وزارة التعليم العالي والبحث العلمي
BADJI MOKHTAR-ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR-ANNABA
جامعة باجي مختار - عنابة

Faculté sciences de l'ingéniorat
Département : Electronique

T H E S E

Présentée en vue l'obtention de diplôme de
DOCTORAT

Détection et classification de la voix pathologique par approche hybride dans le traitement du siganl de la parole

Présentée par
Fethi AMARA

Directeur de thése : FEZARI MOHAMED

Jury :

Présidente :	Halim BAHI	-	Prof UBMA
Examinateurs :	Rafik DJEMILI	-	Prof SKIKDA
	Mohamed seghir BOUMAZA	-	Prof GUELMA
	Rachid HAMDI	-	Prof ANNABA
	DjamilMESSADEK	-	MCA ANNABA

30 juin 2016

Remerciements

D'abord et avant tout, je remercie dieu le tout puissant de m'avoir guider, de m'avoir aider et de m'avoir donner la force pour surmonter les difficultés et de continuer de travailler durant ces longues années.

Je remercie mon directeur de thèse proffesseur FEZARI Mohamed d'accepter de diriger cette thèse.

Je dois un grand merci au Professeur Noureddine DOGHMANE, notre directeur de laboratoire, pour sa passion pour l'exploration scientifique qui sera toujours mon inspiration dans le futur.

Ma gratitude va également au Docteur Hocine BOUROUBA, Maitre de conférence à l'université de Guelma pour ces conseils, ces orientations et ces encouragements. Il m'a aidé à dépasser les obstacles.

Je tiens également à remercie aussi le docteur Evaldas Vaiciukynas, maitre de conférence à l'université Kaunas à Lituanie pour ces conseils.

Mes remerciements vont également aux membres de jury : Proffesseurs : Mme BAHI, M.DJEMILI, M BOUMAZA, M.HAMDI et M.MESSADAG d'avoir bien voulu juger mon travail.

Je tiens à remercier vivement mes amis : Chouaib, Tarek, Lamine, Mouaad, Bilel, Wassim et Djamel pour leurs soutien et leurs encouragements.

Dédicace

Je dédie ce modeste travail à mes parents, à toute la famille et à tous ceux qui m'aiment et je les aime.

ملخص

إن تحليل الاشارة الكلامية تعتمد عليه العديد من التطبيقات. هذه الرسالة تأتي في إطار الأبحاث التي تهدف إلى تطوير نظام آلي من شأنه الفصل بين الأصوات المرضية و الأصوات العادية إنطلاقا من الإشارة الكلامية.

هذه الرسالة تساهم أساسا في تحسين نسب التعرف على الأصوات المرضية وذلك من خلال التهجين بين المصنفات المستعملة بكثرة في هذا المجال. هذه المصنفات لا بد وأن تكون متكاملة في ما بينها.

لبلوغ هذا الغرض، توجهنا بالدرجة الأولى إلى التمييز النوعي للأصوات المرضية مستعملين في ذلك معاملات MFCC ومشتقاتها بالدرجة الأولى والثانية. الخطوة الثانية تخص مرحلة التصنيف حيث استعملنا المصنف GMM للاستغلال، قدراته على تمثيل البيانات، والمصنف SVM الذي يتميز بقدرته على الفصل بين البيانات. لقد تبينت فعالية كل من الصنفين في هذا المجال حيث تحصلنا على نسبة تعرف 88% و 82% على التوالي.

مجموعة التجارب الثانية تخص النظام الهجين GMM-SVM . لقد تم إنجاز مقاربتين : في المقاربة الأولى تم استعمال المصنف GMM لنمذجة البيانات .ثم يتم تجميع مراكز التوزيعات في شعاع ليشكل هذا الأخير معطيات تلقين المصنف SVM.

في المقاربة الثانية: أستعمل المصنف SVM لفصل نماذج وذلك للإستفادة من جميع المعلومات الإحصائية المعرفة للنموذج GMM . مستغلين في ذلك وظيفة التشابه التي تتميز بها النواة. هذه الوظيفة تم تعويضها بمسافات لها القدرة على قياس مدى التشابه ما بين التوزيعات مثل: Kullback-Leibler و Bhattacharyya.

في الواقع هذه المسافات لا تخضع لجميع البديهيات المعرفة للمسافة خاصة عدم المساواة المثلثية. هذه النقطة هي محور المساهمة الثانية حيث استعملنا نفس المسافات المذكورة آنفا مضافة عليها تغيرات لتحقيق جميع البديهيات. تحصلنا على تحسينات في النتائج من ناحية نسب التعرف وكذلك منحنى ROC.

الكلمات المفتاحية:

الأصوات المرضية، جودة الصوت، المميزات، GMM، SVM، التصنيف، الأنظمة الهجينة، وظيفة التشابه.

Résumé

L'analyse du signal de la parole est à la base de beaucoup d'applications. Cette thèse s'inscrit dans le cadre des travaux de recherche qui ont pour objectif le développement d'un système de discrimination entre la voix normale et la voix pathologique à partir de ce signal.
Cette thèse est destinée à contribuer à l'amélioration des taux de detection par la combinaison des classificateurs les plus répondus dans le domaine.
Pour répondre à cet objectif, nous avons procédé en premier lieu à une caractérisation qualitative des voix pathologiques en utilisant les coefficients MFCC et leurs dérivées premières et deuxièmes. Ensuite, deux groupes d'expériences sont réalisées : dans le premier groupe nous nous intéressons à tester l'efficacité du GMM, l'objectif était l'exploitation des capacités génératives, et les SVM reconnus par leurs capacités discriminatives. Les deux classificateurs ont prouvé leurs efficacités dans ce contexte, les taux de reconnaissance sont 88%, 82% respectivement.
Le deuxième groupe d'expériences concerne l'hybridation GMM-SVM. Deux approches sont développées : dans la première approche les GMM sont utilisés dans l'étape de modélisation puis nous construisons un supervecteur GMM en regroupant les centres des gaussiennes. Ce dernier sera l'entrée du SVM. Dans la deuxième approche les SVM sont utilisés pour séparer les modèles GMM afin de bénéficier de toutes les informations statistiques définissant un modèle. Cet objectif est assuré en exploitant la fonction de similarité du noyau RBF. Cette fonction est remplacée par des distances qui ont la capacité de mesurer la similarité entre distributions telles que la divergence de Kullback-Leibler et de Bhattacharyya.
Ces distances en réalité ne sont pas des distances métriques car elles ne vérifient pas tous les axiomes notamment l'inégalité triangulaire. La contribution seconde consiste à utiliser leurs modification qui vérifient tous les axiomes. l'objectif est de tester l'impact de l'inégalité triangulaire lors de la projection des données dans l'espace de redescription. Les résultats sont évaluées en termes spécificité et sensitivité et exactitude. Les courbes ROC sont considérées. La méthode proposée montre une amélioration importante.

Mots clés : Voix pathologique, Qualité de la voix, Paramétrisation, GMM, SVM, Détection et classification, Système hybride, Fonction de similarité.

