système générique qui produit un modèle émotionnel à partir d'informations contenues dans le signal acoustique, il comprend deux phases, la phase d'apprentissage qui permet de créer un modèle associé à chaque émotion et de l'enregistrer sur une base de données, et phase de reconnaissance (test) ; les paramètres du signal acoustique test sont extrait et comparés aux modèles de la base des données pour prendre une décision sur l'émotion testé (détecter l'émotion proclamée).

4.1. Domaine d'application de la reconnaissance automatique des émotions

La reconnaissance automatique de l'émotion permis son application au niveau de domaines très variés au sein des quels l'émotion joue un rôle important. Nous citons à titre d'exemples :

4.1.1. Reconnaissance des émotions pour l'enseignement à distance

Avec le développement des technologies de communication et d'interface homme-machine, l'enseignement à distance a une importance croissante. Il a nombreux avantages par rapport à l'enseignement traditionnel en face à face en classe. L'apprentissage en ligne manque d'interaction entre les enseignants et les étudiants, contrairement a l'apprentissage traditionnel en face à face, où les enseignants peut remarqué les réactions des étudiants et peuvent régler le contenu du cours et sa vitesse. Un système de reconnaissance automatique des émotions vocales est développer en temps

réel en un système éducatif capable de dire si un élève s'ennuie, est frustré ou bouleversé par la matière enseignée et peut ainsi modifier le style et le niveau de la matière enseignée, offrant une compensation et un encouragement émotionnel ou donnant une pause à l'enseignement [7].

4.1.2. Reconnaissance des émotions pour la sécurité

C'est un système automatique de reconnaissance d'émotions qui peut incorporé dans les véhicules qui serrent à detecter si le conducteur est en état de colère extrême, de stress, de la fatigue et de l'influence de l'alcool, et essayer de le réconforter ou de l'alerter sur les conséquences néfastes que cette émotion pourrait avoir sur sa conduite et activer des routines de sécurité [8]. Il peut aussi utiliser dans le cadre de la surveillance dans les lieux publics pour détecter la présence d'émotions extrêmes, principalement la peur [7]. Par example, l'Institut de Technologie de Massachusetts, ils ont créé un système intelligent utilisant la reconnaissance vocale émotionnelle appelé AutoEmotive qui surveille l'état du conducteur et les différentes émotions des passagers et tente également de les réconforter en jouant de la musique douce, en ajustant la température dans la cabine [14].

4.1.3. Reconnaissance des émotions pour la marketing

La reconnaissance émotionnelle vocal est la technologie la plus utilisé pour le marketing et la vente. Par exemple, la société russe Promobot à présenté un robot de service qui était développée par Neurodata pour permet de déterminer les émotions à partir de la voix du client. Aussi, le travail similaire entre les deux sociétés startups Cloverleaf et Affectiva permet d'introduit un système qui appelée « shelf Point» qui collecte des données sur les émotions des acheteurs entre expression faciale et vocale pour la détection de la satisfaction des clients et pour prévoir les produit qui les intéressent au plus [14] [15].

4.1.4. Reconnaissance des émotions dans les banques

Grâce au développement de l'intelligence artificielle dans la reconnaissance vocale, les banques peuvent aujourd'hui déterminer la satisfaction des visiteurs par leurs émotions telles que la tristesse ou la joie dans leurs paroles et aide à recueillir les commentaires des clients et à améliorer les services en analysant la voix du client lorsqu'il appelle le centre de contact et ils utilisent également cette reconnaissance comme empreinte vocale émotionnelle des Asynchronous Transfer Mode (ATM) machines [14].

4.1.5. Reconnaissance des émotions par l'IA dans la médecine

La reconnaissance vocale joue un rôle major en psychiatrie car elle peut détecter l'état d'une personne a partir de ses émotions en analysant ses paroles comme la dépression. Cette étude a été créée par les scientifiques de l'Institut de Technologie de Massachusetts et a eu un taux de réussite de 77%. De plus, grâce à cette technologie, La start-up Beyond Verbal aide les psychologues à

fournir des diagnostics performé de santé mentale aux patients et d'apprentissage émotionnel aux enfants autistes [14].

5. Conclusion

Dans ce chapitre nous avons abordé des généralités sur le domaine qui englobe notre travail. Nous avons commencé par une présentation générale sur l'émotion et la voix et la relation entre eux. Ensuite nous avons passé en revu la reconnaissance automatique des émotions, ainsi que le cadre d'application de cette dernère.

Nous allons, au deuxième chapitre rentré plus en détail au système de reconnaissance et ses étapes qui permet de détecter et d'analyser l'état émotionnel de locuteurs.

1. Introduction

La parole comporte deux formes d'informations permettant de distinguer les catégories d'émotions : les informations linguistiques et les informations paralinguistiques [7]. Le système de reconnaissance automatique des émotions consiste à ne récupérer que des informations paralinguistiques représentatives de l'état émotionnel du locuteur, indépendamment d'autres informations [4].

Dans un système de reconnaissance des émotions à travers la voix, il y a deux phases pour identifier l'état émotionnel du locuteur, la phase d'apprentissage et de reconnaissance (ou test), comme le montre sur la figure 2.1.

La première phase est la phase d'apprentissage, elle permet de créer des modèles associés à chaque catégorie d'émotions puis de construire des frontières de décision pour délimiter ces catégories [7]. Cette phase peut se résumer en quatre étapes :

Prétraitement : Il consiste à traiter le signal d'entrée avant d'extraire les paramètres. On peut citer par exemple l'extraction de canaux d'un fichier audio stéréo, la segmentation du signal en phrases selon l'alignement indiqué dans le fichier de transcription, ou encore la suppression du bruit et du silence [7].

Extraction des paramètres : Elle permet de mesurer les propriétés du signal pour que celui-ci puisse être représenté [4]. Elle repose sur des paramètres spectraux et prosodiques, regroupés en vecteurs de traits [3], ces traits pouvant affecter la précision du système RAE, il est donc nécessaire de sélectionner les meilleures caractéristiques pour la classification [4].

Modélisation : Elle permet de définir différents paramètres ajustables du classificateur choisi en fonction des données apprises pour bien représenter les données [7]. Au bout de cette étape les modèles fixés sont enregistrés dans une base de donnée.

La deuxième phase est la phase de test, la position d'une nouvelle donnée par rapport aux modèles créés détermine à quelle catégorie appartient cette donnée. Elle comprend les deux premières étapes précédentes, c'est-à-dire le prétraitement et l'extraction des caractéristiques qui sont nécessaires pour obtenir les traits acoustique, puis l'étape de comparaison qui permet de comparer les modèles de la base de données et le signal test. Enfin l'étape de la décision permet de détecter l'émotion proclamer [7].

Nous expliquons, dans la suite de ce chapitre, les procédures nécessaires du système de RAE et ces différentes étapes.

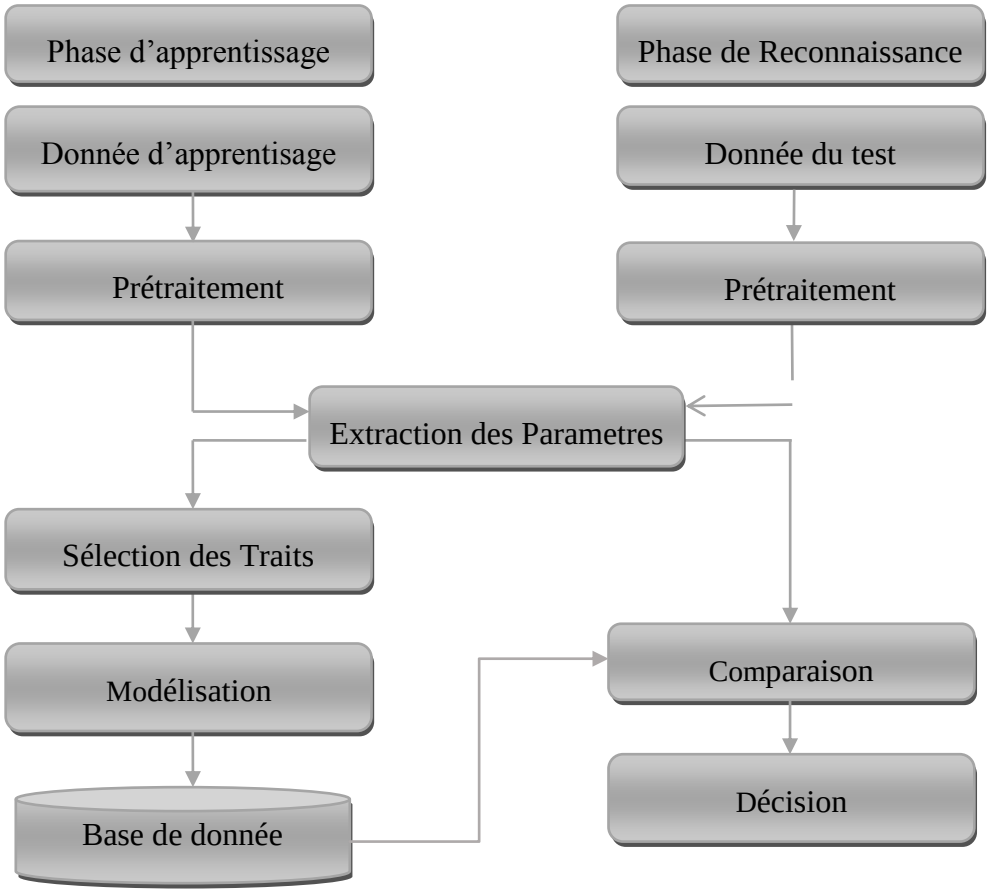


Figure 2.1 : la structure d'un système de RAE.

2. Phase d'apprentissage

2.1. Signal acoustique

Le signal sonore véhicule plusieurs informations telles que des informations linguistiques, l'identité du locuteur, la langue utilisée, ainsi que ses émotions [16]. La reconnaissance de l'état émotionnel d'un locuteur peut être considérée comme un problème de reconnaissance de formes qui reconnaît le signal de parole en entrée et produit en sortie la catégorie d'émotion véhiculée par la voix du locuteur [7].

