UNIVERSITÉ DU QUÉBEC

THÈSE PRÉSENTÉE À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DU DOCTORAT EN GÉNIE ÉLECTRIQUE

PAR
SOUMAYA GHARSALLAOUI

DÉTECTION ET CLASSIFICATION DE TRAITS PARALINGUISTIQUES PAR DES MÉTRIQUES RYTHMIQUES DE LA PAROLE.

AOÛT 2016

UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

DOCTORAT EN GÉNIE ÉLECTRIQUE (PH.D.)

Programme offert par l'Université du Québec à Trois-Rivières

DÉTECTION ET CLASSIFICATION DE TRAITS PARALINGUISTIQUES PAR DES MÉTRIQUES RYTHMIQUES DE LA PAROLE.

PAR

SOUMAYA GHARSALLAOUI

Adel Omar Dahmane, directeur de recherche	Université du Québec à Trois-Rivières
Ismail Biskri, président du jury	Université du Québec à Trois-Rivières
Sid Ahmed Selouani, codirecteur de recherche	Université de Moncton
Habib Hamam, évaluateur externe	Université de Moncton
Douglas O'Shaughnessy, évaluateur externe	INRS-EMTélécommunications

Thèse soutenue le 25 05 16

Résumé

La reconnaissance automatique des traits paralinguistiques de la parole contribue à l'élaboration et au décodage des modèles cognitifs humains. Cette tâche est complexe car les chercheurs étaient et sont encore confrontés à plusieurs défis, particulièrement la dépendance entre les différentes catégories des traits paralinguistiques et la variabilité interlocuteurs.

C'est dans ce contexte que s'inscrit notre objectif qui consiste à proposer une méthodologie capable d'améliorer les performances des systèmes de reconnaissance des traits paralinguistiques actuels. L'idée est d'intégrer les concepts du modèle d'émotion dimensionnel dans la conception de classificateurs d'émotions discrètes. Deux concepts ont été dégagés du modèle dimensionnel : l'existence d'un espace dimensionnel dans lequel les émotions catégorisées peuvent être projetées et l'existence d'une relation de proximité entre ces catégories d'émotion relativement à chacune de ces dimensions. Le premier concept s'est traduit par l'extraction de traits de haut niveau destinés à jouer un rôle similaire à celui incarné par les dimensions du modèle théorique. Le second a motivé l'adoption d'une approche basée sur la similarité pour la représentation et la classification des émotions. Nous avons montré que les scores de vraisemblances générés par les modèles Gaussiens constituent de puissants traits de similarité pour la Reconnaissance Automatique des Émotions (RAÉ) et répondent bien à la contrainte relative à la taille limitée des énoncés.

Dans ce contexte, nous avons présenté une nouvelle métrique rythmique intitulée OPVI. Cette métrique a été conçue pour réaliser la généralisation et l'optimisation des autres métriques de rythme basées sur la variation de l'index (PVI). La normalisation de cette métrique réalisée localement au niveau des syllabes a permis d'atteindre l'objectif de généralisation des métriques standard rPVI, nPVI et les CCI. Cette normalisation a été possible grâce à des coefficients d'optimisation. Un des principaux avantages de la métrique OPVI proposée par rapport aux métriques actuelles est sa capacité à résoudre le problème de la variabilité interlocuteur. Nous avons proposé également de calculer des métriques rythmiques basées sur le paramètre intensité afin d'améliorer la performance des systèmes de classification des traits paralinguistiques en prenant en considération l'aspect prosodique basé sur l'intensité de la parole.

Un modèle d'oreille est proposé pour l'extraction de huit paramètres acoustico-phonétiques les plus pertinents pour la détection des traits paralinguistique. Ces paramètres auditifs sont de nature acoustique et à base d'énergie et permettent selon notre hypothèse d'obtenir des traits phonétiques qui conduisent à la discrimination souhaitée de traits paralinguistiques. L'expérimentation de cette composante a été motivée par la capacité de l'oreille humaine à la distinction entre les différents traits paralinguistiques de la parole.

Différents modèles de combinaison sont proposés et optimisés afin d'améliorer la performance des systèmes de reconnaissance des émotions et la détection des accents natifs et non natifs. Les résultats de ces systèmes sont comparés avec les systèmes de référence. Par ailleurs, nous avons montré que les systèmes conçus basés sur la combinaison des nouveaux paramètres acoustico-phonétiques (OPVI et paramètres auditifs) et les paramètres acoustiques classiques sont puissants et plus performants que les systèmes de

base actuels. Ils constituent de ce fait, des solutions très adéquates au problème de reconnaissance automatique des traits paralinguistiques plus particulièrement la RAÉ et la reconnaissance des accents natifs et non natifs.

Ce travail de recherche s'est également intéressé aux techniques de sélection optimales qui permettent de déterminer les descripteurs acoustiques les plus discriminatoires. La combinaison des méthodes de sélection et des classificateurs adaptés a permis de dépasser les résultats de l'état de l'art. Toutes les expérimentations ont été effectuées sur des corpus de référence à savoir LDC Émotion et le corpus LDC de langue arabe prononcé par des arabophones et anglophones.

Remerciements

Au terme de ce travail, je tiens à remercier mon directeur de recherche Adel Omer Dahmane, professeur de Département de génie électrique et génie informatique et mon co-directeur Sid Ahmed Selouani professeur à l'Université de Moncton, campus de Shippagan de m'avoir donné l'opportunité de faire cette recherche de doctorat, pour leurs conseils, encouragements, orientations et aussi pour leur confiance pour me laisser libre d'explorer de nouvelles idées.

J'adresse mes remerciements également aux laboratoires universitaires de l'UQTR laboratoire des microsystèmes et télécommunications (LMST) et le Laboratoire de recherche en Interaction Humain-machine (LARIHS) de l'UMCS pour avoir permis et encadré cette recherche.

Enfin, j'adresse mes remerciements à ma grande famille, et en particulier à ma mère, pour leurs encouragements et leurs supports moraux durant mes années d'études. À la mémoire de mon père, je dédie ce travail.

Table des matières

Liste des tableaux

Liste des figures

Liste des symboles

ACP	: Analyse en composantes principales
AE	: Algorithme évolutionnaires
ANOVA	: Analyse des variances
ASR	: système automatique de reconnaissance
CCI	Control and Compensation Index
dB	: Décibel
DE	: évolution différentielle
DBM	: Deep Bolmazann multimodale
DBN	: Deep Belief Network
EIHP	: Ensemble Interval Histogram processing
EM	: Expectation Maximization
F0	: Fréquence fondamentale
FS	: Fréquence d'échantillonnage
GFCC	: Gammatone frequency cepstral coefficients
GMM	: Modèle de Mélange Gaussienne
HMM	: HMM Hidden Markov Model

HNR : Harmonics to Noise Ratio

INTERSPEECH : International Speech Communication Association

KNN : *k* plus proches voisins (k-Nearest Neighbor)

LDA : Analyse discriminante linéaire

LDC : Linguistic Data Consortium

LFPC : Log Frequency Power Coefficients

LP : Linear Prediction

LPC : LPC Linear Predictive Coding coefficients

LPCC : LPCC Linear Prediction Cepstral Coefficients

MFCC : Mel Frequency Cepstral Coefficients

ML : Maximum likelihood

MSA : Modern Standard Arabic

OEP : Optimisation par Essaim Particulaire

OPVI : Optimised Pairwise Variability Index

PDF : Probability Density Function

PLP : Perceptual Linear Prediction

PNCC : Power Normalized Cepstral Coefficients

PVI : Pairwise Variability Index

RAE : Reconnaissance automatique des émotions

RBF : Radial basis function

RBM : Restricted Boltzmann Machine

RL : Régression logistique

SVM : Support Vector Machine

Chapitre 1 - Introduction

1.1 Contexte et motivations

La communication exploite plusieurs canaux perceptifs : l'audition, la vision, l'olfaction et le toucher. L'être humain à accès aux formes visuelles, aux sifflements, aux gestes, aux peintures, à l'écriture et surtout à la forme vocale de la communication. Grâce à la parole, qui est le moyen primordial de la communication, les humains échangent et partagent des informations et des idées pour établir des relations qui les relient les uns aux autres. Cet outil de communication leur permet de se rencontrer, de se comprendre et d'aller plus loin dans leurs relations.

> « Nous ne sommes Hommes et nous ne tenons les uns aux autres que par la parole. »[1]

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. Également, nos discours transmettent deux natures d'information à nos interlocuteurs : verbales et non verbales.

La forme verbale, émise par la parole, est constituée par des mots d'une langue donnée. L'intention principale derrière cette forme de communication est la transmission du message que nous voulons communiquer à notre interlocuteur. Les linguistes se sont

intéressés à la transmission du sens du message entre deux individus. C'est la raison pour laquelle au début les chercheurs et surtout les linguistes considéraient que la langue était le centre de la communication humaine. Selon eux, elle était capable de transmettre tous les types d'information simples ou complexes. D'ailleurs, leur concentration était sur la forme vocale de la communication, car la langue dans la forme orale était apparue bien avant l'écriture dans l'histoire humaine. En conséquence, la linguistique s'intéressait beaucoup plus à la forme sonore ou orale du langage qu'à sa forme écrite.

À partir des années 50, Trager a constaté que la capacité de la langue à transmettre de l'information était limitée. Cette limitation découlait des facteurs externes se rapportant à la langue. Ces derniers l'accompagnaient pour déchiffrer les informations explicites intégrées dans le discours [2] [3]. D'ailleurs, d'après Saussure, la parole était l'utilisation personnelle de la langue avec toutes les influences du personnage parlé telles que : le style, le rythme, la syntaxe et la prononciation [4]. Nos paroles n'étaient plus un message sémantique à transmettre à l'interlocuteur parce qu'elles étaient porteuses d'informations personnelles telles que : la prononciation de la langue du message, le ton de la voix, le genre et l'état émotionnel de la personne lors de la communication. Ces renseignements de natures personnelles ne pouvaient pas être identifiés par la communication verbale. On parlait alors de la communication non-verbale.

À l'antithèse de la communication écrite, la communication vocale n'offrait pas la chance à l'auditeur d'écouter une deuxième fois cette dernière dans l'intention de mieux comprendre le message. C'était la raison pour laquelle l'orateur devait connaître et comprendre les indices non verbaux de l'auditeur. On parlait alors de traits

paralinguistiques de la parole. Ceux-ci ont contribué à l'élaboration et au décodage des modèles cognitifs humains.

L'analyse automatique des traits paralinguistiques de la parole représentait le sujet de challenge d'INTERSPEECH durant les années 2009 à 2013. Cet intérêt démontre que cette discipline est devenue un sujet grandissant aux yeux du grand public dans le domaine de traitement automatique des langues et de la parole. La connaissance de traits paralinguistiques est vue comme le moyen d'améliorer l'efficacité des interfaces d'interaction vocale entre l'homme et la machine en offrant la possibilité d'interagir avec n'importe quel utilisateur et d'atteindre son but sans blocage. C'est dans ce cadre que se situe notre travail de recherche, qui vise à développer une méthode de détection et de classification de traits paralinguistiques par des métriques rythmiques de la parole. Cette thèse s'intéresse particulièrement aux émotions et à l'accent non natif comme traits paralinguistiques de la parole.

1.2 Problématique

La détection des traits paralinguistiques de la parole est une tâche complexe avec laquelle les chercheurs étaient et sont encore confrontés à plusieurs défis. Ces derniers s'articulent autour de plusieurs aspects incluant jusqu'au choix du terme « paralinguistique », de sa définition et de sa modélisation automatique. Ces défis vont constituer la problématique majeure à laquelle sont confrontés les chercheurs du domaine. Nous présentons dans ce qui suit les principaux défis.

Définition d'un trait « paralinguistique » : Ce nouveau phénomène, au début de son apparition, était considéré comme complémentaire au trait linguistique même

par l'attribution de son nom. Cet aspect du « paralangage » signifie « à côté de la langue ». Après son affectation dans les autres domaines, sa définition est devenue une question reliée à une discipline. Par la suite, il y a eu absence de consensus pour la définition du terme paralinguistique. Un groupe de chercheurs a donné des définitions élaborées et l'autre groupe a donné des définitions restreintes telles que celles énoncées par Crystal et Abercrombie [151]. En dépit des résultats de recherches par ces deux groupes, il n'y a eu aucune définition bien déterminée concernant les éléments qui appartenaient à ce domaine ainsi qu'à la nature de descripteurs utilisés ;

La catégorisation des traits paralinguistiques. Celle-ci a causé un grand défi pour les chercheurs dans ce domaine, en raison de l'existence de plusieurs bases de classification en catégories. Par exemple, en se basant sur le critère de temps, on pouvait distinguer trois catégories existantes comme : les traits à court, à moyen et à long terme. Cette divergence sur le critère temps révélait le problème de segmentation du signal de la parole et menait au questionnement de la longueur de l'unité à utiliser comme base de segmentation et de traitement. De plus, il faut noter que les traits paralinguistiques sont subjectifs vu que chaque locuteur a des traits de caractères individuels ;

L'ambiguïté au niveau de l'utilisation du terme « paralinguistique ». Diverses significations ont été données à ce terme. En effet, ce terme a été utilisé pour désigner les traits acoustiques de la parole comme : la prosodie, la qualité de la voix et le rythme ; et selon d'autres études, il a été également utilisé pour définir les traits comme : l'âge, le sexe, l'état émotionnel [5], [3];

L'interdépendance entre les catégories paralinguistiques. Plusieurs études ont démontré qu'il existait une interdépendance entre plusieurs catégories de traits paralinguistiques. À titre d'exemple, nous citerons celles qui caractérisaient la relation entre la catégorie biologique et la catégorie état, et qui expliquait que les locuteurs non natifs étaient moins crédibles que les locuteurs natifs [6]. Également, le niveau d'éducation qui pouvait influencer l'identification des locuteurs non natifs d'une langue ;

Les corpus actés vs. spontanés. L'enregistrement et l'utilisation de corpus de données pour effectuer les expérimentations d'analyse des traits paralinguistiques ont conduit à un questionnement touchant à la fiabilité des locuteurs et des données lors de l'expression de leurs traits dans le cas des corpus émotionnels actés. Cependant il faut noter que les corpus spontanés sont très difficiles à construire car ils sont affectés par le changement de scénarios lors de la collecte de données, ce qui explique d'ailleurs leur rareté ;

Le choix des descripteurs pertinents était un défi dû à l'existence de plusieurs descripteurs qui n'étaient pas tous significatifs. Certaines études se sont concentrées particulièrement aux métriques du rythme. En 2011, Loukina a démontré que l'utilisation d'une métrique de rythme ne pouvait pas faire la distinction entre les langues et que les rythmes étaient jusqu'à certains points dépendants du locuteur [7]. Cette dépendance a créé un autre défi qui était relié à la variabilité interlocuteurs [8]. La plupart des recherches focalisait sur les métriques rythmiques comme descripteurs en les calculant sur la base de la durée, mais certaines études

ont déduit que les rythmes avaient des aspects autres que la durée, comme par exemple l'intensité [9];

Inégalité des ressources et des recherches selon la langue. Les aspects touchant spécifiquement une langue donnée tels que les variétés linguistiques ou accents régionaux ont été abordés par différentes recherches. Cependant on note des lacunes lorsqu'il s'agit d'étudier des langues peu dotées. Par exemple, l'étude de l'identification des accents natifs et non natifs de la langue arabe est confrontée à un manque flagrant de ressources en comparaison aux autres langues.

1.3 Objectif principal

L'objectif principal de cette thèse consiste à concevoir un système de reconnaissance automatique innovant de traits paralinguistiques à partir du signal acoustique de la parole tout en exploitant des techniques avancées de traitements acoustiques du signal. L'intérêt est porté particulièrement sur les informations acoustiques basées sur le rythme de la parole et la modélisation auditive pouvant contribuer efficacement à la reconnaissance de l'état émotionnel et de l'accent du locuteur. Pour valider nos résultats, nous ciblons en premier lieu l'identification des classes d'émotions du corpus de données utilisées LDC Emotional Prosody et en deuxième lieu, l'accent des locuteurs natifs et non natifs en langue arabe en utilisant le corpus MSA LDC West Point.

1.4 Objectifs spécifiques

Pour atteindre l'objectif principal, nous avons défini les objectifs spécifiques suivants pour lesquels nous avons apportés des contributions originales :

a) Intégration d'un modèle auditif pour extraire des vecteurs de descripteurs pour la détection et l'identification de traits paralinguistiques ;

b) Proposition d'une nouvelle métrique rythmique pouvant généraliser et optimiser d'autres métriques rythmiques PVI ;

c) Exploitation du paramètre prosodique de l'intensité par l'élaboration d'un algorithme pour le calcul du rythme standard à base d'intensité ;

d) Intégration originale de paramètres segmentaux à court terme et suprasegmentaux modélisant les variations temporelles, fréquentielles et énergétiques pour caractériser les traits paralinguistiques ;

e) Conception d'une approche basée sur les algorithmes évolutionnaires pour la réduction optimale des paramètres acoustiques ;

f) Optimisation *front-end* et *back-end* du système de reconnaissance de traits paralinguistiques de la parole en utilisant les algorithmes évolutionnaires.

1.5 Domaines d'application de la reconnaissance des traits paralinguistiques

La reconnaissance automatique de traits paralinguistiques recèle d'un potentiel d'applications très important à cause de la valeur ajoutée apportée à l'interaction humain système. Dans ce qui suit nous donnons quelques exemples illustratifs.

Interprétation de l'intention des locuteurs : c'est la compréhension des paroles exprimées à partir de la façon utilisée pour le dire. La clarté du langage naturel peut bénéficier et exploiter des éléments paralinguistiques, comme lorsqu'on veut connaître l'état émotionnel de l'individu [10] [11];

Analyse des conversations et la transmission : on procède à une analyse, assistée par ordinateur, de la communication humain-humain, qui inclut l'enquête de la synchronie dans la prosodie des couples [12] ou bien des types spécifiques de discours en psychologie [13], ou encore l'analyse et la synthèse de réunions pour ceux qui ont des implants cochléaire [14] [15]. Les enfants atteints d'autisme qui peuvent tirer profit de l'analyse des signaux émotionnels, étant donné qu'ils peuvent avoir des difficultés de compréhension ou d'expression de leurs sentiments [16]. Également, la transmission d'informations paralinguistiques avec d'autres éléments du message peut être exploitée dans le but d'animer des avatars (personnages virtuels) [17]. Finalement, afin d'enrichir les appels des boîtes vocales, celles-ci sont étiquetées par des symboles tel que des émoticônes [18];

Gestion de la qualité de services des centres d'appels : les gestionnaires de centres d'appels ont un vif intérêt concernant le suivi et l'optimisation touchant la qualité des services fournis par leurs agents [19]. L'identification automatique de l'émotion à travers la voix permettra de faire le suivi de la qualité de la relation avec les clients ;

Interaction avancée dans les jeux : des applications de RAÉ sont intégrées dans les jeux afin de leur apporter plus de réalisme tel que le détecteur d'amour. Celui-ci essaie de classer la parole en se basant sur combien « d'amour » elle véhicule. La détection de stress et de l'interaction vocale permettent la conception de jeux thérapeutiques ayant comme objectif une meilleure interaction verbale et une amélioration des techniques de jeux de rôle ;

Application dans le domaine de la santé : afin d'aider les aînés à vivre plus longtemps dans leurs résidences, on peut se servir d'une surveillance acoustique détectant les douleurs et classant automatiquement les appels de détresse par la reconnaissance automatique de la voix et de traits paralinguistiques [20]. De plus, on peut exploiter cette classification pour diagnostiquer les maladies et les troubles de la parole [21], comme : la maladie de Parkinson [22], l'ablation du larynx pour cause de cancer et les effets pathologiques [23];

Incorporation dans les systèmes de tutorat : Les éléments paralinguistiques sont essentiels pour les tuteurs et les étudiants afin que l'apprentissage soit une réussite [24] [25]. En outre, ces éléments sont utilisés dans l'intention de fournir un meilleur discours public ou tout simplement pour avoir l'intonation correcte lors de l'apprentissage de langues étrangères [26];

Applications en robotique : l'analyse des états affectifs (émotions) et de la personnalité est encore très rudimentaire en robotique. En ayant une meilleure modélisation de ces états et de ces traits, nous serons en mesure d'ajouter aux robots des compétences sociales humaines dans le but d'avoir des robots interactifs et communicants [27] pouvant jouer le rôle de robots d'assistance [28];

Surveillance et monitoring : les traits paralinguistiques soutiennent de nombreuses situations comme la gestion de crises liées à la sécurité [29], la surveillance du niveau de stress, la somnolence, l'ivresse et autres. Dans le but de détecter et de résoudre ces états, on utilise la parole comme modalité d'analyse. En outre, les systèmes de lutte contre le terrorisme peuvent être plus efficaces en analysant les

éléments paralinguistiques tels que : l'agressivité d'agresseurs potentiels [30], ou la peur des victimes potentielles [31];

Utilisation dans les médias: l'information paralinguistique est intéressante pour de multiples types de recherche dans les médias, tels que la mesure du niveau d'excitation dans le discours d'un journaliste [32] ou d'intérêts des auditeurs et téléspectateurs;

1.6 Organisation du manuscrit

Cette thèse est structurée en sept chapitres. Son contenu touche à la définition du domaine de recherche et de ses différents défis. Une étude approfondie de l'état de l'art a permis de fixer les objectifs et les hypothèses afin de répondre à la problématique, d'évaluer la validité des solutions proposées, et enfin de comparer nos résultats avec ceux des systèmes de base.

Dans le chapitre 1, l'introduction porte sur le contexte et la motivation de l'étude du domaine de traits paralinguistiques. Nous y décrirons les problèmes auxquels sont confrontés les chercheurs. Ceux-ci sont expliqués afin d'éclaircir les objectifs et les stratégies pour les atteindre.

Le chapitre 2 présente la terminologie spécifique associée aux traits paralinguistiques notamment l'émotion ainsi que les accents. On y présente les définitions, les théories, les modèles, les corpus ainsi qu'une étude de la littérature en relation avec ces derniers.

Le Chapitre 3 décrit le système automatique de la détection et de la classification de traits paralinguistiques. On y présente les différentes étapes en commençant par l'extraction

des paramètres jusqu'à la classification. Ces étapes sont détaillées en rapport avec la revue de la littérature du domaine.

Le chapitre 4 introduit les nouveaux descripteurs proposés et utilisés au cours de cette recherche, tels que l'OPVI, le calcul des mesures rythmiques à base d'intensité et du modèle auditif.

Le chapitre 5 décrit l'approche multivariable segmentale et suprasegmentale, le classificateur utilisant cette approche, l'optimisation utilisée pour la réduction des descripteurs acoustiques au niveau de l'étape de la classification ainsi que les résultats de leur application.

Le chapitre 6 discute l'amélioration apportée par les nouvelles métriques rythmiques et les méthodes au niveau du système de reconnaissance de traits paralinguistiques.

Le chapitre 7 conclut cette thèse et donne les perspectives de ce travail de recherche.

Chapitre 2 - Traits paralinguistiques de la parole

2.1 Introduction

Ce chapitre présente une revue de la littérature sur l'analyse des traits paralinguistiques. Afin de circonscrire le champ d'étude de cette thèse, nous commençons par préciser le concept de paralangage qui couvre de nombreux domaines. L'étude des traits paralinguistiques sera focalisée principalement sur les deux aspects qui nous intéressent dans cette thèse à savoir l'émotion et l'accent natif et non natif du langage véhiculé par le signal de parole.

2.2 Historique et définition de la paralinguistique

La paralinguistique touche à la discipline qui traite de différents éléments qui accompagnent la parole et qui contribuent à la communication. Ces derniers ne font pas partie du système linguistique. Par contre, ils peuvent être présents sous forme acoustique (vocale et non verbale) ou linguistique. Ce nouveau terme «paralinguistique» a été présenté comme composante de la parole par Trager [3] et a été développé plus tard par Crystal [33] [34].

Trager a déterminé le paralangage comme la partie qui accompagne la langue et qui a la forme sonore. Les gestes du corps humain ne font pas partie du paralangage selon lui, puisque c'est la communication silencieuse. Le paralangage touche aux éléments non verbaux de la communication utilisés pour modifier le sens et l'émotion. Ils peuvent être

exprimés consciemment ou inconsciemment. Le paralangage complète la partie linguistique de la conversation, car la compréhension totale de la langue ne peut être atteinte que par la considération de ces éléments[2] [3].

De même, en 1945, Pike a introduit l'existence d'autres composantes de la parole qui devançaient la linguistique et qui appuyaient la compréhension du message du locuteur [35] mais il ne les a pas identifiées par un terme spécifique comme Trager.

Trager considérait le secteur du paralangage spécifiquement pour la communication humaine par l'analogie avec l'expression vocale de plusieurs animaux. Par opposition, Abercrombie jugeait que le terme paralangage pouvait être exploité pour la communication humaine et animale [36].

Le paralangage a progressé au cours des années, grâce à son utilisation dans des domaines autres que la linguistique et en raison des études effectuées par d'autres chercheurs que les linguistes. Plusieurs descriptions et définitions sont ressorties de ce progrès touchant à ce nouveau domaine. Celles-ci ont été regroupées par Crystal en deux grands groupes : le premier était étroit et le deuxième était large. Il s'est servi de la définition large de ce nouveau domaine. Effectivement, selon lui le paralangage était la communication humaine vocale non segmentale qui excluait tous les phénomènes autres que vocaux [37] [34].

Selon Abercrombie la définition de Crystal, relative aux aspects larges et étroits de ce domaine, n'était pas encore juste et évidente. En effet, il considérait que la paralinguistique est large lorsqu'il y avait inclusion des descripteurs linguistiques de la parole dans son processus et étroite lorsque ceux-ci sont exclus [36].

Crystal estimait que le domaine du paralangage avait trois tendances fondamentales d'évolution. La première se rapportait à la découverte et au développement du domaine par Trager, ainsi que son application en linguistique. La seconde, touchait à l'application du paralangage dans des sphères autres que celles mentionnées par Trager, comme le domaine de la psychothérapie. En plus, l'évolution de la langue a donné de nouvelles définitions au terme paralangage. La troisième tendance s'intéressait au développement d'une confusion théorique causée par les deux premières tendances. Cette dernière connue au niveau du domaine était le résultat de l'absence d'études évolutives du nouveau terme qui permettraient de faire face aux changements que la langue a connus. Cette ambigüité a été mentionnée par Crystal lors de son étude du domaine [33]. D'après lui, plusieurs facteurs étaient à l'origine de cette équivoque liée au domaine de paralinguistique telle que : l'absence d'une description ou définition standard et unique du domaine : au début le paralangage était un domaine marginal pour Trager. En effet, il a été sous l'influence de sa spécialité comme linguiste lors de la description et définition du nouveau domaine. Cette influence est une partie de la description du domaine, de l'attribution du terme « paralangage » et aussi de la définition du terme comme partie complémentaire de la langue. Trager a allié le terme paralangage seulement à la phonétique et à l'orphique qui sont deux domaines propres à la linguistique. Cette insuffisance de clarté a conduit à l'apparition de plusieurs descriptions du paralangage et des signes paralinguistiques. Crystal les a classés en sept descriptions :

a) Communication humaine et animale ;

b) Communication humaine seulement, incluant ce qui est vocal et non vocal ;

c) Communication vocale seulement, incluant segmentale et non segmentale ;

d) Communication vocale non segmentale, incluant la qualité de la voix ;

e) Non segmentale, autre que la qualité de la voix ;

f) Non segmentale, autre que la qualité de la voix et la prosodie ;

g) Une partie très petite du non segmental et autre que la prosodie et la qualité de la voix ;

Contrairement à la communication écrite, la communication orale ne peut pas être « relue ». Les auditeurs n'ont qu'une seule opportunité de tout capter et de comprendre le message véhiculé par la parole. D'ailleurs, la compréhension du message peut seulement être atteinte par la considération des éléments paralinguistiques. Ces éléments du paralangage ne sont pas synchronisés avec les mots de la communication verbale. Ils peuvent arriver antérieurement ou ultérieurement selon le contexte, mais ils sont toujours intégrés dans la conversation de façon consciente ou non consciente. Les buts des signes paralinguistiques de la parole sont :

- Répéter ce qui a été exprimé verbalement ;
- Compléter ce qui a été exprimé verbalement ;
- Contredire ce qui a été exprimé verbalement ;
- Substituer à ce qui sera exprimé verbalement ;
- Réglementer et gérer l'événement de communication.

2.3 Catégories de traits paralinguistiques

L'étude paralinguistique de la parole indique que les traits paralinguistiques sont généralement en relation avec l'état émotionnel, la personnalité, l'identification du groupe

social ainsi que la variété du langage. La classification par catégorie de ces éléments s'effectue selon plusieurs critères menant à plusieurs types de classification. Les éléments paralinguistiques de la parole peuvent être classés en trois catégories selon leur rapidité de changement [38]

a) **Des traits paralinguistiques à long terme** : ce groupe contient les primitives de caractéristiques biologiques (le poids, l'âge, le sexe), sociales, culturelles, traits de personnalité et dialecte ;

b) **Des traits paralinguistiques à moyen terme** : cette catégorie touche à l'état de santé, l'humeur, les attitudes négatives et positives et les signaux sociaux ;

c) **Des traits paralinguistiques à court terme** : cette famille de traits est aussi connue sous le nom « état ». Il inclut l'émotion, le stress, l'intimité, l'intérêt, l'incertitude et la douleur.

La Figure 2-1 représente le domaine de la paralinguistique avec ses différentes composantes et la classification de celles-ci selon les critères de temps. Cette classification n'est pas l'unique qui existe dans la littérature, car si on change le critère de classification, les catégories changeront.

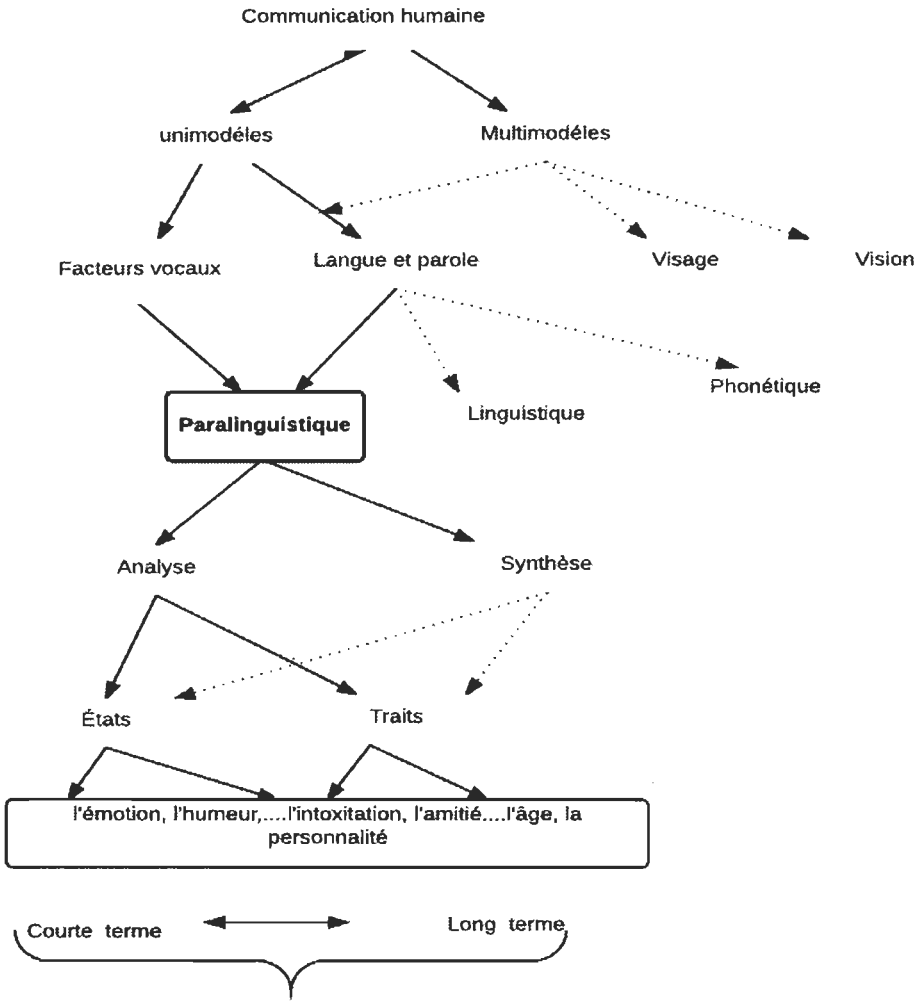

Figure 2–1:La description du domaine des traits paralinguistiques[38].