Abstract

Speech signal processing is the basis of many applications . This thesis is part of researches that have the objective to develop pathological and normal voice discrimination system using this kinds of signals.
The main contribution of this thesis is to improve the enhancement of the recognition rates of pathological voices by the mean of the combination between the most popular classifiers.
To attain this objective, firstly, we have proceed by the characterization of pathological voices using Mel frequency cepstral coefficients (MFCC) and their derivatives. Then, two groups of experimentations were realized : In the first group we are interesting to test the effectiveness of the classifiers : Gaussian mixture model (GMM), which presents an important generative capacities, and the Support Vectors Machine (SVM) recognized by their discriminative capacities. The obtained results confirm the effectiveness of both classifiers for this task, recognition rates are 88% and 82% respectively.
The second group of experiences concerns the hybridization GMM-SVM. Two approaches are realized : in first approach, the GMM are used in the modeling step, GMM supportvector are extracted to be the input of the SVM. In the second approach, SVMs are used to separate GMM models in order to beneficiate from all statistic informations. This objective is ensured by exploiting the similarity function of the RBF kernel. This function is replaced by a distance which have the ability to measure the similarity between probability density function (Gaussian distribution). In this aspect, Kullback-Leibler divergence and Bhattacharyya distance are proved to be effective. Unfortunately, those distances are not a metric since they do not satisfy all metric axioms in particular, they violate the triangle inequality. The second contribution is to use a modified version for both distances, this modification transforms them into distances metric. The goal is to test the impact of this violation during the mapping in the redecription space. The performances are evaluated in term of accuracy, sensitivity and specificity. ROC curves are considered. The proposed scheme shows a significant enhancement.

Key words : Pathological voice , Voice quality, Parametrisation, GMM, SVM, Detection and classification.

Liste des abréviations

ASR	Automatic Speaker recognition.
TMP	temps maximum de phonation.
GMM	Gaussian Mixture Model.
SVM	Support Vectors machine.
MFCC	Mel Frequency Cepstral Coefficients.
LPC	Linear Predictive Coefficients.
LPCC	Linear Predictive Cepstral Coefficients.
WLPCC	Weighted Linear Predictive Cepstral Coefficients.
HNR	Harmonic to noise ratio.
GNE	Glottal to Noise Excitation.
NNE	Normalized Noise Energy.
MDVP	Multi-Dimensional Voice Program.
ANN	Artificial Neural Network.
HMM	Hidden Markov Model.
LDA	Linear Discriminant Analysis.
K-NN	K-Nearest Neighbours.
MEEI	Massachusetts Eye and Ear Infirmary.
SVD	Saarbrucken Voice Database.
MLP	Multi-Layer Perceptron.
PCA	Principal Component Analysis.
EGG	Electoglottography.
TFD	Transformée de Fourier discrete.
ML	Maximum Likelihood.
EM	Expectation Maximisation.
ROC	Receiver Operating Curve.
MCS	Monte- Carlo Simulation.

Table des figures

Liste des tableaux

Table des matières

Introduction générale

Contents

0.1 Contexte et motivation

Le signal de la parole est un signal compliqué et très riche en informations dont plusieurs facteurs sont mis en jeu lors de sa production. L'un des facteurs les plus important est la structure ou la physiologie de l'appareil phonatoire. Cette dernière varie d'une personne à une autre en fonction de son sexe et son âge. Cette variabilité traduit la variabilité inter locuteurs, entre hommes et femmes, entre un jeune homme et un homme âgé. Plusieurs applications de reconnaissance de locuteur tels que l'identification et la vérification sont basées sur ce principe dont nous pouvons reconnaitre la personne à partir de son signal de la parole.

L'évaluation et l'identification de la voix pathologique à partir du signal de la parole ne fait pas exception à la règle. En effet, Les maladies de l'appareil phonatoire sont multiples , quelques-unes changent la structure et la morphologie, d'autres perturbent le fonctionnement de cet appareil. En résultat nous avons un signal de parole de qualité dégradée. D'ailleurs la majorité des patients décident de visiter le médecin quand ils ont des troubles de phonation.

L'évaluation de la qualité de la voix et la perception des causes de sa dégradation à travers différents indices acoustiques a toujours été la préoccupation clinique principale des phoniatres. Le diagnostic de la voix à deux objectifs principales. En premier lieu, le diagnostic précoce permet un traitement immédiat, réduisant le risque de graves problèmes de santé. En deuxième lieu, l'évaluation de la voix

permet d'évaluer l'efficacité d'un traitement ou d'une intervention chirurgicale.

Il existe deux méthodes d'évaluation : la première est basée sur l'évaluation perceptuelle, dont le système perceptif humain est le système le plus performant. Cette méthode consiste à écouter le patient par un jury d'experts sans connaitre ni le patient ni son diagnostic. Ensuite, chacun des membres donne son jugement. L'inconvénient majeur de cette technique est qu'elle est subjective dont le jugement varie d'un expert à un autre. Cette variabilité est connue sous le nom variabilité inter-auditeur. Occasionnellement, le jugement d'un expert n'est pas le même lors de plusieurs séances d'écoute (variabilité intra-auditeur). La deuxième technique est fondée sur l'évaluation instrumentale, à l'aide des équipements, le médecin peut mesurer des paramètres acoustiques et aérodynamiques tels que le débit d'air, le début d'air glottique, temps maximum de phonation (TMP) etc. Cette technique est considérée comme objective par rapport à la première méthode mais elle possède une fiabilité limitée due principalement au coût des instruments, relativement élevé, et l'état des patients qui ne permet pas parfois de prendre des mesures exactes.

Les systèmes de détection et de classification automatique de la voix pathologiques viennent pour répondre aux insuffisances des deux méthodes citées auparavant. Ces systèmes sont très similaires aux systèmes de reconnaissance automatique de locuteur (RAL). En commençant par l'extraction des paramètres acoustique qui ont pour objectif de quantifier la dysphonie vocale. Ces paramètres caractérisent la différence entre un cas normal et un cas pathologique.
Notre système est basé sur l'apprentissage automatique, en psychologie, l'apprentissage signifie tout acquisition d'un nouveau comportement ou habituation. Cette tache nécessite l'utilisation des données étiquetées c'est -à- dire leur diagnostic est connu à l'avance (apprentissage supervisé). Ces données seront l'entrée du classificateur, cette étape est connue par l'étape de modélisation. Ensuite ce système sera capable de classer des nouvelles données dans leur catégorie appropriée.
Cette idée représente l'ossature de notre travail de recherche et de celle-ci se dégagent plusieurs lignes directrices à savoir :

- Choix des données à traiter : ce type de sujet nécessite l'utilisation d'une base de données performante du point de vue matériels et environnement d'enregistrement. La base de données sur laquelle nous avons travaillé est décrite en détail dans le chapitre 3.
- Choix des paramètres : il est évident qu'une bonne caractérisation de l'anomalie permet d'avoir une bonne discrimination.
- Choix des classificateurs.

Grâce aux réponses à ces questions, nous serons en mesure de concevoir un outil complémentaire dans le domaine de la médecine.

0.2 Contributions

L'objective de cette thèse est la discrimination entre la voix pathologique et la voix normale en utilisant un système de détection automatique. L'avantage principal de cette méthode est qu'elle est non invasive, rapide et non couteuse.
Le travail contient les tâches suivantes :

- Etat de l'art sur la détection et la classification de la voix pathologique à partir du signal de la parole.
- Extraction des paramètres et évaluation de l'effet du nombre de coefficients MFCC et leurs dérivées.
- Exploiter les capacités génératives du modèle à mélange de gaussiennes (GMM) et les capacités discriminatives du séparateur à vaste marge (SVM).
- Fusionner les deux classificateurs GMM-SVM.

Cette thèse contient les contributions suivantes :

1- Travailler sur une nouvelle base de données de la voix pathologique et qui présente quelques avantages par rappport aux anciennes bases de données.

2- Séparer des modèles GMM afin de bénéficier de toutes les informations statistiques. En d'autres mots, bénéficier de l'information contenue dans la matrice de covariance, les centres des gaussiennes et les poids. A l'aide du noyau RBF,

qui posséde une fonction dite fonction de similarité, la séparation des GMM est assurée dans l'espace de redescription.