2.2. Prétraitement

Il s'agit d'une étape appliquée avant l'extraction des paramètres du signal parole. Elle passe généralement par trois étapes : filtrage, segmentation et de chevauchement et le fenêtrage. Elle vise à préparer le signal acoustique et à améliorer la précision et l'efficacité du processus d'extraction des paramètres.

2.2.1. Filtrage

Le filtrage peut être utilisé pour supprimer les composants indésirables des signaux audio, comme la réduction du bruit dû aux conditions environnementales ou à d'autres perturbations lors de l'enregistrement d'échantillons audio [2].

2.2.2. Segmentation et chevauchement

Le but de cette étape est de découper le signal de parole en séquences de segments de courte durée appelées trames, sur lesquelles le signal peut souvent être considéré comme quasi-stationnaire, le signal est découpé en trames de N échantillons, de sorte que typiquement N est fixe, donc chaque trame correspond à environ 20 à 30 millisecondes [16]. En général les trames se chevauchent de 10 ms [4].

2.2.3. Fenêtrage

Selon la segmentation du signal, on obtient des trames qui seront multipliées dans une fenêtre temporelle pour l'analyse, cette fenêtre peut prendre plusieurs formes : Gaussienne, Triangle, Hamming, Uniforme. Mais la fenêtre de Hamming est la fenêtre la plus utilisée en traitement de la parole [16].

2.3. Extraction de paramètres

Les paramètres jouent un rôle très important dans le système global. L'ensemble des caractéristiques acoustiques les plus pertinents en RAE reste ouvert et nécessite d'avantage de recherches pour trouver les caractéristiques optimales pour la reconnaissance des émotions. Ces paramètres permettent de réduire la quantité d'information contenue dans un signal de parole échantillonné. Ils sont sélectionnées afin de distinguer une forme appartenant à une classe par rapport aux formes des autres classes dans le domaine de RAE. Pour la reconnaissance du style d'émotion «triste, en colère, dégoûté, ennuyé, neutre ou heureux», il existe différentes techniques d'extraction de caractéristiques et les études montrent que les paramètres spectraux et prosodiques jouent un rôle majeur dans la distinction entre les classes d'émotions [7].

2.3.1. Paramètres Prosodiques

La prosodie est un canal parallèle au contenu sémantique du message parlé dans les conversations quotidiennes, à travers lesquels l'auditeur peut percevoir les intentions et l'état émotionnel de l'orateur. C'est aussi par la prosodie que le locuteur peut donner à l'énoncé le ton d'une déclaration, d'une question ou d'une commande [3]. Ils peuvent impliquer des segments de parole plus longs (syllabe, mot, phrase). En tant que paramètres prosodiques, ils sont caractérisés

par des indices essentielles tels que : la fréquence fondamentale (pitch), l'énergie, la fréquences des formant, la durée, le débit de la parole...etc [17].

2.3.1.1. Fréquence fondamentale

C'est le phénomène prosodique le plus expressif, il exprime la hauteur perçue par un humain. La fréquence fondamentale (pitch) est directement liée à l'anatomie des cordes vocales (poids, taille, rigidité...) durant la phonation des sons voisés. Les cordes vocales des femmes sont généralement plus petites que celles des hommes, ce qui rend leur fréquence fondamentale (entre 250 et 400 Hz) plus haute en comparaison avec celle des hommes (entre 60 Hz et 150 Hz). De ce fait, dans le domaine de la reconnaissance émotionnelle la fréquence fondamentale est utilisée pour faire la classification entre les voix de locuteurs. Elle est appelée F0 (par définition c'est inverse de la période de vibration des cordes vocal) [17]. Il existe plusieurs technique temporelle (auttoccorelation, AMDF) pour l'estimation de pitch.

2.3.1.1.1. Détection de pitch par autocorrelation

Le domaine de traitement du signal digital définie l'autocorrelation comme une convolution d'un signal avec lui même. Cette méthode distingue les son voisée et non voisée qui permet de déterminer la fréquences fondamental du signal de parole. La fonction d'autocorrélation du signal est définie par :

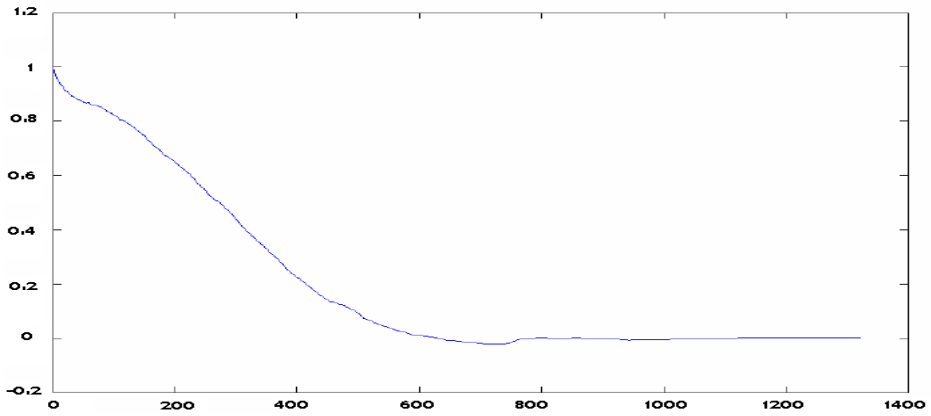
$$X(k) = \sum_{l=0}^{N} u(t).u(t+K) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.1)$$

Avec : u(t) signal d'origine et u (t+K) signal retardée par la valeur K

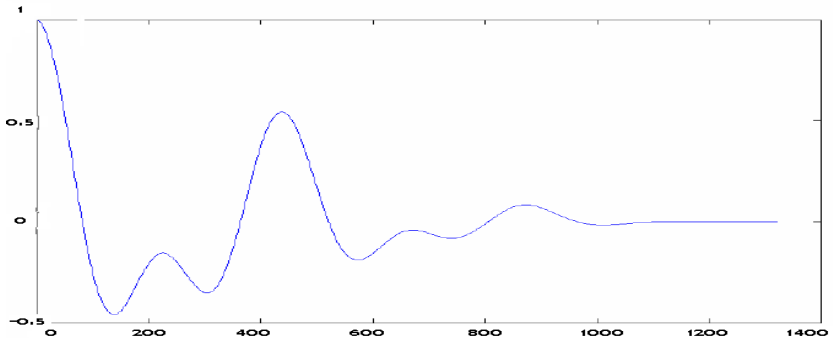
Elle mesure la ressemblance du signal u(t) avec sa version retardée. Le résultat d'autocorrélation dans le cas d'un son voisé est une suite de lobes espacés de n0 échantillons et si le son n'est pas voisé, la courbe d'autocorrélation décroît sans pics jusqu'à zéro et la fréquence fondamentale calculée par l'inverse de la distance entre les deux premiers et donner par [18] :

$$F0 = \frac{1}{n0} = \frac{1}{T0} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.2)$$

Avec n0 : la distance entre deux lobes successifs et équivalente avec la période T0



a) La courbe d'autocorrélation d'un son non voisée.



b) La courbe d'autocorrélation d'un son voisée.

Figure 2.2 : Courbes d'autocorrélation comparatives entre son voisée et non voisée [18].

2.3.1.1.2. Détection de pitch par AMDF

L'AMDF est l'abréviation de "Average Magnitude Difference Function" c'est un algorithme temporel utilisé pour la détection de la fréquence fondamentale grâce à sa simplicité de calcul, il est également utilisé dans le traitement du signal. La fonction AMDF correspond à une analyse d'autocorrélation car la différence entre le signal de la voix retardée et le signal d'origine forme un signal appelée le signal de déférence, la fonction de magnitude absolue est calculé pour chaque valeur de retard et définie par la loi suivante [19]:

$$\text{AMDF}(k) = \frac{1}{N}\sum_{t=1}^{N} |u(t) - u(t+k)| \qquad 0 \le k \le M \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.3)$$

u(t) le signal d'origine.

u(t+K) le signal décalée par la valeur de K dans le temps.

M c'est le nombre des points de la fonction AMDF.

La fonction AMDF décompose en cinq étapes essentielles pour estimer le pitch représentées sur le schéma bloc suivant:

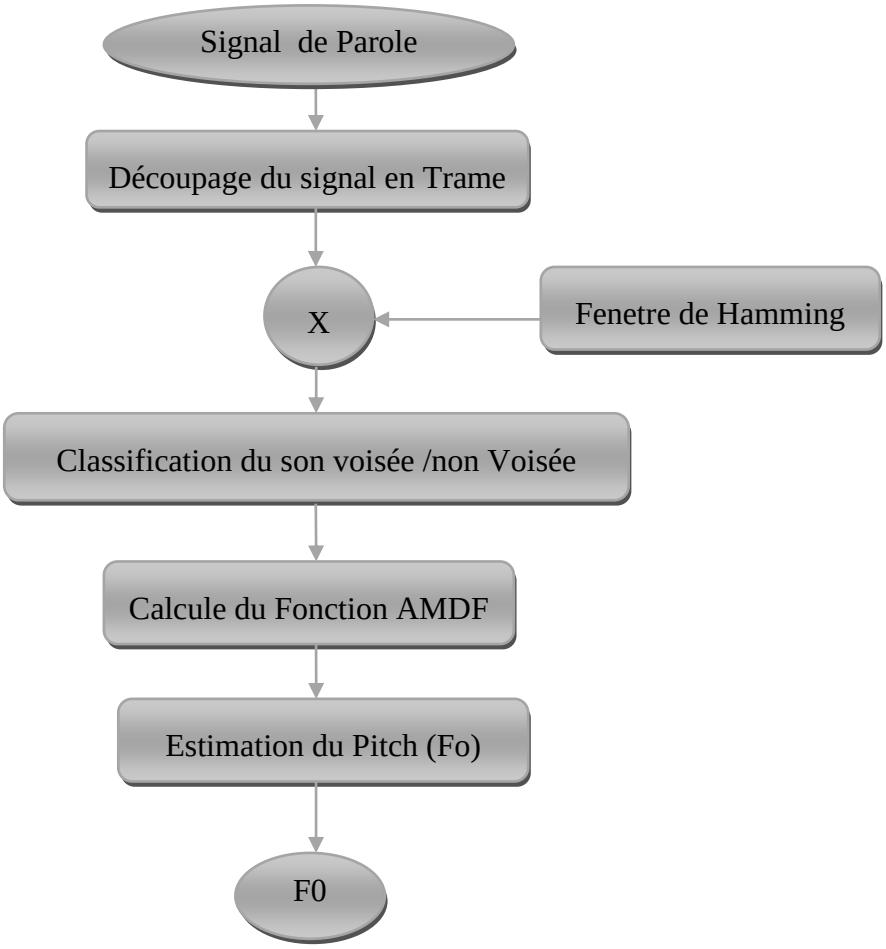


Figure 2.3 : Schéma bloc d'algorithme AMDF.