Les traits et les états paralinguistiques diffèrent au niveau de la durée. En réalité, la variation en fonction du temps varie d'un trait paralinguistique à un autre. Pour le trait âge par exemple, cela prend 10 ans entre l'âge de 20 à 30 ans pour percevoir un changement. Par contre, pour certains états, le changement peut se produire en l'espace de quelques secondes, comme de passer de l'état triste à l'état heureux.

L'âge et le genre peuvent influencer l'expression de l'état émotionnel, car on ne retrouve pas la même expression chez les hommes, les femmes et les adultes. Également, les traits paralinguistiques ont la capacité d'affecter les caractéristiques paralinguistiques à moyen terme, comme dans le cas de l'alcool, l'effet de celui-ci est différent selon le sexe et l'âge des personnes.

La segmentation du signal de la parole en unités, pour l'analyse et le traitement, dépend de cette catégorisation des éléments paralinguistiques en trait ou état. Du fait que les traits sont lents et les états rapides dans leurs changements, cela peut guider une représentation de la longueur de l'unité de la segmentation lors du processus du traitement automatique des traits paralinguistiques. Cependant, si on tient compte de la relation de corrélation entre les traits et les états, cette hypothèse n'est plus valable. D'ailleurs un des défis dans la détection et la classification des éléments paralinguistiques de la parole est l'absence d'un paradigme de segmentation du signal. Néanmoins, il faut préciser que la segmentation du signal est une étape primordiale dans le processus de traitement automatique des éléments paralinguistiques de la parole.

Le changement de traits paralinguistiques exige beaucoup de temps comparé aux états paralinguistiques. Sans compter qu'il diffère d'une personne (locuteur) à une autre, même le changement pour une même personne est souvent inconnu ce qui est des traits paralinguistiques. En revanche, les états paralinguistiques dépendent, dans leurs changements, de facteurs personnels et interpersonnels ce qui entraine lors des traitements automatiques la complexification de la segmentation. D'après le Tableau 2-1, on remarque que les éléments paralinguistiques diffèrent bien entre eux par le degré d'intensité de la

rapidité du changement et de la corrélation avec les autres éléments paralinguistiques, et avec la sensibilité aux différents facteurs d'après Scherer [39].

En plus de la catégorisation à base de temps, et malgré son importance, il existe d'autres types de classification pour les éléments paralinguistiques qui impliquent d'autres critères de classification: spontanée et actée, complexe et simple, quantifiable et qualifiable, intentionnelle et instinctive, ressentie et vue, perceptible et imperceptible, discrète et contenue, prototypique et périphérique, privée et sociale, universelle et culturelle, unimodale et multimodale, et enfin état et trait.

L'objectif derrière tous ces types de classification ainsi que les méthodes de catégorisation consiste à bien définir et décrire un phénomène paralinguistique. Nous pensons qu'il est très important de connaitre préalablement la catégorie de l'élément paralinguistique avant de lancer le processus de sa reconnaissance automatique afin de sélectionner le contexte et les scénarios des données expérimentales les plus adéquats.

Table 2-1:Caractéristiques de démarcation entre les différents états affectifs [80]

Type des états affectifs	Intensité	Durée	Synchronisation	Dépendance	Évaluation de stimulation	Rapidité de changement	Impact sur le Comportement
Émotion	+++++	+	+++	+++	+++	+++	+++
Humeur	++++	++	+	+	+	++	+
Sentiment	++++	+++	+	++	+	++	++
Attitude	0++	++-+++	0	0	+	0++	+
Trait interpersonnel	0-+	+++	0	0	0	0	+
0 : bas, + : moyen, ++ : élevé, +++ : très élevé, - : indique une étendue							

Cependant il faut noter que la majorité des méthodes et théories de catégorisation sont en corrélation les unes aux autres. Gillik a analysé la relation entre les traits démographiques comme le sexe, l'âge, le niveau d'éducation et la région, ainsi qu'entre l'état émotionnel et le sexe qui se retrouvent également dans les travaux de Vogt et Andre [40] [41]. Cette corrélation conduit les chercheurs à bien choisir l'unité de base de la segmentation dans le but de ne pas être privés d'informations et afin de fournir des explications à certains phénomènes lors d'études de ce domaine, et aussi au niveau du choix des descripteurs qui peuvent être significatifs pour la détection et la classification des éléments paralinguistiques. Durant nos recherches, on s'est intéressé aux éléments

paralinguistiques de la parole se rapportant aux états émotionnels et à l'accent natif et non natif de la langue arabe.

2.4 Corpus de données de traits paralinguistiques

La qualité et la fiabilité du système de reconnaissance automatique de traits paralinguistiques dépendent considérablement des données d'apprentissage et d'analyse. l'étape de collecte de données doit définir les paramètres d'ordre technique tels que la fréquence d'échantillonnage, le type de microphone utilisé, ainsi que d'autres considérations reliées au contexte expérimental et à la configuration telles que celles proposées par Batliner [42] :

a) Le choix du type d'enregistrements qui peut être un enregistrement existant ou complètement original. Dans ce qui est existant, on y trouve les enregistrements télévisuels et ceux diffusés sur Internet. Pour la conception originale des enregistrements, divers scénarios sont possibles tel qu'un environnement tranquille impliquant des humains ou l'établissement d'un scénario élaboré entre locuteurs et robots ;

b) Types de corpus spontanés ou actés. Pour le discours acté qui est le plus facile à construire on se basera sur un texte, des phrases isolées, des mots ou juste des syllabes ;

c) La décision sur le nombre et la qualité des locuteurs et le type d'équipement d'enregistrement ;

d) La méthode et l'expertise d'annotation des traits paralinguistiques ainsi que l'établissement d'une norme pour les annotations et l'évaluation de la qualité de ces annotations ;

e) La définition et l'extraction des unités de base retenues pour l'analyse.

Il est nécessaire d'effectuer ces étapes pour l'opération de la collecte des corpus, par contre il existe plusieurs autres étapes complémentaires comme : la correction manuelle de l'annotation automatique, la documentation et la possibilité d'avoir d'autres formes d'enregistrement autre que l'audio.

La création d'une base de données dans ce domaine est une étape très sensible dans le processus de reconnaissance automatique des traits paralinguistiques. Celle-ci exige beaucoup de concentration, de description des détails techniques et du contexte, comme cela est le cas pour plusieurs corpus. Gibbon et son équipe dans [43] ont offert une vision large et approfondie de tous les aspects d'une base de données touchant aux normes et aux ressources pour les systèmes de langues parlées. Cowie et son équipe dans [44] et [45] ont abordé les principes et les questions fondamentales de la collection de bases de données émotionnelles. En 2011, Cowie dans [45] a traité du problème de l'étiquetage de bases de données de traits paralinguistiques tels que les états émotionnels.

2.4.1 L'annotation des corpus

Pour un système de reconnaissance automatique de la parole, il est nécessaire d'avoir la transcription des mots exprimés oralement. Les autres informations telles que le bruit, les pauses et les éléments non-verbaux ne sont pas pertinents et ces derniers sont souvent exclus du processus d'annotation. Par contre, pour le système de reconnaissance de traits

paralinguistiques, il est nécessaire d'enrichir d'avantage les transcriptions afin de caractériser plus efficacement le phénomène paralinguistique. De plus, les informations additionnelles touchant les locuteurs comme l'âge, l'origine, etc. permettront de traiter d'autres éléments paralinguistiques ou de traiter la corrélation entre le phénomène paralinguistique actuel et les informations additionnelles.

L'étiquette donnée à un phénomène paralinguistique peut être binaire (0 ou 1) ou multi-catégories. Cette attribution s'effectue par un ensemble d'annotateurs experts pour cette opération. Les annotations obtenues peuvent diverger car l'annotation est basée sur un processus de perception [46] [47] [42].

2.4.2 Évaluation de l'annotation

Deux critères psychométriques standards sont appliqués afin d'évaluer l'opération d'annotation : ce sont la validité et la fiabilité qui sont applicables selon le type du corpus.

La validité détermine si l'enregistrement a été correctement annoté, c'est-à-dire que le trait paralinguistique perçu correspond au trait vécu.

Plusieurs mesures sont utilisées pour évaluer la fiabilité et l'une d'elles est l'identification de l'accord entre les annotateurs impliqués dans l'opération de l'annotation. Pour l'établissement d'un étalon standard permettant une fiabilité proche de la réalité une approche consiste à utiliser plusieurs annotateurs. Mower a étudié certaines méthodes d'évaluation d'annotateurs [48]. De plus, une méthode basée sur l'entropie, pour l'évaluation des cas de confusion dans le processus de l'annotation d'expressions émotionnelles, a été proposée par Steidl. De même, Hönig et son équipe ont présenté une méthode d'évaluation de la prosodie des accents non-natifs [49] [50].

Étant donné que l'annotation est essentielle et dispendieuse, elle a nécessité la recherche d'une nouvelle approche. Schuller et son équipe ont utilisé le mélange de bases de données sans annotation, ainsi que la technique d'apprentissage semi-supervisée. En 2011 et 2012, cette dernière a permis d'obtenir des résultats satisfaisants tels que présentés dans les travaux de Zhang et Schuller 2012 [51] [52].

2.4.3 Corpus de données de traits paralinguistiques

Il existe un ensemble de corpus de traits paralinguistiques exploités aussi bien en psychologie qu'en ingénierie. Les corpus les plus proches du standard de l'annotation fiable et valide sont expérimentés dans les divers défis proposés dans des conférences reliées au domaine telles que la conférence d'INTERSPEECH. Nous nous restreignons aux corpus employés dans les quatre défis tenus à la Conférence annuelle de l'ISCA, d'INTERSPEECH 2009 – 2012. Ces derniers ont couvert certains des principaux thèmes, ainsi que la plupart des catégories en paralinguistique. Ces corpus sont qualifiés de références dans ce domaine.

2.5 Parole émotionnelle

L'émotion joue un rôle majeur dans les processus cognitifs humains, on doit considérer le facteur émotion pour l'interaction humain-machine. Il est nécessaire, pour la conception d'une interaction fluide entre un humain et la machine, que cette interaction soit fondée sur la maîtrise des états émotionnels par la machine. En effet, celle-ci doit être dotée de capacités à reconnaître et à exprimer des émotions lors de la communication. Cette partie étudie l'émotion sous l'angle du domaine de la psychologie, la description de l'émotion et les modèles théoriques sur lesquels se sont appuyées les études réalisées dans ce domaine.

Nous aborderons notamment les incertitudes, qui caractérisent la définition des émotions, les catégories ainsi que leurs modèles.

Tableau 2-1:Exemple de corpus de traits paralinguistiques [149]

Corpus	Durée (heures)	Nombre de locuteurs	Nombre de classes	Type	Langue	Audio
FAU AEC	8.9	51	5	S	DE	Lab
TUM AVIC	2.3	21	4	S	UK	Lab
aGender	50.6	945	-	P	DE	Tel
ALC	43.8	162	-	P	DE	Lab
SLC	21.3	99	3	P	DE	Lab
SLD	0.7	800	32	P	DE	Lab
TIMIT	4.4	630	-	P	US	Lab
DE : Deutsch	UK : British English	FR : Français	US : American English	Lab : Laboratoire	Tel : Téléphone	

2.5.1 Définition et théories des émotions

« Chacun sait ce qu'est une émotion jusqu'à ce qu'on lui demande d'en donner une définition. À ce moment-là, il semble que personne ne le sache plus ».

C'est ainsi que Willam James a posé la question de la définition de l'émotion « What is an emotion ?» [53]. Sa réponse ainsi que celles des autres chercheurs ont créé un grand débat concernant la définition exacte et simple de l'émotion. Ce débat continue encore aujourd'hui, afin d'avoir un consensus scientifique à propos de cette définition. Une des définitions qui se veut consensuelle est la suivante : «L'émotion est un phénomène rapide,

déclenché par un évènement, qui engendre une réponse émotionnelle cohérente à plusieurs composantes» [54]. Cette définition distingue l'émotion des autres types d'états affectifs à cause de son besoin d'avoir un évènement déclencheur. L'émotion est aussi définie comme un évènement rapide à cause de sa relation avec le temps et de sa rapidité de changement. Afin de mettre l'accent à cette propriété de l'émotion Scherer a utilisé le mot 'épisode émotionnel' au lieu d'état émotionnel [55]. Le phénomène de l'émotion est la source d'un grand défi dans le domaine de la psychologie. En raison de l'absence d'un consensus concernant la définition de différents types d'émotions, et de la différence entre ces derniers et les autres phénomènes affectés de la parole, ce défi est encore présent.

L'intérêt de la recherche pour circonscrire l'émotion humaine a conduit à la naissance de plusieurs théories de celle-ci. Les trois théories les plus répandues et identifiées dans la littérature sont : l'émotion discrète, l'émotion dimensionnelle et l'émotion à composantes. La théorie à composantes de Scherer est jugée la plus adéquate pour l'étude de ce domaine en utilisant l'analyse automatique de la parole. Les trois théories sont différentes, mais elles respectent la définition psychologique de l'émotion.

La théorie de l'émotion discrète a été développée par Darwin et a été utilisée par Ekman [56] [57]. Cette approche se fonde sur le postulat de l'existence des catégories séparées en supposant qu'un évènement spécifique déclenche un nombre limité d'émotions de base qui conduisent à une réponse spécifique. Ces émotions basales ou fondamentales sont caractérisées par des modèles de réponses très spécifiques en physiologie, et sur des expressions faciales et vocales. Une partie des partisans de cette théorie, comme Ekman et Izard [57], se sont concentrés sur la différence entre une émotion simple et complexe, ses effets sociaux culturels et ses différentes composantes et interactions. Dans notre

recherche, nous avons utilisé la joie, la tristesse, la peur, la colère, le dégoût et la surprise comme modèles d'émotion dans l'analyse du signal de parole. Ces six émotions sont considérées comme les bases de l'émotion [56].

La théorie dimensionnelle : Cette théorie trouve ses origines à partir des recherches de William James [53]. Elle met l'accent sur l'importance de deux facteurs principaux ou dimensions valence et intensité de l'activité (actif-passif). Cependant, plusieurs considèrent que l'émotion peut être présentée en trois dimensions. La troisième dimension est le contrôle ou la puissance intellectuelle. La fusion des trois dimensions correspond à l'expérience émotionnelle. Ainsi d'après Osgood qui se base sur l'analyse sémantique de l'émotion caractérise l'émotion par les dimensions suivantes : l'évaluation (négative ou positive), l'activation (faible ou forte) et la puissance (faible ou forte) [46].

Le modèle d'émotion à composantes se base sur l'affirmation que l'émotion est une entité composée. On ne peut pas étudier cette entité comme un tout unitaire, mais il faut la décomposer et l'étudier unité par unité. Scherer en 2010, s'est concentré sur l'évaluation subjective de l'émotion complexe suite à un évènement déclencheur qui a conduit à une réaction [58]. Cette théorie de l'émotion ne cesse de gagner de l'influence. L'émotion selon ce modèle est divisée en cinq composantes qui sont touchées par les changements. Ces cinq composantes sont :

a) L'évaluation qui consiste à donner une valeur à l'émotion ;

b) L'expression qui reflète l'effet corporel tels que la voix, le frisage ;

c) La tendance à une action qui est l'action souvent approchée ou évitée ;

d) La réponse périphérique qui touche au système nerveux ;

e) Le sentiment qui est considéré comme la « conscience » de l'émotion.

2.5.2 Les corpus de la parole émotionnelle

Afin d'étudier l'émotion véhiculée par la parole, les chercheurs ont souvent recours aux corpus actés et induits car la collecte de données de ces corpus est avantageuse pour le processus de reconnaissance automatique des émotions de la parole. En effet, pour ce type de corpus on retrouve trois méthodes pour la collecte de données : la première méthode est l'enregistrement utilisant des acteurs professionnels. Ceux-ci reçoivent un scénario, se préparent et décident comment jouer leurs rôles et les émotions qu'ils vont utiliser pour ceux-ci. Les recherches concernant cette méthode mentionnent que les données de celle-ci sont d'excellente qualité. La deuxième méthode implique des acteurs professionnels ou non. On dit aux acteurs quoi faire, comme par exemple de répéter plusieurs fois la même phrase ou bien de lire un texte et c'est selon un état émotionnel demandé. Le niveau de contrôle de cette méthode est très élevé, car on exerce un contrôle sur l'état l'émotionnel des participants. La troisième méthode est de jouer un rôle spontanément. À titre d'exemple, la simulation peut être de téléphoner à un centre d'appel et de réserver des billets d'avion sans spécifier quel état émotionnel à utiliser. Les corpus actés collectés avec cette troisième méthode sont dits induits. Une liste des corpus est présentée dans le Tableau 2-2.

Les corpus des émotions naturelles sont quant à eux difficiles à obtenir. Ils sont classés en trois groupes. Le premier groupe concerne les émotions naturelles spontanées qui arrivent à tout moment et dans n'importe quel contexte. Le deuxième groupe introduit les émotions naturelles sans scripts, c'est comme par exemple être dans un groupe formé pour effectuer un travail. Chacun a un rôle à jouer, mais il peut sortir à tout moment de ce rôle

pour parler à une personne présente. Ce type est plus ou moins proche du jeu de rôles et des corpus induits dépendamment des paramètres de l'expérience. Le troisième groupe démontre les émotions naturelles avec un script. Les rôles sont prédéfinis et ce n'est pas un comportement naturel. Le contrôle est moyen.

Tableau 2-2: Liste de corpus de la parole émotionnelle disponibles ainsi que leurs descriptions tirées du site web de l'association AAAA (2015) [150]

Identifier	Modalities	Emotional content	Emotion elicitation methods	Size	Language
HUMAINE Database from www.emotion-research.net/download/pilot-db/	Audiovisual + gesture	Categorical and continue	Naturalistic and induced material	50 clips ranging from 5 seconds to 3 minutes	English and some French and Hebrew
Belfast Naturalistic Database (Douglas-Cowie et al 2000, 2003)	Audio- visual	Wide range	Natural	125 subjects; 31 male, 94 female	English
RECOLA - REmote COLlaborative and Affective interactions - Database (Ringeval et al. 2013) ;	Multimodal: audio, video, EDA, ECG	(i) Discrete (5 categories: agreement, dominance, engagement, performance and rapport. (ii) Continuous (arousal and valence).	Natural	34 subjects; 14 male, 20 female	French
Geneva Airport Lost Luggage Study (Scherer & Ceschi 1997; 2000)	Audio-visual	Anger, good humor, indifference, stress, sadness	Natural	109 subjects	
Chung (Chung 2000)	Audio-visual	Joy, neutrality, sadness (distress)	Natural : (television interviews)	77 subjects; 6 1 Korean speakers, 6 Americans	English and Korean
SMARTKOM	Audio-visual and gestures	Joy, gratification, anger, irritation, helplessness, pondering, reflecting, surprise, neutral	Human machine in WOZ scenario	224 speakers ; 4/5 minute sessions	German
Amir et al. (Amir et al, 2000)	Audio + physiological (EM G, GSR, Heart Rate, Temperature, Speech)	Anger, disgust, fear, joy, neutrality, sadness	Induced	140 subjects 60 Hebrew speakers 1 Russian speakers	Hebrew Russian
SALAS database	Audio-visual	Wide range of emotions/emotion	Induced:	Pilot study of 20	English

Identifier	Modalities	Emotional content	Emotion elicitation methods	Size	Language
		related states but not very intense	subjects talk to artificial listener	subjects	
ORESTEIA database (McMahon et al. 2003)	Audio + physiological (some visual data)	Stress, irritation, shock	Induced	29 subjects, 90min sessions per subject	English
Belfast Boredom database (Cowie et al. 2003)	Audio-visual	Boredom	Induced	12 subjects: 30 minutes each	English
XM2VTSDB multi-modal face database	Audio-visual	None	n/a	295 subjects Video	English
ISLE project corpora	Audio-visual+ gesture	None	n/a		
Polzin (Polzin, 2000)	Audio- visual	Anger, sadness, neutrality	Acted	Segment numbers 1586 angry, 1076 sad, 2991 neutral	English
Banse and Scherer (Banse and Scherer 1996)	Audio- visual	Anger (hot), anger (cold), anxiety, boredom, contempt, disgust, elation, fear (panic), happiness, interest, pride, sadness, shame	Acted	12 (6 male, 6 female)	German
TALKAPILLAR (Beller, 2005)	Speech	neutral, happiness, question, positive and negative surprised, angry, fear, disgust, indignation, sad, bore	Contextualised acting	1 actor reading 26 semantically neutral sentences for each emotion (each repeated 3 times in different activation level: low,middle,high)	French
Reading-Leeds database (Greasley et al. 1995; Roach et al. 1998, Stibbard 2001)	Speech	Range of full blown emotions	Natural	Around 4 ½ hours material	English

Identifier	Modalities	Emotional content	Emotion elicitation methods	Size	Language
France et al. (France et al. 2000)	Speech	Depression, suicidal state, neutrality	Natural	115 subjects: 48 females 67 males.	English
Campbell CREST database, ongoing (Campbell 2002; see also Douglas-Cowie et al. 2003)	Speech	Wide range of emotional states and emotion-related attitudes	Natural	Target - 1000 hrs over 5 years	English Japanese Chinese
Capital Bank Service and Stock Exchange Customer Service (as used by Devillers & Vasilescu 2004)	Speech	Mainly negative - fear, anger, stress	Natural: call center human-human interactions		English
SYMPAFLY (as used by Batliner et al. 2004b)	Speech	Joyful, neutral, emphatic, surprised, ironic, helpless, touchy, angry, panic	Human machine dialogue system	110 dialogues, 29.200	German
DARPA Communicator corpus (as used by Ang et al. 2002) See Walker et al. 2001	Speech	Frustration, annoyance	Human machine dialogue system	average length about 2.75 words 13187	English
AIBO (Erlangen database) (Batliner et al. 2004a)	Speech	Joyful, surprised, emphatic, helpless, touchy (irritated), angry, motherese, bored, reprimanding, neutral	Human machine: interaction with robot	1 german children, 51.393 words English (Birmingham): 30 children, 5.822 words	German
Fernandez et al. (Fernandez et al. 2000, 2003)	Speech	Stress	Induced	4 subjects	English
Tolkmitt and Scherer (Tolkmitt and Scherer, 1986)	Speech	Stress (both cognitive & emotional)	Induced	60 (33 male, 27 female)	German
Iriondo et al. (Iriondo et al. 2000)	Speech	Desire, disgust, fury, fear, joy, surprise, sadness	Contextualised acting	8 subjects reading paragraph length	Spanish

Identifier	Modalities	Emotional content	Emotion elicitation methods	Size	Language
				passages (20-40mmsec each)	
Mozziconacci (Mozziconacci, 1998)	Speech	Anger, boredom, fear, disgust, guilt, happiness, haughtiness, indignation, joy, rage, sadness, worry, neutrality	Contextualised acting	3 subjects reading 8 semantically neutral sentences (each repeated 3 times)	Dutch
McGilloway (McGilloway, 1997; Cowie and DouglasCowie, 1996)	Speech	Anger, fear, happiness, sadness, neutrality	Contextualised acting	40 subjects reading 5 passages each	English
Belfast structured Database An extension of McGilloway database above (Douglas-Cowie et al. 2000)	Speech	Anger, fear, happiness, sadness, neutrality	Contextualised acting :	50 subjects reading 20 passages	English
Danish Emotional Speech Database (Engberg et al. 1997)	Speech	Anger, happiness sadness, surprise neutrality	Acted	4 subjects read 2 words, 9 sentences & 2 passages	Danish
Groningen ELRA corpus number S0020	Speech	Database only partially oriented to emotion	Acted	238 subjects reading 2 short texts	Dutch
Berlin database (Kienast & Sendlmeier 2000; Paeschke & Sendlmeier 2000)	Speech	Anger- hot, boredom, disgust, fearpanic, happiness, sadness-sorrow, neutrality	Acted	10 subjects (5 males, 5 female) reading 10 sentences each	German
Pereira (Pereira, 2000)	Speech	Anger (hot), anger (cold), happiness, sadness, neutrality	Acted	2 subjects reading 2 utterances each	English
van Bezooijen (van Bezooijen, 1984)	Speech	Anger, contempt disgust, fear, interest joy, sadness shame, surprise, neutrality	Acted	8 (4 male, 4 female) reading 4 phrases	Dutch
Abelin (Abelin 2000)	Speech	Anger, disgust, dominance, fear, joy, sadness, shyness, surprise	Acted	1 subject	Swedish

Identifier	Modalities	Emotional content	Emotion elicitation methods	Size	Language
Yacoub et al (2003) (data from LDC)	Speech	15 emotions: Neutral, hot anger, cold anger, happy, sadness, disgust, panic, anxiety, despair, elation, interest, shame, boredom, pride, contempt	Acted	2433 utterances from 8 actors	English

2.5.3 Unités de base et descripteurs pour la reconnaissance des émotions

Dans les différents travaux sur la reconnaissance automatique des émotions à partir du signal de la parole, plusieurs types de descripteurs acoustiques, de méthodes de classification, de corpus et d'unités de segmentation ont été expérimentés afin d'avoir une meilleure présentation et analyse de la problématique.

Travaux relatifs à l'unité de base de segmentation : l'unité de base d'analyse représente le segment de données extrait à partir du signal de la parole permettant de déterminer les descripteurs comme dans le cas des paramètres acoustiques par exemple. Ils sont utilisés par le classificateur pour l'identification de l'émotion. Il existe plusieurs types d'unités d'analyse répertoriés dans la littérature parmi lesquels nous pouvons citer l'énoncé, le phonème, le mot, le fragment, la syllabe, la pseudo-syllabe et les régions de voisement.

La majorité des travaux utilisent *l'énoncé* comme unité d'analyse. Les vecteurs de paramètres acoustiques calculés sur cette unité représentent la totalité de la phrase. Parmi les travaux qui se sont basés sur cette unité, citons [58]-[73]

L'unité de base est le *mot*. L'expérimentation de cette unité d'analyse dans les systèmes de RAÉ a produit des améliorations dans ce domaine en comparaison avec l'unité *énoncé*. Également, la position des mots dans la phrase a été testée. En effet, Rao et Koolagudi ont démontré que les mots qui se positionnaient à la fin de la phrase étaient plus significatifs pour la détection des émotions [59].

Des systèmes de RAÉ ont divisé les mots en *syllabes*. L'expérimentation de cette unité a été motivée selon l'hypothèse que l'état émotionnel d'un locuteur affecte la prononciation des syllabes d'un même mot avec différentes intensités. Les systèmes basés sur l'unité syllabe ont

été évalués en comparaison avec le système à base de l'unité énoncé, et aussi par rapport à la position de la syllabe dans le mot. L'étude de la position de la syllabe dans le mot a conduit à l'apparition de trois sous-systèmes, celui de la syllabe en début, au milieu et à la fin du mot. Ceux-ci ont été comparés et les résultats ont montré que les syllabes finales étaient plus déterminantes que les autres concernant la détection des émotions.

L'expérimentation de l'unité *phonème* dans ce domaine a été justifiée par l'hypothèse semblable à celle de l'utilisation de l'unité syllabe. Les systèmes à base de phonèmes sont classés en système à base de voyelles, de semi-voyelles, de nasales, de consonnes occlusives et de fricatives. Ces cinq systèmes ont été expérimentés et la comparaison de leurs résultats a démontré que celui à base de voyelles était le plus déterminant pour les catégories des émotions. La combinaison de ces cinq systèmes a été testée et a indiqué un progrès au niveau de la détection des émotions.

D'autres unités d'analyse ont été utilisées telles que le *pseudo syllabe* dans les recherches de Attabi [60], ou *les fragments* dans les travaux de Schuller et son équipe[61]. Ces deux unités ont obtenu des résultats meilleurs que ceux du système basé sur l'unité syllabe. Notons aussi que certains travaux ont utilisé les régions voisées et/ou non voisées en se basant sur la division d'un énoncé en des segments voisés et d'autres non voisés selon la fréquence fondamentale comme dans les recherche de Shami [62] et [63]. La combinaison de paramètres extraits de l'unité voisée et non voisée avec celle de l'énoncé a révélé un meilleur taux de classification des émotions en comparaison avec les résultats de systèmes à base de l'énoncé. Le résultat de cette comparaison selon l'unité de segmentation est présenté dans le Tableau 2-3.

Tableau 2-3 : Les systèmes de RAÉ selon l'unité d'analyse

Unité d'analyse	Les Travaux de RAÉ
Énoncée	Beritelli [64], El Ayadi [65], Grimm et Kroschel [66], Inanoglu et Caneel [67], Li [68] , Lin [69], Petrushin [70], Seppänen[71] , Sethu [72] et Vlasenko [73]
Mot	Rotaru et Litman[74], Schuller [61] et Rao et Koolagudi [75]
Phonème	Lee [5], Bitouk[76], Koolagudi et Krothapalli [77]
Syllabe	Schuller[78] et Koolagudi [75]
Pseudo-syllabe	Dumouchel [79]
Fragment	Schuller et al [78]
Voisé/ non voisé	Shami et Kamel [62] et [63]

Descripteurs en reconnaissance des émotions : Les descripteurs utilisés et listés dans la littérature sont de deux types : acoustique et linguistique. L'extraction de ces descripteurs est l'étape la plus cruciale dans le processus de détection des émotions de la parole. Ces deux descripteurs doivent être informatifs et significatifs. Fréquemment, les méthodes d'extraction de ceux-ci sont différentes. Les descripteurs acoustiques sont les plus utilisés vu que leurs extractions sont plus simples que celles des descripteurs linguistiques. Les descripteurs acoustiques peuvent être catégorisés en trois groupes : prosodique comprenant la durée, l'intensité et le pitch, les descripteurs spectraux comme les MFCCs et enfin les descripteurs basés sur la qualité de la voix comme le Jitter, le Shimmer et le HNR.