3- Tester l'éfficacité des différentes divergences telles que la divergence de Kullback-leibler et de Bhattacharyya.

4- Tester l'effet de l'inégalité triangulaire lors de la projection des données dans l'espace de redescription sur la capacité de discrimination du système, en utilisant des nouvelles versions des deux divergenges citées auparavant respectant tous les axiomes notamment l'inégalité triangulaire.

Publications et communications

Les publications et les communications relatives à cette thèse sont les suivantes :

Publications :

1. M. El Emarya, M. Fezari, and F. Amara "Towards Developing a Voice Pathologies Detection System". ISSN 1064 2269, Journal of Communications Technology and Electronics, 2014, Vol. 59, No. 11, pp. 1280-1288 Pleiades Publishing, Inc., 2014.

2. Amara. F, Fezari. M and Bourouba. H "An Improved GMM-SVM System based on Distance Metric for Voice Pathology Detection. Appl. Math. Inf. Sci. 10, No. 3, 1061-1070 (2016) ".

Communications

1. Amara.F , Fezari. M "Voice Pathologies Classification Using GMM And SVM Classifiers." ICESTI 2012 Annaba, Algerie.

2. Fezari. M, Amara. F "Acoustic Voice Analysis, a Non-Invasive Tool for Detection of Voice Disorders Using Adaptive Features. ICIST'13, March 22-24, 2013. Tangier, Morocco.

3. AMARA. F, Bourouba. H, Fezari. M "Pre-computed kernel for detection of spasmodic dysphonia from human voice.ICSIP may 12-14, 2013, Guelma, Algeria.

0.3 Organisation du manuscrit

Le manuscrit est composé, essentiellement, de cinq principaux chapitres :

Chapitre1 : Ce chapitre rappelle les éléments essentiels à la compréhension des mécanismes de la production de la parole, les pathologies de l'appareil phonatoire et leurs influence sur la qualité de la voix. Nous discutons aussi les différentes méthodes d'évaluation.

Chapitre 2 : Dans ce chapitre nous présentons un état de l'art détaillé sur les différentes méthodes réalisées. L'état de l'art est présenté sous forme de deux parties. La première partie concerne les travaux qui comportent des contributions sur l'étape de paramètrisation et l' autre partie traite les travaux qui s'appuient sur l'importance des classificateurs.

Chapitre 3 : Ce chapitre est réservé pour présenter l'architecture typique d'un détecteur de la voix pathologique. Nous décrivons en détails les différentes étapes qui composent ce système.

Dans le **quatrième chapitre** nous examinons l'efficacité du classificateur GMM et du classificateur SVM dans le contexte de la classification de la voix pathologique. Deux groupes d'expérimentation sont réalisés dont l'objectif est de trouver le meilleur modèle qui représente les données. En d'autre mot, nous voulons exploiter les capacités génératives des GMM et les capacités discriminatives des SVM.

Le dernier chapitre est réservé à la présentation de la méthode proposée. Deux systèmes hybrides GMM-SVM sont développés. Dans le premier système nous utilisons les SVM pour séparer les centres des gaussiennes dites GMM supervecteurs. Dans le deuxième système, les SVM sont utilisés pour séparer tout le modèle GMM.

Nous terminons par une conclusion générale et nous présentons par la suite les perspectives de notre travail de recherche.

Chapitre 1

Généralités sur la voix pathologique

Contents

1.1 Introduction

Dans ce chapitre, nous présentons des généralités sur le signal de la parole, y compris l'anatomie de l'appareil phonatoire et le mécanisme de sa production. Ensuite nous discutons les principales pathologies qui atteignent cet appareil et leurs influence sur l'aspect acoustique du signal de la parole. La dernière section est réservée à présenter les différentes méthodes d'évaluation de ces changements acoustiques.

1.2 Appareil phonatoire

La parole, très souvent considérée comme activité propre de l'homme, elle présente un moyen d'expression de l'instinct, des sentiments et des idées. C'est à travers ce signal que nous pouvons communiquer, convaincre, et se défendre. La chose qui traduit la complexité du signal de la parole.

Parce qu'il n'y a pas un organe responsable de la parole, sa production met en jeu plusieurs organes, en quelque sorte c'est une collaboration entre ces organes. Ce système est appelé souvent appareil phonatoire. Cette dernière se compose de trois parties essentielles : les poumons, le larynx et les trois cavités résonantes. Nous présentons dans ce qui suit brièvement le rôle de chaque partie :

1.2.1 Poumons

Appelés aussi soufflerie, le rôle principal du poumon est d'assurer l'air suffisant pour vibrer les cordes vocales. Lors de l'inspiration les poumons se remplissent suite aux mouvements des muscles intercostaux et du diaphragme et que par conséquent la cage thoracique est écartée. Ensuite, L'air est chassé du poumon lorsque les muscles mentionnés auparavant se relâchent, c'est en cour de cette opération ; connue par l'expiration , que nous pouvons produire de la parole. La figure 1.1 nous montre une vue shématique des poumons extrait du livre du gray 1918 [H.Gray 1918]

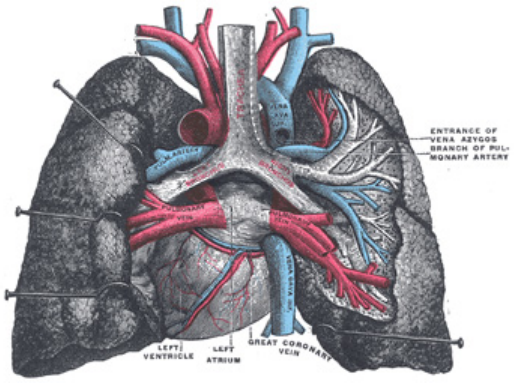

FIGURE 1.1 – Vue shématique des poumons [H.Gray 1918]

1.2.2 Larynx

Le larynx est un tube rigide constitué par une armature cartilagineuse solide, à l'intérieur de laquelle se trouvent les organes mobiles qui vont permettre au larynx d'assurer ses différentes fonctions, toutes basées sur la possibilité de mouvements, d'ouverture et de fermeture. Nous connaissons au larynx trois fonctions essentielles :

- Une fonction phonatoire par l'émission des sons.
- Une fonction de protection des voies aériennes inférieures lors de la déglutition.
- Enfin un rôle actif dans la respiration.

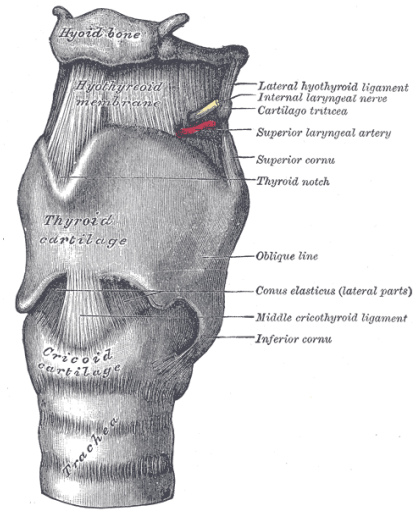

FIGURE 1.2 – Vue de face du larynx [H.Gray 1918]

Les organes mobiles du larynx sont les cordes vocales qui sont présentées par la figure 1.3 :

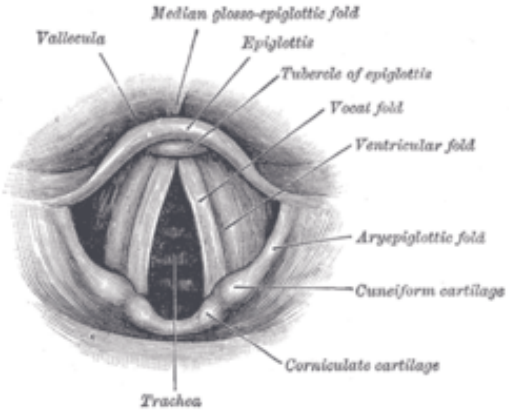

FIGURE 1.3 – Cordes vocles [H.Gray 1918]

Le mouvement des cordes vocales est assuré par le contrôle des muscles et du cartilage qui constituent le larynx. Le plus important est le cartilage thyroïde qui forme le relief de la pomme d'Adam. Ces muscles sont illustrés dans la figure 1.4.