La figure 2.3 montre le schéma bloc de la technique AMDF. Le signal d'origine va être découpé en blocs (trames), chacun d'eux correspond à 20 ms. Ensuite, le signal échantillonné sera multiple par une fenêtre de Hamming pour l'augmentation du rapport signal sur bruit. Dans le troisième bloc le signal vise à classifier le son voisé du son non voisé, selon le calcul de l'énergie de chaque segment et le seuil de comparaison sert à prendre la décision de la classification du son :

-Si l'énergie ≥ la valeur 0.4 le son est voisé .

-Si l'énergie ≤ la valeur 0.4 le son est non voisé .

Pour les sons non voisés la fonction AMDF attribue comme une fréquence nulle.

Par contre, les son voisés la fonction AMDF sera calculée pour chaque segment qui sera décalé de 1 à N/2 fois et après chaque décalage, la somme de la différence entre le signal original et celui décalé sera calculée. Ceci génère un vecteur de dimension N/2. Par la suite, le minimum de ce vecteur donne l'indice de position de la valeur minimale [19]. Dans le dernier bloc on calcule la fréquence fondamentale F0 en déterminant les minima de la fonction AMDF(k). La distance entre deux minima successifs constitue la période fondamentale [20].

2.3.1.2. Intensité (Energie)

L'intensité d'un son ; est définie comme une sensation auditive vocalisée sur la perception de la force d'un signal audio [7]. Elle permet de distinguer un son fort d'un son faible. Elle représente l'amplitude de l'onde sonore provoquée par une énergie plus ou moins forte provenant du diaphragme et provoquant une variation de la pression de l'air sous la glotte [21].

L'intensité est calculée en décibels (dB) comme suit :

$$I = 10 \log \sum_{n=1}^{N} S_p^2(n) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.4)$$

2.3.1.3. Formant

Un formant est un résonateur spectral créé par les cavités du conduit vocal. Les formants décrivent les structures spectraux des sons voisés qui nous permettent d'identifier le type de son prononcé (voyelles, consonnes ou autres sons vocaliques) [10].

D'autre part, les formants ne sont que les pics spectraux du spectre sonore |P(f)| de la voix. C'est une résonance acoustique du conduit vocal humain [2]. Le formant se caractérise par la présence de maxima spectraux, c'est-à-dire de zones où les harmoniques sont intenses. Il est généralement utilisé en phonétique ou en acoustique pour décrire les vibrations des tractus vocaux ou des instruments de musique [20].

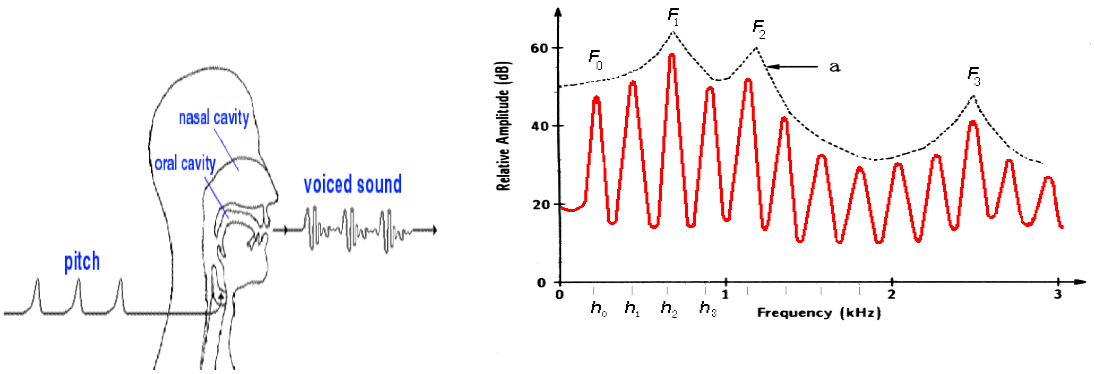


Figure 2.5 : structure spectrale des sons voisés et son spectre [10].

2.3.2. Les paramètres spectraux

Le signal vocal résulte de la convolution de la source par le conduit vocal. Son analyse spectrale permet de convertir un vecteur acoustique constitué d'un ensemble réduit de paramètres, c'est-à-dire, il extrait des coefficients représentative du signal parole.

Ces paramètres doivent être discriminants en rendant les sons de base facilement séparables. Il existe différentes méthodes spectrales pour transformer une trame de signal fenêtrée en un vecteur acoustique telles que les méthodes paramétriques basées sur un modèle de production comme LPC et LPCC et les méthodes basées sur un modèle de perception comme MFCC et PLP [16].

2.3.2.1. Coefficients cepstraux sur l'échelle Mel (MFCC)

Les MFCC (Mel-Frequency Cepstral Coefficients) sont des coefficients spectraux qui ont été intensivement utilisée comme des vecteurs de traits acoustique dans la reconnaissance basée sur la parole. Les coefficients MFCC sont définie comme étant la transformée en cosinus inverse de logarithme du spectre d'énergie du segment de la parole. L'énergie spectrale est calculée en appliquant un banc de filtres uniformément espacés sur une échelle fréquentielle modifiée, appelée échelle Mel [7]. La principale caractéristique de cette échelle est sa simulation du mécanisme perceptuel non linéaire de l'oreille humaine car l'échelle des Mel est un échelle biologique et modélisant l'oreille humaine [1].

Nous décrirons brièvement les étapes principales du processus d'extraction des vecteurs acoustiques de type MFCC :

Prétraitement : Ce sont deux étapes (segmentation et fenêtrage) sont déjà mentionnée précédement. Le signal se découpe en trames chevauchées de faible durée (une plage qui varie entre 20 à 30ms) pour être considéré comme un signal quasi stationnaire. Ensuite, chaque trame chevauché est multiplié par la fenêtre de Hamming pour atténuer les discontinuités du signal au bout de ces trames [1]. La fenêtre de Hamming est donnée par la formule suivante :

$$W(n) = 0.54 - [0.46 * \cos(\frac{2\pi n}{N-1})] \quad \text{..... (2.5)}$$

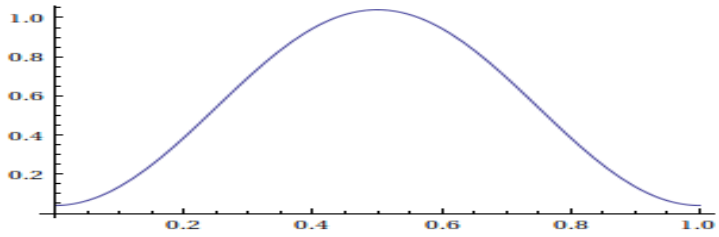


Figure 2.6 : La Fenetre de Hamming [1].

Transformée de Fourier rapide : Cette étape représente l'application de la FFT (Fast Fourier Transfer) sur les trames fenêtrés. La transformé de Fourier rapide est un algorithme permettant de calculer rapidement la transformation de Fourier discret, elle sera appliquée pour chaque trame fenêtrée [1]. Pour réaliser le passage et la conversion du domain temporelle au domain fréquentiel (spectral).

Filtrage : Il se fait selon l'échelle de Mel, qui est une echelle biologique qui imite l'audition humaine et pour cela le signal sera tracé par ce spectre [2]. Il est construit à partir d'une série de filtres passe-bandes de formes triangulaires positionnées d'une façon linéaire, pour les basses fréquences et logarithmique pour les hautes fréquences [1]. Chaque filtre donne la Somme des composantes spectrales filtrées. Ensuite, pour calculer le Mel pour une fréquence on utilise l'équation suivante:

$$F(Mel) = 2595 * \log_2(1 + \left|\frac{f}{700}\right|) \quad \text{..... (2.6)}$$

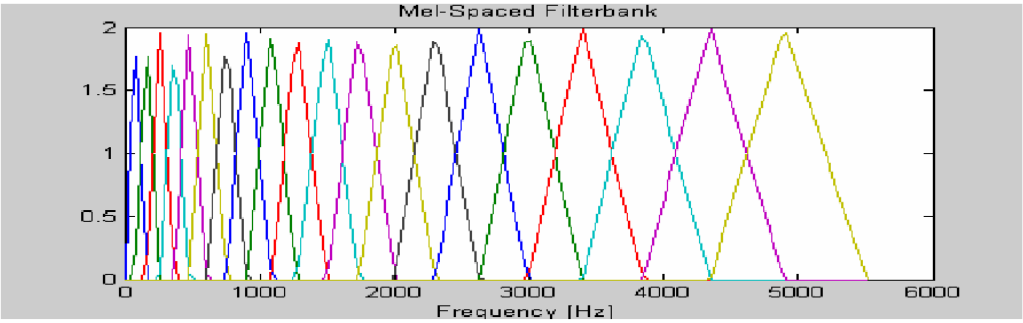


Figure 2.7 : Filtrage par echel de Mel [2].