Les descripteurs prosodiques **:** La prosodie est l'étude de traits phonétiques suprasegmentaux. Elle est aussi liée à l'impression musicale que fournit un locuteur lorsqu'il parle. La prosodie s'intéresse à la relation entre la durée, l'amplitude et le pitch du

son. Parmi les caractéristiques prosodiques de la parole on trouve le pitch qui est estimé par la fréquence fondamentale (F0). Celle-ci est variable. Elle est estimée à 300 Hz chez les jeunes femmes et les enfants et peut baisser à 60 Hz chez les hommes. Cette fréquence correspond à la vibration des cordes vocales lors de la prononciation d'un son voisé. Deux autres paramètres prosodiques (intensité et durée) ont été utilisés dans la détection des émotions tel que mentionné dans les travaux de Johnstone [80]. Les paramètres prosodiques ont été les premiers utilisés dans ce domaine par McGilloway dans [81]. Une multitude d'études ont démontré que les paramètres prosodiques étaient pertinents dans le cas de deux d'émotions ayant différents niveaux d'excitation. Toutefois, ces paramètres n'arrivent pas à bien distinguer les catégories d'émotions ayant un même niveau d'excitation [16]. Parmi les recherches qui ont utilisé les paramètres prosodiques, nous citons [82], [83],[60].

Les descripteurs spectraux : selon certaines études, les descripteurs prosodiques corrèlent avec l'axe activation (dimension de l'émotion), cependant ils ne permettent pas une bonne modélisation de la dimension représentée par l'axe valence[84]. Cette lacune a motivé l'intérêt pour d'autres types de descripteurs afin de régler ce problème, notamment les descripteurs spectraux qui ont démontré un plus grand pouvoir discriminatif entre des catégories d'émotions. On retrouve dans ce groupe de descripteurs acoustiques spectraux : MFCC, PLP, LPC, LPCC, LFPC et les formants[84] [80].

Les descripteurs de la qualité de la voix : Les propriétés de simulation de la pulsion glottale, tels que l'articulation et la prosodie, sont représentées par les descripteurs de qualité de la voix, qui rend leur utilisation importante dans le cas d'émotions qui se distinguent par le type de phonation. Ces descripteurs sont le Jitter, le Shimmer et le HNR. Parmi les travaux qui ont testé la pertinence de ces descripteurs nous citons [83] [85] [86] [87]. Plus récemment, un autre descripteur suscite de l'intérêt c'est celui relatif au **rythme** de la parole. Le rythme est

défini comme un effet impliquant la récurrence isochrone, c'est-à-dire qu'un certain type d'unité de discours répété à des intervalles réguliers. Généralement le rythme des langages est associé au battement des phonèmes.

2.6 Variété native et non native d'une langue

La linguistique et la paralinguistique font partie de la communication vocale, qui est une activité sociale. Selon le groupe social, la langue et le dialecte, le paralangage va différer. Le paralangage, qui est un phénomène non linguistique, avec la langue forme une communication vocale humaine complète.

L'introduction et le développement de la partie paralangage de la parole a encouragé les chercheurs à utiliser cet aspect de la parole comme point de départ pour l'étude de plusieurs questions concernant la langue employée pendant une communication orale. Les questions d'intérêt sont l'identification et l'identification de la langue, la différence entre la prononciation interlangue et intralangue, le classement de la langue, la différence entre la prononciation native et non-native, la différence entre l'homme et la femme concernant la prononciation du même dialecte et la différence entre les âges concernant la prononciation du même dialecte.

L'identification de l'accent natif et non natif appartient à la catégorie de la parole déviante dans le domaine de la paralinguistique. L'accent non natif est considéré comme un problème de prononciation de la parole tout comme la pathologie qui peut être temporairement déviante. Il nécessite de l'amélioration parce que la prononciation de la langue native par une personne non native semble différente. La volonté de surmonter cette déviance marque la différence entre l'accent non natif et les autres traits de cette catégorie.

La parole non native est caractérisée par trois propriétés principales : accents multiples, les différences non-phonémiques et la confusion phonémique. En effet, pour la même langue

parlée, on peut trouver plusieurs accents non natifs, en raison de la langue maternelle du locuteur. À titre d'exemple, pour la langue arabe, la personne anglaise forme un accent non natif et la personne française forme un autre accent non natif.

Les locuteurs et locutrices ont un vocabulaire et un savoir des structures grammaticales limités de la langue seconde parlée. Cette limite oblige les personnes non natives à s'exprimer avec des mots de base, rendant leurs discours quotidiens très typé pour les personnes natives. Par contre, cette limitation nécessite un besoin de temps supplémentaire, d'hésitation, d'arrêt pour trouver le mot à utiliser afin d'être sûr de la prononciation d'un mot ou bien de penser à la suite de la phrase. En général, ce besoin conduit à un choix de mots non reconnus. Les hésitations sont autorisées par le modèle de langage. Par contre, une longue hésitation peut conduire ce dernier à déclarer la fin d'une phrase, ce qui résulte la plupart du temps à un discours démuni de sens. L'hésitation, l'arrêt et le besoin supplémentaire de temps font partie des problèmes non phonétiques qui caractérisent le discours non natif. Étant une propriété du discours non natif, la confusion phonétique est reconnue par la difficulté de distinction au niveau de la prononciation de quelques phonèmes de la langue native. La prononciation non native d'une langue présente deux niveaux de variation. Le premier est au niveau des phonèmes et le deuxième au niveau des mots.

Selon la littérature, le paralangage de la parole est toujours défini en référence à un ou plusieurs paramètres parmi les paramètres suivants : l'intensité, la durée, le rythme, la pause et l'articulation nasale. Ces paramètres sont classés comme des paramètres suprasegmentaux car on ne peut pas les décrire en faisant référence à un seul segment ou phonème du signal de la parole, puisqu'ils sont dépendants de plusieurs segments du signal.

Au niveau segmental, dans le domaine de l'identification des accents non natifs, une multitude de travaux se sont concentrés sur les segments dans le but de clarifier l'interaction

entre deux phonèmes de la langue maternelle (L1) et de la langue seconde (L2). L'influence de la langue maternelle dans la perception et la production de la langue seconde a été abordée dans un certain nombre d'études psycholinguistiques tels que dans [88] [89]. La comparaison de systèmes phonémiques de L1 démontre l'existence de phonèmes communs entre les deux langues, tandis que d'autres phonèmes peuvent être spécifiques pour chacune d'elles. Dans cette situation, le locuteur non natif peut remplacer des phonèmes de langue native par d'autres de sa langue maternelle.

Au niveau suprasegmental, de nombreuses études ont démontré que la prosodie de la langue maternelle persistait dans la production et la prononciation de la seconde langue. Cette affirmation se retrouve parmi les exemples suivants : la langue anglaise prononcée par les allemands, la langue allemande prononcée pas les anglais [90] ou lorsque la langue espagnole prononcée par les italiens et la langue italienne prononcée par les espagnols. D'après Ricard, cette influence diminue lorsque les locuteurs non natifs suivent une formation spécifique dans la prosodie de la langue seconde [91].

Le rythme est également utilisé pour faire la distinction entre les langues en les classant en trois catégories différentes : langue syllabique, langue accentuelle et la langue morique. Est-ce que l'association d'une catégorie des trois classes à une langue reste la même dans le contexte des accents non natifs de cette langue ? Parmi les chercheurs qui ont essayé de trouver une réponse à cette question citons [92], [93], [94] et [96]. Dans leurs recherches, les métriques de rythmes proposés par Ramus, Grabe et Low ont été expérimentées sur des corpus segmentés et étiquetés pour valider ou nier l'existence de classes rythmiques. L'expérimentation de ces métriques pour l'identification des accents natifs et non natifs a conduit à une distinction pertinente comme mentionné dans [93], [94], [95], [96], [97].

Récemment, l'accent non natif a été traité dans le domaine de la reconnaissance automatique de la parole dans le but de réduire l'impact de la prononciation non native sur le taux d'erreur [95], [96], [97], [98]. Le Tableau 2.5 présente quelques corpus de langues avec leurs variétés natives et non natives.

Tableau 2-4: Exemples de corpus pour les accents natifs et non natifs de la parole

Corpus	Native accent	Non native accent
MSA	Arabic	English
IBM-Fischer	English	Spanish, Frensh, German, italian
ISLE	English	German
MIST	English, frensh	German
NATO HIWIRE	English	Frensh, German, English, Spanish

2.7 Conclusion

Dans ce chapitre, nous avons effectué une revue de la littérature sur les traits paralinguistiques tout en se concentrant plus particulièrement sur deux d'entre eux à savoir l'émotion et l'accent natif et non natif d'une langue parlée.

Pour la reconnaissance automatique des émotions à partir de la parole, plusieurs définitions et théorèmes ont été présentés dans ce chapitre. Également, la stratégie pour la collecte de bases de paroles émotionnelles et les aspects reliés à la segmentation en unités de base. Ces unités permettront de circonscrire le calcul des paramètres acoustiques pour les utiliser comme entrées pour le système reconnaissance des traits paralinguistiques. Dans le prochain chapitre des détails sont fournis sur les approches utilisées dans l'analyse

acoustique des traits paralinguistiques. Nous présenterons également les méthodes de sélection et de classification automatique de traits paralinguistiques de la parole.

Chapitre 3 - Reconnaissance automatique de traits paralinguistiques de la parole

3.1. Introduction

L'objectif de l'analyse acoustique est d'extraire les informations caractéristiques du signal de la parole en éliminant au maximum les parties redondantes. Un tel système prend un signal d'entrée, et par la suite il émet un vecteur de paramètres. Ces derniers doivent être pertinents et discriminants, afin que le système de reconnaissance de traits paralinguistiques soit robuste et précis. C'est dans ce contexte que nous détaillerons le paramètre du rythme de la parole et proposerons une nouvelle métrique rythmique (L-OPVI) qui a l'avantage de généraliser un ensemble de métriques conventionnelles. Nous présentons également dans ce chapitre des méthodes d'optimisation de sélection des paramètres d'entrée ainsi que les classificateurs de base utilisés dans notre expérimentation.

3.2. Rappel sur les descripteurs prosodiques

Ayant un rôle important dans l'échange linguistique, les phénomènes prosodiques permettent à l'interlocuteur d'anticiper et de décoder le message du locuteur plus efficacement en effectuant un certain découpage syntaxique et sémantique qui facilitera la compréhension de l'énoncé pour le locuteur. En effet, les différentes informations, linguistiques ou paralinguistiques sont exprimées simultanément par la prosodie tels que :

a) Le sens véritable de certains mots ;

b) La nature de la phrase, si elle est interrogative ou affirmative ;

c) Les états émotionnels du locuteur ;

d) L'accent de la prononciation de la langue du locuteur.

Les informations paralinguistiques nécessaires au locuteur, afin de comprendre et de décoder le message du discours, sont transmises par les signaux prosodiques qui se manifestent par des variations : au niveau de la fréquence fondamentale du signal de la parole (pitch), de la hauteur, de l'intensité et/ou de la durée des sons. La variation de ces paramètres sont perçus par l'auditeur.

3.2.1 Le pitch

Dans le domaine acoustique, la fréquence fondamentale (F0) est définie par la fréquence des vibrations des cordes vocales, sous l'effet du passage de l'air à travers la glotte. Toutefois, dans le domaine perceptif, elle est représentée par la hauteur de la voix. La variation de la fréquence dépend de l'âge et du sexe du locuteur. L'extraction de la F0 utilise deux représentations qui sont : temporelle ou spectrale du signal. La première estimation, la temporelle, se fait directement du signal par l'utilisation de la similarité de celui-ci d'une période à l'autre, pour identifier la période fondamentale. Cependant la deuxième, la spectrale, présente la fréquence fondamentale (F0) comme l'inverse de la période fondamentale T0 :

$$F0 = \frac{1}{T0} \tag{3-1}$$

Les harmoniques de la fréquence fondamentale sont utilisées par les algorithmes, dans le domaine fréquentiel, dans le but de trouver la F0. Elles peuvent être visualisées sur un spectrogramme du signal.

Plusieurs méthodes de détection de la fréquence fondamentale existent dans la littérature. Cependant, l'algorithme basé sur la fonction d'autocorrélation du signal de la parole, dans le domaine temporel, demeure toujours le plus connu et robuste. Son fonctionnement, qui est simple, se base sur la recherche des maximums locaux de la fonction d'autocorrélation. Ceci se fait en trois phases successives :

a) Le prétraitement du signal de la parole ;

b) L'extraction de la fréquence fondamentale ;

c) Le post-traitement pour la correction des erreurs.

L'objectif de la première phase est l'optimisation des caractéristiques du signal afin d'entamer la phase d'extraction en utilisant, par exemple un filtrage passe-bas ou un filtrage non linéaire. La deuxième phase est l'extraction de la fréquence fondamentale qui dépend de l'algorithme utilisé. La troisième phase est le post-traitement suivi du pitch. Dans le cas d'une fausse estimation, cette phase s'occupe de la correction du contour du pitch. Cette dernière se fait par l'élimination des pics isolés et par la reconstruction des régions qui possèdent des périodes nulles. Cette méthode de correction touche aux trames passées, par rapport à la trame en cours d'analyse. D'après Rabiner dans [99], la fonction d'autocorrélation r_s du signal temporel *s(n)* est donnée par la relation:

$$r_s(m) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} s(n)s(n+m) \qquad (3\text{-}2)$$

Si le signale *s(n)* est périodique, sa fonction d'autocorrélation est aussi périodique. Sa période est égale à celle du signal *s(n)* :

$$\forall n \ : \ s(n) = s(n+p) \qquad \Rightarrow r_s(m) = r_s(m+p) \tag{3-3}$$

3.2.2 L'intensité

La variation de l'amplitude du signal de la parole, causée par une force plus ou moins forte dérivant du pharynx et provoquant une variation de la pression de l'air sous la glotte, résulte en énergie ou en intensité du signal qui donne une mesure de la force sonore de la voix. L'énergie à court terme d'un signal échantillonné sur une fenêtre de longueur T, $(s_t)_{t=1,T}$, est définie par :

$$E = \frac{1}{T}\sum_{t=1}^{T} s_t^2 \tag{3-4}$$

Pour respecter l'échelle perceptive, elle est généralement exprimée en décibel :

$$E_{db} = 10 \times \log_{10}(\frac{1}{T}\sum_{t=1}^{T} s_t^2) \tag{3-5}$$

3.2.3 La durée

L'aspect temporel du signal de la parole est représenté par la durée, autrement dit c'est l'intervalle de temps nécessaire pour émettre un signal. Elle comprend : le débit de la parole, les pauses et la durée des phonèmes qui forment le message. Généralement, la durée est en corrélation avec les informations linguistiques de la parole tels que : les phonèmes, les syllabes, les mots et les phrases.

3.3. Le rythme de la parole

La définition du rythme, comme caractéristique prosodique de la parole, a conduit à plusieurs interprétations. D'ailleurs l'étude du rythme, réalisée par Zellner dans le domaine linguistique, est en relation avec les langues parlées, et a donné un certain nombre d'interprétations que l'on pouvait associer à ce terme, comme par exemple : le rythme est une suite de stimulus, une structuration d'une suite de stimulus, une structuration d'une suite de stimulus accent exclu, une répartition d'accents, une répétition régulière d'accents, une répétition d'accents avec impression de retour régulière, une répétition d'accents et de pauses en fonction du temps, une alternance de syllabes accentuées vs les syllabes non accentuées, une synchronie de syllabes inaccentuées, une synchronie de groupes rythmiques, un nombre de syllabes par groupes rythmiques et les différents débits ou tempo [100].

3.3.1 L'hypothèse de Pike et Abercrombie relative au rythme

L'utilisation du rythme est très connue dans le domaine des traits paralinguistiques, particulièrement en classification des langues. Pour cette dernière, Pike a introduit les termes syllabiques et accentuels dans le but de classer les langues en deux groupes [35]. Cette classification a été faite en 1945, mais n'a pas été prouvée expérimentalement due à l'absence de mesures à ce moment-là. Pourtant, bon nombre de chercheurs du domaine, ont continué à utiliser ce type de classement dans leurs récents travaux.

En effet, Pike a affirmé, que pour les langues accentuelles, les syllabes pouvaient avoir des durées différentes, mais la durée entre deux syllabes était approximativement constante. Par contre, pour le groupe de langues syllabiques, les durées des syllabes étaient

approximativement constantes, alors que la durée des intervalles entre les syllabes dépendait du nombre de syllabes qui formaient les intervalles.

Avec la même vision que Pike, Abercrombie a continué l'étude de classement des langues en se basant sur les classes rythmiques de la parole tout en reformulant l'hypothèse de Pike. Cette nouvelle hypothèse a introduit le concept d'*isochronie,* et elle a insisté sur la classification des langues en deux classes rythmiques qui étaient : la langue accentuelle et syllabique [36] [101].

Selon la citation d'Abercrombie, la durée des syllabes pour les langues accentuelles était variée, alors qu'elle était constante ou égale pour les langues syllabiques. Le terme isochronie désigne la répétition régulière des unités phonétiques dans le même laps de temps. Cette unité phonétique peut être l'accent, et dans ce cas on parle de langues accentuelles, ou syllabes pour les langues syllabiques. Cette distinction, selon Abercrombie, se fait au niveau de la production de la parole. En effet, la première (l'accent) est liée à l'expiration de l'air et la seconde (la syllabe) au mouvement de la contraction et de la relaxation alternative des muscles de la respiration.

L'hypothèse de Pike et d'Abercrombie a été vérifiée par plusieurs auteurs pour des langues autres que l'anglais, afin d'en établir la classification. Les expérimentations de cette hypothèse ont toutes été effectuées avec la mesure de la durée des intervalles syllabiques et non syllabiques, comme par exemple dans les études d'Allen [102].

Dès les années 80, l'hypothèse d'Abercrombie a été rejetée. Les recherches remettaient en question la théorie de l'isochronie et ont conduit à l'apparition de nouvelles hypothèses pour le classement des langues, selon leur classe rythmique. Les métriques rythmiques sont

devenues le moyen qui a conduit à une nouvelle classification des langues et de leurs variétés.

3.3.2 Les métriques du rythme de la parole

Une métrique du rythme est une formule mathématique appliquée pour calculer et mesurer les durées des voyelles et des consonnes ainsi que leurs déviations d'une langue à une autre.

Cette approche trouve ces origines dans les recherches de Daouer et Bertinoti qui ont démontré l'existence d'une propriété spécifique dans la structure des langues [103] [104] . Cette propriété concernait les voyelles et les consonnes qui avaient des durées différentes. Cette approche a eu plusieurs évolutions qui ont commencé avec Ramus en 1999[105] et se base sur trois catégories de métriques. La première avec Ramus qui a développé les métriques Deltas, la deuxième avec Grabe pour les métriques *Pairwise Variability Index* (PVI) [106] et la troisième est la normalisation des deux premières. La normalisation est réalisée dans le but d'empêcher la dépendance entre la métrique et le débit de la parole.

Les Deltas : L'approche de Ramus marque un nouveau départ pour les études basées sur le rythme pour le classement des langues. En effet, Ramus a commencé avec l'idée que les langues étaient différenciées par leurs structures, et plus précisément par les propriétés phonétiques de la structure de la syllabe. Ce dernier a défini cette propriété pour les voyelles et les consonnes. En plus, il a appliqué une formule mathématique pour mesurer cette propriété, ce qui a donné naissance à trois mesures : Delta-V, Delta-C et %V. Le Delta représente l'écart-type des intervalles (C pour les consonnes et V pour les voyelles). Sa valeur permet d'attribuer une classe, parmi les trois classes, à la langue étudiée. À titre

d'exemple, si la mesure Delta-C est élevée, on est dans le contexte de syllabes complexes, alors la langue est classée syllabique. Par contre, si la valeur de Delta-V est élevée, alors la langue est accentuée. Le %V est la troisième mesure acoustique liée au rythme des langues. Cette mesure est en relation avec l'absence ou la présence de voyelles.

Les PVI : Ces mesures ont été introduites par Grabe et Low. Les PVI ressemblent beaucoup aux Deltas, mais, la différence réside dans la prise en considération de la variation des intervalles des différentes paires de phonèmes tels que les voyelles ou les consonnes, ainsi que la succession de voyelles et de consonnes [106].

Le premier PVI introduit est le rPVI est sa formule mathématique est la suivante :

$$rPVI = \frac{1}{N-1} \times \sum_{K=1}^{N-1} \left| d_K - d_{K+1} \right| \tag{3-6}$$

Cette formule met l'accent sur la différence de durée, notée d_k, entre toutes les paires de consonnes ou de voyelles successives (au nombre de N) et calcule la moyenne de cette différence.

La normalisation de ce paramètre a mené à la naissance d'une nouvelle métrique acoustique nommée nPVI qui est représentée par la formule suivante :

$$nPVI = \frac{100}{N-1} \times \sum_{K=1}^{N-1} \frac{\left| d_K - d_{K+1} \right|}{\frac{d_K + d_{K+1}}{2}} \tag{3-7}$$

Dans les deux équations d_k et d_{k+1} représentent la durée d'un intervalle de la consonne ou de la voyelle à la position k *et* $k+1$ successivement ; N est le nombre de segments (phonèmes) dans chaque signal de la parole. Par la suite, ces deux paramètres de la famille PVI ont été pris en considération pour les consonnes et les voyelles.

Les rythme et le débit : Ramus a proposé d'augmenter le nombre de locuteurs comme solution pour réduire la sensibilité des Deltas au débit de la parole. Par contre, Dellow [94] a proposé la division des Deltas (ΔC ou ΔV) par la durée moyenne des consonnes ou des voyelles, ce qui a donné naissance à des Varcos (Varco-V, Varco-C et Varco-VC). VC dénote une syllabe composée d'une voyelle est d'une consonne. Le calcul s'effectue comme suit :

$$VarcoC=100\times\frac{\Delta C}{meanC} \tag{3-8}$$

$$VarcoV = 100 \times {}^{\Delta V}\!/_{meanV} \tag{3-9}$$

3.4. Qualité de la voix

D'après Abercrombie, la qualité de la voix est le timbre ou la coloration qui caractérise l'ensemble de sons produits lorsqu'un locuteur s'exprime. Cette coloration est le résultat de l'ajustement des organes qui participent à la production du son [101]. Par ailleurs, selon Mackenzie, la qualité de la voix résulte de l'ajustement des organes qui contribuent à la production de la parole, ainsi que ceux qui contribuent à la phonation [107]. La qualité de la voix a des fonctions linguistiques et paralinguistiques dans la communication vocale.

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. Elle est très utile pour les émotions où on y retrouve une vraisemblance, comme celle entre la colère et l'impatience. Également, elle sert à identifier les accents de différentes langues. En effet, selon Honikman, toutes les langues du monde sont caractérisées par un positionnement

particulier des organes articulateurs, ce qui explique que chaque langue est caractérisée par une qualité de voix spécifique lors de sa prononciation. Aussi, chaque dialecte de même langue possède une qualité de voix spécifique, par rapport aux autres. Ce paramètre acoustique de la voix est pertinent et significatif lors de l'identification des accents non natifs [108]. Parmi les paramètres de la qualité de la voix on trouve : le Jitter, le Shimmer et le Rapport Signal Harmonique sur Bruit.

Le Jitter représente la variation de la fréquence fondamentale dans l'évolution temporelle de l'énoncé. Il indique la variabilité ou la perturbation du laps de temps (T0) à travers plusieurs cycles d'oscillations. Il est principalement affecté par une insuffisance de contrôle de la vibration du champ vocal. Cette information est particulièrement nécessaire, pour identifier l'âge des locuteurs à partir de la voix, et pour la détermination du degré de pathologie. Les Jitters brutes et normalisés sont définies respectivement comme :

$$jitter = \sum_{i=1}^{N-1} |T_i - T_{i+1}| \Big/ N - 1 \qquad (3\text{-}10)$$

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

De façon similaire, **le Shimmer** 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. Le Shimmer est estimé de façon similaire à celle du Jitter, sauf qu'il utilise l'amplitude comme paramètre. Les deux paramètres : le pitch et l'intensité de la parole deviennent difficiles à contrôler lors de l'arrivée d'air aux cordes vocales. Le Shimmer est généralement mesuré sur une échelle logarithmique en décibel. Selon Haji et al [109], une personne saine possède un Shimmer situé entre 0,05 et 0,22 dB et peut être calculé par la formule suivante :

$$shimmer(\mathrm{dB}) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\log(\mathrm{A}_{i+1}-\mathrm{A}_i)\right| \tag{3-11}$$

Rapport signal harmonique sur bruit (HNR) représente le degré de périodicité acoustique. L'harmonicité est mesurée en dB, et calculée comme le ratio de l'énergie de la partie périodique (E_p) avec l'énergie du bruit (E_n) :

$$HNR = 10\log(\frac{E_p}{E_n}) \tag{3-12}$$

Durée vocalique : décrit la séparation entre la libération et la constriction dans un contexte vocalique. La proportion de la durée vocalique est la fraction de la durée de l'énonciation composée d'intervalles vocaliques. Les difficultés à maintenir la voix au-dessus d'une voyelle soutenue peuvent être considérées comme un signe de pathologie.

Degré des pauses de la voix est la durée totale des pauses sur le signal divisé par la durée totale, en excluant le silence du début et de la fin de la phrase. Une pause de la voix peut se produire par l'arrêt soudain du flux d'air, en raison d'une insuffisance transitoire dans le contrôle du mécanisme de la phonation.

3.5. Les paramètres acoustiques basés sur une analyse dans le domaine spectral

L'utilisation de filtres dans la représentation spectrale permet une séparation de l'excitation glottique et des résonances du conduit vocal. Le passage du domaine temporel au domaine spectral mène à cette division. Ce passage permet seulement la conservation de paramètres pertinents.

La représentation temporelle du signal S_n selon le principe d'excitation $\{g_n\}$ et de la réponse impulsionnelle du conduit vocal $\{h_n\}$ est donnée par la formule suivante :

$$S_n = g_n \times h_n \qquad 3\text{-}13)$$

Dans la représentation spectrale, la convolution du signal est l'addition de deux logarithmes, l'un pour l'excitation et l'autre pour la réponse vocale :

$$\log|S_k| = \log|G_k| + \log|H_k| \qquad (3\text{-}14)$$

Où $\{s_n\}$, $\{g_n\}$ et $\{h_n\}$ possèdent respectivement les spectres $\{S_k\}, \{G_k\}$ et $\{H_k\}$.

On fait l'interprétation de s_n par l'utilisation d'une échelle de fréquence nommée Mel et présentée par :

$$M = \frac{1000}{\log_2} * \log(1 + \frac{f}{1000}) \qquad (3\text{-}15)$$

Où f est la fréquence en Hz.

Cette échelle est linéaire dans le cas de basses fréquences, et logarithmiques pour les hautes fréquences, afin de prendre en considération les propriétés de la perception humaine du son. Son rôle est de définir le banc de filtres triangulaires.

Les coefficients cepstraux sont obtenus par plusieurs méthodes, telles que la récursivité faite à partir des coefficients LPC et qui donne les coefficients LPCC, la transformée de Fourier ou de la transformée de Fourier inverse permet de calculer les coefficients MFCC, LFCC et PLP.

Les paramètres MFCC (Mel Frequency Cepstral Coefficients) sont des coefficients cepstraux calculés par la méthode de transformée Fourrier, par le filtrage des énergies et par un filtre banc en échelle de fréquence Mel.

Sur l'échelle de Mel, on fait la partition de l'échelle en fréquence entre 0 et $\frac{f_e}{2}$ en N bandes. Si f_e est la fréquence d'échantillonnage du signal, alors le filtrage se fait par la multiplication du spectre du signal S_n par le gabarit des filtres. La représentation logarithmique des énergies du signal en sortie de ces filtres triangulaires est X_k pour $k = 1,2,\ldots,N$. La transformée en cosinus discret de ces logarithmes donne la définition des coefficients spectraux Mel (MFCC) :

$$MFCC_i = \sum_{k=1}^{N} X_k \cos\left(i(k-1)\frac{\pi}{N} \right) \tag{3-16}$$

Une déconvolution entre la source de sons et le conduit vocal est fournie par les MFCC, ces coefficients sont fortement décorrélés. L'énergie de contribution du conduit vocal est contenue dans les premiers coefficients MFCC. Ces derniers sont indépendants de l'excitation, alors ça résulte en une indépendance de la prosodie du locuteur. Cette propriété conduit à l'utilisation massive de ces paramètres dans le domaine des traits paralinguistiques.

3.6. Classification des traits paralinguistiques

Le processus d'automatisation de la classification des éléments paralinguistiques se base sur des étapes principales, ainsi que sur d'autres étapes complémentaires à celles-ci. Elles sont présentées dans Figure3–1.

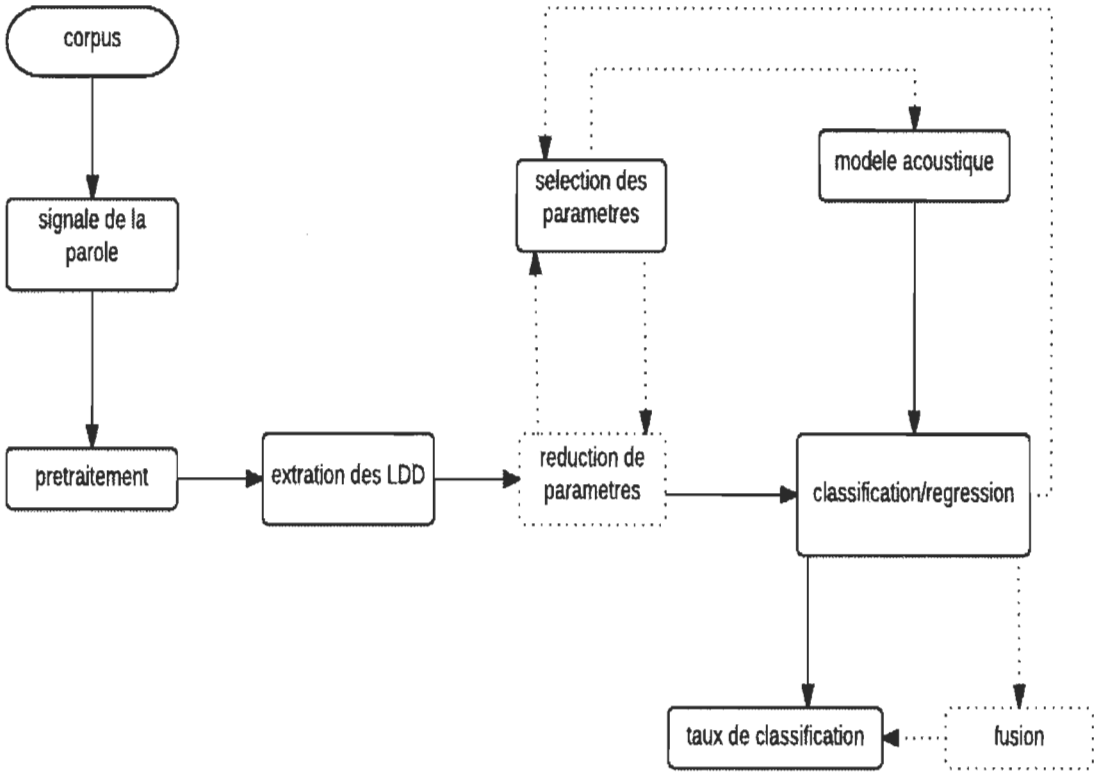

Figure3–1: Système d'analyse/classification automatique des traits paralinguistiques

3.6.1 Méthodes de sélection des paramètres acoustiques

La phase de sélection est une étape fondamentale lors du processus de traitement des traits paralinguistiques. En effet, elle a un rôle de premier plan dans la prévision de l'efficacité globale du système de classification. Les classificateurs sont sensibles à la taille de données d'entrées, spécialement lorsqu'elles sont corrélées. En effet, si les paramètres sont fortement corrélés, il y a de forte chance d'avoir une inefficacité du classificateur. Pour remédier à cette limitation, nous avons utilisé différentes méthodes de sélection telles que, l'analyse en composantes principales (ACP), l'analyse discriminante linéaire (LDA) et l'analyse de variance (ANOVA).