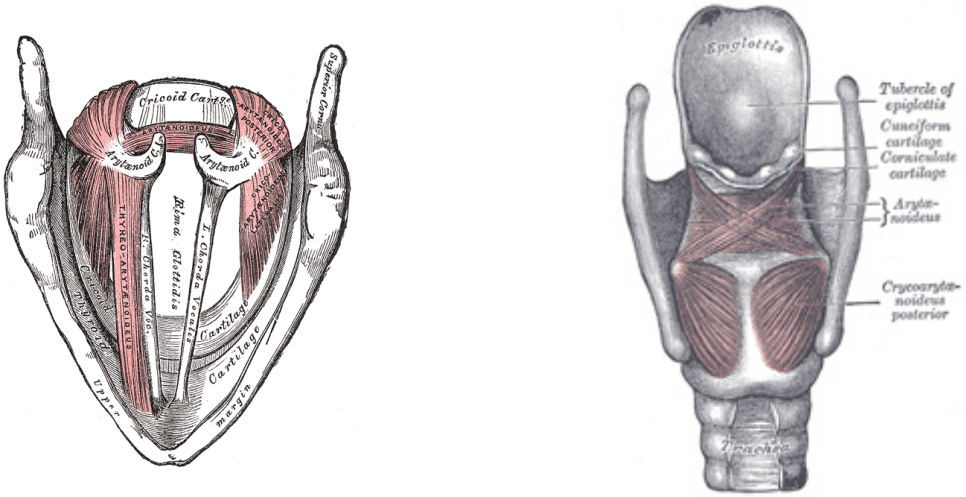

FIGURE 1.4 – Vue shématique des muscles formants le larynx [H.Gray 1918]

1.2.3 Conduit vocal

Le conduit vocal s'étend du larynx jusqu'a l'ouverture bucco-nasale. Son rôle principal consistera à amplifier les sons produits par les plis vocaux. Par ailleurs, grâce aux différentes cavités résonantes qui le constituent. Ces cavités sont :

- Le pharynx : Le mot pharynx vient du grec ancien qui veut dire gorge. Situé entre la cavité buccale et le larynx. Ce conduit musculaire et membraneux joue le rôle d'un carrefour aéro-digestif qui sépare les voies aériennes et les voies digestives. En d'autres mots, le pharynx intervient dans l'opération de déglutition dont il permet aux aliments de passer de la bouche vers l'osophage et éviter ce qu'on appelle fausse route (passage des aliments à la trachée). Comme il participe dans l'opération de respiration par laisser l'air de passer dans le larynx.
  Il joue aussi un rôle crucial dans la phonation, c'est l'espace où les sons produits par la glotte sont modifiés.

- La cavité buccale : En plus de sa fonction biologique dans le système digestif, la cavité buccale joue un rôle important dans la production de la parole. Sa forme change grâce aux mouvements de la langue, les mâchoires et le voile du palais. Elle intervient dans l'étape de l'articulation.

- La cavité nasale : Les cavités nasales ou fosses nasales sont deux espaces creux de volume fixe. L'aire expiré pendant la phonation s'oriente grâce à la position du voile du palais. Pendant la prononciation des voyelles orales le voile est relevé, l'air passe uniquement par la bouche. Quand il s'agit d'une voyelle nasale, le voile du palais est partiellement abaissé, l'air passe par la bouche et le nez. Le dernier cas concerne les occlusives nasales où le voile est complétement abaissé. L'air s'échappe totalement à travers le nez comme dans le son /on/.

1.3 Production de la parole

La parole est produite lorsque l'air des poumons est chassé à travers la trachée grâce aux mouvements des muscles intercostaux et du diaphragme. L'air expulsé alimente l'organe principal de la phonation qui est le larynx qui contient les cordes vocales. Ces dernières peuvent être en deux positions étendues ou relâchées. Dans la première position, les cordes vocales se mettent en vibration et un train d'impulsion est généré. Ces impulsions sont périodiques de période T, ces périodes sont appelées aussi cycles glottiques. L'inverse de T (1/T) définit la fréquence fondamentale F0 du signal de la parole. Les cordes vocales peuvent vibrer à une fréquence de 400 Hz. Ces types de sons , qui font intervenir les cordes vocales sont dits les sons voisés. Le deuxième type de sons est dit non voisé. L'onde sonore générée au niveau de la glotte (espace entre les cordes vocales) est connue sous le nom source glottique ou excitation passe par les cavités résonnantes. Comme leur nom l'indique un résonateur vibre lorsqu'il est excité. L'ors du passage de cette onde, les parois de ces résonateurs entrent en vibration et que par conséquent la fourniture glottique est modifiée. Sa forme exacte dépend de la position des articulateurs que sont les lèvres, les dents, la mâchoire, la langue et le voile du palais. Des formes distinctes du conduit vocal modifient le champ sonore qui s'y propage, produisant ainsi les différents sons de parole. L'appareil phonatoire est présenté dans la figure 1.5. Cette figure est tirée de la thèse de [H.Thomas 2009].

1.4 Pathologies du larynx

Comme tout système biologique, l'appareil phonatoire peut être atteint de nombreuses maladies. L'étude des pathologies de cet l'appareil est un domaine très large, il est pratiquement impossible d'aborder ce sujet dans son intégralité dans le cadre de cette thèse. Selon l'origine de la maladie, nous pouvons distingué trois types de maladies :