Transformée en cosinus discrète : Enfin, le cepstre sur l'échelle de fréquence Mel est obtenu par une transformée en cosinus discrète DCT (Discret Cosinus Transform) à partir des logarithmes des énergies issues du banc de filtres donc c'est reconversion du log-Mel-spectre vers le domaine temporel. La DCT est donnée par la formule suivante : [2]

$$Cn = \sum_{k=1}^{N} \left(\log S_K \cos(n * (k - \frac{1}{2})\frac{\pi}{N}\right) \quad \text{..... (2.7)}$$

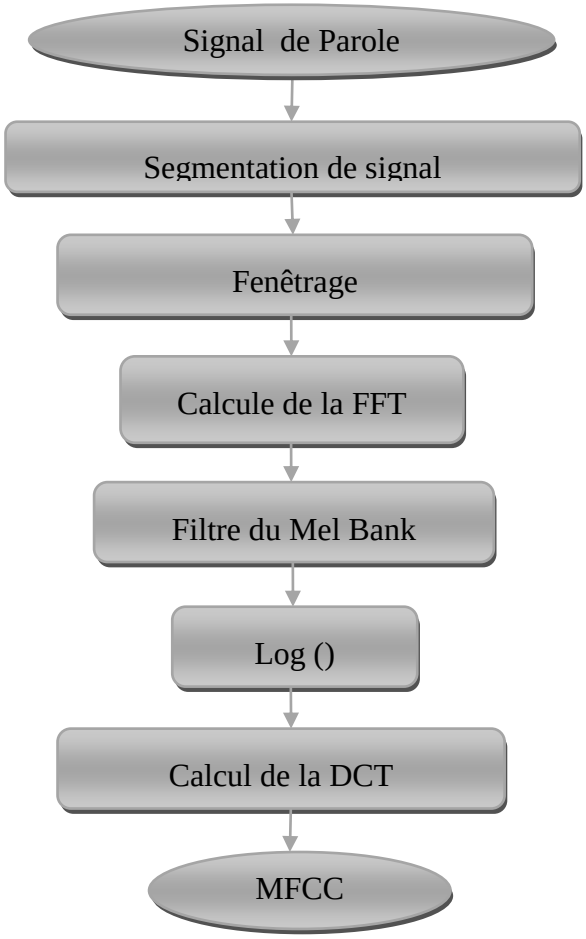


Figure 2.8: Schema bloc de MFCC.

2.3.2.2. Codage prédictif linéaire (LPC)

Les coefficients LPC (Linear Prediction Cepstral) sont définis comme une représentation de l'enveloppe spectrale du signal. Ils sont également un candidat pour rechercher l'ensemble de paramètres efficaces. Ils se concentrent sur le modèle de production de la parole.Il considère que l'appareil de production se compose d'une source et d'un filtre [6].

Le filtre de prédiction linéaire reflète les pics du spectre du signal audio par sa réponse en fréquence, donc cette analyse est largement utilisée pour déterminer les formants [17].

Le signal de parole est modélisé comme une sortie d'un filtre H(z) dont la source d'excitation à l'entrée du filtre u(t) est une série d'impulsions quasi-périodiques ou une source de bruit aléatoire.

Le LPC se base sur l'hypothèse que le filtre est un filtre tous-pôles. La fonction de transfert associée à ce filtre linéaire de prédiction est donnée par :

$$H(z) = \frac{\mathrm{S(z)}}{\mathrm{G.U(z)}} = \frac{1}{1-\sum_{k=1}^{p} \mathrm{a_k} z^{-k}} = \frac{1}{\mathrm{A(z)}} \quad \text{(2.8)}$$

G : le coefficient de gain a_k : les coefficients LPC p : l'ordre du filtre

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

$$s(n) = \sum_{k=1}^{p} a_k . s(n-k) + G.u(n) \quad \text{(2.9)}$$

Les coefficients a_k et le gain G sont calculés grâce à des méthodes fondées sur le calcul de la matrice de covariance ou grâce à des méthodes fondées sur le calcul de la matrice d'autocorrélation [6].

2.3.2.3. Coefficients cepstraux prédictifs linéaires (LPCC)

Les coefficients LPCC abrivé par « Linear Prediction Cepstral Coefficients » est l'une des méthodes des coefficients cepstraux utilisés dans la reconnaissance automatique des émotions. Ils sont calculés à partir des coefficients LPC du signal [4].

Considèrons le signal de parole *x* comme la conséquence de l'excitation du conduit vocal par un signal provenant des cordes vocales. La prédiction est basée sur le fait que les échantillons de parole adjacents sont fortement corrélés et elle permet d'obtenir une estimation du signal[22] :

$$\hat{x}_n = \sum_{k=1}^{p} a_{k.}\, x_{n-i} \quad \text{(2.10)}$$

a_k : sont des coefficients constants sur une fenêtre d'analyse.

La définition devient précise si l'on inclut le terme excitation :

$$x_n = \sum_{k=1}^{p} a_k . x_{n-i} + G e_n \quad \text{(2.11)}$$

e : est le signal d'excitation G : est le gain de l'excitation

La transformée en Z de cette égalité donne :

$$GE(z) = \left(1 - \sum_{k=1}^{p} a_k\, z^{-k}\right)X(z) \quad \text{..... (2.12)}$$

D'où :

$$H(z) = \frac{\mathrm{X(z)}}{\mathrm{E(z)}} = \frac{\mathrm{G}}{\left(1-\sum_{k=1}^{p} a_k z^{-k}\right)} = \frac{G}{A(z)} \quad \text{..... (2.13)}$$

Le signal X : est le résultat de l'excitation

H(z) : est un filtre multipolaire

L'erreur quadratique moyenne est minimiser par les coefficients a_i :

$$E_n = \sum_m (G.e_{n+m})^2 = [\sum_m (x_{n+m} - \sum_{k=1}^{p} a_k x_{n+m-k})]^2 \quad \text{..... (2.14)}$$

Les paramètres cepstraux peuvent être calculés à partir de ces échantillons prédis, par le résultat de la transformée de Fourier inverse appliquée au logarithme de la transformée de Fourier du signal parole.

Les paramètres cepstraux C_k sont les coefficients du développement de Taylor du logarithme de filtre multipolaire:

$$\ln[G^2/|A(z)|^2] = \sum_{-\infty}^{+\infty} C_k z^{-k} \quad \text{..... (2.15)}$$

D'où :

$$C_m = a_m + \sum_{=k-p}^{k-1} \frac{j}{k} C_j a_{k-j} \quad \text{..... (2.16)}$$

avec : k=p,……,N_c et $C_{-m} = C_m$ et $C_0 = \ln(G^2)$

2.3.2.4. Prédiction Linéaire Perceptuelle (PLP)

Les coefficients PLP (Perceptual Linear Prediction) sont l'un des coefficients spectraux utilisés dans la reconnaissance automatique des émotions, Ils sont calculés à partir d'une amélioration des coefficients LPC par une transformée de Fourier inverse appliquée à la racine cubique du spectre de puissance suivie d'une analyse par prédiction linéaire [4]. La technique d'analyse par PLP basée sur la perception humaine de la parole [17].

Comme les coefficients LPC dans PLP, le signal est analysé sur une fenêtre glissante de courte durée. En général, la longueur de cette fenêtre est comprise entre 10 et 30 ms avec un décalage de 10 ms pour chaque trame [22].

2.4. Modélisation

Dans le système de la reconnaissance automatique de l'émotion par la voix, plusieurs approches ont été étudiées afin de modéliser la structure complexe qui caractérise la catégorie de l'émotion. Il existe deux types d'approches de classification telle que l'approche paramétrique et l'approche non paramétrique.

l'approche paramétrique, dans ce modèle d'appariement est probabiliste (évaluation des probabilités). Il considère la classe(dans notre cas l'émotion) comme étant une source probabiliste et la modélise par une densité de probabilité connue. Ses principales techniques sont le modèle de mélange gaussien (GMM), le modèle de Markov caché (HMM)… [2]

l'approche non paramétrique est considérée comme la plus simple approche [2], telle que l'appariement est déterministe (évaluation des distances). Dans ce cas on considère que la classe est représentée par un ensemble de vecteurs de paramètres dans l'espace acoustique. Cette aproche compte deux grands algorithmes, la quantification vectorielle VQ et l'algorithme temporel dynamique DTW [7].

2.4.1. Quantification vectorielle (QV)

QV est l'abréviation de Vector Quantisation, il s'agit de décomposer l'espace acoustique (X) en un nombre fini de vecteurs acoustiques. L'ensemble des sous-espaces (M) seront représentés par leur vecteurs centroides C={c1,c2, …,cM}. Ces vecteurs centroides constituent un dictionnaire (de taille M) qui modélise chaque classe. Pour déterminer la distance du vecteur acoustique à cet espace, une mesure de distance est effectuée avec chacun des centroïdes des régions et la distance minimale est conservée. Cette mesure est donnée par l'équation suivante :

$$D(X,C) = \frac{1}{T}\sum_{t=1}^{T} \min_{1<m<M} \mathrm{d}(x_t, c_m) \quad \text{(2.17)}$$

Avec : $D(X,C)$: est une mesure de distance.

La classe représentera à travers son dictionnaire de quantification, pour une meilleure représentation en augmentant la taille du dictionnaire, mais le système sera plus lent et plus gourmand en mémoire. La distance sera généralement moindre si le vecteur acoustique provient dela même émotion que si ce vecteur provient d'une autre émotion [17].