L'analyse en composantes principales (ACP) : Est une méthode descriptive multidimensionnelle. Ses fonctions sont similaires aux autres méthodes d'analyse de données, car elle essaye d'établir des relations entre les observations, entre les variables, et entre les observations et les variables. Durant son fonctionnement elle ne se base que sur un modèle géométrique. La nouvelle présentation de données servira à faire la distinction entre deux groupes de variables : les variables très corrélées et celles non corrélées, des unités ayant des ressemblances et d'autres n'en ayant aucune. L'ACP est un outil permettant une meilleure visualisation de nos données. En effet, il suffit de faire la projection sur deux axes orthogonaux des valeurs des variables de nos observations, pour voir le lien qui relie les deux variables.

L'ACP se base sur des vecteurs et des valeurs propres dans son fonctionnement mathématique. En effet, les vecteurs propres sont des propriétés qui marquent la grande efficacité de l'algorithme ACP. Cette efficacité est le résultat de la sélection des meilleurs vecteurs propres. Ces derniers vont définir le nouvel espace vectoriel de la représentation de données. De plus, cette méthode ne nécessite aucune connaissance *a priori* sur les classes de celles-ci.

L'analyse des variances (ANOVA) : Est utilisée pour déterminer l'absence ou la présence de différence significative entre les moyennes de deux ou de plusieurs groupes indépendants (non liées). Le résultat d'un test d'ANOVA est une valeur significative « *valeur-p ou p-value.* » Celle-ci indique la probabilité d'obtenir une différence moyenne entre les groupes plus élevés de tous ceux observés au hasard. Plus la valeur *p-value* est petite plus la différentiation est significative entre les groupes. Lors de l'analyse des

résultats de l'ANOVA, trois paramètres sont caractéristiques : le degré de liberté (df); la statistique; la valeur p (*p-value*).

L'ANOVA formule les deux hypothèses : H0 et H1 pour le problème à traiter. L'hypothèse nulle H0 ne suppose aucune différence entre les moyennes des groupes de variables, alors que l'hypothèse H1 est à l'opposée, car elle propose qu'il existe une différence entre les moyennes des groupes. Les sommes des carrés suivantes sont ensuite calculées :

$$\underbrace{\sum_{i=1}^{I}\sum_{n=1}^{N_i}\left(X_{jin}-\overline{X_j}\right)}_{Total}=\underbrace{\sum_{i=1}^{k}N_i\left(\overline{X_{ji}}-\overline{X_j}\right)^2}_{entre-groupe(SSR)}+\underbrace{\sum_{i=1}^{I}\sum_{n=1}^{N_i}\left(X_{jin}-\overline{X_{ji}}\right)^2}_{itra-groupe(SSE)} \tag{3-17}$$

Les moyennes des carrés sont ensuite calculées et la moyenne quadratique pour les traitements est définie par :

$$\text{MSE} = \frac{SSE}{n - k} \tag{3-18}$$

$$\text{MSE} = \frac{SSE}{n - k} \tag{3-19}$$

Par la suite la moyenne quadratique pour l'erreur est définie par :

$$\text{MSR} = \frac{SSR}{k - 1} \tag{3-20}$$

La statistique (rapport F) utilisée pour tester l'hypothèse nulle est définie par :

$$\text{F} = \frac{MSR}{MSE} \tag{3-21}$$

L'analyse discriminante linéaire (LDA) : Est une méthode de sélection qui ne modifie pas l'emplacement ou la structure des caractéristiques originales, contrairement à celle de l'ACP. Elle utilise les vecteurs propres de la matrice de dispersion afin de créer des régions de séparation entre les différentes classes en maximisant le rapport entre celles-ci dans un ensemble particulier de données. Ceci nécessite l'organisation préalable des données en classes pour la phase d'apprentissage. Elle garantit ainsi la séparation maximale des classes.

Le principe de la LDA consiste à cartographier les caractéristiques importantes sur une combinaison linéaire en associant les coefficients optimaux à chaque fonctionnalité, afin d'obtenir les différentes fonctions linéaires. Ces fonctions sont discriminatoires entre les classes. Dans ce travail de recherche, l'analyse discriminante a été utilisée avec l'approche Lambda de Wilks dans le but de sélectionner les caractéristiques les plus importantes. La sélection de la fonction commence par le choix des caractéristiques les plus importantes. Ensuite, elle les relie avec la première fonction, ce qui assure la séparation des classes de données en groupes grâce une équation de prédiction discriminante. Ainsi la LDA permet de déterminer le pourcentage de variance de la variable dépendante expliquée par les variables indépendantes ; d'évaluer l'importance relative des variables indépendantes pour classer la variable dépendante et enfin de de supprimer les variables qui sont peu liées à des distinctions de groupe. Le fonctionnement mathématique de LDA dépend des variables discriminantes (variables indépendantes, également appelés prédicteurs) ; de la variable de catégories (variable dépendante, également appelée la variable de regroupement) et de la fonction discriminante. Cette dernière est aussi appelée fonction canonique et est créée comme une combinaison linéaire des variables discriminantes. Elle

$$L = b_1 X_1 + b_2 X_2 \cdots b_n X_n + c \tag{3-22}$$

b_n sont les coefficients de discrimination, **X** est le vecteur de variables et c est une constante.

Généralement, le choix de la méthode de sélection de données repose sur les critères suivants :

- Le nombre de variables : binaires ou multivariables.
- Le type de tests : comparaison de moyennes, de variances, de corrélations.
- Le type d'échantillons : appariés ou indépendants.
- Le type de données : qualitatives, quantitatives, continues.
- La nature de loi de distributions : normales ou non. Par exemple, l'ANOVA n'est applicable que pour les distributions normales.
- L'hypothèse à vérifier.
- La décision à obtenir à la fin du processus d'une méthode.

La LDA et l'ACP diffèrent pour certains points et se ressemblent pour d'autres. En effet, la LDA possède des variables X et Y ainsi que des groupes prédéterminés, alors que l'ACP n'a qu'un seul ensemble de variables. Par contre, les deux résultent en une nouvelle représentation de variables. Cette dernière est une combinaison linéaire des variables d'origine.

La LDA et l'ANOVA sont très similaires et reposent sur plusieurs aspects théoriques communs. En effet, le fonctionnement de LDA repose sur le principe de l'ANOVA. Elle utilise cette dernière pour prendre sa décision concernant l'ajustement du modèle final des variables par la séparation à deux niveaux : intergroupe et intragroupe. Toutes les deux ont

des catégories X et des variables continues Y. Les calculs statistiques sont les mêmes pour les deux, même si leurs buts sont différents. L'ANOVA vise la vérification de l'existence d'une différence importante entre les groupes, alors que LDA vise principalement le développement des fonctions discriminantes, afin de mieux classer les objets en groupes.

3.6.2 Les approches statiques de la classification

La classification statique désigne le processus d'attribution d'une étiquette de classe discrète à un vecteur caractéristique inconnu de la dimension fixe. Un bon système de classification doit avoir les caractéristiques suivantes :

- L'utilisation de toutes les informations disponibles ;
- Un taux d'erreurs de classification faible ;
- Une minimisation des effets négatifs des erreurs de classification.

On utilise fréquemment trois classificateurs dans la reconnaissance des traits paralinguistiques qui sont : les machines à vecteurs de support (SVM), les modèles de mélange gaussien (GMM) et la régression logistique linéaire (RL)

La machine à vecteurs de support ou le séparateur à vaste marge (SVM) : Est un algorithme qui a été développé pour la classification, mais a récemment été adapté à d'autres usages, comme la recherche, l'estimation de la régression et la distribution. Il a été utilisé dans de nombreux domaines tels que la détection des traits paralinguistiques. Cet algorithme a été inventé par Vapnik en 1970. La SVM est probablement considérée comme le classificateur le plus fréquemment utilisé dans le domaine des traits paralinguistique.

L'idée principale des SVMs est construite autour des classificateurs linéaires binaires optimisés pour fournir la meilleure séparation possible entre les classes, dans l'espace des

variables. Dans cette thèse, on expérimente l'espace des paramètres acoustiques. Les SVMs multiclasses peuvent être conçus à partir d'une combinaison des SVMs binaires par le biais de diverses stratégies. On peut former des SVMs pour chaque paire de classes en additionnant les "votes" de chaque classe lors de la reconnaissance, ou en formant un arbre de décision binaire avec le seuil de décision remplacé par la classification de la SVM.

La SVM est bien adaptée pour la discrimination à deux classes (L1/L2). Pour le classement de base de la SVM, nous avons formé un modèle de cible pour l'accent natif versus la classe de l'accent non natif. L'expression conventionnelle d'une SVM est donnée par f (x) [110]:

$$f(\mathrm{x}) = \sum_{n=1}^{N} \alpha_n K\left(x, x_i\right) + d \tag{3-23}$$

$$\text{Avec} \quad \sum_{n=1}^{N} \alpha_n = 0 \tag{3-24}$$

Où le K (.,.) est une fonction du noyau, le $\boldsymbol{x}_i$ est le vecteur de support, et le d obtenue par l'optimisation. La méthode directe de classification des traits paralinguistiques par les SVMs consiste à utiliser des noyaux non linéaires permettant de comparer les séquences de vecteurs de la fonctionnalité. Les fonctions du noyau non linéaire visent la meilleure performance de la projection de données d'entrées dans un espace de dimension supérieure (cartographie) dans laquelle la classification linéaire peut être effectuée. Plusieurs fonctions du noyau ont été développées afin de s'en servir principalement avec la SVM appliquée à la vérification du locuteur. La fonction radiale de base du noyau (RBF) a été utilisée avec succès. Son équation est :

$$K\left(x_i, x_j\right) = e^{\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}} \qquad (3\text{-}25)$$

Le paramètre σ définit la largeur du RBF.

Un modèle de mélange gaussien (GMM) : Est de type de classificateur probabiliste. Ce dernier suppose que tous les points de données sont générés à partir d'un mélange d'un nombre fini de distributions gaussiennes avec des paramètres inconnus. Il incorpore des informations sur la structure de la covariance des données, ainsi que sur les moyennes des gaussiennes.

Pour une dimension d la distribution gaussienne du vecteur $x = \left(x^1, x^2, x^3, \ldots x^d\right)^T$ est définie par :

$$N(\mathrm{x}; \mu_{mk}, \Sigma_{mc}) = (2\pi)^{-d/2} \left|\Sigma_{mc}\right|^{-1/2} e^{\frac{1}{2}(x-\mu_{mc})^T \sum_{mc}^{-1} x - \mu_{mc}} \qquad (3\text{-}26)$$

Où $N(\mathrm{x}; \mu_{mk}, \Sigma_{mc})$ est le mélange de distributions gaussiennes de dimension d -M pour la classe c ; $\mu_{mk} \in \Re^d$ et $\Re^{d \times d}$ représentent la matrice de moyenne et le vecteur et covariance, respectivement, pour la mth distribution gaussienne, de la *classe* c^{th}.

La distribution P (x | c) peut être écrite comme : La probabilité donnée dans une distribution gaussienne est :

$$\mathrm{P(x/\,c)} = P(\mathrm{c}) \sum_{m=1}^{M} \lambda_{mc} N\left(\mathrm{X}; \mu_{mc}, \Sigma_{mc}\right) \qquad (3\text{-}27)$$

Avec λ_{mc} qui est la probabilité préalable de la m^{th} gaussienne est considérée comme la pondération définie par :

$$\sum_{m=1}^{M} \lambda_{mc} = 1 \qquad \text{et } 0 \le \lambda_{mc} \le 1 \tag{3-28}$$

Pour un ensemble de données $x = \{x_1, x_2, x_3 \cdots x_N\}$ d'une distribution inconnue, l'estimation du paramètre θ de GMM est obtenue par la maximisation de la vraisemblance de $p(x|\theta)$. Elle est représentée par :

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, p(X|\theta) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p(x_i|\theta) \tag{3-29}$$

L'approche la plus connue pour la maximisation de la vraisemblance est l'algorithme de l'espérance-maximisation (Expectation-Maximization) noté EM.

La régression logistique : Est une technique de modélisation qui permet de prédire et d'expliquer les valeurs d'une variable catégorielle binaire *Y*, à partir d'une collection de variables *X* continues ou binaires. Elle fait partie des méthodes d'apprentissage supervisées ; elle peut s'inscrire dans le cadre de la régression linéaire généralisée.

Elle est considérée comme une alternative intéressante aux GMM et aux SVMs dans le cadre de l'identification des locuteurs [111] [112]. Le principe de la LR consiste à modéliser la probabilité postérieure de l'appartenance de classe à l'aide d'une fonction linéaire [112]. Cette fonction est définie par :

$$f(x) = \beta^T x \tag{3-30}$$

Où $\boldsymbol{\beta}$ est le vecteur de poids, et $\boldsymbol{x}$ est le vecteur de l'échantillon. L'équation 3.31 a été réarrangée pour estimer la probabilité P (x, β). En utilisant la fonction de transfert logistique, ce qui donne :

$$\log it\left(P\left(x,\beta\right)\right)=\log\frac{P\left(x,\beta\right)}{1-P\left(x,\beta\right)}=\beta^{T}x \tag{3-31}$$

Ce qui conduit à :

$$P\left(x,\beta\right)=\frac{1}{1+e^{-f(x)}} \tag{3-32}$$

Ainsi, l'estimation de la probabilité postérieure a été effectuée en vertu du principe que les données d'apprentissage proviennent d'une loi de Bernoulli. Cette hypothèse a été justifiée par cette recherche, car nous avons traité un problème de classification binaire des (natifs et non natifs). Par conséquent, la probabilité postérieure est définie par cette équation :

$$P\left(x,\beta|c\right)=\begin{cases}\dfrac{1}{1+e^{-f(x)}}\\[2ex] 1+\dfrac{1}{1+e^{-f(x)}}\end{cases} \tag{3-33}$$

Selon la Loi de Bernoulli, la dernière équation a été reformulée comme suit :

$$P\left(x,\beta|c\right)=P\left(x,\beta|\right)^{c}\left(1-P\left(x,\beta|\right)\right)^{1-c} \tag{3-34}$$

Le log de la vraisemblance négative, dénotée par $\mu\left[\beta\right]$, a été formulé comme suit :

$$\mu\left[\beta\right]=\sum_{i=1}^{N}-c\beta^{T}x_{i}+\log\left(1+e^{T}x_{i}\right) \tag{3-35}$$

Dans la plupart des implémentations, un terme de pénalité a été ajouté à $\mu[\beta]$ afin d'éviter le surapprentissage (*overlearning)*. Alors l'équation se récrit comme suit :

$$\mu[\beta]+\frac{\theta}{2}\|\beta\|^2 \tag{3-36}$$

Où θ est appelé le paramètre de régularisation. L'optimisation vise à trouver le meilleur jeu de poids β_i en réduisant au minimum le $\mu[\beta]_p$. Elle a été faite en mettant la dérivée de $\mu[\beta]_p$ à zéro. Pour obtenir les poids optimaux β_{opt}, nous avons a utilisé les relations suivantes :

$$\beta_{opt} = \left(X^T W X + \theta I\right)^{-1} X^T W z \tag{3-37}$$

Où z est exprimé par :

$$z = X\beta_{old} + W^{-1}\left(C - P\right) \tag{3-38}$$

Où ***I*** est la matrice d'identité ; ***p*** est le vecteur composé par des éléments de probabilité filtrés. L'élément j^{th} est defini par $P(\beta_{old}, x_j)$; ***W*** est la matrice NxN avec les probabilités $P(\beta_{old}), x_j)(1-P(\beta_{old}, x_j))$ sur la diagonale.

3.7. Conclusion

Dans ce chapitre, nous avons présenté les descripteurs les plus usités dans la littérature. Ces descripteurs étaient de natures acoustiques et linguistiques. On s'est concentré sur la nature acoustique des descripteurs, plus particulièrement celles relatives aux métriques rythmiques. Cette étude nous conduira au développement de nouveaux descripteurs acoustiques, qui seront détaillés au chapitre suivant afin d'améliorer la classification des paralinguistiques de la parole. La problématique de la reconnaissance automatique des traits

paralinguistiques a été décrite tout en insistant sur la nécessité d'effectuer une sélection optimale des paramètres d'entrée. Notons que le choix des classificateurs est toujours motivé par la nature des descripteurs et la dimension du problème à résoudre. Au prochain chapitre, nous proposons des solutions originales à toutes les étapes de classification.

Chapitre 4 - Nouveaux descripteurs pour la reconnaissance des émotions et des accents

4.1. Introduction

Historiquement parlant, en ce qui concerne la reconnaissance automatique des traits et des états paralinguistiques de la parole, il y a toujours eu plusieurs efforts et recherches faits pour améliorer les performances de reconnaissance. Le but était toujours d'obtenir de l'information la plus précise issue du message véhiculé par la parole, mais celle-ci était toujours incertaine, versatile et imprécise, surtout dans le cas de l'émotion des accents. Pour ce faire, nous commencerons en présentant de nouvelles métriques rythmiques. Il s'agit notamment d'une métrique généralisant les PVI et d'une métrique à base d'intensité. Nous présentons également le modèle d'oreille pour l'extraction d'un ensemble de paramètres acoustiques pertinents. Finalement, des résultats d'expérimentation accompagnent la représentation de chacun de descripteurs proposés.

4.2. État de l'art sur les rythmes des familles pairwise (PVI)

Diverses métriques du rythme ont été utilisées afin d'identifier les différents traits et opérer la classification de différentes variétés de différentes langues captant ainsi certaines caractéristiques pertinentes de la parole [113].

Plusieurs études ont montré que l'utilisation de mesures normalisées pouvait améliorer la discrimination entre les variétés linguistiques. Par exemple, Wiget et al ont prouvé que les VarcoV étaient plus performants dans la discrimination entre les différentes langues que les VarcoC et les nPVI [114]. De plus, ils ont recommandé l'utilisation d'une combinaison d'au moins deux métriques rythmiques. Loukin a partagé la même vision que Wiget et al [7]. En effet, dans son étude qui concernait la classification de cinq langues, Loukin a signalé qu'aucune métrique du rythme seule ou en combinaison avec d'autres ne permettrait pas une séparation optimale de toutes les paires de ces langues. D'autres études ont constaté que l'utilisation d'au moins trois mesures était nécessaire pour réaliser une telle séparation. De plus, il a été déterminé que les mesures normalisées vocaliques favorisaient la possibilité d'avoir une bonne séparation [115].].

Notons qu'Arvaniti a déterminé une grande variabilité au niveau des rythmes métriques. Elles étaient causées par plusieurs facteurs qui étaient : les méthodes d'enregistrement, le matériel et les variations entre les locuteurs [8].

Gut a annoncé que la plupart des métriques n'avait pas obtenu le même résultat après différentes études de même langue, ce qui a fait douter de la fiabilité de ces métriques [116].

Malgré toutes les limitations mentionnées par les différentes recherches, plusieurs chercheurs continuent d'utiliser les métriques du rythme dans leurs travaux. Prenons l'exemple de Nolan et Jeon qui ont justifié l'utilisation de ces métriques par l'absence d'une autre alternative qui permettait un classement des langues aussi performant [117].

Comme solution à ces lacunes, Bertinetto et Bertini ont proposé l'indice de contrôle (CCI), qui tient compte du degré de la complexité phonétique des langues naturelles. Cette

métrique effectue la normalisation des PVI en prenant en considération le nombre de segments (ou phonèmes) qui compose chaque intervalle vocalique ou consonantique [118]. Cette métrique rythmique est définie par ce qui suit :

$$CCI = \frac{100}{N-1} \times \sum_{K-1}^{N-1} \left| \frac{d_K}{n_K} - \frac{d_{K+1}}{n_{K+1}} \right| \tag{4-1}$$

- d_k et d_{k+1} représentent la durée d'un intervalle de la consonne ou de la voyelle à la position k *et* $k+1$ successivement;
- n représente la complexité phonétique de la structure, il est défini comme le nombre de phonèmes qui composent chaque intervalle vocalique ou consonantique;
- N est le nombre de segments dans chaque signal de la parole

Dans les études de l'influence de la langue maternelle sur la prononciation de la langue seconde, Gut a utilisé le DeltaC, le pourcentage V (%V) et le PVI-V dans le cas de l'impact des langues maternelles chinoises, anglaises, françaises, italiennes et romaines sur la prononciation de la langue allemande [119]. Carter a étudié l'influence de l'espagnol mexicain sur la prononciation de l'anglais américain en utilisant le nPVI-V et le rPVI-C[120]. Alors Chen, dans ses recherches à partir de 2010 et de 2013, a utilisé les métriques du rythme pour la conception d'un système automatique d'évaluation de la parole, pour les non natifs, qui a démontré que l'ajout du rythme à d'autres paramètres avait amélioré ce système [121]. D'ailleurs, la distinction entre la prononciation native et non native est un sujet important où le rythme trouve une place de choix. White et Mattys en 2007 ont affirmé que parmi toutes les métriques, le VarcoV était le meilleur concernant la distinction entre L1 et L2 [122]. De plus, Mok et Dellwo en 2008, ont déterminé que le VarcoC et le pourcentage (%V) donnait la bonne classification dans le cas de la distinction

entre la prononciation de l'anglais par des personnes natives et non natives de cette langue [123].

En se basant sur le taux de classification obtenu par le classificateur SVM, dans les travaux de Tortel et Hirst, le %V, le VarcoV et le nPVI-V ont été déclarés comme les meilleurs discriminants pour la détection et l'identification des accents natifs et non natifs de la parole. Dans le même contexte, nos recherches visent l'expérimentation des métriques rythmique pour la discrimination des accents natifs et non natifs ainsi que des émotions [124].

4.3. Proposition d'une nouvelle métrique : OPVI

La nouvelle métrique que nous proposons est désignée par OPVI, ce qui signifie qu'elle est une optimisation des métriques PVI, afin de faire face aux lacunes de cette famille de métriques cités dans les recherches et plus particulièrement ceux de Arvaniti en 2012, et de Loukina en 2011 en [8] et [7]. Dans ce qui suit, on présente la définition mathématique de notre nouvelle métrique et la technique de détermination des paramètres qui lui sont associés.

4.3.1 Définition de OPVI

L'optimisation de l'index de variabilité par paires (OPVI-Optimized Pairwise Variability Index) se définie comme suit:

$$OPVI = \frac{\alpha}{N-1} \times \sum_{K=1}^{N-1} \frac{\left| \frac{d_K}{\theta n_K} - \frac{d_{K+1}}{\varepsilon n_{K+1}} \right|}{\left(\frac{d_K + d_{K+1}}{2} \right)^{\beta}} \tag{4-2}$$

- $\alpha, \beta, \varepsilon, \theta \in \Re$ sont les paramètres d'optimisation pour l'OPVI, avec $1 \leq \alpha \leq 100$, $0 < \beta \leq 1$, $0 \leq \varepsilon \leq 1$, et $0 \leq \theta \leq 1$;
- d_k et d_{k+1} représentent la durée d'un intervalle de la consonne ou de la voyelle à la position k *et* $k+1$ successivement;
- n représente la complexité phonétique de la structure, il est défini comme le nombre de phonèmes qui composent chaque intervalle vocalique ou consonantique;
- N est le nombre de segments dans chaque signal de la parole

Les quatre paramètres d'optimisation de la métrique OPVI soient (α, β, ε, θ) peuvent être calculés avec un algorithme qui maximise la classification du trait à l'étude. Cette recherche applique l'optimisation par essaims particulaires (PSO-Particle Swarm Optimization), qui est une technique de calcul basée sur l'intelligence des essaims, proposé par Kennedy et Eberhart en 1995 [125]. PSO peut se converger mathématiquement assez rapidement. Le fonctionnement de cette méthode d'optimisation est plus détaillé dans le chapitre cinq. Dans ce travail, les quatre paramètres, (α, β, ε, θ) forment les individus pour la PSO. La meilleure position obtenue par ces coefficients est considérée comme la meilleure pour l'individu et celle obtenue par la population est considérée comme la meilleure pour toute la population. L'algorithme de calcul de ces paramètres est donné à la Figure 4–1.

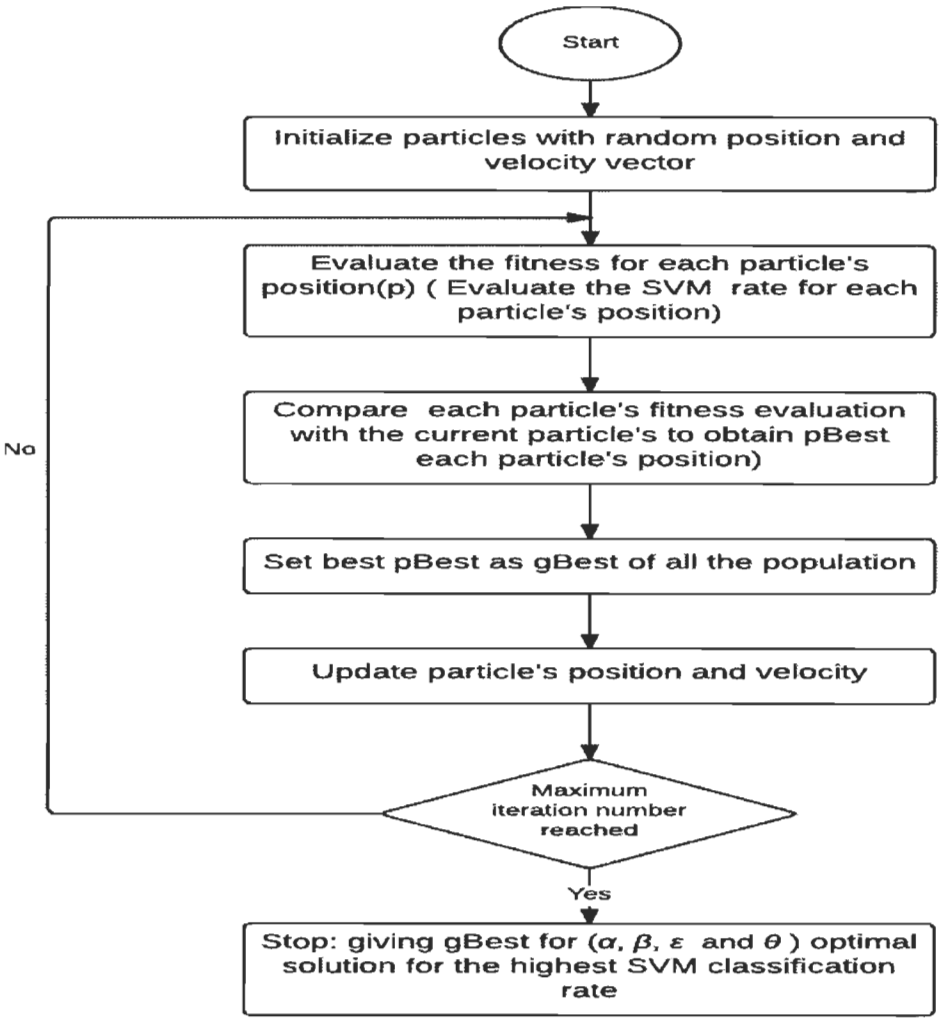

Figure 4–1:L'algorithme du calcul des coefficients par PSO

Une des caractéristiques de la métrique OPVI est sa relation avec les métriques de rythme du PVI. Dans ce qui suit, on démontre comment les trois rythmes classiques, le nPVI, le rPVI et le CCI, peuvent être dérivés de l'OPVI. La métrique de rythme OPVI intègre chacun de ces paramètres comme un cas particulier.

D'après la formule 4-3 d'OPVI on observe que lorsque $\alpha = 100$, $\beta = 1$, $\varepsilon = 0$ et $\theta = 0$, l'OPVI devient :

$$OPVI = \frac{100}{N-1} \times \sum_{K=1}^{N-1} \frac{\left| \frac{d_K}{n_K^0} - \frac{d_{K+1}}{n_{K+1}^0} \right|}{\left(\frac{d_K + d_{K+1}}{2} \right)^1} \tag{4-3}$$

La formule précédente devient donc celle du nPVI qui est définie par :

$$nPVI = \frac{100}{N-1} \times \sum_{K=1}^{N-1} \frac{|d_K - d_{K+1}|}{\frac{d_K + d_{K+1}}{2}} \tag{4-4}$$

Quand on utilise la formule (4-1) avec $\alpha = 1$, $\beta = 0$, $\varepsilon = 0$ et $\theta = 0$, O-PVI devient :

$$OPVI = \frac{1}{N-1} \times \sum_{K=1}^{N-1} \frac{\left| \frac{d_K}{n_K^0} - \frac{d_{K+1}}{n_{K+1}^0} \right|}{\left(\frac{d_K + d_{K+1}}{2} \right)^0} \tag{4-5}$$

Qui est la définition du rPVI dont la formulation est :

$$rPVI = \frac{1}{N-1} \times \sum_{K=1}^{N-1} |d_K - d_{K+1}| \tag{4-6}$$

Également, si $\alpha = 100$, $\beta = 0$, $\varepsilon = 1$ et $\theta = 1$, OPVI devient la métrique CCI :

$$OPVI = \frac{100}{N-1} \times \sum_{K=1}^{N-1} \frac{\left| \frac{d_K}{n_K^1} - \frac{d_{K+1}}{n_{K+1}^1} \right|}{\left(\frac{d_K + d_{K+1}}{2} \right)^0} = CCI = \frac{100}{N-1} \times \sum_{K-1}^{N-1} \left| \frac{d_K}{n_K} - \frac{d_{K+1}}{n_{K+1}} \right| \tag{4-7}$$

Ces équations démontrent que la métrique du rythme de l'OPVI généralise les PVI et intègre chacune des trois métriques du PVI qui sont : le nPVI, le rPVI et le CCI, comme un cas particulier.

On observe une caractéristique importante qui distingue la relation entre les métriques décrites ci-dessus ; c'est la question de la normalisation. Tel que mentionné dans les sections précédentes, différentes normalisations ont été proposées afin de réduire les effets, tel que le débit de la parole sur les métriques. Dans le cas de la métrique nPVI, qui est une normalisation du rPVI, la normalisation a été réalisée en divisant la différence de durées des intervalles adjacents (vocaliques et consonantiques) par la durée moyenne des deux intervalles, comme c'est représenté par les dénominateurs de l'équation 4.7.

Concernant le rythme métrique CCI, qui est aussi une normalisation des rPVI, cette normalisation inclut le nombre de segments dans chaque intervalle mesuré, la durée de chaque intervalle est divisée par le nombre de segments de cet intervalle. Également, comme une généralisation du nPVI, du rPVI et du CCI, la métrique OPVI fournit la normalisation en deux points : le premier est au niveau de chaque intervalle où ε et θ sont deux paramètres qui abordent la structure phonétique de chaque intervalle, le deuxième est la normalisation dans les deux intervalles qui est représentée par la comparaison par paire, où le paramètre β est associé à l'expression $\left(\frac{d_K + d_{K+1}}{2}\right)$, et il traite de la durée des intervalles.

4.3.2 Performance de la métrique OPVI dans la classification des accents natifs et non natifs

Le corpus de la Langue Arabe West Point du standard LDC a été utilisé [126]. Ce corpus contient les enregistrements de la parole arabe prononcée par des locuteurs arabes natifs et non natifs ayant l'anglais comme première langue. On a utilisé plus précisément 16 locuteurs, huit parmi eux étaient des natifs et les autres étaient des non natifs. Les deux groupes étaient composés de 4 femmes et de 4 hommes.

Chaque locuteur de ces groupes devait prononcer dix phrases très similaires dont les sujets incluaient : le nom, la résidence et l'occupation du locuteur. Chaque phrase de la liste choisie a été analysée acoustiquement par Praat [127].