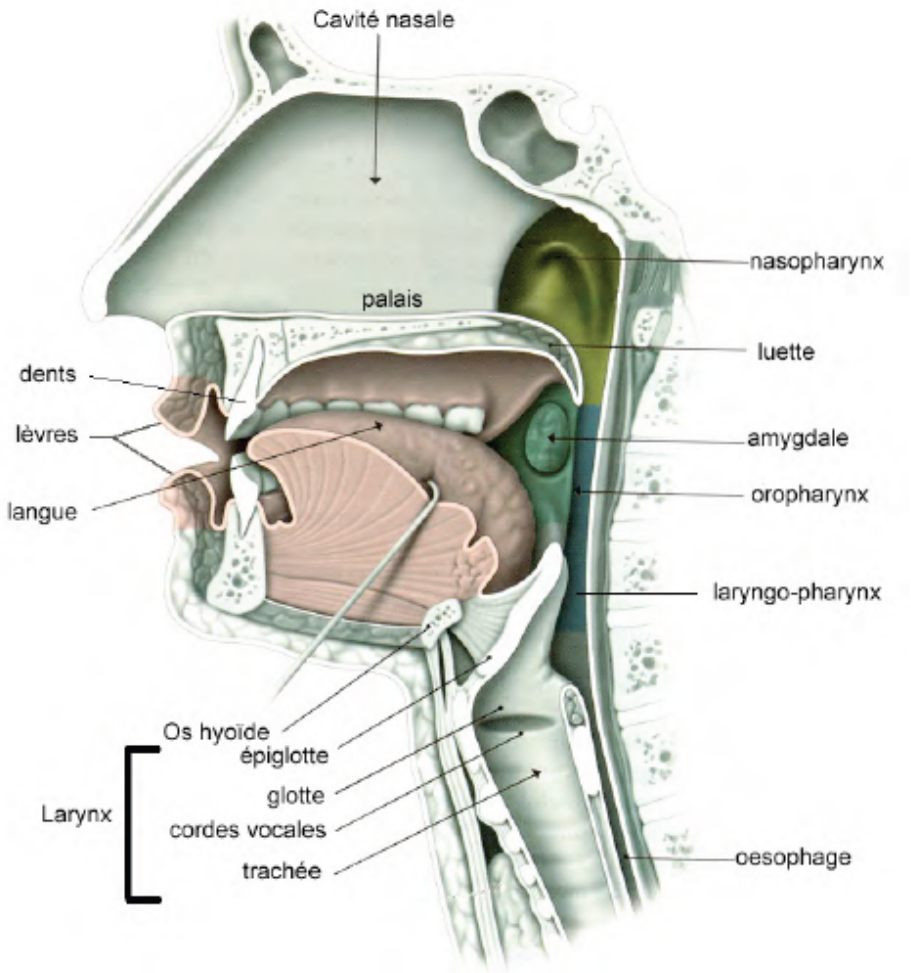

FIGURE 1.5 – Vue shématique de l'appareil phonatoire

1.4.1 Les pathologies organiques

Ces maladies sont dues essentiellement aux changements de structure et de la morphologie de l'organe.

Du point de vue étiologique, les causes de ces pathologies sont multiples. La famille la plus courante est la famille connue par les atteintes inflammatoires telles que la laryngite. Il s'agit d'une infection aigue du larynx, d'origine virale ou bactérienne, comme elle peut être d'origine allergique. En conséquence, cette inflammation empêche les plis vocaux à vibrer librement. Parler devient douloureux et la voix produite est rauque, il peut aller progressivement à l'extinction totale.

Les polypes et les nodules sont parmi les pathologies les plus fréquentes dans la société. Ce sont des petites excroissances charnus, Sessile ou pédicule, à la surface des cordes vocales. Ces pathologies modifient les caractéristiques des cordes vocales et empêchent leur vibrations et leur rapprochements.

Dans le cas des polypes et des nodules, la voix est altérée dans ses trois paramètres :

- Hauteur : La présence du nodule, en déformant le bord libre, gène l'accolement des cordes et diminue l'efficacité du geste vocal. La personne va dans un premier temps tenter d'améliorer l'accolement en augmentant la surface de contact entre les deux cordes, c'est à dire en passant en mécanisme lourd, d'où l'aggravation du fondamental.
- Timbre : éraillé, soufflé, Le caractère soufflé du timbre est lié au mauvais accolement des cordes, entraînant une fuite d'air permanente au cours du cycle vibratoire, très facilement audible par l'oreille. D'autre part, la déformation du bord libre par le nodule modifie le contact entre les deux cordes qui ne peuvent plus avoir des mouvements parfaitement symétriques au cours du cycle vibratoire : elles se comportent donc non plus comme un seul mais deux sources vibratoires, avec 2 F0 audibles, d'où l'éraillement.
- Intensité : elle augmente, surtout à l'attaque. En augmentant la force d'accolement des cordes, le patient améliore, du moins transitoirement, l'accolement des bords libres et réduit le souffle de son timbre. Mais l'augmentation de la pression d'accolement augmente la pression sous glottique nécessaire au main-

tien du cycle vibratoire : l'amplitude vibratoire augmente d'où l'augmentation de l'intensité.

Un exemple d'un polype et un exemple d'un nodule sont illustré dans la figure 1.6 et 1.7.

FIGURE 1.6 – Exemple d'une pathologie organique : polype

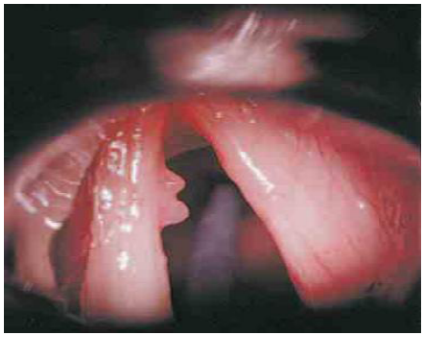

FIGURE 1.7 – Exemple d'une pathologie organique :nodule

Nous présentons dans la figure 1.8 un exemple d'une cas pathologie organique (laryngite)

1.4.2 Les pathologies fonctionnelles

Dans ce groupe de pathologies, appelées aussi non organique, le larynx est anatomiquement normal. Il s'agit d'une perturbation du geste vocal qui peut être d'origine respiratoire, psychologique.

Généralement, les patients souffrent d'un débit sous glottique insuffisant pour faire vibrer les cordes vocales, il existera une fuite d'air à la phonation, source de fa-

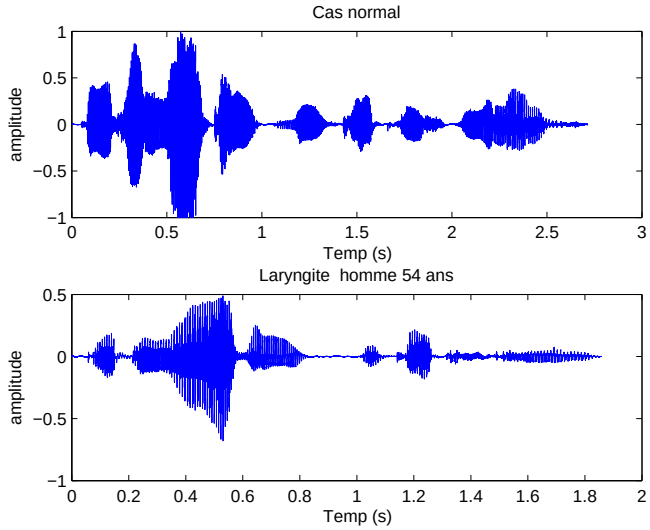

FIGURE 1.8 – Exemple du signal de parole : Cas normal(homme agé de 54), Cas pathologique (homme agé 52 ans)

tigue vocale et de baisse de l'intensité vocale ou de difficulté à l'augmenter. La voix produite est dite voilée ou soufflée. Nous trouvons également l'appellation voix hypokinétiques. La compensation de cette faiblesse s'accomplit par le forçage vocale, effort supplémentaire fournit pour élever la voix. Cet effort peut se résumer en :

- Une prise d'air importante.
- Une vibration des plis supplémentaires (bandes ventriculaires).

Tout forçage vocale est suivi par une fatigue vocale. Dans le langage des cliniciens ce cercle est appelé cercle vicieux fatigue-forçage vocale. L'apparition de lésions organiques conséquences d'une perturbation fonctionnelle du comportement phonatoire est fréquente.

1.4.3 Les pathologies neurologiques

Cette famille regroupe les dysphonies provoquées par l'état neuromoteur du patient. La dysphonie peut être provoquée schématiquement par de multiples facteurs tels que l'hypotonie ou, à contrario, l'hypertonie de la musculature laryngée et respiratoire ou encore des tremblements, qui ont pour conséquence de moduler la hauteur,

l'intensité et le timbre de la voix. Elle peut également être provoquée par un mauvais contrôle de la fermeture de la glotte, conséquence de spasmes ou de paralysies. L'hypotonie a pour conséquence une faible intensité de la voix et un abaissement de la F0. L'hypertonie, qui se manifeste par la difficulté à initialiser un acte volontaire du larynx, se traduit par des hésitations au démarrage du voisement, des émissions vocales discontinues, une augmentation de la F0, un timbre sourd par manque d'harmonique, est voilé par suite d'un mauvais accolement des cordes vocales. Les tremblements, qui peuvent être de fréquence variable en fonction de leur origine, rendent la voix chevrotante (tremor), on peut leur associer des instabilités de F0 en voix tenue Les dysphonies spasmodiques (ou dystonies laryngées) provoquent des changements brutaux de la hauteur de la voix qui peut s'interrompre, repartir, glisser et chevroter (tremor). Elle peut avoir un timbre désagréable et être, au pire, inintelligible. Dans les paralysies laryngées, une corde vocale demeure en position plus ou moins ouverte à la suite d'un mauvais contrôle neuromoteur. La voix est monotone avec perte de mélodie et de nombreuses désonorisations. Elle est soufflée et rauque avec une fuite d'air importante, entraînant un essoufflement en fin de phrase et une voix projetée continue impossible. Un exemple de signal de parole est illustré dans la figure 1.9. Nous pouvons remarquer que le signal dans le cas pathologique est interompu et prend une durée plus longue que celle du cas normal.