2.4.2. Programmation dynamique (DTW)

DTW est l'abréviation de Dynamic Time Warping. Elle est considéré comme une méthode efficace pour reconnaître les formes de parole, mais son inconvénient est qu'elle nécessite un long temps de traitement et une grande capacité de stockage, en particulier pour la reconnaissance en temps réel. La déformation temporelle dynamique est un algorithme largement utilisée pour résoudre les problèmes d'alignement temporel permettant de mesurer la similarité entre deux séquences qui peuvent évoluer dans le temps. La mesure de la distance entre deux vecteurs caractéristiques peut être calculée à l'aide de l'échelle de distance euclidienne. Ainsi, la distance locale entre les vecteurs d'attributs x et y est donnée par [2]:

$$\mathrm{d}(\mathrm{X},\mathrm{Y}) = \sqrt{\sum_{\mathrm{j}=1}^{\mathrm{p}} (\mathrm{Xj} - \mathrm{Yj})^{\,2}} \quad \text{(2.18)}$$

2.4.3. Modèle de mélanges gaussiens (GMM)

Un modèle de mélange gaussien (abrégé par l'acronyme anglais GMM pour Gaussian Mixture Model). C'est une méthode statique qui fait une analyse discriminante sur la densité de probabilité et a été utilisée dans de nombreux domaines tels que la reconnaissance du locuteur, le traitement d'images, la reconnaissance automatique des émotions [7]. La plupart des systèmes de reconnaissance des émotions utilise les Modèles de Mélange de Gaussiennes (GMM) dans les taches d'identification et de vérification à l'étape de la modélisation car c'est un modèle de base [17].

Ce modèle consiste à générer les données par une somme de plusieurs courbes gaussiennes avec une détermination de la moyenne, la variance et l'amplitude de chaque gaussienne. Il est défini sous la forme mathématique suivante :

$$p\left(\frac{x}{\lambda}\right) = \sum_{i=1}^{M} w_i \, b_i(x) \quad \text{..... (2.19)}$$

Avec : x : Le vecteur de données de dimension d.

λ : Le modèle GMM

w_i : Les pondérations de GMM.

$b_i(x)$: La densité normale multidimensionnelle qui est donnée par l'équation suivante :

$$b_i(x) = \frac{1}{(2\pi)^{d/2}.|\sum_i|^{1/2}} exp\left(-\frac{1}{2}\left(x-\mu_i\right)^r \sum_i^{-1}(x-\mu_i\right) \quad \text{..... (2.20)}$$

$\sum_i$: La matrice de covariance. μ_i : La moyenne. π : Le poids du mélange.

Le modèle GMM (λ) est représenté par les trois paramètres : la pondération, la moyenne et la matrice de covariance de chacune des M composantes gaussiennes.

$$\lambda = \{w_i, \mu_i, \sum_i\} \quad \text{..... (2.21)}$$

Ces trois paramètres sont estimés par plusieurs algorithmes dont le plus couramment utilisé est l'algorithme EM (Expectation Maximisation) dans la RAE. Il donne des résultats plus robustes quand on ajoute une grande quantité de données [7].

EM est défini comme un algorithme itératif partant d'une estimation initiale de θ (c.-à-d. Valeur aléatoire), ensuite il continue à mettre à jour itérativement θ jusqu'à la détection de la convergence. Chaque itération contient deux étapes [23] :

L'étape E : Qui permet de calculer l'espérance conditionnelle en fonction des données examinées et les paramètres actuels. Elle s'écrit sous la forme suivante :

$$Q(\theta; \theta^{(c)}) = E(L_c(\theta, z) \mid x, \theta^{(c)}) \quad \text{..... (2.22)}$$

L'étape M : Permet de faire une mise à jour sur les paramètres en maximisant la fonction Q :

$$\theta^{(c+1)} = \arg\max_{\theta} Q(\theta; \theta^{(c)}) \quad \text{..... (2.23)}$$

2.4.4. Technique GMM-UBM

En réalité, la modélisation de l'émotion du locuteur nécessite de nombreux paramètres GMM.Par conséquent, une quantité relativement importante de données est nécessaire pour parvenir à une estimation suffisamment robuste du modèle d'émotion. Pour surmonter ce problème, ils ont un modèle très bien formé qui utilise un grand ensemble d'émotions différentés appelé modèle du monde UBM (Universal Background Model). En utilisant une quantité faible de données pour chaque classe. L'adaptation des paramètres du modèle UBM permet la création d'un modèle robuste quiest connue sous l'appellation GMM-UBM. La méthode d'adaptation MAP (Maximum à posteriori) est une méthode d'adaptation bayésienne et qui la plus connue et parmi les méthodes les plus performantes dans l'adaptation des données entre les paramètres du modèle GMM et les données du modèle UBM [1]. Le schéma suivant représente l'architecture du modèle GMM-UBM :

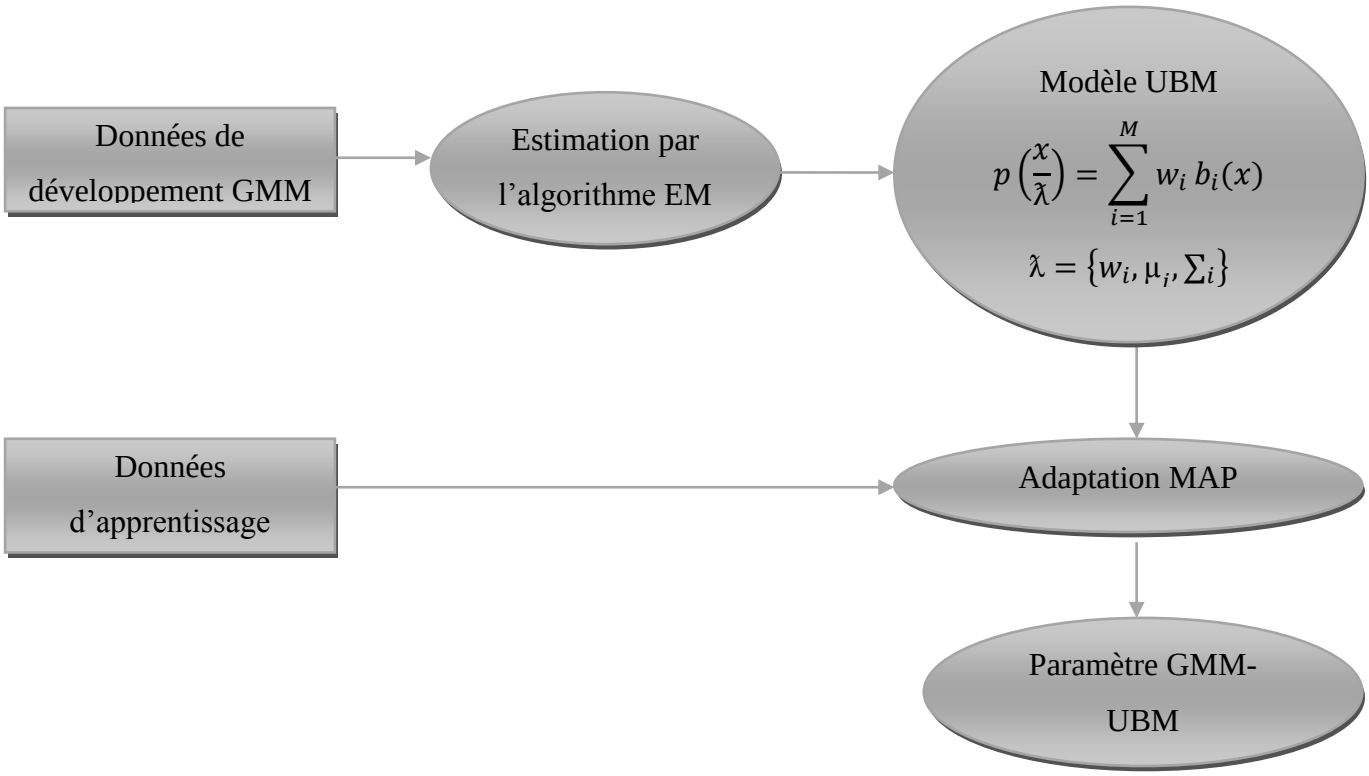


Figure 2.9 : Schema bloc de l'architecture du modèle GMM-UBM.

2.5.Base de données

C'est une entité dans laquelle sont stockées les données des categories emotionnelles fournis a partir de l'étape précédente (modélisation) de façon structurée et avec le moins de redondance possible. Elle stocke les modéles traités pour une meilleure détection et reconnaissance du modèle proclamé dans la phase de test.

3. Phase de test

Comme précédemment expliqué, cette phase se compose de quatre parties pour aboutir une détection et analyse de l'état émotionel du locuteur tel que le prétraittement et l'éxtraction de paramétres, la comparaison et la prise de décision.

3.1. Comparaison

Après l'examination des modèles stockés dans la base de données, dans cette étape, le système collecte un certain nombre de modèles d'émotions qui ressemblent le plus au cas émotionnel testé. Pour le faire une comparaison entre eux, par des méthodes de probabilités conditionnelles ou par dictionnaire est effectuée pour créer un seuil limite pour permettre l'accès à l'étape suivante.

3.2. Décision

Dans cette étape le système de la reconnaissance automatique d'émotion permet de reconnaître le modèle qui correspond le mieux au modèle stocké dans la base de données. Les résultats de la comparaisons donnent des scores de reconnaissance qui permettent de faire une identification d'émotion proclamé.

4. Conclusion

Dans le présent chapitre,nous avont présenté les deux phases du système de reconnaissance émotionels. La phase d'apprentissage et la phase de test.

Cette étude nous a permis de parcourir diverses méthodes appliquées sur le signal de parole (segmentation et fenetrage...) et d'extraire les coefficients les plus essentiels dans la voix par des techniques spectraux et prosodiques. Ces derniers sont essentiellement basés sur l'estiomation du pitch par des algorithmes connues ainsi que les formants et l'intensité (ou l'énergie).

En terme de modélisation ce chapitre presente les approches paramétriques comme le GMM-UBM et les approches non paramétriques telles que la DTW et la QV. Enfin, nous avons vu les étapes de comparaison et de prise de décision.

1.Introduction

Tout au long des deux précédents chapitres nous avons expliqué toutes les notions théoriques nécessaires à notre travail et avec un seul but ; le développement d'un système de reconnaissance automatique de l'émotion à partir de la voix. Afin de mettre en œuvre ce système en pratique, nous avons utilisé une base de données qui comporte des échantillons audio enregistrés, et on a extrait leurs traits acoustiques par différentes méthodes spectrales et prosodiques. Une partie est utilisée pour créer des modèles des états émotionnels par DTW et GMM-UBM. Une autre partie sert à tester le système via un critère de performance qui est le taux de reconnaissance correct.