Pour les 160 phrases choisies, la nouvelle métrique rythmique OPVI a été calculée, ainsi que l'optimisation de ses paramètres (α, β, ε, θ) et ça incluait la version vocalique (OPVI-V) et consonantique (OPVI-C). Nous avons appliqué l'ANOVA pour avoir une idée sur la pertinence de cette métrique rythmique dans la discrimination entre l'accent natif et non natif. Les résultats ont démontré que cette métrique était discriminatoire dans sa version vocalique où on a noté un p (0,003) < 0. Le tableau 4-1 présente les résultats de la signification statistique de l'OPVI dans sa version vocalique et consonantique pour les accents natifs et non natifs.

Tableau 4-1: La moyenne, l'écart-type et les résultats du test de signification (valeur-p) de l'ANOVA des deux versions de l'O-PVI (O-PVI-V et O-PVI-C). La métrique est significative lorsqu'elle obtient une valeur-p<0.05.

Métrique rythmique	Natif	Non natif	Valeur- p
O-PVI-V	3.18 (1.02)	3.44 (0.87)	0.507
O-PVI-C	17.64 (4.34)	16.26 (4.93)	**0.003**

Les SVMs ont été utilisées pour tester la capacité de cette nouvelle métrique qui touche à la classification des locuteurs natifs et non natifs de parole arabe. Pour réussir cette étape du système de reconnaissance automatique du trait accent, les données ont été divisées en deux groupes : l'un pour l'apprentissage qui représentait 70% des données choisies et l'autre pour le test qui représentait le reste des données soit 30%. Également, on a utilisé deux SVM, l'un pour l'OPVI-C et l'autre pour l'OPVI-V. Les résultats obtenus ont démontré que OPVI-C était plus performante pour la classification des accents natifs et non natifs pour la langue arabe en atteignant un score de 89.8 % comme taux de classification correct.

4.4. Métriques rythmiques à base d'intensité

Notre motivation à utiliser l'intensité comme paramètre de base du rythme de la parole découle du fait que plusieurs études ont montré que la variation de l'intensité était corrélée au rythme de la parole. Cette relation, ouvrait la possibilité que le rythme est représenté

dans un espace multidimensionnel et que la durée représentait une parmi d'autres dimensions. Cette relation a permis donc de classer les langues en deux groupes : les langues ayant des variabilités importantes en intensité, mais moins en durée, et les langues qui étaient caractérisées par une variabilité importante dans la durée, mais moins au niveau de l'intensité. Ceci nous amène à penser qu'une langue donnée peut être caractérisée par une coexistence entre différents aspects, incluant l'intensité, qui contribuent de façon différente dans la représentation du rythme de la parole.

Nazzi a démontré que les enfants étaient capables de différencier plusieurs langues en se basant seulement sur les propriétés prosodiques de la parole. Il a justifié cette méthode par le manque de connaissances de la sémantique des langues parlées chez les enfants[128]. Kurodo, lui, a calculé les rythmes à base d'intensité pour les adultes [129]. Avec Ferrage, les PVI, en version vocalique et consonantique à base d'intensité, ont amélioré la séparation entre 13 dialectes de la langue anglaise en BRETAGNE [130].

Chacune des études du rythme à base d'intensité a utilisé une méthode de calcul différente. Notre approche a été de procéder à la généralisation à toutes les métriques similaires à celle utilisée pour le calcul du rythme à base de durée. Cette méthode a conduit à l'obtention des mêmes métriques que la durée, bien qu'ils étaient à base d'intensité (voir le tableau 4.2). Nous avons également effectué, tel que montré au Tableau 4.3, une étude statistique (ANOVA) des métriques rythmiques à base d'intensité.

Tableau 4-2: Description des métriques à base d'intensité

Métrique	Description
Delta-C_i	L'écart-type des intensités d'intervalles des consonnes par phrase
Delta-V_i	L'écart-type des intensités d'intervalles vocaliques par phrase
%V_i	La moyenne des intensités d'intervalles vocaliques,
Varco-V_i	La normalisation des écarts-types des intensités d'intervalles vocaliques par phrase
Varco-C_i	La normalisation des écarts-types des intensités d'intervalles des consonnes par phrase
Varco-VC_i	La normalisation des écarts-types des mesures syllabiques successives dans la phrase
rPVI-V_i	La différence moyenne des intensités entre deux intervalles vocaliques successifs dans la phrase
rPVI-C_i	La différence moyenne des intensités entre deux intervalles de deux consonnes successives dans la phrase
nPVI-C_i	La normalisation de la différence moyenne des intensités entre deux intervalles des consonnes successives dans la phrase
nPVI-V_i	La normalisation de différence moyenne des intensités entre deux intervalles vocaliques successifs dans la phrase

Tableau 4-3 : La moyenne (l'écarte-type) et la signification de l'ANOVA pour les rythmes à base d'intensité d'accent arabe natif et non natif. Le rythme métrique est considéré significatif lorsqu'on a p-value<0.05.

Métrique	Natif	Non natif	p-value
Delta-C_i	42.79 (2.43)	41.98 (7.50)	0.38
Delta-V_i	4.5 (1.99)	5.22 (3.15)	0.45
%V_i	5.47 (1.92)	6.18 (2.92)	**0.034**
Varco-V_i	6.10 (2.82)	7.77 (5.02)	0.23
Varco-C_i	7.92 (2.91)	9.68 (4.59)	0.15
Varco-VC_i	8.21 (2.34)	8.85 (3.05)	0.15
rPVI$_i$	5.96 (2.55)	8.40 (6.48)	0.82
rPVI-C_i	8.35 (2.67)	10.52 (5.87)	0.29
nPVI$_i$	5.67 (1.6)	6.58 (3.44)	0.19
nPVI-C_i	8.22 (1.84)	8.94 (2.29)	0.17
nPVI-Vi	5.74 (1.09)	5.82 (1.14)	0.38

Des expériences de classifications utilisant les classificateurs SVM et GMM ont été effectuées pour classer l'accent de la prononciation de la langue arabe de la base de données MSA. Ces classificateurs ont admis comme entrées des rythmes métriques à base d'intensité. Le classificateur SVM a obtenu un taux d'erreur de 18.67%, par contre le GMM a obtenu un taux d'erreur égal à 46.79% pour 16 gaussiennes.

4.5. Un modèle d'audition pour extraire des descripteurs distinctifs

L'oreille capte les sons et les transforme en influx nerveux qui sont envoyés au cerveau et celui-ci s'occupe de les interpréter. Le système auditif contient trois parties importantes qui sont : l'oreille externe, l'oreille moyenne et l'oreille interne. Elles jouent un rôle bien défini dans le processus de la transformation des ondes sonores en impulsions nerveuses envoyées au cerveau.

L'oreille externe est le seul moyen de contact avec le milieu extérieur. Le pavillon et le conduit auditif sont les deux parties de l'oreille externe. La fonction du pavillon est de transmettre les sons au tympan.

L'oreille moyenne est un espace rempli d'air connecté à la trompe d'eustache. Cet espace est composé de trois petits os qui sont : le marteau, l'enclume et l'étrier. Ils permettent de transmettre le son de la membrane tympanique, qui est la partie externe de l'oreille, vers la membrane basilaire.

L'oreille interne est responsable de la perception, de l'accélération et de la position angulaire de la tête. Ces mouvements sont transmis à la cochlée, qui est un organe creux rempli d'un liquide appelé endolymphe et qui est couverte par des cellules ciliées sensorielles. Ces dernières se trouvent le long de la membrane basilaire. Les vibrations

transmises au travers de l'oreille médiane mettent la membrane en mouvement. Chacune des cellules ciliées répond à un type bien déterminé de fréquences, afin que le cerveau soit capable de différencier la hauteur des sons. Alors, les cellules ciliées les plus proches de la base de la cochlée répondent aux fréquences aiguës, alors que celles situées dans son apex répondent aux basses fréquences. Les cellules ciliées transforment un mouvement en signal nerveux par le nerf auditif, cette opération de transformation est connue par la transduction mécano-électrique.

Étant donné que l'oreille humaine est capable de comprendre, d'analyser et de déchiffrer la parole, cette grande capacité du système auditif a encouragé les chercheurs à l'exploiter dans leurs travaux dans ce domaine.

Par exemple, Flanagan a utilisé un modèle automatique (assisté par ordinateur) afin d'évaluer le fonctionnement de la membrane basilaire, celui-ci a été efficace pour les rapports entre : la relation subjective du comportement du système auditif, le système acoustique et la mécanique de l'oreille.

Dans le domaine de la modélisation auditive, plusieurs modèles ont connu du succès dont les plus populaires sont ceux de [131], Seneff [132], [133] et Ghitza [134].

Les développements remarquables, dans le domaine de la vitesse de calcul des algorithmes complexes, ont mené à de l'implémentation de nouvelles techniques de modélisation basées sur la compréhension du système auditif physiologique et pyschoacoustique. Également, les observations de Stern et Morgan ont provoqué une renaissance concernant le développement des paramètres issus de modèle auditif pour l'extraction de descripteurs robustes [135]. De plus, après l'apparition des MFCC, des PLP [136], des GFCC [137] et des PNCC [138] , la modélisation du système auditif est devenue

un outil de grande importance pour l'extraction des paramètres acoustiques robustes. Dans notre travail, nous avons utilisé le modèle auditif conçu par Caelen, pour effectuer le décodage du signal de la parole afin d'en extraire les paramètres pertinents [139].

4.5.1 Modèle d'oreille de Caelen

Le modèle auditif de Caelen se compose de trois parties (externe, moyenne et interne) et leurs fonctions sont de simuler le fonctionnement du système auditif.

L'oreille externe et moyenne sont représentées par des filtres passe-bande qui s'adaptent aux signaux énergiques en tenant compte de la variation des vibrations des osselets. L'oreille interne, la partie la plus importante, a comme fonction de simuler le comportement de la membrane basilaire. Celle-ci se comporte en grande partie comme un filtre non-linéaire. La membrane basilaire est sensible aux sons, qui ont une fréquence différente, à cause de la variation de sa rigidité. En ce qui concerne la base de la membrane basilaire, elle est en général dure et mince, mais plus sensible et moins rigide aux signaux de basse fréquence à l'apex. On reconnaît une fréquence spécifique, à chaque endroit sur le long de la membrane basilaire, lorsqu'elle vibre au maximum pour un signal de son. Ce comportement est simulé par un banc de filtres. Plus le nombre de filtres est élevés plus le modèle et efficace. La figure 4.2 illustre par un diagramme le modèle auditif de Caelen utilisé dans notre recherche.

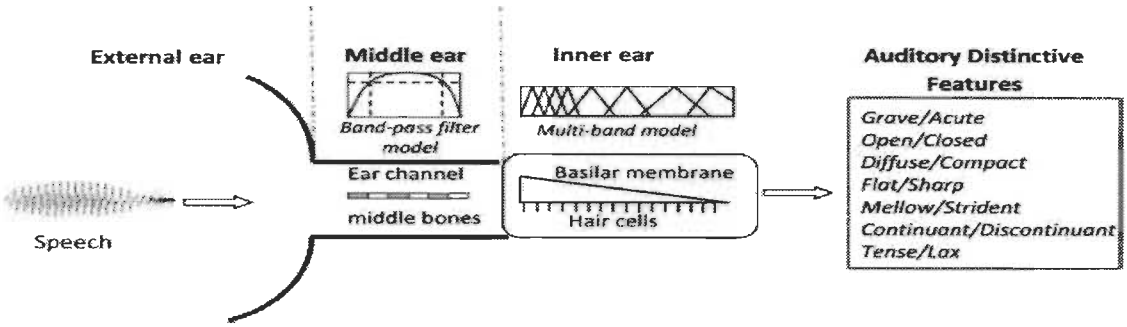

Figure 4–2: Diagramme du modèle d'oreille humain utilisé qui représente les trois parties de l'oreille : l'oreille externe, l'oreille moyenne et l'oreille interne

Le fonctionnement du filtre passe-bande modélisant l'oreille externe et moyenne est donné par la formule récurrente suivante :

$$s'(k) = s(k) - s(k-1) + \alpha_1 s'(k-1) + \alpha_2 s'(k-2) \tag{4-8}$$

Où s *(k)* est le signale de la parole, s *'(k)* est le signal filtré, le $k=1...K$ est l'index de temps et le K représente le nombre d'échantillons par trame. Les coefficients α_1 *et* α_2 dépendent de la fréquence d'échantillonnage FS, de la fréquence centrale du filtre et de son Q-facteur (facteur de qualité).

L'énergie correspondant à cette partie est calculée et définie par la formule suivante :

$$W_{ME}(T) = 20\log\sum_{k=1}^{K}\left|s'(k)\right| \tag{4-9}$$

W_{ME} est le premier paramètre acoustique calculé par le modèle d'oreille.

Le fonctionnement de la membrane basilaire est simulé par des filtres. Chaque réponse d'une cellule ciliée est simulée par un filtre. Chaque cellule ciliée est excitée par un type spécifique de fréquence selon son placement ou positionnement dans la membrane. Cette propriété est prise en considération dans notre modèle. En effet, chaque filtre (canal) est

associé à un type bien défini de fréquence comme c'est présenté dans le Tableau 4-4 Celui-ci contient les 24 filtres utilisés avec une fréquence centrale associée à chaque filtre.

Tableau 4-4 La résonnance de la fréquence de la membrane simulée par le modèle d'oreille exprimée sur 24 canaux.

Canal	Fréquence	Canal	Fréquence
1	180	13	1340
2	215	14	1550
3	260	15	1790
4	320	16	2060
5	380	17	2350
6	450	18	2700
7	540	19	3100
8	650	20	3550
9	760	21	4000
10	880	22	4500
11	1000	23	5000
12	1130	24	5600

Le résultat de la transformation de chaque trame du signal par la partie de l'oreille externe et moyenne est envoyé au filtre de la cochlée. Ce filtre répond par des fréquences qui représentent la simulation de la membrane pour cette trame du signal. Cette réponse est représentée par :

$$y_i(k) = \beta_{1i}\, y_i(k-1) - \beta_{2i}\, y_i(k-2) + G_i\left[s'(k) - s'(k-2)\right] \tag{4-10}$$

Cette dernière formule admet une fonction de transfert définie par :

$$H_i(z) = \frac{G\left[1 - z^{-2}\right]}{1 - \beta_{i,i} z^{-1} + \beta_{2,i} z^{-2}} \tag{4-11}$$

Dans notre modèle d'oreille, les dernières parties du modèle s'occupent de la simulation de la partie électromécanique de la transduction des cellules ciliées et des fibres. Dans le but d'éviter la complexité, seuls les effets qui sont engendrés par ces deux derniers (cellules ciliées et fibres) ont été pris en considération.

L'énergie absolue de sortie de chaque canal a été calculée comme suit

$$W_i' = 20\log\sum_{k=1}^{k}\left|y'_i(k)\right| \tag{4-12}$$

Une fonction de lissage a été appliquée afin de réduire la variation d'énergie. Cette fonction est définie par l'équation suivante :

$$W_i(T) = c_0\, W_i(T-1) + c_1\, W_i(T) \tag{4-13}$$

Où $W_i(T)$ est l'énergie de lissage et c_0 et c_1 sont les coefficients de moyenne $W_i(T-1)$ et $W_i(T)$, tels que leurs somme est égale à 1. Dans nos expériences, nous avons fixé la valeur des deux coefficients à 0.5.

Les sorties des filtres modélisant la membrane basilaire a permis de calculer des descripteurs distinctifs auditifs en effectuant la combinaison linaire des énergies de sorties des filtres. Sept paramètres ont été sélectionnés comme étant les plus pertinents et distinctifs pour le domaine de la reconnaissance des traits paralinguistiques. Alors, la description avec les formules de calcul de ces paramètres est donnée comme suit :

Grave/Aigu qui mesure la différence d'énergie entre les basses fréquences (50 à 400 Hz) et hautes fréquences (3800-6000 Hz). Cette différence correspondant à la combinaison linéaire suivante :

$$G/A = W_1 + W_2 + W_3 + W_4 + W_5 - W_{20} - W_{21} - W_{22} - W_{23} - W_{24} \tag{4-14}$$

Ouvert/Fermé qui permet de caractériser un son fermé si l'énergie des basses fréquences (230-350Hz) est supérieure à l'énergie des fréquences moyennes (600-800Hz). Ce paramètre est calculé comme suit :

$$O/C = W_8 + W_9 - W_3 - W_4 \tag{4-15}$$

Diffus/Compact qui est un paramètre de compacité qui reflète l'importance de la région centrale (800-1050 Hz) par rapport aux régions des fréquences de la plage (300-700 Hz) et (1450-2550 Hz). Ce paramètre est donné par la différence d'énergie suivante :

$$D/C = W_{10} + W_{11} - (W_4 + W_5 + W_6 + W_7 + W_8 + W_{13} + W_{14} + W_{15} + W_{16} + W_{17})/5 \tag{4-16}$$

Bémolisé/Diésé où un phonème qui est considéré comme fort si l'énergie est entre (2200-3300 Hz) et important si l'énergie est entre (1900-2900 Hz). Ce paramètre est donné par :

$$F/S = W_{17} + W_{18} + W_{19} - W_{11} - W_{12} - W_{13} \tag{4-17}$$

Doux/Strident où les phonèmes stridents sont caractérisés par la présence de bruit, à cause d'une turbulence à leur point d'articulation, qui mène à plus d'énergie entre (3800-5300 Hz) qu'entre (1900-2900 Hz).

$$S/M = W_{21} + W_{22} + W_{23} - W_{16} - W_{17} - W_{18} \tag{4-18}$$

Continu/Discontinu quantifie la variation de l'amplitude du spectre en comparant l'énergie des cadres actuels et précédents. Ce paramètre est bas dans le cas d'une variation faible et élevé dans le cas contraire. Il est représenté par la formule suivante :

$$C/D = \sum_{i=1}^{N_c} \left| W_i(\mathrm{T}) - Wa(\mathrm{T}) - W_i(\mathrm{T}-1) - Wa(\mathrm{T}-1) \right| \tag{4-19}$$

Tendu/Lâche qui est défini par la mesure de la différence d'énergie entre les fréquences moyennes (900-2000 Hz) et les fréquences relativement élevés (2650-5000 Hz). Le T/L est calculé par :

$$T/L = W_{11} + W_{12} + W_{13} + W_{14} + W_{15} + W_{16} - W_{18} - W_{19} - W_{20} - W_{21} - W_{22} - W_{23} \tag{4-20}$$

4.5.2 Classification des accents natifs et non natifs par le modèle d'audition

Lors de l'expérimentation du modèle d'oreille, les classificateurs SVM, GMM et GMM/SVM ont été utilisés pour classer l'accent de la prononciation de la langue arabe de la base de donnés West Point de LDC pour les natifs ou non natifs arabes. Avec la SVM, un taux d'erreur égale à 5.21% a été atteint avec un σ =28.1. Par contre, avec le GMM le taux était égal à 7.56% pour 8 gaussiennes et avec le GMM/SVM, ce taux était égal à 15.92% pour 4 gaussiennes et pour σ =13.5.

4.6. Conclusion

Dans ce chapitre, nous avons proposé une nouvelle métrique de rythme, O-PVI, qui est une généralisation de trois mesures de variabilité d'index par paire classique rythme : nPVI, rPVI et CCI. Cette nouvelle mesure a optimisé la classification des groupes de locuteurs en adaptant les quatre coefficients (α, β, ε et θ) associés à deux points de normalisation et

calculés par un algorithme d'optimisation par essaim particulaire PSO. Nous avons également présenté des métriques à base d'intensité avec des résultats préliminaires de statistiques et de classification, pour l'identification et la détection des accents natifs et non natifs de langue arabe. Toujours dans le but de trouver de nouveaux paramètres acoustiques plus robustes et pertinents pour le système de reconnaissance des traits paralinguistique de la parole, un modèle d'oreille qui simule le fonctionnement de l'oreille humaine a été présenté. Ce modèle a permis l'extraction de huit paramètres acoustiques. Ces derniers ont été pertinents pour la reconnaissance des traits paralinguistiques de la parole. Dans le prochain chapitre nous proposerons de nouvelles approches dans la sélection et la classification automatiques des traits paralinguistiques de la parole.

Chapitre 5 - Nouvelles approches de classification

5.1. Introduction

Dans ce chapitre, nous présenterons de nouvelles approches d'optimisation du système de reconnaissance automatique des traits et états paralinguistiques de la parole. Pour ce faire, nous débuterons en soulignant l'utilité de combiner les différents descripteurs et en présentant la notion de la trame acoustico-phonétique multi-variable et les composants de celle-ci. L'optimisation de l'extraction, la sélection et la classification des émotions et des accents seront présentées dans le détail.

5.2. Combinaison des approches : une voie prometteuse

La reconnaissance des traits et états paralinguistiques est un domaine caractérisé par de l'ambiguïté au niveau de la distinction des différents traits. Cette ambiguïté n'est pas causée uniquement par un système non fiable, mais également par les personnes qui cachent parfois, de façon explicite, leurs états émotionnels ou leurs traits de personnalité, ce qui rend très difficile la tâche du système au niveau de la classification. Dans ce contexte, la combinaison de différentes approches devient nécessaire pour avoir une amélioration du taux de classification des traits paralinguistiques. Plusieurs sources d'information peuvent être mises à profit telles que l'audio et le visuel ou vidéo, l'audio et le texte. Dans notre cas nous nous restreignons à la parole comme seule et unique source de données et

d'information et par conséquent la combinaison des approches se limite seulement au signal de la parole. Plusieurs stratégies de combinaison peuvent être adoptées :

La combinaison des descripteurs qui est la fusion de plusieurs paramètres acoustiques calculés à partir de différentes unités de segmentation. Ainsi, l'analyse acoustico-phonétique multivariable est basée sur la fusion de plusieurs sources d'information. Ces dernières sont utilisées simultanément de façon parallèle. Dans cette thèse, la différence entre les flux n'est pas au niveau des sources, mais au niveau de la représentation du signal de la parole, car les flux venant de différentes représentations du signal engendrent des erreurs à des niveaux différents. Les paramètres acoustiques se complètent afin d'avoir des systèmes de reconnaissance des traits paralinguistiques robustes et fiables.

Généralement, on a trois types de systèmes multivariable : le premier est défini par différentes natures sensorielles comme source d'information, où on trouve l'exemple visuelle et audio ou l'audio et la vidéo. Le deuxième est défini lorsqu'on a différents types de processus d'extraction des descripteurs, comme dans le cas de cette thèse. On a utilisé différentes unités d'analyse pendant le calcul des paramètres acoustiques qui étaient différents au niveau de la portée temporelle. Comme résultat, on a obtenu des informations à court et à long terme. Pour le troisième type de système multivariable, il est défini par la combinaison des deux types cités antérieurement.

Au niveau du système multivariable à processus d'extractions différentes, on a deux catégories : l'intégration précoce et l'intégration probabiliste. L'intégration précoce (prématurée) est aussi connue par le terme concaténation. Elle se fait antérieurement à la phase de modélisation acoustique. Cette approche de vecteur multivariable est la plus populaire d'après la littérature. Par contre, au niveau de la combinaison probabiliste, la

fusion des paramètres acoustiques se produit seulement au niveau de la modélisation acoustique : c'est la combinaison des modèles acoustiques.

Les descripteurs qui nous concernent sont les coefficients MFCC, la qualité de la voix, les métriques du rythme à base d'intensité et à base durée, les paramètres auditifs acoustiques et les paramètres prosodiques. Cette fusion sera également nommée analyse multivariable (multistream analysis).

La combinaison au niveau de la sélection se sert de différentes méthodes de sélection combinées. Dans cette thèse, nous avons expérimenté trois méthodes de sélection : l'ACP, la LDA et l'ANOVA. Chacune des méthodes de sélection possède des propriétés et des caractéristiques exploitables, afin de faire simultanément la réduction et la sélection optimale des descripteurs et d'augmenter ainsi la chance que toutes les données soient représentatives et pertinentes. Par exemple, pour les vecteurs propres de l'ACP, chacun présente toutes les données d'entrées. Lorsqu'on applique la LDA à ces vecteurs propres (résultant de l'application de l'ACP), il sera possible de déterminer les vecteurs propres les plus pertinents et de donner une présentation linéaire de ces derniers.

Au niveau de la classification ce sont les classificateurs qui sont combinés. D'après Lugger dans [140], les méthodes de sélection sont très importantes dans le cas de l'utilisation d'une masse de paramètres acoustiques calculés à partir de la parole. Par contre, les paramètres éliminés lors de la sélection peuvent contenir des informations utiles pour la classification. Cette possibilité l'a conduit à garder tous les paramètres et à remplacer la sélection par l'utilisation de l'approche de la combinaison des différents classificateurs, ce qui a donné la possibilité de conserver toutes les informations jusqu'à la phase finale du système. On distingue trois types de combinaisons à ce niveau : la

combinaison hiérarchique, la combinaison en série et la combinaison en parallèle qui ont toutes été expérimentées par de nombreux chercheurs [141] [142] [143] [140].

Dans la combinaison en série les classificateurs sont organisés à la file. Chacun passe un modèle acoustique au suivant de la file, le taux de classification est délivré par le dernier classificateur de la file. Cette stratégie a été utilisée, afin d'obtenir des descripteurs de haut niveau plus pertinents, pour la représentation des traits paralinguistiques. Ce type de combinaison a abouti à une architecture à deux ou trois niveaux de classificateurs, comme par exemple les DBN (Deep Belief Network). Les descripteurs à haut niveau sont connus sous le nom de supervecteurs. Plusieurs chercheurs ont démontré que ces derniers avaient une capacité discriminatoire plus élevée que les descripteurs acoustiques de bas niveau [140]. Dans notre approche nous avons utilisé la combinaison GMM/SVM. Le GMM a calculé les supervecteurs à partir de la distribution gaussienne des paramètres acoustiques, et en se basant sur ces supervecteurs la SVM a effectué la classification. Dans ce type de combinaison, on retrouve aussi les architectures à base de réseaux de neurones profonds. Généralement, ces réseaux sont composés d'un ensemble de RBM (Restricted Boltzmann Machine). Pour ces types de réseaux, l'apprentissage se fait en deux étapes : une pour les couches visibles et l'autre pour les couches invisibles en se basant sur la fonction énergie de ces dernières. On détaille cette architecture dans la section 5.4 de ce chapitre.

Dans la combinaison en parallèle, chaque classificateur réalise sa modélisation et prend sa décision indépendamment des autres classificateurs de cette combinaison, puis il se produit une fusion des taux de classification.

Dans la combinaison hiérarchique les classificateurs sont organisés sous forme d'arbre en passant d'un nœud à un autre, alors la discrimination des traits paralinguistiques devient plus précise.

Au niveau des méthodes d'optimisation qui est utilisée pour les mêmes raisons qui ont conduit à la combinaison des méthodes de sélection. Cette combinaison peut être en parallèle ou en série et permet de retenir les poids optimaux à assigner aux descripteurs et/ou aux classificateurs. Souvent il est fait recours à une combinaison de méthodes d'optimisation pour bénéficier de leur complémentarité. Par exemple, le manque de diversité dans le cas de PSO et la difficulté de convergence pour les algorithmes évolutionnaires différentiels a poussé les chercheurs à une combinaison des deux méthodes pour échapper à leurs restrictions.

5.3. Les algorithmes d'optimisation par évolution différentielle (DE)

L'évolution différentielle (DE) fait partie de la famille des algorithmes évolutionnaires (AE) utilisés pour des tâches d'optimisation[144]. Les AE disposent d'une terminologie spécifique qui peut être résumée par les points suivants :

- chaque point de l'espace de recherche *d* est un individu ;
- les individus de l'espace de recherche forment une population ;
- la fonction à optimiser est appelée : fonction objective ou fonction fitness ;
- l'évaluation est le processus du calcul des performances des individus afin de trouver ceux qui sont optimaux ;
- la génération est un nombre défini d'itérations pour chaque population ;

- les opérateurs de variation sont utilisés pour la génération de nouveaux individus qui sont : le croisement et la mutation ;
- la sélection permet le choix des individus qui se reproduisent en se basant sur leurs performances ;
- le remplacement est l'opération de génération d'une nouvelle population à partir de parents et d'enfants.

L'évolution différentielle a été développée par Ston et Price [145]. Son rôle est d'optimiser des problèmes continus en se basant sur le principe d'une population de solutions aléatoirement initialisée, qui est constituée de plusieurs individus. Pour appliquer cet algorithme, un nombre de paramètres a été désigné nécessaire afin d'obtenir des résultats satisfaisants. La qualité du choix de ces paramètres a conditionné le succès de l'application de ces algorithmes pour l'optimisation. Ces paramètres sont cités dans ce qui suit avec une brève description :

- Un mécanisme de codage pour représenter les individus. Parmi les codages utilisés on distingue le codage binaire réel et le codage réel.
- La création de la population initiale. Cette étape d'initialisation est primordiale, car elle produit une population d'individus non homogènes qui sera la base des générations futures. Ce mécanisme d'initialisation est important, car la population initiale peut influencer la rapidité de convergence.
- La fonction objective qui fait la correspondance de la solution de problèmes en valeurs « fitness ».

- Les paramètres à définir sont la taille de la population, le nombre de générations et la dimension du problème.
- Le critère d'arrêt peut être le nombre maximal de générations, une solution satisfaisante (un bon taux de classification pour un problème de classement), l'absence de changement au niveau des solutions après un certain nombre de générations, un temps maximal de calcul et la combinaison de ces différents critères.

Les opérateurs de DE sont nécessaires pour la diversité de la population, ainsi que pour l'exploration de l'espace de recherche. On distingue trois opérateurs : l'opérateur de croisement, l'opérateur de mutation et l'opérateur de sélection. La mutation est l'opération d'application de la différenciation vectorielle entre les membres actuels de la population pour déterminer le degré et la direction de perturbation. Le processus de mutation s'applique à chaque génération et commence par la sélection au hasard des individus de la population sujets à cette mutation. Le croisement est utilisé afin d'augmenter la diversité des vecteurs de paramètres. La sélection permet de choisir la population avec le meilleur fitness constituer la prochaine génération.

Le principe de fonctionnement d'un algorithme d'évolution différentielle commence par la création d'une population d'individus choisis de façon aléatoire. Pour faire le passage d'une génération k à la génération k+1, les trois opérations mentionnées antérieurement sont appliquées pour tous les individus de la population k.