1.5 Troubles de la voix

La présence d'une pathologie au niveau du larynx change le fonctionnement et/ou la morphologie des plis vocaux. Les conséquences directes de ce changement sont les troubles de la voix appelés également dysphonies vocales. Généralement, ces troubles sont audibles. Ces troubles se résument dans les irrégularités des cycles glottiques qui connaissent plusieurs causes. Cette partie est inspirée du rapport de l'expertise INSERM [ins 2006]

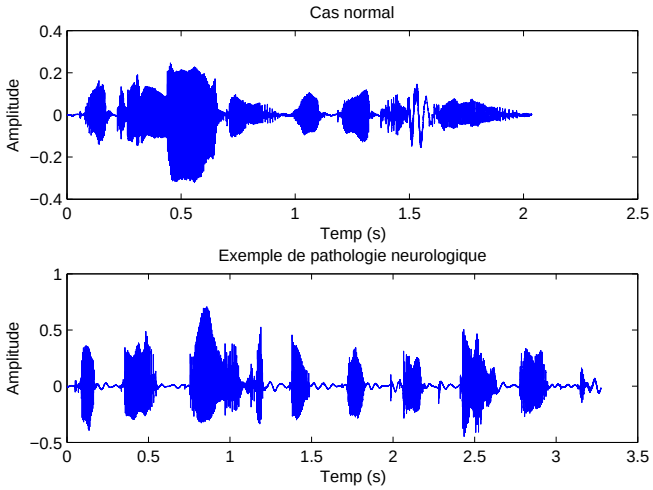

FIGURE 1.9 – Exemple de signal de la parole dans deux cas : cas normal : femme agée de 54 ans, cas pathologique (dysphonie spasmodique femme agée 62 ans)

1.5.1 Vibrations non modales

Les plis vocaux peuvent vibrer dans différent régimes. Le régime peut se caractériser par les cycles glottiques, un régime est dit modal lorsque ces cycles glottiques sont réguliers en terme de durée, forme et amplitude. Il existe trois régimes principaux :

- Diplophonie : ce régime se caractérise par l'inégalité des cycles glottiques, c'est-à-dire nous pouvons trouver plusieurs fréquences fondamentales, leur rapport est rationnel. Le spectre contient plusieurs séries d'harmonique.
- Bi-phonation : à l'inverse de la diplophonie, le spectre d'harmonique est discret et le rapport des fréquences fondamentales n'est pas rationnel. La parole est dite apériodique.
- Irrégulier : dans certains cas, la durée, l'amplitude et la forme des cycles glottiques se changent aléatoirement.

1.5.2 Jitter et shimmer

La gigue vocale (jitter) est définie par la perturbation de la durée des cycles glottiques. autrement dit c'est la perturbation de la fréquence fondamentale. Les

causes sont multiples, les plus connues sont : causes neurologiques, répartition inégale de la mucus sur les plis vocaux. Le jitter est donné par la formule :

$$jitter = \frac{\frac{1}{N-1}\sum_{k=1}^{N}|T_k - T_{k+1}|}{\frac{1}{N}\sum_{k=1}^{N}|T_k|} \tag{1.1}$$

Où T_k est la période de la k^{ime} cycle,N est le nombre de cycle.
Le shimmer est définie par la perturbation de l'amplitude des cycles glottiques. Les cause sont les mêmes que pour la gigue. Le shimmer est donné par la formule :

$$shimmer = \frac{\frac{1}{N-1}\sum_{k=1}^{N}|A_k - A_{k+1}|}{\frac{1}{N}\sum_{k=1}^{N}|A_k|} \tag{1.2}$$

A_k est l'amplitude de la k^{ime} cycle.

1.5.3 Tremblement vocal

Le tremblement vocale est défini par des oscillations de basse fréquence qui contribue à l'altération de la fréquence fondamentale et/ou l'amplitude instantanée. Le tremblement vocal est due non seulement aux pathologies mais aussi aux contraintes physiologiques (respiration, battement cardiaque).

1.5.4 Bruit de turbulence

A cause de la fermeture non complète de la glotte, l'écoulement de l'air devient turbulent. La présence d'un obstacle tels que polype, nodule augmente l'intensité de ce flux turbulent et génère un signal acoustique appelé bruit de turbulence.

1.5.5 Vibration de structure

Le larynx contient des plis supplémentaires appelés plis vestibulaires qui ne vibrent pas. Ces plis peuvent entrer en vibration dans certains cas pathologiques

surtout dans les cas de dysphonie fonctionnelle. La voix produite est de qualité rauque.

1.6 Différentes méthodes d'évaluation

Dans le domaine de la phonétique clinique, l'évaluation de la qualité vocale apparait nécessaire pour établir un bilan vocal. analyser des cas pathologiques, évaluer un bilan thérapeutique, comparer et distinguer entre différents cas pathologiques. Il existe plusieurs méthodes :

1.6.1 Evaluation perceptive

L'évaluation perceptive repose sur le jugement humain et sur les capacités de l'auditeur à évaluer la qualité d'une voix. Dans la mise en place d'un protocole d'évaluation perceptive, le choix du jury est également fondamental. Sa qualité est évaluée en termes de fiabilité, laquelle correspond à la reproductibilité du jugement entre les auditeurs (variabilité interauditeur) et par l'auditeur lui-même, lors de plusieurs sessions d'écoute (variabilité intra-auditeur). Afin d'améliorer la fiabilité de l'évaluation, plusieurs sessions d'écoute sont organisées pendant lesquelles les voix sont présentées dans un ordre aléatoire. Il convient en effet de contrôler le contexte dans la mesure où ce dernier est un facteur d'influence d'importance : une voix moyennement dysphonique paraît plus altérée si elle est présentée après une voix normale qu'après une dysphonie sévère. Menée dans des conditions expérimentales dûment contrôlées, l'analyse perceptive facile à mettre en oeuvre, accessible à tout clinicien et peu coûteuse. En revanche, elle est chargée de plusieurs biais intrinsèques qui la rendent imparfaite voire insuffisante. En effet, de nombreux facteurs influençant le jugement perceptif ne peuvent être tout à fait contrôlés parmi lesquels : l'état émotionnel de l'auditeur au moment de l'évaluation, ses valeurs esthétiques, sa langue maternelle et/ou son dialecte d'origine, la manière dont il conçoit l'échelle de mesure, etc. Concrètement, une analyse perceptive de la qualité de la voix est menée par un panel de spécialistes (thérapeutes de la voix, phoniatres et orthophonistes), lesquels effectuent une description analytique des caractéristiques vocales

sur l'échelle GRBAS Hirano, 1981 [Hirano 1981]. La GRBAS est une échelle perceptive basée sur l'évaluation de cinq paramètres acoustiques (Hirano, 1981 ; 1989) : le grade global de la dysphonie (Grade), le degré de raucité de la voix (Roughness), le souffle (Breathiness), la faiblesse vocale ou asthénie vocale (Aesthenia) et le forçage vocal (Strain). Chacun de ces paramètres est côté selon quatre degrés de sévérité : 0 (normal), 1 (légèrement altéré), 2 (altéré) et 3 (sévèrement altéré). Cette échelle peut être appliquée lors de la production d'une voyelle tenue, d'une phrase ou d'un texte généralement lu.