Le travail qu'on a réalisé sur logiciel MATLAB, vise à trouver le meilleur système permettant la détection émotionnelle avec le taux de reconnaissance correct le plus élevé.

2. Base de données

Dans notre étude, nous utilisons un ensemble de données en accès libre appelé RAVDESS (Ryersonaudio-visual Data base of emotion speech and song). C'est une base de données audiovisuelle de la parole et de la chanson émotionnelles de l'Université Ryerson qui contient 7356 fichiers de taille totale égale à 24,8 Go. Elle est disponible en trois formats de modalité : audio-uniquement, vidéo uniquement et audio-vidéo. RAVDESS à été créé par 24 acteurs professionnels (12 femmes, 12 hommes), qui prononcent deux déclarations lexicalement assorties avec un accent nord-américain. La parole comprend des expressions calmes, heureuses, tristes, en colère, peur, surprises et dégoûtées, et chaque expression est produite à deux niveaux d'intensité émotionnelle (normal, fort), avec une expression neutre supplémentaire.

Nous utilisant les fichiers audio-uniquement pour notre simulation. Cette partie de RAVDESS contient 1440 dossiers dont 60 essais par acteur. Chacun de ces fichiers a un nom de fichier unique qui est composé d'un identificateur numérique en 7 parties

par exemple : 03-01-06-01-02-01-12.wav.

Ces identificateurs définissent respectivement les caractéristiques de l'acteur comme suit :

-Modalité : 01 = audio-vidéo, 02 = vidéo uniquement, 03 = audio uniquement.

-Canal vocal : 01 = parole, 02 = chanson.

-Émotion : 01 = neutre, 02 = calme, 03 = heureux, 04 = triste, 05 =colère ,06 = effrayé 07=dégoûté, 08 = surpris.

-Intensité émotionnelle : 01 = normal, 02 = fort. « pour émotion neutre y a pas une intensité forte ».

-Énoncé : 01 = « Les enfants parlent près de la porte », 02 =«Les chiens sont assis près de la porte».

-Répétition : 01 = 1ère répétition, 02 = 2ème répétition.

-Les acteurs (1 à 24) : les acteurs paires sont des femmes et les impaires sont des hommes.

3. Protocole

Notre travail consiste à réaliser un système de reconnaissance d'émotion de l'apprenant par le canal vocal. Pour cela, on utilise une base de données de 24 acteurs, Chacun des ces derniers a enregistré 8 émotions en plusieurs fois. Le système est composé de trois parties principales ; une plateforme de développement pour la création d'un modèle du monde UBM utilisant douze premiers acteurs (de 1 à 12). Pour les deux autres parties on prendra les douze (12) derniers acteurs (13 à 24) afin de créer des modèles GMM et DTW. En phase d'apprentissage, nous avons pris la première répétition d'enregistrement pour créer les 96 modèles représentant toutes les émotions avec tous les acteurs. Enfin, dans la dernière partie, à savoir la phase de reconnaissance nous prenons la deuxième répétition pour réaliser 96 tests qui seront ensuite comparer avec les 96 modèles déjà créés. Les résultats de la comparaison sont donnés par des scores qui sont la probabilité conditionnelle que l'émotion testée provient du modèle créé pour la technique GMM_UBM, et des scores qui sont des distances pour la technique DTW. L'évaluation du système se calcule sous la forme d'un taux de reconnaissance correct.

4. Résultats et discussions

4.1. Utilisation de DTW sur le système de reconnaissance

4.1.1. Paramètres spectraux

Dans cette partie nous avons testé individuellement les différents types de paramètres spectraux. On a varié le nombre de coefficients de tous les paramètres utilisés (MFCC, LPCC, et PLP) de 12 à 22 coefficients. Dans le tableau 3.1, on montre seulement le meilleur taux de reconnaissance pour chaque paramètre qui correspond à 14 coefficients pour le MFCC, 16 pour le LPCC et PLP.

Paramètres / Acteurs	MFCC	LPCC	PLP
Acteur 13	100,00	75,00	87,50
Acteur 14	50,00	75,00	50,00
Acteur 15	62,50	87,50	62,50
Acteur 16	87,50	87,50	62,50
Acteur 17	87,50	75,00	75,00
Acteur 18	87,50	100,00	100,00
Acteur 19	75,00	75,00	75,00
Acteur 20	62,50	62,50	62,50
Acteur 21	100,00	100,00	100,00
Acteur 22	87,50	100,00	87,50
Acteur 23	100,00	100,00	100
Acteur 24	75,00	75,50	87,50
Taux global(%)	81,25	**84,38**	79,17

Tableau 3.1 : Taux correct de reconnaissance en fonction des paramètres spectraux.

On observe que le meilleur taux de reconnaissance est obtenu pour le paramètre LPCC avec un taux correct de **84,38%**.

4.1.2. Paramétrés Prosodiques

Le tableau suivant montre l'utilisation individuelle des paramètres prosodiques.

Paramètre	Pitch	Energie	Formant
Taux global (%)	40,62	**56,25**	52,08

Tableau 3.2 : Taux correct de reconnaissance en fonction des paramètres prosodiques.

Les résultats montre que le paramètre prosodique énergie donne le taux de reconnaissance le plus élevé.

4.1.2.1. Fusion des paramétrés prosodiques

Dans cette expérience nous présentons les différents tests effectués avec une combinaison des paramètres prosodiques.

Paramètres	Pitch et Energie	Pitch et Formant	Energie et Formant	Pitch et Energie et Formant
Taux global(%)	43,75	40,62	52,08	40,62

Tableau 3.3: Effet de la fusion des paramètres prosodiques sur le système RAE.

D'après ces résultats, Il est clair, que la fusion des paramètres prosodiques n'améliore pas le taux de reconnaissance.

4.1.3. Fusion des paramètres spectraux et prosodiques

Dans la section précédente, nous avons testé individuellement les différents types de paramètres. En raison de faire une étude d'amélioration sur le taux de reconnaissance, il est important d'étudier la combinaison des caractéristiques (prosodique-spectral). Pour cela nous avons varié le nombre de coefficients de tous les paramètres utilisés (MFCC, LPCC, et PLP) de 12 à 22 coefficients avec chaque fusion.

4.1.3.1. Fusion de paramètre MFCC avec les paramètres prosodiques

Dans le tableau 3.4 Nous rapportons seulement le meilleur résultat de reconnaissance pour chaque fusion qui correspond à 12 coefficients pour toutes les fusions.

Paramètres	MFCC et Pitch	MFCC et Energie	MFCC et Formant	MFCC et Pitch et Energie	MFCC et Energie et Formant	MFCC et Pitch et Formant	MFCC et Pitch et Energie et Formant
Taux global(%)	46,88	81,25	52,08	47,92	52,08	40,62	40,62

Tableau 3.4 : Taux correct de reconnaissance en fonction des coefficients MFCC et paramètres prosodiques.

Le meilleur résultat obtenu dans ce cas correspond à la fusion de l'énergie avec les coefficients MFCC. Cependant, cette fusion n'améliore pas le taux de reconnaissance.

4.1.3.2. Fusion de paramètre LPCC avec les paramètres prosodiques

D'après tous les résultats obtenus nous exposons seulement les meilleurs taux de reconnaissance qui correspond comme suite :12 coefficients avec le formant et énergie-formant, 14 coefficients avec le pitch et pitch-formant,et 20 coefficients avec la fusion des trois paramètres prosodiques.

Paramètres	LPCC et Pitch	LPCC et Energie	LPCC et Formant	LPCC et Pitch et Energie	LPCC et Energie et Formant	LPCC et Pitch et Formant	LPCC et Pitch et Energie et Formant
Taux global (%)	54,17	**84,38**	57,29	54,17	57,29	61,46	62,50

Tableau 3.5 : Taux correct de reconnaissance en fonction des coefficients LPCC et paramètres prosodiques.

De même que précédemment, le tableau 3.5 montre que la fusion entre LPCC et énergie donne le plus grand taux de **84,38%** mais sans améliorer le système sans fusion.

4.1.3.3. Fusion de paramètre PLP avec les paramètres prosodiques

De même que précédemment, on représente sur le tableau 3.6 les taux les plus élevés pour chaque fusion qui correspond a : 12 coefficient avec les fusions suivant :Formant, (Pitch- Energie), (Energie-Formant),(Pitch-Formant) et fusion des trois paramètres, ainsi 16 coefficients avec le pitch et 22 coefficients avec l'énergie.

Paramètres	PLP et Pitch	PLP et Energie	PLP et Formant	PLP et Pitch et Energie	PLP et Energie et Formant	PLP et Pitch et Formant	PLP et Pitch et Energie et Formant
Taux global (%)	**79,17**	78,75	53,12	36,46	53,12	57,12	57,29

Tableau 3.6: Taux correct de reconnaissance en fonction des coefficients PLP et paramètres prosodiques.

Pour cette expérience on constate que la fusion de PLP avec le pitch donne le meilleur résultat avec un taux correct de **79,17%**. La fusion n'a aucun avantage par rapport au paramètre spectral seul.

4.1.4. Taux de reconnaissance par émotion

Dans cette expérience nous avons fait une détection par émotion qui est montré sur le tableau 3.7, en utilisons le paramètre spectral LPCC avec 16 coefficients qui nous donner un taux optimal de 84.38%.

Emotions	Neutre	Calme	Heureux	Triste	Colère	Effrayé	Dégoût	Surpris
Taux global (%)	**91,67**	83,33	**91,67**	75,00	**91,67**	**91,67**	75,00	75,00

Tableau 3.7: Taux correct en fonction de reconnaissance par émotion.

On observe que le système de RAE par la technique DTW arrive à mieux détecté les émotions neutres, heureux, en colère, effrayé avec un taux **91.67%.**

4.2. Utilisation du système GMM-UBM sur le système de reconnaissance

4.2.1. Paramètres spectraux

Dans les expériences suivantes nous allons tester individuellement les paramètres spectraux (MFCC, LPCC, PLP) par une variation sur le nombre de coefficients de 12 à 22 et sur le nombre de gaussienne de 32 à 1024.