Deux individus a_{r_3} et a_{r_2} sont sélectionnés de façon aléatoire avec $r_2 \neq r_3$. La différence entre ces deux individus est additionnée au troisième individu a_{r_1} *avec* $r_1 \neq r_2 \neq r_3$. Ce troisième est choisi aléatoirement pour obtenir le vecteur muté v_i qui peut être choisi

comme individu dans la nouvelle population. Cette opération est définie par l'équation suivante :

$$V_{ig} = a_{gr_1} + F(a_{gr2} - a_{gr_3}) \tag{5-1}$$

Le vecteur muté V_i est combiné avec les individus x_i de la population, $i \neq r_2 \neq r_3 \neq r_1$, le résultat de cette combinaison est U_i. On parle ici de l'opération de croisement qui est définie par :

$$U_{jig} = \begin{cases} V_{jig} rand(0,1) \leq \mathrm{CR} \\ a_{jig} \end{cases} \tag{5-2}$$

Dans l'étape de sélection, une comparaison est faite entre le nouveau vecteur généré et le vecteur de la cible. Le vecteur ayant la meilleure aptitude est choisi pour remplacer le vecteur de la cible. Dans notre cas, la fonction objective est le taux de classification. Par conséquent, l'individu avec le taux élevé de classification a été conservé pour la prochaine étape. La sélection est définie comme suit :

$$a_{jig} = \begin{cases} U_{jig} f(U_{jig}) \square \ f(a_{jig}) \\ a_{jig} \end{cases} \tag{5-3}$$

Dans les équations ci-dessus, *le* U_{jig} est le vecteur généré par la combinaison du vecteur muté V_{jig} et le vecteur cible X_{jig}. Le CR est un coefficient de croisement qui est une probabilité utilisée pour contrôler la fraction des valeurs du vecteur muté. F est le facteur de la mutation et r_2, r_3 *et* r_1 sont des paramètres aléatoires de distribution entre 0 et 1

Algorithme DE

1. Initialisation:
   - Définir la taille de la population ;
   - Définir la dimension de l'espace de recherche ;
   - Définir le nombre de générations ;
   - Définir le facteur de la mutation $F \in [0.1, 0.9]$
   - Définir CR : le taux de croisement, $\in [0.1, 1]$
   - Générer la population aléatoire.
2. Calcul du vecteur muté par l'utilisation de l'équation de mutation ;
3. Si un élément du vecteur muté est au-dessous du minimum ou au-dessus du maximum, alors il est remplacé par le minimum ou le maximum respectivement ;
4. Générer le vecteur de remplacement par l'opération de croisement ;
5. Calculer de la fonction objective pour chaque vecteur de remplacement;
6. Utiliser de la fonction de sélection des individus après avoir évalué les performances ;
7. Répéter les étapes 2 à 6 jusqu'à ce que le nombre maximum de générations soit atteint ;
8. Obtenir la valeur optimale de la fonction objective.

Figure 5–1: L'algorithme de fonctionnement de DE

5.4. L'optimisation par essaims particulaires (OEP)

L'optimisation par essaims particulaires a été proposée par Kennedy et Eberhart en 1995 [146]. Cette méthode s'est inspirée du comportement social des animaux évoluant en essaim. En effet, on peut observer chez eux des dynamiques de déplacement relativement complexes, alors qu'individuellement chacun a une intelligence limitée et seulement une connaissance locale de sa situation dans l'essaim. Sa seule connaissance est au niveau de la

position et de la vitesse de ses plus proches voisins. Chacun utilise non seulement sa propre mémoire, mais aussi l'information locale de ses plus proches voisins pour décider de son propre déplacement. Des règles simples, telles que « aller à la même vitesse que les autres », « se déplacer dans la même direction » ou encore « rester proche de ses voisins » sont des exemples de comportements qui suffisent à maintenir la cohésion de l'essaim, et qui permettent la mise en œuvre de comportements collectifs complexes et adaptatifs. L'intelligence globale de l'essaim est donc la conséquence directe des interactions locales entre les différentes particules de l'essaim. La performance du système entier est supérieure à la somme des performances de ses parties. Kennedy et Eberhart se sont inspirés de ces comportements socio-psychologiques afin de créer la PSO.

L'optimisation par essaims particulaires est basée sur le principe que le travail d'un groupe d'individus améliore la performance globale du groupe et aussi celle de l'individu appartenant à ce groupe. Un essaim de particules, est une solution potentielle au problème d'optimisation, *survolent* l'espace de recherche en quête de l'optimum global. Le déplacement d'une particule est influencé par les trois composantes suivantes :

- **Une composante physique** : la particule tend à suivre sa direction courante de déplacement ;
- **Une composante cognitive** : la particule tend à se diriger vers le meilleur site par lequel elle est déjà passé ;
- **Une composante sociale** : la particule tend à se fier à l'expérience de ses congénères, et ainsi à se diriger vers le meilleur site déjà atteint par ses voisins.

Les individus peuvent être regroupés sous plusieurs formes ou topologies. Le type de topologie va décrire la nature de communication entre les individus de la population. On trouve quatre types de topologie : en maille, en étoile, en anneau et en carré. L'algorithme de fonctionnement de cette méthode d'optimisation est présenté dans la Figure 5-2.

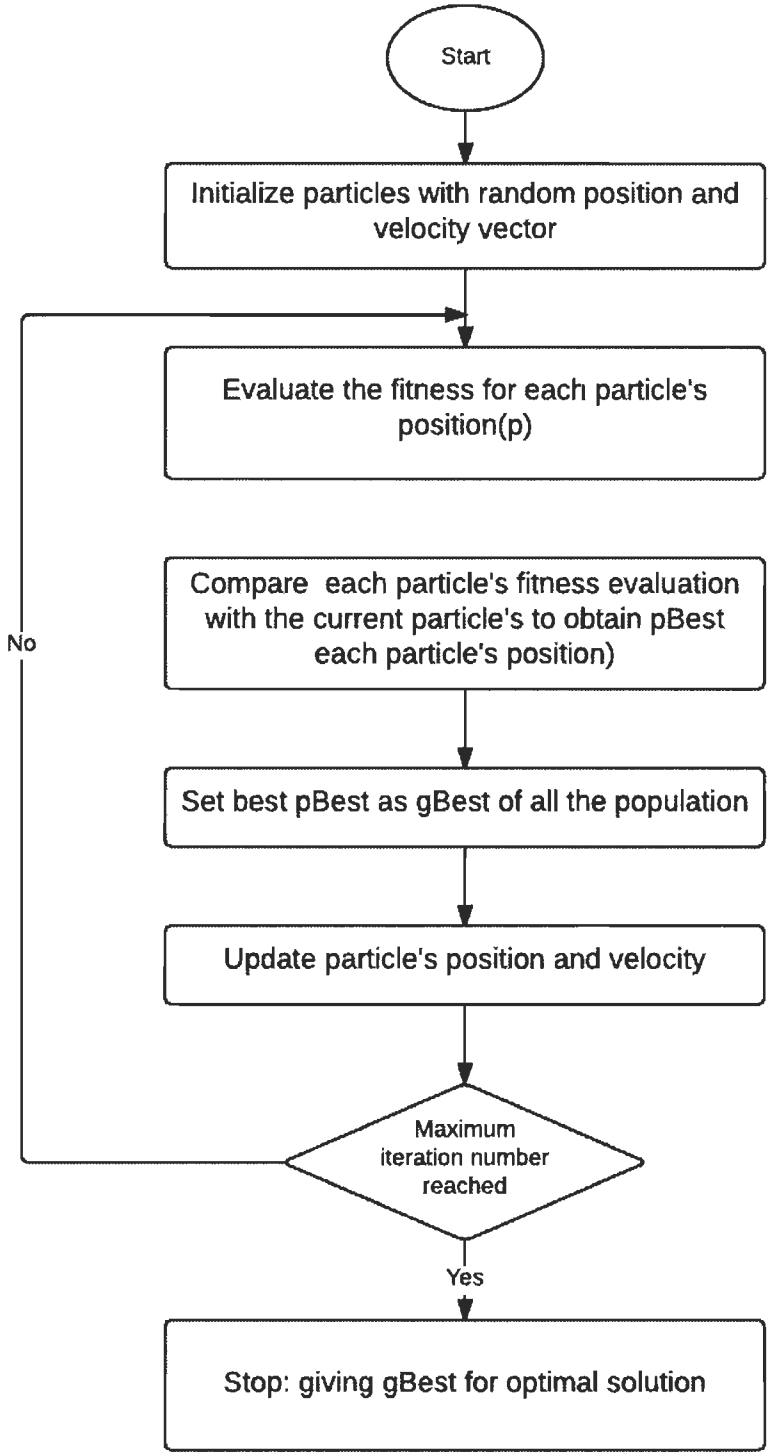

Figure 5–2: L'algorithme de fonctionnement de la méthode d'optimisation PSO

5.5. Apprentissage supervisé profond des traits paralinguistiques

Les classificateurs à architectures simples tels que : SVM, GMM, HMM et la régression logistique sont les plus utilisés. Si on se base sur un concept de couches, les systèmes précédents réalisent de bonnes performances grâce à leurs simples architectures composées d'une seule couche qui est responsable de faire correspondre les paramètres acoustiques d'entrées en un autre espace spécifique. Ce type d'architecture a réussi à résoudre plusieurs problèmes de classification. Par contre, cette architecture simple peut devenir une source de limitation surtout dans le contexte des applications réelles qui intègrent les signaux naturels tels que : les signaux de la parole, la langue naturelle et les signaux nerveux.

La reconnaissance des traits paralinguistiques de la parole dépend d'une architecture profonde afin de s'adapter à la complexité des signaux de la parole et d'exploiter au maximum leurs richesses en information. Également, les systèmes sensoriels humains sont connus par une architecture complexe et profonde. Plusieurs recherches démontrent que commencer l'extraction par un modèle d'apprentissage profond, qui prend en considération les aspects dynamique et temporel des traits paralinguistiques, peut améliorer le taux de classification. Nous pouvons citer quelques architectures implémentant des réseaux d'apprentissage profonds tels les DBN (Deep Belief Networks) et DBM qui sont composés de multiples machines de Boltzmann restreintes (RBM) organisées en forme de pile. Cette façon d'organisation conduit à une architecture à plusieurs couches.

Les techniques d'apprentissage profond ont un avantage important qui réside dans leurs fortes capacités à représenter les corrélations inséparables et à son pouvoir de mapper les paramètres en entrée de grande dimension à des supervecteurs stochastiques. Ces

supervecteurs sont plus discriminatoires que les vecteurs acoustiques bruts pour les systèmes de reconnaissance automatique des traits paralinguistiques.

5.5.1. Machines de Boltzmann restreintes (RBM)

La Machine de Boltzmann Restreinte est un type particulier de réseaux de Markov à deux niveaux. Le premier niveau est formé d'une couche d'unités visibles dont le rôle est de recevoir les données et l'autre est formé d'unités invisibles. Les unités invisibles sont généralement binaires stochastiques et les unités visibles sont des unités binaires ou gaussiennes stochastiques. La RBM représente la distribution conjointe entre un vecteur visible et une variable aléatoire invisible. Elle a uniquement des connexions entre les unités des deux couches à travers les unités de biais. Une RBM définit une fonction énergie pour toutes les configurations de vecteurs des unités visibles et invisibles, notées *v* et *h* respectivement. Pour les unités binaires, la fonction énergie *E (v, h)* est définie par :

$$E(v,h,\lambda) = -\sum_{i=1}^{v} a_i v_i - \sum_{j=1}^{H} b_j h_j - \sum_{i=1}^{v}\sum_{j=1}^{H} w_{ij} v_i h_j \qquad (5\text{-}4)$$

Où w_{ij} représente l'interaction symétrique entre les v_i et h_j, a_i *et* b_j sont les coefficients de biais et λ désigne les paramètres du modèle $a = \left[a_1,\ldots,a_V\right]^T$, $b = \left[b_1,\ldots,b_H\right]^T$ et $W = \left\{w_{ij}\right\} \in \square^{V\times H}$. La connexion entre la couche visible et la couche invisible est déterminée par la distribution de Boltzmann définie par :

$$p(v,h|\lambda) = \frac{1}{Z_\lambda}\exp\left\{-E(v,h;\lambda)/C_T\right\} \qquad (5\text{-}5)$$

C_T est défini comme le paramètre de température et Z_λ est la fonction de partition. Elle est définie par :

$$Z_\lambda = \sum_{\forall v} \sum_{\forall h} \exp\{-E(v,h;\lambda)\} \tag{5-6}$$

La probabilité de distribution PDF à partir des unités visibles peut être calculée par la fonction suivante :

$$p(v|\lambda) = \frac{1}{Z_\lambda} \sum_{\forall h} \exp\{-E(v,h;\lambda)\} \tag{5-7}$$

Pour un ensemble de données d'apprentissage, le modèle λ, de ces données est estimé par la technique de maximum de vraisemblance qui est définie par :

$$\frac{\partial \log \mathrm{P}(v|\lambda)}{\partial w_{ij}} = E_{PDonné}\left[v_i h_j\right] - E_{PModéle}\left[v_i h_j\right] \tag{5-8}$$

Dans cette équation $E_{PDonné}[.]$ définit la distribution des données et $E_{PModéle}[.]$ définit celle du modèle des données en respectant le modèle de distributions $\mathrm{P}(v|\lambda)$.

La RBM est employée pour la modélisation des données à valeurs réelles, discrètes et des vecteurs composés par les données binaires, réelles et discrètes. Cet emploi se fait par la définition de nombreuses fonctions d'énergie. Dans le contexte d'une distribution gaussienne cette fonction est définie par :

$$E(v,h;\lambda) = \sum_{i=1}^{V} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j=1}^{H} b_j h_j - \sum_{i=1}^{V} \sum_{j=1}^{H} w_{ij} h_j \frac{v_i}{\sigma_i} \tag{5-9}$$

5.5.2. Approche profonde optimisée à base de RBM

Dans notre approche, nous proposons une de recourir à une RBM qui combine en entrée différents paramètres acoustiques dans le but d'améliorer le taux de reconnaissance des traits paralinguistiques de la parole. Ces paramètres acoustiques comprennent les MFCCs, les paramètres relatifs à la qualité de la voix, les paramètres prosodiques et les paramètres auditifs. La configuration de ce système basée sur la RBM s'articule autour des points suivants :

- Le GMM utilise les paramètres auditifs afin de calculer les supervecteurs et pour avoir la même structure que les paramètres de rythme métrique à base de durée et intensité ;
- Un RBM ont été utilisées dans ce système ;
- Le RBM a été optimisée séparément par un algorithme évolutionnaire, dans ce cas-ci c'est PSO qui a été utilisée ;
- Les paramètres optimisés par PSO obtenus de premiers RBM sont passés pour le deuxième RBM ;

Notre système est une machine PSO-RBM (Deep restrict Boltzmann optimisé). Elle représente une autre approche présentée dans cette thèse pour la combinaison multi variable utilisant une fusion des paramètres acoustiques et une combinaison série qui mène à une architecture de reconnaissance à plusieurs niveaux. Dans ce système les classificateurs en série sont le GMM/RBM et le RBM respectivement.

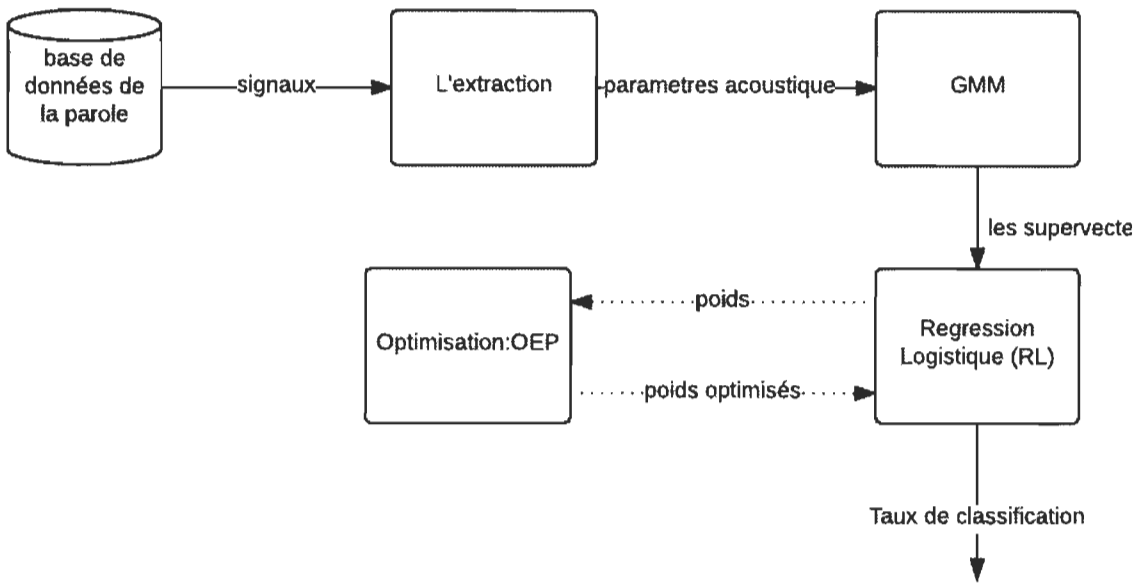

Figure 5–3: L'architecture du système GMM-RL-OEP

5.5.3. Approche complète de combinaison (GMM-RL-PSO)

Notre deuxième approche de combinaison représentée dans la Figure 5-3 est basée sur une de combinaison complète puisqu'elle expérimente une trame phonétique acoustique qui contient tous les paramètres acoustiques calculés pour la détection des traits paralinguistiques. Ces paramètres ont été fusionnés puis ont été dirigés vers le GMM qui a calculé les supervecteurs. On a eu recours au GMM, car on a utilisé des descripteurs non linéaires et parce que les supervecteurs étaient reconnus par leur capacité élevée de détection des traits paralinguistiques comparés à la capacité des vecteurs de bas niveau. En se basant sur ces supervecteurs, la RL a donné une représentation linéaire en plus de sélectionner celles qui étaient pertinentes pour la classification. Avec un algorithme évolutionnaire PSO, on a évalué les poids attribués au supervecteur sélectionné par la Régression Logistique, afin d'en garder les plus pertinents tout en améliorant le taux de classification.

5.6. Résultats

On a utilisé plusieurs paramètres acoustiques calculés à partir de différentes représentations du signal de la parole. À partir de ces paramètres plusieurs trames acoustico-phonétiques ont été formées dans le but d'améliorer la détection des traits paralinguistiques.

Pour l'identification et la classification des accents natifs et non natifs, on a testé l'efficacité d'une trame composée de rythmes métriques à base d'intensité et de durée. Cette trame a donné un taux de classification de 87.6 % lorsqu'un SVM a été utilisé et 59 % lorsqu'on a utilisé le classificateur GMM avec nombre gaussienne égal à 16.

Dans le cas de l'identification des états émotionnels, une trame composée de paramètres acoustiques auditifs, de paramètres prosodiques, de MFCCs et de paramètres de la qualité de la voix a été utilisée pour la discrimination entre cinq classes d'émotions. Cette trame a permis un taux de classification de 85.7 % avec le classificateur LDA et 62.4 % avec la SVM un à un.

5.6.1. Combinaison en série

Les combinaisons en série des classificateurs GMM et SVM, ainsi que GMM et LDA ont été testées, car les paramètres acoustiques utilisés n'avaient pas tous une structure linéaire comme le MFCC et le modèle d'oreille. Notre motivation était aussi pour profiter de la capacité des supervecteurs. L'application de cette combinaison pour la détection des émotions où GMM et LDA a donné un taux de classification de 85.7 %. L'application de la combinaison GMM/SVM pour la détection des accents a donné un taux de classification de

84.8 % lorsque le modèle d'oreille a été utilisé comme paramètre d'entrée pour cette combinaison.

L'approche multivariable dont la trame est composée de paramètres acoustiques auditifs, de paramètres de la qualité de la voix et de MFCC et d'une combinaison en série de GMM et SVM. La trame était l'entrée pour le GMM pour calculer les supervecteurs. Les SVM se sont basés sur ces supervecteurs et le taux de classification de cette combinaison a été de 92.3 % pour la détection des états émotionnels.

5.6.2. Combinaison parallèle

La combinaison parallèle a été effectuée par le classificateur SVM afin d'améliorer le taux de classification. À titre d'exemple, on a eu un taux de 81.30 % lorsqu'elle s'était basée sur les rythmes métriques à base d'intensité et ce taux a atteint 89.7 % en se basant sur les rythmes métriques à base de durée, ainsi que 94.8 % lorsque les paramètres auditifs ont été utilisés pour l'identification et la classification d'accents natifs et non natifs.

Afin d'améliorer notre système tout en se basant sur ces trois types de paramètres avec la fusion des décisions, nous avons utilisé l'équation suivante :

$$T = a.SVM(\text{rythme}_D) + b.SVM(\text{rythme}_i) + c.SVM(auditif) + d \qquad (5\text{-}10)$$

Pour trouver les coefficients de fusion, un algorithme évolutionnaire a été utilisé. Les valeurs des coefficients de fusion *a, b, c,* et *d* sont respectivement les poids attribués au classificateur SVM basé sur les rythme métrique de nature durée, SVM basé sur les rythme métrique de nature intensité, SVM basé sur les paramètres auditifs et la constante *d*. Le taux de classification par cette méthode de combinaison a atteint 96% de taux correct de reconnaissance pour l'accent natif et non natif de la langue arabe.

5.6.3. *Combinaison des différentes méthodes de sélection*

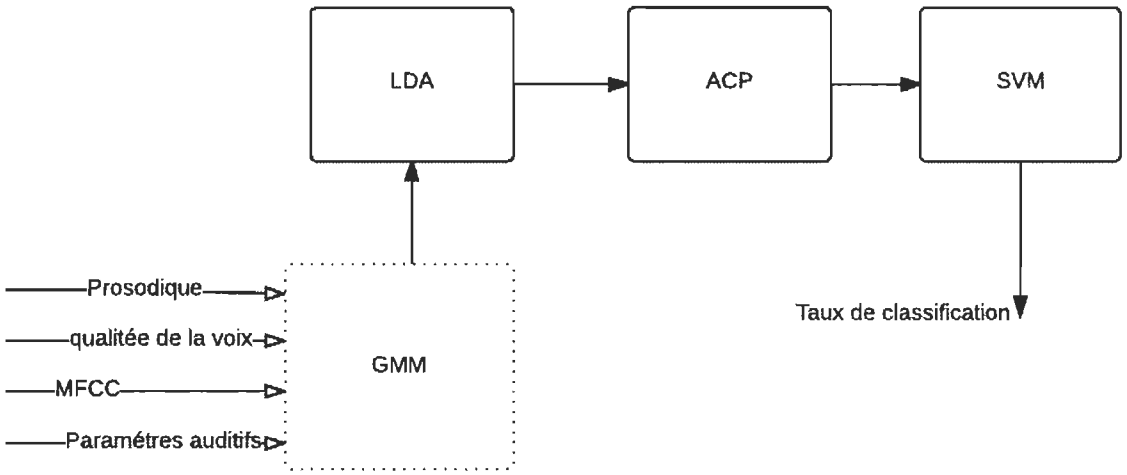

Figure 5–4: Architecture du système GMM-LDA-ACP-SVM

Le principe de cette approche était de bénéficier des atouts de ces différentes méthodes de sélection qui se complétaient ensemble comme c'est expliqué dans le chapitre 3.

L'application de cette approche pour la détection des états émotionnels a conduit à un taux de classification de 94 %, lorsqu'on a utilisé GMM-LDA-ACP-SVM comme système de combinaison (voir figure 4-5). Ce système a utilisé comme entrée une trame acoustique composée de paramètres prosodiques, de paramètres de la qualité de la voix, de MFCCs et de paramètres acoustiques auditifs.

5.6.4. *Approche complète de combinaison (GMM-RL-PSO)*

L'application de cette approche a donné des résultats encourageants pour la détection des accents natifs et non natifs avec un taux de classification de 96.8 %.

Dans cette approche l'algorithme évolutionnaire DE a été expérimenté pour l'optimisation de sélection des paramètres acoustiques afin d'améliorer la classification

effectuée par un classificateur LDA-RL. Le fonctionnement de cette technique se déroule comme suit :

Étape 1. L'initialisation des poids pour chaque composante de la trame a procédé à une combinaison sans sélection avec une la LDA ou RL pour donner seulement une combinaison linéaire des différentes composantes de la trame en leur affectant des poids, et sans faire la sélection des pertinentes parmi celles-ci ;

Étape 2. Évaluation de ces poids à chaque génération de DE ;

Étape 3. Utiliser la fonction objective discriminante de LDA ou bien celle de RL ;

Étape 4. Vérifier la condition d'arrêt.

L'application de cette approche pour la détection des émotions a conduit à l'obtention d'un taux de classification de 91.6 % lorsque cinq émotions (Bonheur, Colère, Panique, Tristesse et Intérêt) ont été étudiées et que LDA a été utilisé comme classificateur linéaire discriminant.

Lors de l'étude d'identification et de la classification des accents natifs et non natifs, l'application de cette approche avec le classificateur linéaire RL se basant sur les rythmes métriques à base d'intensité et à base de durée, qui étaient concaténés dans une seule trame multivariable, a obtenu un taux de classification de 96.7 % qui a été atteint avec seulement huit paramètres considérés comme pertinents.

5.6.5. Optimisation multiobjectif.

Nous avons également expérimenté une approche avec optimisation multiobjectif. Dans nos applications, nous avons identifié deux sous-objectifs :

1. Choisir les paramètres les plus pertinents, c'est-à-dire éliminer ceux qui ne le sont pas donc à minimiser le nombre de paramètres.
2. Parmi les paramètres de (1) on détermine ceux qui minimisent le taux d'erreur de classification.

L'application de cette approche pour la méthode de sélection LDA, qui est en même temps un classificateur discriminant linéaire, a permis de réduire au maximum le nombre de paramètres tout en augmentant le taux de classification de ce classificateur. Concernant la détection des émotions, cette approche a réduit le nombre de paramètres de 86 à 21 et elle a amélioré le taux de classification pour atteindre 92 %.

5.6.6. L'optimisation de RBM

Le test de cette approche pour la détection des émotions de la parole mène à un taux de classification de 77% pour six émotions (le bonheur, la colère chaude, l'intérêt, la neutralité, la panique et la tristesse). Les paramètres optimisés de RBM sont : Taux d'apprentissage, élan pour éviter le surapprentissage (momentum), maxepoch, nombre d'epochs avant le maxpoch pour la rapide de convergence et pénalité.

5.7. Conclusion

Dans ce chapitre, nous avons examiné la relation de complémentarité entre des paramètres acoustico-phonétiques et auditifs ce qui nous a conduits à la conception d'une approche multivariable qui fusionne tous ces paramètres.

Nous avons proposé de nouvelles approches pour cette combinaison qui étaient plus efficaces pour la reconnaissance des traits paralinguistiques de la parole, et plus particulièrement pour l'émotion et l'accent natif et non natif. Ces approches ont été expérimentées sur les corpus LDC West Point, et les émotions. La première approche complète était basée sur les supervecteurs de GMM, la régression logistique linéaire et les algorithmes évolutionnaires d'optimisation. Alors que, la deuxième était basée sur une architecture profonde de classification des supervecteurs et des algorithmes évolutionnaires. Les performances obtenues avec ces deux approches ont dépassé les résultats cités dans la littérature par d'autres approches de combinaison.

La complémentarité des paramètres extraits du signal a été exploitée différemment par la proposition de deux nouvelles approches de sélection. La première approche était basée sur le principe de la combinaison des différentes méthodes de sélection afin de réduire le nombre de paramètres et aussi de garder le plus d'informations utiles pour la phase de classification. La deuxième approche était de modifier le problème de sélection d'un problème à un seul objectif à optimiser, vers un problème d'optimisation multiobjectif.

Les améliorations apportées par ces différentes approches au domaine de la reconnaissance automatique des traits paralinguistiques de la parole sont présentées en détails dans le chapitre six où une comparaison sera établie entre nos résultats obtenus par ces approches et celles citées dans la littérature.

Chapitre 6 - Résultats globaux et discussions

6.1. Introduction

Ce chapitre contient les discussions des résultats obtenus par les nouveaux paramètres et les nouvelles approches proposées pour la reconnaissance automatique des traits paralinguistique de la parole. Pour ce faire, la description de l'environnement de travail utilisé pour obtenir ces résultats est donnée en première partie. Le chapitre inclut également une présentation des corpus et des paramètres acoustiques expérimentés aux cours de cette thèse. Une analyse comparative des résultats obtenus et ceux présentés dans la littérature de y est également effectuée.

6.2. Corpus utilisés et prétraitement du signal vocal

6.2.1. Emotional Prosody Speech and Transcript

Parmi les corpus de la parole émotionnelle, on en trouve plusieurs de langue anglaise tel que celui que nous avons utilisé, à savoir Emotional Prosody Speech and Transcript du Linguistic Data Consortium (EPLDC). Il faut noter que la version annotée de ce corpus n'est pas disponible. Ce corpus contient des énoncés couvrant les 15 émotions suivantes : l'ennui, l'anxiété, le mépris, le dégoût, l'exaltation, le bonheur, la colère chaude, la colère froide, le désespoir, la fierté, l'intérêt, la neutralité, la panique, la tristesse et la honte. Les locuteurs ont été conseillés d'éviter les expressions exagérées, car elles ont des effets sur la

production de ce type de variations contextuelles. Un groupe de sept locuteurs acteurs (3 hommes et 4 femmes) a été choisi et chacun des locuteurs devait prononcer les différents états émotionnels en les répétant 7 fois afin que la simulation de l'état émotionnel soit jugée satisfaisante. Les enregistrements sont sur deux canaux et échantillonnés à 22.05 KHz. Les deux microphones utilisés sont de type Shure SN94 monté sur un pied et un casque Seinnheiser HMD4I0. À chaque fichier audio nous avons associé un fichier de transcription incluant un alignement temporel des énoncés. Dans notre travail, nous avons utilisé 840 phrases dont 525 étaient pour l'apprentissage des classificateurs et 325 pour le test [147]. Un résumé de description de ce corpus est donné dans le Tableau 6-1.

6.2.2. Corpus des accents : Linguistic Data Consortium (LDC) West Point Arabic

Le corpus de Consortium linguistique de données (LDC) West Point Arabic a été utilisé dans notre travail. Il a été conçu pour l'apprentissage des modèles acoustiques pour la reconnaissance de la parole qui a été utilisé pour aider à enseigner l'arabe aux cadets dans les bases militaires américaines.

Le corpus West Point LDC consiste en 8516 fichiers de parole et totalise 1.7 Gigabytes. Chaque fichier a été prononcé par un seul locuteur et lu à partir d'un script parmi les scripts de cette base de données. Ceux-ci ont été enregistrés avec une fréquence d'échantillonnage de 22.05 kHz. Ensuite, ils ont été convertis sous format NIST SPHERE. Les 8516 fichiers sont composés de 7200 pour les locuteurs natifs et les autres pour les locuteurs non natifs.

Tableau 6-1: Description du corpus émotionnel Emotional Prosody Speech and Transcript avec une répartition des données en fonction du locuteur et de la classe d'émotion

Initiales locuteurs →	CC	CL	GG	JG	MF	MK	MM	Total
Anxiété	10	12	10	10	11	11	21	**85**
Ennui	11	10	11	11	19	10	12	**84**
Mépris	11	11	10	11	11	11	10	**75**
Colère froide	11	10	11	11	21	10	11	**85**
Désespoir	11	11	10	11	13	11	11	**78**
Dégoût	11	10	22	13	6	10	11	**83**
Exaltation	10	11	10	10	11	11	10	**73**
Bonheur	11	10	11	11	10	20	11	**84**
Colère	10	11	10	10	11	11	10	**73**
Intérêt	11	10	11	11	10	20	11	**84**
Neutralité	8	7	7	7	9	7	7	**52**
Panique	11	10	10	11	10	10	20	**82**
Fierté	13	10	11	11	10	10	11	**76**
Tristesse	11	10	11	11	12	10	10	**75**
Honte	9	11	10	9	11	11	11	**72**
Total	**159**	**154**	**165**	**158**	**175**	**173**	**177**	**1161**

Le corpus de West Point Arabic contient une collection de quatre principaux scripts arabes. Le premier contient 155 phrases lues par 75 natifs arabophones. Le second est composé de 40 phrases prononcées par 23 locuteurs non natifs. Le troisième contient 41 phrases parlées par des locuteurs non natifs. Enfin, le quatrième script de 22 phrases lues par des locuteurs non natifs. Au total, il y a eu 1131 mots arabes distincts et 110 locuteurs ont été analysés : 66 hommes (41 natifs et 25 non natifs) et 44 femmes (34 natives et 10 non natives). Chaque phrase du corpus LDC West Point Arabic a été étiquetée selon ses intervalles vocaliques et consonantiques. Le tableau 6-2 présente une description globale pour ce corpus. Plus de détails sur ce dernier, se retrouve dans l'article de Alotaibi et Selouani [148].