1.6.2 Evaluation instrumentale

L'analyse instrumentale est conçue pour qualifier et surtout quantifier les dysphonies à partir de mesures acoustiques et/ou aérodynamiques. Ces mesures sont les plus fréquemment réalisées sur une voyelle tenue , en général le /a/, à l'aide de différents capteurs conçus pour enregistrer et étudier de multiples paramètres de la production de parole. Il est toutefois souvent nécessaire de combiner différentes mesures complémentaires afin de tenir compte de la dimension multidimensionnelle de la production vocale. Les mesures acoustiques (i.e. fréquence et amplitude , jitter et shimmer , analyse spectrale) révèlent les caractéristiques audibles de la dysphonie. Il s'agit principalement des mesures de la fréquence fondamentale et de l'intensité, de leur stabilité, ainsi que de l'analyse du spectre du son émis. Les mesures aérodynamiques, sans être à proprement parler des mesures de la voix, permettent d'évaluer les caractéristiques biomécaniques du système pneumo-phonatoire. Il s'agit principalement de mesures de débit, de pression et d'efficacité glottique. Dans le cadre de l'évaluation des dysphonies, l'approche instrumentale multiparamétrique, rendue possible par l'existence d'outils d'expertise fiables, offre une approche complémentaire à l'approche perceptive qui est, par nature, considérée comme 'l'étalon-or', dans la mesure où la voix est avant tout un phénomène perceptif tant du point de vue de celui qui l'émet (le sujet parlant) que de celui qui l'évalue (le clinicien, mais également l'entourage).

1.7 Conclusion

Dans ce chapitre nous avons présenté des généralités sur le signal de la parole y compris la physiologie de l'appareil phonatoire, mécanise de la production de la parole, les pathologies les plus fréquentes et leurs influences sur la qualité de la voix .

Il est important de comprendre que la présence d'une pathologie au niveau de l'appareil phonatoire peut provoquer un dysfonctionnement de tout le mécanisme vocal. En d'autre terme le comportement vocal sera changé, par conséquent, le fruit de la phonation qui est le signal de la parole comportera des changements acoustiques. Ces changements se traduisent par la dégradation de la qualité de la voix. Cette dernière est évaluée par l'oreille clinicienne appelée évaluation perceptive comme elle peut être évaluée par des instruments permettant de mesurer et de quantifier les paramètres considérés comme indicateurs sur la nature et le degré de la dégradation. Le premier propos de cette thèse, est de bénéficier de ces informations dans la caractérisation des voix pathologiques par rapport aux voix normales.

Chapitre 2

Etat de l'art : reconnaissance de la voix pathologique à partir du signal de la parole

Contents

2.1 Introduction

La détection et la classification automatique de la voix pathologique à partir du signal de la parole est la convergence de plusieurs disciplines, aussi variées que

celles de la médecine, la linguistique, la phonétique physiologique, la phonétique acoustique, et le domaine de l'intelligence artificielle. La médecine permet de définir la nature de la maladie, déterminer le grade de la dysphonie, établir la différence par rapport aux autres maladies, fournir des consignes lors de la création de la base de données. La linguistique permet de définir les unités linguistiques, phonèmes, mots ou énoncés, contenant des informations pertinentes sur la qualité de la voix. En d'autres mots, dans certains cas pathologiques, la dysphonie apparait clairement lorsque le patient prononce une unité linguistique bien déterminée. La phonétique physiologique est la discipline qui s'intéresse à définir comment un son est produit, quels sont les organes intervenants, durée de phonation, comme elle s'intéresse aussi à reconnaitre la manière dont ce son est perçu. Puis vient la phonétique acoustique pour déterminer les paramètres caractérisant la dysphonie. Enfin, un problème de classification de la voix pathologique est peut être résolu en utilisant les techniques de l'apprentissage automatique, commençons par le choix des paramètres, la réduction de dimension et la conception des modèles en utilisant les classificateurs les plus appropriés afin d'assurer l'efficacité et la robustesse du système.
Dans la majorité des travaux, un système de détection automatique de la voix pathologique suit une architecture générale. La figure 2.1 présente les étapes essentielles pour concevoir un tel système.

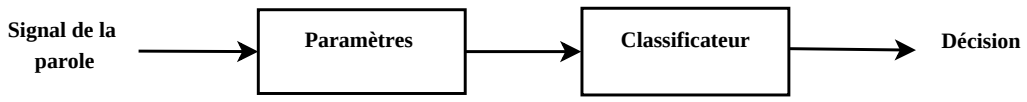

FIGURE 2.1 – Architecture générale d'un détecteur.

Après avoir déterminer le corpus de données sur lequel nous voulons travailler, Un corpus est défini par la nature des maladies qu'il contient, le nombre des patients et leur sexe ainsi l'unité ou l'énoncé linguistique utilisé.
La première étape concerne l'extraction des paramètres à partir du signal de la parole. Ces paramètres ont le rôle de caractériser la qualité de la voix, la chose qui permet de discerner la voix normale de la voix pathologique dans le cas d'une

classification binaire. Ces derniers seront l'entrée du classificateur, qui a le rôle de créer un modèle pour chaque classe et que sera par la suite capable de donner une décision sur la présence ou la non présence d'une pathologie pour une nouvelle entrée. Dans la majorité des travaux la contribution est peut être apportée au niveau de la première étape comme elle peut être au niveau de la deuxième étape. Selon la contribution l'état de l'art sur la reconnaissance de la voix pathologique est présenté en deux grandes parties : travaux selon les paramètres et travaux selon l'approche de classification.

2.2 Etat de l'art : travaux selon les paramètres

Dans la littérature il existe une grande variété des travaux y compris la caractérisation, la détection et la classification des voix pathologiques. Afin de mieux présenter les motivations et les intêrets des travaux que nous présentons. Cette section est réservée à la description des travaux selon le paramètre utilisé. Il existe trois types de paramètres :

- Paramètres du domaine temporel.
- Paramètres du domaine cepstral.
- Mesures de bruit.
- Paramètres calculés en utilisant le logiciel MDVP (Multi dimensional voice parameters)et Praat.

2.2.1 Paramètres du domaine temporel

- **Fréquence fondamentale :**

  La fréquence fondamentale F_0 est définie comme le nombre de vibrations du cordes vocales par seconde. Dans le domaine temporel, c'est la période d'un signal voisé à un instant donné. Pour le signal de la parole, sa fréquence fondamentale n'est rien d'autre que la fréquence du cycle d'ouverture/fermeture des cordes vocales. Elle dépend de la masse vibrante, de la tension des cordes vocales déterminée par les muscles qui contrôlent celles-ci. Elle varie d'un

locuteur à un autre en fonction de son âge et de son sexe. Elle s'étend approximativement de 80 à 200 Hz chez les hommes, de 150 à 450 Hz chez les femmes, et de 200 à 600 Hz chez les enfants [R.Boite 2000] La détection de F_0 joue un rôle essentiel dans le domaine de traitement de la parole et notamment pour l'évaluation de la voix pathologique. C'est un témoin des propriétés biomécaniques des cordes vocales et la configuration laryngée. L'estimation exacte de la F_0 est un problème basique qui a étè la préoccupation des chercheurs. La discussion des algorithmes dédiés pour ce problème ne rentre pas dans les objectives de cette thèse. Le lecteur peut se référer au survey de [T.Drugman 2014] qui contient une description de la majorité des algorithmes de détection de la F_0. Comme nous avons déjà mentionné la fréquence fondamentale donne une information cruciale qui contribue à la distinction entre voix normale et voix pathologique.