4.2.1.1. MFCC

Dans ce tableau, on mentionne seulement les résultats qui correspondent à 12 coefficients MFCC car il donne les meilleurs taux de reconnaissance.

Nb GMM / Acteurs	32	64	128	256	512	1024
Acteur 13	87,50	87,50	100,00	100,00	100,00	100,00
Acteur 14	75,00	75,00	87,50	75,00	87,50	62,50
Acteur 15	50,00	37,50	50,00	62,50	62,50	62,50
Acteur 16	100,00	100,00	100,00	100,00	100,00	100,00
Acteur 17	75,00	62,50	75,00	62,50	62,50	62,50
Acteur 18	75,00	62,50	87,50	75,00	62,50	100,00
Acteur 19	75,00	87,50	75,00	87,50	100,00	100,00
Acteur 20	75,00	62,50	62,50	75,00	62,50	75,00
Acteur 21	100,00	100,00	100,00	87,50	87,50	100,00
Acteur 22	75,00	75,00	75,00	75,00	75,00	75,00
Acteur 23	100,00	100,00	100,00	100,00	100,00	100,00
Acteur 24	100,00	100,00	100,00	100,00	100,00	100,00
Taux global (%)	82,29	79,17	84,38	83,33	83,33	**86,46**

Tableau 3.8 : Taux correct de reconnaissance par MFCC en fonction de nombre de GMM.

A partir des résultats obtenus, on remarque que le meilleur taux correct correspond à 12 MFCC et 1024 gaussiennes. Il est de **86,46%.**

4.2.1.2. LPCC

De même, le tableau 3.9 expose les résultats de reconnaissance obtenus par 22 coefficients LPCC :

Nb GMM	32	64	128	256	512	1024
Taux global (%)	65,63	66,67	65,63	70,83	**72,92**	70,83

Tableau 3.9: Taux correct de reconnaissance par LPCC en fonction de nombre de GMM.

On constate cette fois que le 512 GMM donnent la meilleure valeur en termes de taux de reconnaissance correct qui est égal à **72,92%**.

4.2.1.3. PLP

Comme les deux cas précédents, on montre sur la tableau suivant les résultats de reconnaissance qui correspond aux meilleurs taux correct avec les16 coefficients PLP :

Nb GMM	32	64	128	256	512	1024
Taux global (%)	77,08	77,08	82,29	85,42	**87,50**	**87,50**

Tableau 3.10: Taux correct de reconnaissance par PLP en fonction de nombre de GMM.

On remarque que parmis tous les nombres de GMM que nous avons utilisés, les deux nombres 512 et 1024 donne le meilleur résultat qui correspond à **87,50%**.

4.2.2. Paramétres Prosodiques

Nous avons fait une étude sur l'utilisation des paramètres prosodiques par une variation sur le nombre de GMM et sur le tableau suivant nous rapportons les taux élevés qui correspond à 2 GMM pour le pitch et l'énergie, et1024 GMM pour les formants :

Paramètre	Pitch	Energie	Formant
Taux global (%)	14,58	27.08	**55,21**

Tableau 3.11 : Taux correct de reconnaissance en fonction des paramètres prosodiques.

Les résultats montrent cette fois que les formants sont les meilleurs paramètres prosodiques car le système réalise le taux de reconnaissance correct le plus élevé.

4.2.2.1. Fusion des paramétrés prosodiques

Nous avons combiné les paramètres prosodiques entre eux avec une variation de nombres de GMM et les resultats de cette fusion sont obtenus avec : 256 GMM pour (pitch-énergie) et 1024 GMM pour (pitch –formant), (énergie-formant) et fusion des trois paramètres. Ces résultats sont mentionnés sur le tableau suivant en fonction de taux correct :

Paramètres	Pitch et Energie	Pitch et Formant	Energie et Formant	Pitch et Energie et Formant
Taux global (%)	42,71	64,58	66,67	**67,71**

Tableau 3.12 : Effet de la fusion des paramètres prosodiques sur le système RAE.

Ces taux obtenus montrent que la fusion des trois paramètres donne la meilleure détection émotionnelle par rapport aux autres fusions avec un taux de **67,71%**.

4.2.3. Fusion des paramètres spectraux et prosodiques

Dans cette expérience, nous avons fait une fusion entre les deux paramètres spectraux et prosodiques dans le but de trouver des taux optimaux. Pour cet objectif nous avons variéle nombre de GMM de 32 à 1024 et le nombre de coefficients de tous les paramètres utilisés (MFCC, LPCC, et PLP) de 12 à 22 coefficientspour chaque fusion.

4.2.3.1. Fusion de paramètre MFCC avec les paramètres prosodiques

Nous présentons ici les meilleurs résultats obtenus par la fusion des coefficients MFCC avec les paramètres prosodiques, qui sont illustrés par la suite : 512 GMM et 12 MFCC pour la combinaison (MFCC-Pitch-Energie-Formant), 1024 GMM et 14 MFCC pour (MFCC-Energie), aussi 1024 GMM et 12 MFCC pour (MFCC-Energie-Formant), 256 GMM et 12 MFCC pour tous les autres fusion qui restes.

Paramètres	MFCC et Pitch	MFCC et Energie	MFCC et Formant	MFCC et Pitch et Energie	MFCC et Energie et Formant	MFCC et Pitch et Formant MFCC	MFCC et Pitch et Energie et Formant
Taux global (%)	84,38	87,50	87,50	85,42	87,50	86,46	**88,54**

Tableau 3.13 : Effet de la fusion entre les coefficients MFCC et les paramètres prosodiques.

Une amélioration est obtenue par la fusion. La combinaison entre le MFCC avec les trois paramètres prosodiques nous a permis une meilleure détection émotionnelle avec un taux de reconnaissance correct de **88,54%.**

4.2.3.2. Fusion de paramètre LPCC avec les paramètres prosodiques

De même que l'expérience précédente, on a étudié l'effet de la fusion entre le paramètre LPCC et les paramètres prosodiques.Pour cela, on a varier le nombre de gaussiennes et le nombre de coefficients LPCC.Nous prenons seulement les meilleurs taux ;qui correspondent à : 512 GMM et 12 LPCC pour (LPCC-Pitch),512 GMM et 20 LPCC pour (LPCC-Formant) ainsi 1024 GMM et 18 LPCC pour (LPCC-Energie), (LPCC-Energie-Formant),(LPCC-Energie-Formant) et 1024 GMM 20 LPCC pour(LPCC-Pitch-Energie) et (LPCC-Pitch-Energie-Formant). Ces résultats sont mentionnés sur le tableau suivant :

Paramètres	LPCC et Pitch	LPCC et Energie	LPCC et Formant	LPCC et Pitch et Energie	LPCC et Energie et Formant	LPCC et Pitch et Formant	LPCC et Pitch et Energie et Formant
Taux global (%)	69,79	73,96	78,13	76,04	78,12	81,25	**82,29**

Tableau 3.14 : Effet de la fusion entre les coefficients LPCC et les paramètres prosodiques.

D'après le tableau3.14, on constante que la fusion de LPCC avec tous les paramètres prosodiqueshisse le taux de reconnaissance à **82.29%**

4.2.3.2. Fusion de paramètre PLP avec les paramètres prosodiques

Dans cette tache,nous rapportons les résultats les plus intéressants de la fusion entre le paramètre PLP et les paramètres prosodiques qui sont obtenus comme suite : 256 GMM et 14 PLP pour les cas (PLP-Pitch-Energie) et (PLP-Pitch-Formant). Ensuite 512 GMM et 16 PLP pour (PLP-Formant). 512 GMM et 20 PLP pour (PLP-Energie-Formant). 1024 GMM et 12 PLP pour (PLP-énergie). 1024 GMM et 14 PLP pour le PLP avec les trois paramètres prosodiques. Enfin, 1024 GMM et 18 PLP sont utilisés pour (PLP-Pitch).

Paramètres	PLP et Pitch	PLP et Energie	PLP et Formant	PLP et Pitch et Energie	PLP et Energie et Formant	PLP et Pitch et Formant	PLP et Pitch et Energie et Formant
Taux global (%)	84,38	86,46	86,46	85,42	**89,58**	86,46	86,46

Tableau 3.15 : Effet de la fusion entre les coefficients PLP et les paramètres prosodiques.

Le tableau 3.15, montre que le système de reconnaissance automatique d'émotion est amélioré par la fusion de (PLP-Energie-Formant) par un taux de reconnaissance correct de **89.58%.**

4.2.4. Taux de reconnaissance par émotion

Pour montrer la détection individuelle par émotion obtenue avec notre système, nous avons considéré le meilleur castrouvé avec 20 coefficients PLP et 512 GMM qui donneun taux optimal de 89.58%. Les résultats sont illustrésdans le tableau 3.16

Emotions	Neutre	Calme	Heureux	Triste	Colère	Effrayé	Dégoût	Surpris
Taux global (%)	**100,00**	75,00	83,33	91,67	91,67	91,67	**100,00**	83,33

Tableau 3.16 : Taux correct de reconnaissance par émotion

D'après les résultats rapporter sur le tableau, l'utilisation du modèle GMM-UBM permet d'améliorer le taux de la détection émotionnelle individuel. On constate aussi, que le système RAE détecte la neutralité et le dégoût par un taux parfait de **100%.**

5. Conclusion

Dans ce chapitre, nous avons présente notre système de reconnaissance des émotions a travers la voix. Ce système a été validé par ces phases d'apprentissage et de test, en utilisant les deux techniques de modélisation DTW et GMM-UBM. Une extraction des trais, est effectuée par lesparamètres spectraux (MFCC, LPCC, PLP) et les paramètres prosodiques (énergie, Pitch, Formant). Dans notre système de RAE, on a testé la détection des émotions par ces paramètres individuellement puis avec différentes combinaisons entre eux. Différentes configurations du nombre de coefficients et de GMM ont été testées dans le modèle GMM-UBM dans le but d'amélioration ce système. Le meilleur résultat donne un taux de reconnaissance de 89,58% en utilisant la combinaison de PLP-Energie-Formant.