6.3. Paramètres acoustiques expérimentés

Le Tableau 6-2 donne une brève description des paramètres acoustiques utilisés dans nos expériences : les MFCCs, les paramètres auditifs, des métriques rythmiques les plus standards et les plus utilisés dans la littérature. Ces métriques rythmiques sont calculées sur la base de durée et d'intensité.

Tableau 6-2 Description du corpus West Point Arabic

	Mâle	**Femelle**	**Total**
Nombre de locuteurs			
Natifs	41	34	75
Non natifs	25	10	35
Total	66	44	110
Heures des données			
Natifs	6	4.4	10.4
Non natifs	0.74	0.28	1.02
Total	6.74	4.68	11.42
Données en Mégabytes			
Natifs	913	663	1576
Non natifs	111	42.4	153.4
Total	1024	705.4	1729.4
Nombre de fichiers audio			
Natifs	4107	3163	7270
Non natifs	883	363	1246
Total	4990	3526	8516

Tableau 6-3: Les paramètres acoustiques de la parole expérimentés

Paramètres	**Description**
MFCC	Les coefficients spectraux calculés par la méthode de transformée Fourrier.
Jitter	Représente la variation de la fréquence fondamentale dans l'évolution temporelle de l'énoncé.
Shimmer	Indique la perturbation ou la variabilité de l'amplitude sonore.
Intensité	La variation de l'amplitude du signal de la parole.
Pitch	La fréquence des vibrations des cordes vocales.
Durée	L'intervalle de temps nécessaire pour émettre un signal.
HNR	Représentant le degré de périodicité acoustique.
DVB	Le degré de voix

DUV	Le degré de voyelles
Delta-C	L'écart-type des durées d'intervalles des consonnes par phrase
Delta-V	L'écart-type des durées d'intervalles vocaliques par phrase
%V	La moyenne des durées d'intervalles vocaliques,
Varco-v	La normalisation de l'écart-type des durées d'intervalles vocaliques par phrase
Varco-C	La normalisation de l'écart-type des durées d'intervalles des consonnes par phrase
Varco-VC	La normalisation de l'écart-type des mesures successives dans la phrase
Rpvi	La différence moyenne des durées entre deux mesures successives dans la phrase
rPVI-V	La différence moyenne des durées entre deux intervalles vocaliques successifs dans la phrase
rPVI-C	La différence moyenne des durées entre deux intervalles de deux consonnes successives dans la phrase
Npvi	La normalisation de la différence moyenne des durées entre deux mesures successives dans la phrase
nPVI-C	La normalisation de la différence moyenne des durées entre deux intervalles des consonnes successives dans la phrase
nPVI-V	La normalisation de différence moyenne des durées entre deux intervalles vocaliques successifs dans la phrase
OPVI-V	C'est la version vocalique de nouveau métrique OPVI
OPVI-C	La version consonantique de nouveau métrique OPVI
Delta-C_i	L'écart-type des intensités d'intervalles des consonnes par phrase
Delta-V_i	L'écart-type des intensités d'intervalles vocaliques par phrase
%V_i	La moyenne des intensités d'intervalles vocaliques
Varco-V_i	La normalisation de l'écart-type des intensités d'intervalles vocaliques par phrase
Varco-C_i	La normalisation de l'écart-type des intensités d'intervalles des consonnes par phrase
Varco-VC_i	La normalisation d'écart-type des mesures successives dans la phrase
rPVI$_i$	La différence moyenne des intensités entre deux mesures successives dans la phrase
rPVI-V_i	La différence moyenne des intensités entre deux intervalles vocaliques successifs dans la phrase
rPVI-C_i	La différence moyenne des intensités entre deux intervalles de deux consonnes successives dans la phrase

nPVI$_i$	La normalisation de la différence moyenne des intensités entre deux mesures successives dans la phrase
nPVI-C$_i$	La normalisation de la différence moyenne des intensités entre deux intervalles des consonnes successives dans la phrase
nPVI-V$_i$	La normalisation de différence moyenne des intensités entre deux intervalles vocaliques successifs dans la phrase
G/A : Grave/Acute	C'est ce qui mesure la différence d'énergie entre les basses fréquences (50 à 400 Hz) et les hautes fréquences (3800-6000 Hz)
O/C : Open/Closed	Un événement est considéré fermé si l'énergie des basses fréquences (230-350Hz) est supérieure à l'énergie des moyennes fréquences (600-800Hz)
D/C : Diffuse/Compact	C'est un paramètre de compacité qui reflète l'importance de la région centrale formante (800-1050 Hz) par rapport aux régions des fréquences de la plage (300-700 Hz) et (1450-2550 Hz),
F/S : Flat/Sharp	C'est un phonème qui est considéré comme fort si l'énergie est entre (2200-3300 Hz) et important si l'énergie est entre (1900-2900 Hz)
M/S : Mellow/Strident	Les phonèmes stridents sont caractérisés par la présence de bruit, à cause d'une turbulence, à leur point d'articulation qui mène à plus d'énergie entre (3800-5300 Hz) qu'entre (1900-2900 Hz)
C/D : Continuant/Discontinuant	Il quantifie la variation de l'amplitude du spectre en comparant l'énergie des cadres actuels et précédents. Ce paramètre est bas dans le cas d'une variation faible et élevé dans le cas contraire.
T/L : Tense/Lax	Ce paramètre est défini par la mesure de la différence d'énergie entre les fréquences moyennes (900-2000 Hz) et les fréquences relativement élevés (2650-5000 Hz).

6.4. Évaluation de la métrique OPVI et les autres métriques de rythme

Dans cette expérience 15 métriques de rythme ont été calculées sur un ensemble de 160 phrases de notre corpus MSA. Le calcul a été fait pour les deux versions de chaque rythme, vocalique et consonantique. Le Tableau 6-4 fournit des statistiques descriptives pour chaque métrique de rythme et les résultats d'une analyse de la variance préliminaire qui évalue la pertinence de chaque métrique dans la distinction entre les locuteurs natifs et non natifs.

Tableau 6-4: La moyenne (les écarts-types) des métriques de rythme. La signification statistique (valeur-*p*) de l'ANOVA, des locuteurs L1 versus les L2. Les valeurs-p significatives en caractères (gras) ont atteint une valeur de valeur-p <0.05

Rythme métrique	**L1**	**L2**	***p*-valeur**
%V	41.01 (7.02)	35.58 (5.40)	**0.001**
DeltaV	0.061 (0.02)	0.053 (0.02)	0.090
DeltaC	0.060 (0.01)	0.055 (0.02)	0.067
VarcoV	58.43 (10.80)	63.30 (15.59)	0.07
VarcoC	51.094 (7.85)	48.06 (10.90)	0.095
nPVI-V	70.11 (17.13)	73.56 (16.65)	0.206
nPVI-C	61.97 (14.19)	56.91 (17.8)	**0.033**
rPVI-V	0.0828 (0.029)	0.073 (0.026)	0.4137
rPVI-C	0.078 (0.022)	0.07 (0.026)	**0.040**
CCI-V	6.42 (2.65)	6.80 (2.46)	0.730
CCI-C	4.35 (1.41)	4.38 (1.31)	0.354
nCCI-V	47.60 (24.56)	66.78 (21.10)	**0.001**
nCCI-C	48.27 (9.937)	46.53 (9.935)	0.052
O-PVI-V	3.18 (1.02)	3.44 (0.87)	0.507
O-PVI-C	17.64 (4.34)	16.26 (4.93)	**0.003**

L'ensemble de phrases choisies du corpus a été divisé en deux sous-ensembles : un sous-ensemble pour l'apprentissage contenant 70 % des phrases et un sous-ensemble de test contenant 30 % des phrases. Le sous-ensemble d'apprentissage inclut les sept premières phrases prononcées /lues par les locuteurs (un total de 56 phrases par des locuteurs natifs arabes et 56 phrases par des locuteurs non natifs) ; l'ensemble du test contient les trois dernières phrases prononcées par chaque locuteur (un total de 24 phrases prononcées par les locuteurs natifs et 24 phrases par locuteurs non natifs). La division en deux groupes

d'apprentissage et de test a été faite dans le but de réaliser la classification des accents natifs et non natifs par la machine à Vecteur Support (SVMs).

Nous avons exécuté deux séries d'expériences de classification. Dans la première série, nous avons classé les deux groupes de locuteurs à l'aide des métriques consonantiques et vocaliques. Dans la deuxième série d'expériences, nous avons procédé à la classification en utilisant les combinaisons de paramètres suivantes : toutes les mesures PVI (pour les intervalles vocaliques et consonantiques), toutes les mesures à base d'intervalle, une combinaison des deux mesures à base PVI et à base intervalle. Chaque combinaison a servi comme point de référence à l'évaluation de la métrique O-PVI que nous avons proposée.

Les résultats de la première série d'expériences sont donnés dans le Tableau 6-5 qui énumère les taux de classification pour chacune des métriques. Parmi les métriques vocaliques, VarcoV a obtenu les meilleures performances. Les métriques %V et l'O-PVI-V ont eu des résultats très proches qui étaient entre 70 % à 72 % de taux correct. Parmi les mesures consonantiques, l'O- PVI-C a dominé les autres métriques. En général, les métriques consonantiques ont été plus performantes et la nouvelle métrique O-PVI-C a obtenu le meilleur taux de classification (89,8 %).

Tableau 6-5 : Taux de classification de SVM des rythmes métriques à base de durée

Rythme métrique	Vocalique	Consonantique
Delta	51.0	64.6
rPVI	51.0	70.8
CCI	54.1	62.5
nPVI	59.0	77.1
nCCI	65.0	58.3
%V	70.8	--
Varco	71.8	75.0
O-PVI	70.3	89.8

Les résultats de la deuxième série d'expériences sont présentés dans le Tableau 6-5. Notons que l'ajout de la métrique O-PVI aux deux familles des métriques classiques, PVI et à base d'intervalle, a conduit à améliorer la performance de classification. L'amélioration a été remarquable dans le cas des métriques de la famille des PVI où une augmentation de 6.25 % a été notée. Cette amélioration est de l'ordre de 1.9% dans le cas où les PVIs et les intervalles étaient combinés.

Tableau 6-6: Comparaison des performances des différentes combinaisons des métriques de rythme avec et sans OPVI pour la classification d'accent natif et non natif de la langue arabe.

Les métriques	**Taux de classification (%)**
(1) PVI (CCI-V, CCI-C, nCCI-V, nCCI-C, nPVI-V, nPVI-C, rPVI-V, rPVI-C)	83.33
(1) + O-PVI-V, O-PVI-C	89.58
(2) métriques à base d'intervalle (%V, deltaV, deltaC, VarcoV, VarcoC)	87.50
(2) + O-PVI-V, O-PVI-C	89.58
(3) = (1) + (2) = PVI et métriques à base d'intervalle	85.42
(3) + O-PVI-V, O-PVI-C	87.34

Les résultats des deux séries d'expériences montrent une grande performance atteinte par OPVI lors de la discrimination de l'accent natif et non natif.

Deux tendances générales se dégagent des résultats de classification de la parole arabe présentées au Tableau 6-6 Tout d'abord, les métriques normalisées ont tendance à obtenir de meilleures performances de classification que les métriques non normalisées. Des

comparaisons entre des paires de paramètres appuient cette observation. Par exemple, comparez le VarcoV qui est une métrique normalisée avec DeltaV qui est une métrique non-normalisé (71.8 % contre 51.0 %) ; le même résultat est constaté pour la métrique normalisée nPVI-C et la métrique non normalisé rPVI-C (77.1 % contre 70.8 %). Cette tendance, où les métriques normalisées effectuent une meilleure classification, est conforme à la tendance générale mentionnée dans d'autres comparaisons de métriques de rythmes citées dans la littérature tels que celles de Liss et al., Loukina et al. et Wiget et al respectivement dans [115], [7] et [114].

La deuxième tendance observée est que les métriques consonantiques ont fourni une meilleure discrimination dans le cas de la discrimination L1/L2 que les métriques vocaliques. Ceci a été constaté dans presque toutes les paires de métriques comme par exemple : DeltaC et DeltaV (64.6 % contre 51.0 %), VarcoC et VarcoV (85.0 % vs.71.8%) et l'O-PVI-C et l'O-PVI-V (89.8 % contre 70.3 %). Cependant, ce résultat était inattendu à la lumière des études telles que celle de Wiget et al. qui ont affirmé que toutes les rythmes métriques basées sur les intervalles de consonnes n'étaient pas significatifs au niveau de la discrimination. En revanche, Mok et Dellwo ont démontré que le VarcoC et %V, donnaientt la meilleure classification des accents cantonais et mandarin vs. accents anglais [123]. Cela met les chercheurs devant une situation où ils doivent examiner une plus grande diversité de langues afin d'établir un modèle robuste et sûr.

Tableau 6-7 :Les coefficients de l'OPVI

	α	B	E	θ
nPVI	100	1	0	0
CCI	100	0	1	1
rPVI	1	0	0	0
OPVI-C	41.856	0.824	0.008	0
OPVI-V	11.207	0.644	0.507	0.105

Nos résultats laissent supposer que les bonnes performances de classification de la métrique OPVI-C est en relation directe avec la nature de la normalisation. Le nombre de segments dans les intervalles consonantiques ne semble pas jouer un rôle dans cette optimisation, ce qui est représenté par l'association d'une valeur de 0 ou près de zéro aux paramètres ε et θ de l'OPVI-C tel que montré dans le Tableau 6.7 Toutefois, la valeur du paramètre β (0,824), qui est associé à l'expression $\left(\frac{d_K + d_{K+1}}{2}\right)$ suggère que la pondération accordée à la durée moyenne des intervalles contrastives est importante dans le contexte de classification. Généralement, les chercheurs utilisaient la métrique normalisée nPVI, pour mesurer le rythme basé sur des intervalles vocaliques et la métrique non normalisée rPVI pour les intervalles consonantiques. Les résultats expérimentaux suggèrent que ni nPVI-C (où β = 1) ni rPVI-C (où β = 0) peut parfaitement faire une distinction franche entre les variétés de langage. Au contraire, une normalisation pondérée semble offrir une avenue de recherche plus utile sur les différences entre les variétés de langues parlées. Cela nous a conduits à constater que non seulement les modèles de durée consonantique, mais aussi les rythmes métriques relatifs à des intervalles de consonnes adjacentes fournissent un haut degré de défi pour les apprenants anglais de langue arabe.

L'interprétation générale des quatre paramètres d'optimisation (α, β, ε et θ) nécessitera la poursuite de l'investigation. Une caractéristique de ces paramètres est leur capacité d'adaptation à la classification des différents types d'accents. Les paramètres peuvent être considérés non seulement comme les valeurs optimales pour la classification, mais aussi ils peuvent être comme des coefficients associés à une variété particulière de langage (accent) ou un groupe de variétés ; autrement dit, peut-être qu'elles seront interprétées comme des variétés particulières. Une autre caractéristique de l'approche de normalisation-avec-optimisation que sous-entend O-PVI est l'idée que cette métrique est capable de normaliser considérablement au moins la partie de la variabilité de l'interlocuteur que les chercheurs Loukina et Arvaniti ont souligné pour les rythmes métriques dans [8] et [7].

6.5. Évaluation des paramètres auditifs et les métriques rythmiques à base d'intensité

Les résultats obtenus prouvent que l'exploitation des métriques de rythme seules n'ont pas la capacité de performer efficacement pour la classification dans un environnement variable tel que dans le cas des accents natifs et non natifs. Les métriques classiques à base d'intensité et de durée ont énormément de variance et ont un niveau de succès modéré pour la classification des locuteurs L1 et L2 dans le Corpus de la parole arabe, West Point Arabic. Généralement, les taux d'erreurs se situent entre un taux élevé de 46.79 % (rythme à base d'intensité avec le classificateur de GMM) et un taux d'erreur raisonnable de 10.33 % (rythme à base de durée avec le classificateur SVM). Cet intervalle, assez large du taux d'erreur, reflète les difficultés qui ont été rapportées dans la littérature concernant l'utilisation du rythme pour le classement des langues. Nous avons observé que l'ajout des

métriques à base d'intensité aux métriques à base de durée a contribué positivement à la classification faite par le système GMM/SVM.

Le système basé sur les paramètres auditifs a été notablement meilleur en comparaison avec les trois systèmes étudiés (SVM-rythmes métriques à base de durée, SVM-rythmes métriques à base d'intensité et SVM-combinaison de deux rythmes métriques). Effectivement, il a atteint un faible taux d'erreur de 5.21 % avec le classificateur SVM. Ce résultat appuie l'idée que l'utilisation de fonctionnalités qui modélisent des mécanismes perceptifs peut renforcer la modélisation de la structure rythmique de la parole.

La combinaison des métriques du rythme et des paramètres auditifs peut clairement améliorer la classification. Cette combinaison avec l'application d'un classificateur GMM/LR, qui prend avantage des différentes sources d'informations offertes et des supervecteurs calculés, a atteint un taux d'erreur de 3.70 %.

Notre approche de combinaison complète a ainsi atteint le meilleur classement dans les expériences réalisées, avec un taux d'erreur de seulement 3.01 %. Ce résultat suggère également que le système GMM/LR optimisé par un algorithme évolutionnaire présente une alternative puissante aux classificateurs reconnus tels que GMM, SVMs et l'hybride GMM/SVM.

Tableau 6-8: Comparaison des performances des 5 systèmes de classification testés pour la classification des accents arabe natifs et non natifs

Systèmes	**Paramètres**	**Taux d'erreur %**
SVM (Paramètres auditifs acoustiques)	$\sigma = 28.1$	05.21
SVM (Durée des rythmes métriques)	$\sigma = 2.1$	10.33
SVM (Intensité des rythmes métriques)	$\sigma = 10.3$	18.67
SVM (Durée et intensité des rythmes métriques)	$\sigma = 6.2$	12.42
GMM (Paramètres auditifs acoustiques)	8 Gaussiennes	07.56
GMM (Durée métriques)	16 Gaussiennes	35.25
GMM (Intensité métriques)	16 Gaussiennes	46.79
GMM (Durée et intensité métriques)	16 Gaussiennes	41.27
GMM/SVM (Paramètres auditifs acoustiques)	4 Gaussiennes, σ =13.52	15.92
GMM/SVM (Durée métriques)	16 Gaussiennes, $\sigma = 0.3$	27.58
GMM/SVM (Intensité métriques)	4 Gaussiennes, $\sigma = 7$	37.50
GMM/SVM (Durée et intensité métriques)	4 Gaussiennes, $\sigma = 4.84$	26.44
GMM/RL (métriques à base durée et intensité, et paramètres auditifs)	12 meilleurs paramètres	03.70
GMM/RL/PSO (métriques à base durée et intensité, et les paramètres auditifs)	8 paramètres	03.2

Enfin, les résultats présentés au tableau Tableau 6-8 montrent la variation de rythme des variétés natives et non natives de langue arabe. Les différences rythmiques ont été liées

principalement aux durées avec un faible impact touchant l'intensité des voyelles. En termes de durées vocaliques, les locuteurs L2 ont des voyelles plus courtes (petit % V) et ces voyelles sont moins variables (petits V) que les locuteurs L1. En fait, le petit % V (34.53 %) des locuteurs L2, quand ils parlaient arabe est considérablement plus petit que les valeurs moyennes de % V trouvées dans l'anglais américain, qui étaient de l'ordre de 45 à 54 %. Ces résultats suggèrent que les locuteurs de L2, qui étaient des locuteurs natifs de l'anglais américain, ont eu de la difficulté avec les voyelles courtes et longues comparativement aux locuteurs natifs arabe. En plus, l'intensité des voyelles chez les locuteurs L2 est inférieure (petit % V_i) à celles des locuteurs de L1. Il est important de signaler qu'une différence significative a été atteinte par le niveau de segmentation, comme c'est démontré par les métriques à base de durées qui basent leurs calculs sur les syllabes. Les locuteurs L2 ont une plus grande variabilité concernant ces métriques (VarcoVC, nPVI−VC et rPVI−VC) que les locuteurs L1. Toutefois, il semble que le rôle de la durée consonantique à la variation de rythme, plus précisément à l'intérieur de l'unité VC, nécessite une étude plus approfondie. En effet, en regardant le Tableau 6-9 on trouve que cette métrique n'est pas significative pour la discrimination entre les accents natifs et non natifs, d'après les résultats de l'ANOVA. Alors que, le même rythme, rPVI−C_d, joue un rôle significatif dans le modèle GMM/RL, selon la liste des paramètres pertinents choisi par RL.

Tableau 6-9 : Les moyennes et les écarts-types pour les métriques à base de durée et d'intensité. La signification statistique (p-valeur) basée sur l'ANOVA pour les locuteurs L1 et L2 comme variables indépendantes.

Rythme métrique	**L1**	**L2**	***p*-valeur**
%Vi	42.79 (2.43)	41.98 (7.50)	**0.034**
Δ Vi	4.52 (1.99)	5.22 (3.15)	0.45
Δ Ci	5.47 (1.92)	6.18 (2.92)	0.38
Varco Vi	6.10 (2.82)	7.77 (5.02)	0.23
Varco Ci	7.92 (2.91)	9.68 (4.59)	0.15
Varco VCi	8.21 (2.34)	8.85 (3.05)	0.38
nPVI-Vi	5.96 (2.55)	8.4 (6.48)	0.17
nPVI-Ci	8.35 (2.67)	10.52 (5.87)	0.12
rPVI-Ci	5.67 (1.60)	6.58 (3.44)	0.29
nPVI-VCi	8.22 (1.84)	8.94 (2.29)	0.19
rPVI-VCi	5.74 (1.09)	5.82 (1.41)	0.82

6.6. Combinaison et optimisation des méthodes de sélection de paramètres

Pour l'évaluation de cette approche le corpus des émotions a été utilisé. Sept locuteurs (3 hommes et 4 femmes) ont été choisis. Chaque locuteur a prononcé 8 phrases pour chaque émotion et chaque phrase a été prononcée avec les différents états émotionnels. Les huit phrases étaient les mêmes pour toutes les émotions. Le nombre total de phrases utilisées dans les expériences était 840. L'ensemble d'apprentissage était composé de 525 phrases et l'ensemble du test était composé de 315 phrases. Le nombre de paramètres de la trame d'entrée (avant la réduction) était 86. Elle était composée de 12 MFCCs, 8 paramètres auditifs représentés par 4 gaussiennes, qui fournissaient 80 paramètres (20 * 4). Puis, six prosodiques et paramètres de qualité de la voix : le Pitch, le Jitter, le Shimmer, le degré de

voix (DVB), le degré de voyelles (DUV) et les harmoniques à bruit Ratio (HNR) ont été ajoutés pour constituer une trame acoustique de dimension originale de 86.

Les techniques de l'ACP et de LDA ont été combinées pour réduire le nombre de paramètres et pour améliorer le taux de classification des traits paralinguistiques. Par le biais de la première expérience de validation croisée (ACP-LDA), l'ACP a tout d'abord été appliquée aux paramètres originaux et nous a permis de conserver les 29 composantes qui avaient les plus hautes valeurs propres. Puis, LDA a pris les 29 composantes dans l'ordre pour faire la sélection des plus discriminantes.

Dans la deuxième expérience de validation croisée (LDA-ACP), la LDA a été appliquée directement aux paramètres originaux et 24 des 86 paramètres ont été sélectionnés. Ensuite, les paramètres sélectionnés par LDA ont été transmis à l'ACP comme paramètres originaux et ont été mappés sur un nouvel espace ayant une dimension inférieure. Par le biais de l'ACP, seulement un certain nombre de composantes (15 dans notre cas) ont été conservées comme paramètres pertinents. Le nombre final des paramètres pertinents sélectionnés par chaque système est présenté dans le Tableau 6-10

Le système hybride GMM/SVM s'est avéré plus efficace que la SVM seule pour la classification des émotions. Le système de reconnaissance de l'émotion est basé sur la SVM « un contre tous ». On a fait l'apprentissage de la SVM plusieurs fois afin d'améliorer le taux de classification. Comme il est présenté dans le Tableau 6.10, le système à base de SVM a atteint 92.3 % de reconnaissance correcte en utilisant tous les paramètres combinés. Ce taux a diminué à 90.6 % lorsque les paramètres auditifs acoustiques étaient absents.

Nous avons également évalué les systèmes de la GMM-LDA-SVM et GMM-ACP-SVM. Nous avons constaté qu'ils étaient meilleurs que le système de base, et ont atteint la

même performance lorsque le modèle d'oreille était considéré (93.6 %). Enfin, les systèmes GMM-ACP-LDA-SVM et la GMM-LDA-ACP-SVM ont été comparés aux configurations précédentes. Notons que le système GMM-ACP-LDA-SVM a vu ses performances dégradées par rapport au système de référence. Cependant, le système GMM-LDA-ACP-SVM a permis d'obtenir les meilleures performances. Un taux de reconnaissance de 94 % a été atteint lorsque tous les paramètres ont été considérés. Également, il est important de mentionner que ce résultat a été obtenu avec un nombre restreint de 9 paramètres par rapport à la SVM seule qui a utilisé 86 paramètres. Un résumé du taux de classification de la reconnaissance des émotions par différents systèmes est présenté dans le Tableau 6 11.

Le test d'analyse de la variance (ANOVA) a également été appliqué aux 86 paramètres du vecteur original afin de déterminer la signification statistique des variables. Les résultats ont déterminé que seulement 45 des 86 paramètres étaient significatifs (statistiquement pertinents). Cependant, en gardant ces 45 paramètres, nous avons constaté que le taux de classification avait diminué de 15 % pour les mêmes émotions alors que le taux était de 94 % lorsque le système GMM-LDA-ACP-SVM avait été utilisé.

Tableau 6-10 : Performance des systèmes : GMM-SVM, GMM-LDA-SVM, GMM-PCA-SVM, GMM-PCA-LDA-SVM et GMM-LDA-PCA-SVM pour la reconnaissance des émotions de la parole

Système de classification	Les paramètres prosodiques, les paramètres de qualité de la voix, MFCCs et les paramètres auditifs		Les paramètres prosodiques, les paramètres de qualité de la voix et MFCCs	
	Classification correcte (%)	Nombre de paramètres	Classification correcte (%)	Nombre de paramètres
GSVM	92.3	86	90.6	54
GMM-LDA-GSVM	93.6	24	91.2	20
GMM-PCA-GSVM	93.6	29	91.8	22
GMM-PCA-LDA-SVM	83.7	15	80.0	10
GMM-LDA-PCA-SVM	**94.0**	9	89.3	8

6.7. Optimisation de la sélection des paramètres par l'évolution différentielle (DE)

Afin d'évaluer les résultats de l'approche d'optimisation évolutionniste, des expériences ont été effectuées en tenant compte de deux configurations statistiques des méthodes de sélection des paramètres. La première utilise l'ACP pour réduire la dimensionnalité des paramètres en supprimant le bruit et les informations redondantes, et en localisant la structure la plus importante des vecteurs propres qui maintient le plus grand pourcentage de l'information d'origine.

La deuxième méthode est une méthode supervisée qui sélectionne les paramètres importants sans modifier leur emplacement (projection) basée sur l'ANOVA. Cette

dernière a été appliquée sur le vecteur des paramètres originaux en effectuant le test statistique de signification.

Suite aux résultats de test les douze premières composantes disposant des plus hautes valeurs propres avec l'ACP ont été gardées comme les plus optimales. Les tests de l'ANOVA nous ont permis de retenir 83 sur 128 paramètres comme étant les plus pertinents.

L'application de la DE a permis de déterminer le nombre de paramètres optimaux qui était de 50 parmi 128. Par la suite, l'ensemble des résultats finaux de l'application de l'APC et de l'ANOVA a été présenté à la LDA afin d'évaluer leurs capacités de discrimination. Les résultats ont démontré que l'ACP était la meilleure des trois méthodes de sélection des paramètres utilisés dans notre travail pour la réduction de leur dimension. Les résultats de l'application des méthodes de sélection sont présentés dans le Tableau 6-11.

Tableau 6-11: Performance des systèmes : GMM-DE-LDA, GMM-LDA, GMM-ANOVA-LDA, GMM-SVM et GMM-ACP-LDA pour la reconnaissance de 5 classes d'émotion.

Systèmes de classification	**Nombre de paramètres**	**Taux de classification (%)**
GMM-DE-LDA	50	**91.6**
GMM-LDA	128	85.7
GMM-ANOVA-LDA	83	72.5
GMM-SVM "un contre tous"	128	62.4
GMM-ACP-LDA	12	57.5

Comme c'est présenté dans le tableau 6-11, le système à base de LDA a réalisé 85.7 % de reconnaissance correcte en utilisant tous les paramètres originaux. Ce taux aurait pu

descendre à 72.5 % si on avait seulement utilisé les éléments pertinents sélectionnés par le test d'ANOVA. L'utilisation de l'ensemble des composantes optimales obtenues par la réduction des dimensions réalisées par l'ACP, comme entrée pour LDA a atteint un taux de 57.5 %. Lorsque tous les paramètres sont utilisés par SVM ce taux de classification est égal à 62 %. La plus importante amélioration pour le système de reconnaissance automatique des émotions de la parole a été atteinte par la configuration DE-LDA avec un taux de classification de 91.6 %. Il est important de mentionner que ce résultat a été obtenu avec une réduction des paramètres d'entrée par rapport au vecteur de paramètres originaux.

6.8. Conclusion

L'objectif principal de ce chapitre était l'évaluation et la discussion des résultats obtenus par les différentes approches proposées. Pour effectuer la discrimination L1/L2, nous avons testé les classificateurs standards dont GMM, SVM et hybride GMM/SVM. Ces trois classificateurs ont été plus performants en intégrant les paramètres auditifs.

Nous avons également proposé un cadre qui intégrait les métriques rythmiques qui opèrent au niveau suprasegmental et les paramètres auditifs basés sur des indices trouvés au niveau de la trame. Le classificateur basé sur GMM/LR/PSO utilisant ce modèle intégré a démontré les meilleures performances de tous les systèmes testés. Cela suggère que l'approche basée sur la GMM/LR/PSO offre une alternative puissante aux méthodes standards. D'autre part, nous avons démontré qu'un modèle basé sur le rythme et le modèle d'oreille peut être intégré avec succès dans un cadre unifié.

Notre travail confirme qu'une discrimination des accents ne se limitent pas seulement à l'utilisation des durées des segments comme les voyelles et les consonnes, mais sur les

effets qui sont situés à un niveau au-dessus du segment, telles que la syllabe et d'autres caractéristiques prosodiques comme l'intensité.

Dans ce chapitre nous avons présenté les résultats obtenus la nouvelle métrique O-PVI, qui se veut une généralisation de la plupart des métriques de la famille PVI : nPVI, rPVI et CCI. Cette nouvelle mesure optimise la classification des groupes de locuteurs en adaptant les quatre coefficients (α, β, ε et θ), qui sont associés à deux points de normalisation et qui sont calculés conformément à la particularité de l'algorithme d'optimisation par essaim particulaires. Pour tester la performance de cette métrique des expériences ont été effectuées au niveau de la classification de l'accent natif et non natif de langue arabe. Les résultats montrent que O-PVI dépasse les deux mesures basées sur le PVI (nPVI, rPVI, CCI, nCCI) et celles basées sur l'intervalle (%V, Delta, Varco) dans cette tâche de classification.