Conclusion Générale

Dans notre travail, nous avons d'abord proposé un système complet de reconnaissance automatique de l'état émotionnel d'un locuteur. Pour atteindre cet objectif, nous avons commencé par une étude théorique, où nous avons présenté dans le premier chapitre les différentes notions liées aux émotions et la parole. Ensuite, dans le deuxième chapitre nous avons étudié avec plus de détails les étapes qui permettent de réaliser le système de RAE tels que l'extraction des trais acoustique par les paramètres spectraux (MFCC, LPCC, PLP), et les paramètres prosodiques (énergie, formant et pitch). Aussi, on a étudié la création des modèles d'états émotionnels par les techniques de modélisation DTW et GMM_UBM.

Dans le dernier chapitre on s'est penché à la mise en pratique du système RAE par le logiciel MATLAB. Comme tout système de reconnaissance, un premier travail consistait à organiser les données utilisées en trois parties distinctes : partie de développement, partie d'apprentissage, et partie de test. Ensuite, une extraction des trais acoustiques par les paramètres prosodique et spectraux, et une création des modèles d'état émotionnel de chaque apprenant par les deux méthodes GMM-UBM et le DTW sont effectuées. Enfin on test les émotions et on obtient la performance du système via le taux de reconnaissance correct.

Les différentes expériences réalisées sur le système nous ont permis de constater que le système le plus performant en DTW est obtenu pour le paramètre LPCC avec un taux de 84.38%. Alors qu'en GMM-UBM le paramètre PLP donne le meilleur résultat qui est de 87.50%. En outre la fusion des paramètres prosodiques et spectraux nous a permet d'obtenir une amélioration de performance en terme de taux de reconnaissance, puisque nous avons obtenu 89,58 % comme taux correct. Dans ce cas on a fusionné le paramètre PLP avec l'énergie et les formants.

Comme perspectives de travaux futurs pour améliorer les performances du système RAE, nous indiquons ; l'utilisation d'autres méthodes d'extraction de paramètres permettant d'améliorer l'ensemble des caractéristiques existantes comme la methode de passage to zero ,i-vector …, ainsi que les techniques de modélisation tellque HMM,VQ . En particulier, il est intéressant de tester la reconnaissance de plusieurs canaux à la place qu'un seul canal en utilisant d'autres sources d'information, comme les expressions faciales, les mouvements du corps ou les interactions physiologiques. Ainssi utiliser des corpus qui collecte plus d'enregistrements pour chaque locuteur et la fusion de classifieurs peut également apporter des améliorations significatives.

Références bibliographiques

[1] M.SENOUSSAOUI,"Amélioration de la robustesse des systèmes de reconnaissance automatique du locuteur dans l'espace des i-vecteurs", Thèse de doctorat, École de technologie supérieure, Université du Québec, 2014.

[2] H. BOUSIBA,T .HADJ ALI, "speaker recognition", mémoire de magister, Institut de génie électrique et électronique, Université M'Hamed Bougara – Boumerdes, Algérie, 2012.

[3] M.AKAK, " la détection des émotions du locuteur à partir des caractéristiques vocales extra-linguistiques", mémoire de magister, Faculté d'Electronique et d'Informatique, Université des Sciences et de la Technologie Houari Boumediene, Alger, Algérie, 17/10/2012.

[4] L. Kerkeni, " Analyse acoustique de la voix pour la détection des émotions du locuteur. Vision par ordinateur et reconnaissance de formes",thèse de doctorat, Université du Maine, Université du Centre,Tunisie, 2020.

[5] https://www.formation-ressources-humaines.com/emotions-definition-types-et-comment-lescontroler/,Consulté le 22.01.2022.

[6] X .Hung Le, "Indexation des émotions dans les documents audiovisuels a partir de la modalite auditive", thèse de doctorat, Institut National Polytechnique de Grenoble,Institut Polytechnique de Hanoı, 2009.

[7] Y.ATTABI, "Reconaissance automatique des émotions à partir du signal acoustique", mémoire de Magister, l'École de technologie supérieure,université de Québec,2008.

[8] A. Trabelsi, "Configuration et exploitation d'une machine émotionnelle", mémoire de Magister, Département d'informatique et de recherche opérationnelle Faculté des arts et des sciences, Université de Montréal, Novembre 2010.

[9] R. Boite, H. Bourlard, T.Dutoit, J. Hancq et H. Leich, " Traitement de la Parole ", Presses Polytechniques Universitaires Romandes, Lausanne, 2000.

[10] I,Jemaa ,"Suivi de Formants par analyse en multirésolution",Thèse de Doctorat, Université de Lorraine Nancy 1,19/02/2013.

[11] https://www.sfu.ca/fren270/phonetique/page3_1.html#start, consulté le 02/01/2022.

[12] A.Ghio, S.Pinto, "Résonance sonore et cavitées supralaaryngées", 2007, hal-01616691.

[13] Le classement des sons de parole — Au son du fle - Michel Billières (verbotonale-phonetique.com). consulté le17/02/2022.

[14]https://mbamci.com/la-reconnaissance-des-emotions-par-lia-bonne-ou-mauvaise-idee/, consulté le 23/02/2022.

[15] S. Gharsalli, "Reconnaissance des émotions par traitement d'images", thèse de doctorat, école doctorale mathematique, informatique, phisique theorique et ingenierie des systemes, université d'orléans, 12 juillet 2016.

[16] A. Hacine Gharbi,"Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole", thèse de doctorat, à la Faculté de Technologie Département d'Électronique, Université Ferhat Abbas-Sétif Algérie, Soutenue le 09 Décembre 2012.

[17] R. JOURANI, "Reconnaissance automatique du locuteur par des GMM à grande marge",thèse de doctorat, Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier), 6septembre2012.

[18] K. Bensafia, M.Chabane, "reconnaissance automatique du locuteur par la méthode du taux de passage par zero", thèse de Magister, Faculté des sciences et sciences appliquées,Université de Mouloud mammri Tizi ouzou, algérie, 2007/2008.

[19] Kh.NEKAB,"Etude comparative des différentes méthodes d'estimation de la fréquence fondamentale des cris des nouveau-nés", Thèse de Doctorat, École de technologie supérieure, Université du Québec, 2013.

[20] R.AJGOU, "Techniques de détection de la période du pitch par les méthodes temps fréquence et temps échelle", Thése de Magister, Université de Biskra, 2010.

[21] C.Clavel, "Analyse et reconnaissance des manifestations acoustiques des émotions de type peur en situations anormales", Télécom ParisTech, 2007.

[22] K.BERBECHE, "Modèles de Markov Cachés : Application à la reconnaissance automatique de la Parole",Mémoire de Magister,Université Mouloud Mammeri,Tizi-Ouzou,2014.

[23] H.EL ASSAAD, "Modélisation et classification dynamique de données temporelles non stationnaires", Thèse de doctorat, Université Paris-Est,Décembre 2014.

ملخص

يعد الكلام أحد أكثر الوسائل اللغوية المعروفة التي يستخدمها البشر للتعبير عن الحالات العاطفية الداخلية.ومنه فإن النظام الذي يسمحبالتعرف التلقائي على المشاعر البشرية يعد أمرًا مثيرًا للاهتمام. ولذلك تهدف مذكرتنا إلى تصميم نظام للتعرف العاطفي عن طريق الصوت. يرتكز هذا النظام علي استخدام ثلاثة معلمات طيفية و هي MFCC و LPCC و PLP بالإضافة إلى المعلمات العروضية؛ الطاقة والصيغ والتردد الأساسي. سيتم تصميم كل نوع من أنواع المشاعر بتقنيتين للنمذجة : GMM-UBM و DTW.و من اجل ذلك لقد تم إجراء العديد من الاختبارات من أجل العثور على أفضل معدل للتعرف الصحيح. و قد حققت تقنية PLP أفضل نتيجة بنسبة 87.50٪ ثم تم تحسين النظام من خلال إدماج المعلمات الطيفية و العروضية و التي تقابل نسبة 89.58٪.

الكلمات المفتاحية: التعرف ، العواطف ، المعلمات العروضية ، DTW ، المعلمات الطيفية ، GMM-UBM.

Résumé

La parole est l'un des moyens linguistiques les plus couramment utilisés par les humains pour transmettre les états émotionnels internes. Par conséquent, un système capable de reconnaissance automatiquement des émotions humaines serait intéressant. Notre mémoire vise, donc, à concevoir un système de reconnaissance émotionnelle par la voix. Le système consiste àutiliser les trois paramètres spectraux MFCC, LPCC, PLP ainsi que les paramètres prosodiques ; énergie, formants et pitch. Chaque type d'émotion sera modélisé par deux techniques : le modèle GMM-UBM et DTW. Plusieurs tests ont été effectués afin de trouver le meilleur taux de reconnaissance correct. En termes de paramètre, la technique PLP a réalisé le meilleur résultat avec 87.50%. Une amélioration additionnelle des performances du système a été obtenue par la fusion entre les paramètres prosodiques et spectraux qui correspond à un taux de 89,58%.

Mots clés: Reconnaissance, émotions, paramètres prosodiques, DTW, paramètres spectraux, GMM-UBM.

Abstract

The speech is one of the most common linguistic means used by humans to convey internal emotional states. Therefore, a system capable of automatically recognizing human emotions would be interesting. Our work aims, therefore, to design a system of emotional recognition by voice. The system consists of using the three spectral parameters MFCC, LPCC, PLP as well as the prosodic ones; energy, formants and pitch. Each type of emotion will be modeled by two techniques: the GMM-UBM modeland DTW. Several tests were carried out in order to find the best correct recognition rate. In terms of parameter, the PLP technique achieved the best result with 87.50%. An additional improvement of system performance was obtained by the fusion between the prosodic and spectral parameters, which corresponds to a rate of 89.58%.

Keywords: Recognition, emotions, prosodic parameters, DTW, spectral parameters, GMM-UBM.