Également nous avons présenté et comparé diverses approches discriminatoires telles que PCA, LDA, ANOVA et SVMs pour la classification des émotions de la parole. Un cadre pertinent d'analyse acoustique a été proposé. Il repose sur une trame acoustique composée par des paramètres auditifs, MFCCs, prosodiques et les paramètres de qualité de la voix. Cette trame a été utilisée comme entrée par le système GMM-LDA-PCA-SVM et a achevé un taux de classification, pour les émotions, très satisfaisant avec un nombre réduit de paramètres par rapport aux valeurs initiales du système.

Une comparaison de diverses approches de sélection des paramètres à l'aide de ACP, ANOVA, DE et LDA a été faite afin de réaliser la classification de l'émotion humaine à partir de la parole. Le système GMM-DE-LDA a obtenu un taux de classification très

satisfaisant pouvant atteindre 91.6 % de la décision correcte avec une réduction drastique des paramètres d'entrée par rapport au vecteur original.

Chapitre 7 - Conclusion générale

La reconnaissance automatique des traits paralinguistiques à partir de la parole est un domaine très vaste. Il touche à toutes les particularités de l'être humain : le comportement, la personnalité, les habitudes, les caractéristiques biologiques et même les relations avec son environnement. Ce champ est tellement vaste qu'on ne peut pas avoir un système qui prenne en considération tous ces éléments. En plus, ces catégories sont confusionnelles et perplexes à tel point que pour plusieurs la différence entre plusieurs phénomènes de ce domaine est absente ou minime, comme pour les émotions et les sentiments. Cette confusion rend la tâche de l'automatisation très complexe et même sensible durant toutes les phases, de l'annotation jusqu'à la décision.

C'était dans ce contexte que s'inscrivait notre objectif qui était de proposer une méthodologie capable d'améliorer les performances du système de reconnaissance automatique des traits paralinguistiques de la parole.

Une étude profonde de la littérature nous a conduit à la conception de nouvelles approches basées sur les métriques du rythme de la parole appliquées à l'identification des accents natifs et non natifs. Ainsi, une nouvelle métrique, qui généralise la plupart des rythmes métriques de la famille de PVI a été développée dans le cadre de cette thèse. Nous avons également généralisé le développement de métriques à base d'intensité qui ont

contribué positivement avec celles à base de durée pour la reconnaissance des accents natifs et non natifs de la parole.

Durant cette thèse, nous avons également proposé d'intégrer un modèle auditif dans un cadre unifié d'analyse segmentale et suprasegmentale pour la distinction des traits paralinguistiques dans la parole. D'une part, l'idée de l'expérimentation du modèle auditif a été motivée par un grand besoin de paramètres très pertinents pour la conception d'un système fiable et efficace. D'autre part, on voulait exploiter le mécanisme de l'oreille qui est un organe complexe capable de faire des traitements perceptifs de phénomènes acoustiques compliqués.

Notre méthodologie s'est traduite par l'extraction des paramètres acoustiques pertinents intégrés en utilisant diverses configurations de combinaison. Les résultats obtenus ont démontré que les systèmes proposés étaient très efficaces pour la détection d'accent natif et non natif de langue arabe. Cette méthode de combinaison se base sur le rythme et les paramètres auditifs, ce qui a permis de remédier au problème lié à l'utilisation des métriques rythmiques car ces derniers n'ont pas la capacité de faire la distinction des classes rythmiques lorsqu' ils sont utilisés seuls.

Notre approche d'optimisation de la sélection de paramètres basée sur les algorithmes évolutionnaires a été très efficace dans la réduction du nombre de paramètres et l'augmentation de la performance de classification du système de reconnaissance des trais paralinguistiques.

7.1. Recommandations

Les nouveaux paramètres calculés tels que l'OPVI, les paramètres auditifs, les rythmes métriques à base de durée ont apporté des améliorations au système de reconnaissance des traits paralinguistiques. C'était particulièrement le cas lorsqu'ils étaient utilisés avec les systèmes proposés de combinaisons multiples. Par contre, cette capacité n'a pas été entièrement expérimentée, car ils n'ont pas été utilisés dans d'autres environnements tels que les bases de données d'émotions spontanées et les bases de données multilingues pour l'accent arabe. Ces environnements seront nos prochains objectifs visés.

Concernant la nouvelle métrique du rythme, l'OPVI, sa capacité au niveau de distinction entre les accents natifs et non natifs a été prouvée et nous pensons qu'elle peut aider à résoudre d'autres problèmes en relation avec ce domaine, tel que la variabilité interlocuteur, grâce à ses coefficients d'optimisation qui permettent un ajustement et une adaptation à un nouveau contexte acoustique. Cette possibilité est encore à l'étude.

La capacité de l'OPVI sera également expérimentée en reconnaissance automatique des émotions spontanées de la parole.

Le modèle profond optimisé proposé dans cette thèse se base sur un RBM qui prend en entrée tous les paramètres acoustiques en même temps pour le RBM en phase d'apprentissage. Une possibilité d'amélioration de ce modèle réside dans son extension de telle sorte qu'on aura pour chaque paramètre un RBM à la phase d'apprentissage qui sera optimisé indépendamment par un algorithme évolutionnaire. Une couche de liaison qui combine tous les modèles générés par les différents RBM de la phase d'apprentissage sera conçue.

Pour la fusion des classificateurs à la phase de décision la méthode de fusion de noyaux pour les SVM n'a pas été testée durant cette recherche. Elle pourra l'être dans le cadre de travaux futurs.

Pour toutes les expériences effectuées dans cette thèse, le facteur bruit n'a pas été pris en considération. L'expérimentation de nos approches dans un environnement bruité est intéressante à prospecter.

Bibliographies

[1] M.E. Montaugne, «Essais1,9 », livreII ,chapitreIV,1588

[2] G. L. Trager, « Paralanguage», Anthropological Linguistics, vol. 2, 1960, pp. 24–30.

[3] G. L. Trager, « Paralanguage: A First Approximation », Studies in Linguistics, 13, 1958.

[4] F. Saussure, «Course in General Linguistics »1916. ed. by M.Ryan and J.Rivkin, «Cours in Literary Theory: An Anthology« Blackwell Publishers,2001

[5] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, et al. « Emotion recognition based on phoneme classes». INTERSPEECH, 2004.

[6] L. Shiri and K. Boaz, «Why don't we believe non-native speakers? The influence of accent on credibility», 2010.

[7] A. Loukina, G. Kochanski, B. Rosner, E. Keane, and C. Shih, « Rhythm measures and dimensionsof durational variation in speech, « J. Acoust. Soc. Am.,, vol. 129, pp. 3258-3270, , 2011.

[8] A. Arvaniti, « 'The usefulness of metrics in the quantification of speech rhythm » ,J. Phon, vol. 40, pp. 351–373, 2012 2012.

[9] L. He, « Syllabic intensity variations as quantification of speech rhythm: Evidence from both» ,in L1 and L2. In: The 6th International Conference, ed Shanghai, China,2012.

[10] A. Chen, « Perception of paralinguistic intonational meaning in a second language », Language Learning, vol. 59, 2009, pp. 367–409.

[11] S. Steidl, A. Batliner, D. Seppi, and B. Schuller, « On the impact of children's emotional speech on acoustic and language models », EURASIP Journal on Audio, Speech, and Music Processing,2010, pp. 1–14.

[12] C. C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, et al., « Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples », in Proc. Of Interspeech, ed, 2010, pp. 793–796.

[13] N. Romanyshyn, « Paralinguistic maintenance of verbal communicative interaction in literary discourse (on the material of W.S. Maugham's novel « theatre «)», 2009, pp. 550–552.

[14] L. Kennedy and D. Ellis, « Pitch-based emphasis detection for characterization of meeting recordings», in Proc. of ASRU, Virgin Islands.Technologies, Harrigan, ed: Springer, Berlin, 2003, pp. pp. 243–248.

[15] K. Laskowski, « Contrasting emotion-bearing laughter types in multiparticipant vocal activity detection for meetings », in Proc. of ICASSP, ed Taipei: Taiwan, 2009, pp. 4765–4768.

[16] J. Demouy, M. Plaza, J. Xavier, F. Ringeval, M. Chetouani, D. Périsse, et al., « Differential language markers of pathology in autism, pervasive developmental disorder not otherwise specified and specific language impairment », Research in Autism Spectrum Disorders, vol. 5, 2011,pp. 1402–1412.

[17] « A multimodal listener behaviour driven by audio input », 2010, pp. 1–4.

[18] C. Biever, «You have three happy messages ». New, 2005.

[19] G. Mishne, D. Carmel, R. Hoory, A. Roytman, and A. Soffer, « Automatic analysis of call-center conversations », 2005, pp. 453–459.

[20] P. Belin, S. Fillion-Bilodeau, and F. Gosselin, « The montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing », Behavior Research Methods, vol. 40, pp. 208–531.

[21] J. Schoentgen, « Vocal cues of disordered voices: an overview », Acta Acustica United with Acustica, vol. 92, 2006, pp. 667–680.

[22] I. Rektorova, J. Barrett, M. Mikl, I. Rektor, and T. Paus, « Functional abnormalities in the primary orofacial sensorimotor cortex during speech in Parkinson's disease », Movement Disorders, vol. 22,2007 pp. 2043–2051.

[23] « PEAKS—a system for the automatic evaluation of voice and speech disorders »,Speech Communication, vol. 51, 2009, pp. 425–437.

[24] L. Price, J. T. E. Richardson, and A. Jelfs, « Face-to-face versus online tutoring support in distance education »,Studies in Higher Education, vol. 32, 2007, pp. 1–20.

[25] H. Boril, S. Sadjadi, T. Kleinschmidt, and J. Hansen, « Analysis and detection of cognitive load and frustration in drivers' speech », 2010, pp. 502–505.

[26] T. Pfister and P. Robinson, « Speech emotion classification and public speaking skill assessment », 2010.

[27] C. A. Martinez and A. Cruz, « Emotion recognition in non-structured utterances for human–robot interaction », in IEEE International Workshop on Robot and Human Interactive Communication, Nashville, ed, 2005, pp. 19–23.

[28] « Associating children's non-verbal and verbal behaviour: body movements, emotions, and laughter in a human–robot interaction, « 2011, pp. 5828–5831.

[29] B. Schuller, M. Wimmer, D. Arsic, T. Moosmayr, and G. Rigoll, « Detection of security related affect and behaviour in passenger transport »,2008, pp. 265–268.

[30] H. Kwon, V. Berisha, and A. Spanias, « Real-time sensing and acoustic scene characterization for security applications »,2008, pp. 755–758.

[31] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, « Fear-type emotion recognition for future audio-based surveillance systems », Speech Communication, vol. 50, 2008, pp. 487–503.

[32] H. Boril, A. Sangwan, T. Hasan, and J. Hansen, « Automatic excitement-level detection for sports highlights generation », 2010, pp. 2202–2205.

[33] D. Crystal, « A perspective for paralanguage », in Le Maître Phonétique, 120, ed Berlin, 1963, pp. 25–29.

[34] D. Crystal, in Paralinguistics, Current Trends in Linguistics, 12, ed: Mouton de Gruyter, The Hague, 1974, pp. 265–295.

[35] K. L. Pike, «The Intonation of American English »,University of Michigan Press, Ann Arbor, 1945.

[36] D. Abercrombie « Paralanguage», International Journal of Language & Communication Disorders, vol. 3, 1968, pp. 55–59.

[37] D. Crystal, «Paralinguistic behaviour as continuity between animal and human communication, Language and Man Anthropological Issues », Mouton de Gruyter, The Hague, 1975.

[38] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, et al., « Paralinguistics in speech and language – state-of-the-art and the challenge»,, Computer Speech and Language Special Issue on Paralinguistics in Naturalistic Speech and Language, 2013, vol. 27, pp. 4.

[39] K. R. Scherer, « Vocal communication of emotion: a review of research paradigms », Speech Communication, vol. 40, 2003, pp. 227–256.

[40] D. Gillick, « Can conversational word usage be used to predict speaker demographics?», 2010, pp. 1381–1384.

[41] T. Vogt and E. Andre, « Improving automatic emotion recognition from speech via gender differentiation », In: Proceedings of the Language, 2006.

[42] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, et al., « The automatic recognition of emotions in speech », in Emotion-Oriented Systems: The Humaine Handbook , Cognitive Technologies, P. Petta, C. Pelachaud, and R. Cowie, Eds., ed: Springer, 2011, pp. 71–99

[43] D. Gibbon, I. Mertins, and R. K. Moore ,Handbook of multimodal and spoken dialogue systems resources, terminology and product evaluation Kluwer Academic, Boston, 2000.

[44] P. Petta, C. Pelachaud, and R. Cowie « Issues in data collection»,in Emotion-Oriented Systems: The Humaine Handbook , Cognitive Technologies, , Eds., ed: Springer, Berlin, 2011, pp. 197–212.

[45] P. Petta, C. Pelachaud, and R. Cowie « Principles and history», in Emotion-Oriented Systems: The Humaine Handbook , Cognitive Technologies, P. Petta, C. Pelachaud, and R. Cowie, Eds., ed: Springer, 2011, pp. 167–196.

[46] C. E. Osgood, « Semantic differential technique in the comparative study of cultures », American Anthropologist, vol. 66, 1964, pp. 171–200.

[47] S. Steidl, M. Levit, A.Batliner, E. Noth, and H. Niemann «Of all things the measure is man': Automatic classification of emotions and inter-labeler consistency»,. In Proc. of ICASSP, ed Philadelphia, 2005, pp. 317–320.

[48] E. Mower, M. Mataric, and S. Narayanan« Evaluating evaluators: a case study in understanding the benefits and pitfalls of multi-evaluator modeling», 2009.

[49] F. Honig, A. Batliner, K. Weilhammer and E.Noth, « Automatic assessment of non-native prosody for English as L2», in Proc. of Speech Prosody, ed Chicago IL, 2010.

[50] F. Honig, A. Batliner, K. Weilhammer and E.Noth, « Automatic assessment of non-native prosody – annotation, modelling and evaluation », In Proceedings of IS-ADEPT, International Symposium on Automatic Detection of Errors in Pronunciation Training, June 6-8 2012 , Stockholm, Sweden

[51] Z. Zhang and B. Schuller, « Semi-supervised learning helps in sound event classification, », in Proc. of ICASSP, ed Kyoto, Japan, 2012, pp. 333–336.

[52] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, « Unsupervised learning in cross-corpus acoustic emotion recognition », 2011, pp. 523–528.

[53] C. E. Williams and K. N. Stevens, « Vocal correlates of emotional states », in Speech evaluation in psychiatry, ed, 1981, pp. 221–240.

[54] D. Sander, K.R.Scherer, «La psychologie des émotions : Survol des théories et débats essentiels ». In D. Sander & K.R. Scherer, Traité de psychologie des émotions 2009 (pp. 1- 39). Paris : Dunod.

[55] K. R. Scherer, « Psychological models of emotion »,The neuropsychology of emotion, vol. 137, 2000, pp. 137–162.

[56] C. Darwin, «The Expression of the Emotions in Man and Animals»,. London: John Murray, 1872.

[57] P. Ekman and W. V. Friesen, « Relative importance of face, body, and speech in judgments of personality and affect, », Journal of Personality and Social Psychology, vol. 38, 1980, pp. 270–277.

[58] K. R. Scherer, «Emotion and emotional competence: conceptual and theoretical issues for modelling agents », A Blueprint for Affective Computing A sourcebook and manual. K. R. Scherer, T. Banziger and E. Roesch. New York, Oxford University Press: 3-20, 2010.

[59] S. Koolagudi and K. S. Rao, « Emotion recognition from speech using source, system, and prosodic features. », International Journal of Speech Technology, vol. 15, 2012, pp. 265–289.

[60] P.Dumouchel, N. Dehak, Y. Attabi, R. Dehak, and N. Boufaden, « Cepstral and long-term features for emotion recognition. », 2009.

[61] B. Schuller, D. Seppi, A. Batliner, and A. M. et S. Steidl, « Towards more reality in the recognition of emotional » in IEEE International Conference on Acoustics, Speech, and Signal Processing (IEEE Cat. vol. 7, ed Honolulu, HI, USA: The Institution of Engineering and Technology, 2007, pp. 941–944.

[62] M. T. Shami and M. S. Kamel, « Segment-based approach to the recognition of emotions in speech. Multimedia and Expo», in IEEE International Conference on, ed, 2005.

[63] M. Shami and W. Verhelst, «Automatic classification of expressiveness in speech: a multi-corpus study. Speaker classification»,. II: Springer, 2007.

[64] F. Beritelli, S. Casale, A. Russo, S. S., and E. D, « Speech emotion recognition using MFCCs extracted from a mobile terminal based on ETSI front end », in Signal Processing, 8th International Conference on, ed: IEEE, 2006.

[65] M. M. El Ayadi, M. S. Kamel, and F. Karray, « Speech emotion recognition using Gaussian mixture vector autoregressive models », in IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, ed, 2007.

[66] M. Grimm and K. Kroschel, « Rule-based emotion classification using acoustic features », in Proc. Int. Conf. on Telemedicine and Multimedia Communication, Citeseer, ed, 2005.

[67] Z. Inanoglu and R. Caneel, « Emotive alert: HMM-based emotion detection in voicemail messages », 2005.

[68] W. Li and Y. Z. et Yingzi Fu, « Speech emotion recognition in E-learning system based on affective computing » 2007, pp. 809–813.

[69] Y.-L. Lin and G. Wei, « Speech emotion recognition based on HMM and SVM. Machine Learning and Cybernetics», 2005.

[70] V. A. Petrushin, « Emotion recognition in speech signal: experimental study, development, and application. international conference on spoken language processing».

[71] E. V. Seppänen, T. and J. Toivanen, « Prosody-based classification of emotions in spoken finnish»,: INTERSPEECH, 2003.

[72] V. Sethu and E. A. et Julien Epps, « Speaker normalisation for speech-based emotion detection », in Digital Signal Processing, 2007 15th International Conference on, ed, 2007, pp. 611–614.

[73] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, « Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing», Affective Computing and Intelligent Interaction,2007, pp. 139–147.

[74] M. Rotaru and D. J. Litman, «Using word-level pitch features to better predict student emotions during spoken tutoring dialogues», INTERSPEECH, 2005.

[75] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, « Emotion recognition from speech using global and local prosodic features, », International Journal of Speech Technology, 2012 pp. 1–18.

[76] D. Bitouk, V. R., and N. A, « Class-level spectral features for emotion recognition », Speech communication, 2010.

[77] S. G. Koolagudi and S. R. Krothapalli, « Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features », International Journal of Speech Technology, vol. 15, 2012, pp. 495–511.

[78] B. Schuller, B. Vlasenko, R. Minguez, G. Rigoll, and A. Wendemuth, « Comparing one and two-stage acoustic modeling», in the recognition of emotion in speech.

Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on, IEEE, ed, 2007.

[79] N. Boufaden and D. P, «Leveraging emotion detection using emotions from yes-no answers »,INTERSPEECH, 2008.

[80] T. Johnstone and K. R. Scherer, « Vocal communication of emotion. », in Handbook of emotions. vol. 2, ed, 2000, pp. 220–235.

[81] S.McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk and S. Stroeve. «Approaching automatic recognition of emotion from voice: a rough benchmark». ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion ,2000.

[82] S. Goronzy and R. Kompe, « A combined MAP+ MLLR approach for speaker adaptation, « 1999.

[83] M. Lugger and B. Yang, Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters. Acoustics, Speech and Signal Processing, 2008.

[84] E.Marchi, B. Schuller, A. Batliner, S. Fridenzon, S. T. and, and O. Golan, « Emotion in the speech of children with autism spectrum conditions: Prosody and everything else, », in Proceedings 3rd Workshop on Child, Computer and Interaction (WOCCI 2012), Satellite Event of INTERSPEECH, 2012.

[85] M. Lugger and Y. Bin, « The Relevance of Voice Quality Features in Speaker Independent Emotion Recognition, », in Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, ed, 2007.

[86] C. M. A. Cummings, K. E., « Analysis of the glottal excitation of emotionally styled and stressed speech, », J Acoust Soc Am, vol. 98, pp. 88–98, 1995 1995.

[87] R. Huang and e. C. Ma, « Toward a speaker-independent real-time affect detection system », in Pattern Recognition. vol. 1, ed, 2006, pp. 1204–1207.

[88] T. Piske, I. McKay, and J. Flege, « Factors affecting degree of foreign accent in an L2: a review, », Journal of Phonetics vol. 29, pp. 191–215, 2001.

[89] J. E. Flege, D. Birdson, Bialystok, M. Mack, H. Sung, and K. Tsukada, « Degree of foreign accent in English sentences produced by Korean children and adults, », Journal of Phonetics vol. 34, pp. 153–175., 2006.

[90] M. Jilka « The contribution of intonation to the perception of foreign accent, » International Journal of Phonetic Science, 2000.

[91] M. Freland-Ricard, « Organisation temporelle et rythmique chez les apprenants étrangers,, », Etude multilingu, 1996, pp. 61-92.

[92] S. Frota, M. D'Imperio, G. Elordieta, P. Prieto, and a. M. Vigario, « The phonetics and phonology of intonational phrasing in Romance, », Linguistic Theory, 2007,pp. 131-153.

[94] V. Dellwo, « Influences of speech rate on the acoustic correlates of speech rhythm: An experimentalphonetic study based on acoustic and perceptual evidence, », 2010.

[95] K. Livescu and J. Glass, «Lexical modeling of non-native speech for for automatic speech recognition,», Acoustics, Speech, and Signal Processing, ICASSP '00. Proceedings, IEEE International Conference on Vol3 ,2000.

[96] G. Silke, R. Stefan, and K. Ralf, «Generating non-native pronunciation variants for lexicon adaptation. ,», Speech Communication., vol. 42, 2004, pp. 109-123.

[97] R. Gruhn, T. Cincarek, and S. Nakamura, « A Multi-Accent Non-Native English Database, », in Proc. Acoust. Soc, Japan, 2004, pp. 195-196.

[98] G. Bouselmi, D. Fohr, I. Illina, and J.-P. Haton, « Multilingual Non-Native Speech Recognition using Phonetic Confusion-Based Acoustic Model Modificationand Graphemic Constraints », in Proc. Interspeech., 2006, pp. 109-112.

[99] L. R. RABINER, « On the use of autocorrelation analysis for pitch detection », IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING,1977, pp. 24–33.

[100] B. Zellner, Prosodie en synthèse de la parole. Actes de l'Ecole d'été européenne en communication Homme Machine, Alpe d'Huez. Isère, France, 1998.

[101] D. Abercrombie, « Elements of General Phonetics, Edinburgh University, », Journal of Phonetics, vol. 1 1967, pp. 219–222.

[102] G. D. Allen, « Speech Rhythm: its Relation to Performance Universals and Articulatory timing, », Journal of Phonetics, vol. 3, 1975, pp. 75-86.

[103] R. M. Dauer, « Phonetic and Phonological Components of Language Rhythm, », 1987, pp. 447-450.

[104] P. M. Bertinetto, Coarticolazione e ritmo nelle lingue naturali: Rivista Italiana di Acustica, 1990.

[105] F. Ramus, Rythme des langues et acquisition: du langage. Ph.D, 1999.

[106] E. Grabe and E. L. Low, « Durational variability in speech and the rhythm class hypothesis, », Papers in Laboratory Phonology, vol. 7, 2002, pp. 515–546.

[107] J. M. MACKENZIE-BECK, « Perceptual Analysis of Voice Quality: the Place of Vocal Profile Analysis », in W. J. HARDCASTLE & J. M. MACKENZIE-BECK (éds.), A Figure of Speech: a Festchrift for John Laver. Londres : Laurence Erlbaum,2005 , pp. 285-322.

[108] B. HONIKMAN, « Articulatory Settings, », 1964.

[109] T.Haji, « Frequency and amplitude perturbation analysis of electroglottograph during sustained phonation», JASA,1986, pp. 58-62.

[110] W. M. Campbell and Z. N. Karam, « A framework for discriminative SVM/GMM systems for language recognition », in INTERSPEECH, ed, 2009, pp. 2195–2198.

[111] N. Landwehr, M. Hall, and E. Frank, « Logistic model trees », Machine Learning vol. 59, 2005.

[112] M. Katz, M. Schaffoner, E. Andelic, S. Kruger, and A. Wendemuth, « Sparse kernel logistic regression using incremental feature selection for text-independent

speaker identification », In: Proceedings of Speaker and Language Recognition Workshop, IEEE 2006.

[113] J. L. Rouas, J. Farinas, F. Pellegrino, and R. Andre-Obrecht, « Rhythmic unit extraction and modelling for automatic language identification», Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, vol. 47, 2005, pp. 436–456.

[114] L. Wiget, L. White, B. Schuppler, I. Grenon, O. Rauch, and S. Mattys, « How stable are acoustic metrics of contrastive speech rhythm? », J. Acoust. Soc. Am, vol. 127, 2010, pp. 1559–1569.

[115] J. Liss, L. White, S. Mattys, K. Lansford, A. Lotto, S. Spitzer, et al., « Quantifying speech rhythm abnormalities in the dysarthrias, », Journal of Speech Language and Hearing Research, vol. 52, 2009, pp. 1334–1352.

[116] U. Gut, Rhythm in L2 speech. Poznan, 2012.

[117] F. N. a. H.-S. Jeon, « Speech rhythm: a metaphor?, , », Phil. Trans. R. Soc. B, 2014.

[118] P. M. Bertinetto and C. Bertini, « On modeling the rhythm of natural languages, », 2008, pp. 6-9.

[119] U. Gut,«Rhythm in L2 speech », In Gibbon, D., editor, Language and Speech Technology, 20120, pp. 83–94.

[120] P. M. Carter, « Quantifying rhythmic differences between Spanish, English, and Hispanic English, in Theoretical and Experimental Approaches to Romance Linguistics, », Selected Papers from the 34th Linguistic Symposium on Romance Languages. R. Gess and E.J. Rubin (eds). Amsterdam, John Benjamins, 2005, pp. 63-75.

[121] S. X. Chen and M. H. Bond, « Two languages, two personalities? Examining language effects on the expression of personality in a bilingual context. , », Personality and Social Psychology vol. 36, 2010, pp. 1514–1528.

[122] L. White and S. Mattys, « Calibrating rhythm: First language and second language studies, », Journal of Phonetics, vol. 35, 2007, pp. 501–522.

[123] M. P.K and V. Dellwo, « Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English », in Proceedings of the 4th International Conference on Speech Prosody, Campinas Brazil, , vol. , 2008, pp. 423-426.

[124] A. Tortel and D. Hirst, « Rhythm metrics and the production of English L1/L2, », in Proceedings of the 5th International Conference on Speech Prosody, 2010.

[125] J. Kennedy and R. C. Eberhardt « Particle Swarm Optimisation », in Proceedings of the IEEE International Conference on Neural Networks, 1995, pp. 1942-1948.

[126] LDC, 2002, LDC Catalog number LDC2002s02. Linguistic Data Consortium, [Online]. Available: http://www.ldc.upenn.edu/.

[127] P. Boersma, Praat, a system for doing phonetics by computer: Glot, 2001.

[128] T. Nazzi, J. Bertoncini, and J. Mehler, « Language discrimination by newborns: Toward an understanding of the role of rhythm, », Journal of Experimental Psychology: Human Perception and Performance, vol. 24,1998, pp. 756–766.

[129] T.Kuroda, S. Grondin, Y. Nakajima, and K. Ueda, « French and English rhythms are perceptually discriminable with only intensity changes in low frequency regions of speech », in In: Proceedings of the 28th Annual Meeting of the International Society for Psychophysics, Ottawa, Ontario,Canada., Canada., 2012.

[130] E. Ferragne and F. Pellegrino, «Le rythme dans les dialectes de l'anglais : une affaire d'intensit´e ? », In Actes de Journ´ees d'Etude de la Parole. Avignon, France, 2008.

[131] J. L. Flanagan, « Models for approximating basilar membrane displacement, », Bell System Technology Journal, vol. 39,1960, pp. 1163.

[132] S. Seneff, « A joint synchrony/mean-rate model of auditory speech processing », Journal of Phonetics, vol. 16, 1988.

[133] R. F. Lyon, « A computational model of filtering, detection, and compression », in the cochlea. In IEEE International Conference on Acoustics, Speech and Signal Processing, ed, 1982, pp. 1282–1285.

[134] O. Ghitza, « Auditory models and human performance in tasks related to speech coding and speech recognition », IEEE Trans. Speech Audio Proc. SAP, vol. 2, 1994.

[135] R. Stern and N. Morgan, « Hearing is believing: Biologically inspired methods for robust automatic speech recognition, », IEEE Signal Process. Mag, vol. 29, 2012, pp. 34–43.

[136] H. Hermansky, « Perceptual linear predictive (PLP) analysis of speech., », the Journal of the Acoustical Society of America, vol. 87, 1990, pp. 1738–1752.

[137] R. Schluter, L. Bezrukov, H.Wagner, and H. Ney, « Gammatone features and feature combination for large vocabulary speech recognition, », IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4, 2007, pp. 649-652.

[138] C. Kim and R.M.Stern, « Power-normalized cepstral coefficients (PNCC) for robust speech recognition », International Conference on Acoustics, Speech and Signal Processing, 2012, pp. 4101–4104.

[139] J. Caelen, « Space/time data-information in the ARIAL project ear model», Speech Communication, vol. 4,1985, pp. 457–467.

[140] M. Lugger, M.-E. Janoir, and B. Yang, « Combining classifiers with diverse feature sets for robust speaker independent emotion recognition», 2009.

[141] H. Cao, V. R., and N. A, Combining Ranking and Classification to Improve Emotion Recognition in Spontaneous Speech: INTERSPEECH, 2012.

[142] C. M. Lee, S. S. Narayanan, and R. Pieraccini, « Combining Acoustic and Language Information for Emotion Recognition», in Seventh International Conference on Spoken Language Processing, ed, 2002.

[143] R. P. D. Pękalska, E. and M. Skurichina, A discussion on the classifier projection space for classifier combining. Multiple Classifier Systems: Springer, 2002.

[144] T. B¨ack., « Evolutionary Algorithms in Theory and Practice», NewYork:Oxford University Press, 1995.

[145] R. Storn and K. Price., « Differential evolution - a simple and efficient heuristic for global

optimization over continuous spaces », Journal of Global Optimization, 1997, pp. 341–359.

[146] W. Berlin. Fisher, G. Doddington, and K. Goudie-Marshall, « The DARPA Speech Recognition Research Database: Specifications and status », in Proc. of the DARPA Workshop on Speech Recognition, ed, 1986, pp. 93–99.

[147] M. Liberman, K. Davis, M. Grossman, N. Martey , J. Bell, and « the Emotion Prosody Speech and Transcript database », 2002.

[148] Y. Alotaibi and S. A. Selouani, « Evaluating the MSA West Point Speech Corpus, », in International Journal of Computer Processing of Languages. vol. 22, ed, 2009, p. 285.

[149] B Schuller, A Batliner «Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing » Wiley,2013

[150] Association — AAAC emotion Online Available: http:// emotion-research.net

[151] B.Schuller, S. Steidl, A. Batliner, F. Burkhardt,L. Devillers, C. Müller, S. Narayanan«Paralinguistics in speech and language—State-of-the-art and the challenge», Computer Speech and Language, vol 27, 2013, pp.4-39.