- **Perturbation de la fréquence fondamentale et de l'amplitude :**
  Les cycles glottiques naturellement ne sont pas parfaitement périodiques et la présence d'une pathologie augmente considérablement ces apériodicités. Deux phénomènes qui apparaissent : le vacillement (jitter) et le tremblotement (shimmer). Le vacillement est défini par les variations trame par trame dans les périodes de F_0. Le tremblotement représente les variations cycle par cycle dans les périodes de l'énergie. La description mathématique des deux paramètres est présentée dans la section 1.5.2. Dans les travaux Moran et al. [Moran 2006], Michaelis et al.[Michaelis 1998] , Alonso et al. [Alonso 2001], le jitter et le shimmer sont utilisés pour l'évaluation de la voix pathologique. Une étude comparative entre différents algorithmes d'évaluation à étè réalisée par Darcio G. Silva et al. [D.G.Silva 2009], L'objective c'était l'estimation exacte de ce paramètre. L'algorithme proposé jitter local (locjitt) présente une meilleur performance par rapport au jitter estimé par le logiciel PRAAT et MDVP voire même par rapport au jitter estimé par l'algorithme STJE (Short Time Jitter Estimation) proposé par M. Vasilakis and Y. Stylianou. [Vasilakis 2007]

2.2.2 Paramètres cepstraux

Le signal de la parole est peut être représenté par le modèle source-filtre, il résulte d'un produit de convolution (dans le domaine temporelle) entre la source (excitation) et le conduit vocal qui joue le rôle d'un filtre :

$$s(n) = e(n) * h(n) \tag{2.1}$$

Le produit de convolution devient une multiplication dans le domaine fréquentiel :

$$S(f) = E(f).H(f) \tag{2.2}$$

L'application du logarithme permet de séparer l'excitation du filtre.

$$\log(|S(e^{jw})|) = \log(|E(e^{jw})|) + \log(|H(e^{jw})|) \tag{2.3}$$

Le cepstre réel est obtenu par la transformée de Fourier inverse :

$$cc(k) = DFT^{-1}\log(|E(e^{jw})|) + DFT^{-1}\log(|H(e^{jw})|) \tag{2.4}$$

k est l'ordre de raie spectrale.
Le nombre de coefficients cepstraux calculés détermine le niveau de lissage de l'enveloppe spectrale estimée. Les coefficients cepstraux d'ordre faible sont très utilisés en reconnaissance automatique du locuteur (RAL). Ils caractérisent un trait anatomique de l'individu, principalement le conduit vocal.

- **Coefficients MFCC :**

  Les coefficients MFCC sont calculés en utilisant l'échelle de mel (linéaire pour les fréquences inférieur à 1000 et logarithmique pour les fréquences supérieure à 1000). Cette échelle est connue pour prendre compte de la perception humaine. Ces coefficients ont la propriété d'être décorrélés. La description de l'algorithme de l'extraction de ces paramètres est bien détaillée dans la section 4.4. Les coefficients MFCC sont utilisés dans plusieurs domaines de traitement du signal de la parole tel que la reconnaissance de la parole [Schafer. 2009],

la synthèse [Tokuda 2002] et la reconnaissance des émotions [Y.Attabi]. La détection et la classification de la voix pathologique ne fait pas exception à la règle, dans les travaux [Godino-Llorente 2004], [Godino-Llorente 2005] et [Godino-Llorente 2006] les coefficients MFCC présentent l'entrée des classificateurs ANN, SVM, GMM respectivement dont l'objectif était la discrimination entre la voix normale et la voix pathologique.

Ces paramètres sont utilisés aussi pour l'évaluation du grade de la dysphonie selon l'échelle de GRBAS (expliquée déjà dans la section 1.6.1). Dans les travaux [Fredouille 2005], [Pouchoulin 2006], 16 coefficients MFCC et leurs dérivées sont proposés d'être l'entrée du GMM. Chaque grade est représenté par un modèle.

- **Coefficients LPC :**

  Le signal de parole présente une corrélation à court terme induite principalement par les cavités buccales et aussi une corrélation à long terme induite par la périodicité du signal d'excitation. La corrélation à court terme est traduite dans le spectre par la structure des formants (l'enveloppe du signal) et la corrélation à long terme est traduite dans le spectre par une structure fine en peigne dite harmonique [A.Fort 1996].

  Dans la littérature nous ne trouvons pas beaucoup de travaux qui utilisent les coefficients à prédiction linéaire LPC comme paramètres pour caractériser la voix pathologique. Ces coefficients sont utilisés dans [Childers 1992] avec la méthode de quantification vectorielle comme classificateur. Ils ont trouvé un taux de reconnaissance de 82.9%.

- **Coefficients LPCC :**

  Les coefficients cepstral à prédiction linéaire LPCC comme les coefficients LPC sont peu utilisés. Dans l'étude de [O.C.Ai 2012] une comparaison entre trois paramètres LPC, LPCC et WLPCC (LPCC pondérés) montre que WLPCC avec le classificateur LDA et le classificateur KNN présente les meilleurs performances. Les mêmes coefficients LPC présentent des taux faibles par rapport aux deux autres.

2.2.3 Mesures de bruit

Dans certains cas pathologiques, les cordes vocales perdent leurs caractéristiques dynamiques et leurs fermeture devient incomplète. L'air provenant des poumons devient turbulent et se superpose sur la composante périodique au niveau de la glotte. En terme évaluation perceptuelle, ce bruit présente la source de la raucité dans la voix selon les auteurs [Kacha 2006]. Dans la littérature nous pouvons trouver trois types de bruit :

- Rapport harmonique/bruit (HNR) :

  Yumoto [Yumoto 1982] a été l'un des premiers qui a proposé une technique simple de calcul du rapport Harmonique/Bruit (H/N en Anglais). Pour le groupe normal, ce rapport est centré sur 12 dB (harmoniques significativement plus énergétiques que le bruit) et peut devenir négatif pour des pathologies sévères (harmoniques noyées dans le bruit). Plusieurs études sont dédiées à l'estimation exacte du HNR [Severin 2005]. Ce paramètre est peut être mieux estimé dans le domaine spectral que dans le domaine temporel.

- Energie de bruit normalisée(NNE) :

  Ce paramètre est proposé [Kasuya 1986] pour quantifier l'énergie de bruit dans le signal de la parole. Les auteurs supposent que la voix pathologique est plus bruitée que la voix normale.

- Energie de bruit glottique (GNE) :

  Ce paramètre est défini comme le rapport entre la quantité du signal produit par les cordes vocales et le bruit de turbulence au niveau de la glotte [D.Michaelis]. Ce paramètre est supposé comme un bon indicateur sur le degré de la raucité.

2.2.4 Paramètres calculés en utilisant le logiciel MDVP et PRAAT

Programme de voix multidimensionnelle (MDVP) comme son nom nous l'indique est un logiciel commercial destiné aux applications cliniciennes développé par Kay Pentax corporation 2010 [1]. Son rôle principal est de calculer différents paramètres

1. http ://www.kayelemetrics.com/

caractérisant la voix. Nous citons ici les paramètres les plus connus :

TABLE 2.1 – Paramètres acoustiques calculés en utilisant le logiciel MDVP

Paramètre	Signification
F_0	Fréquence fondamentale
T0	Période moyenne
F_{hi}	Fréquence fondamentale la plus élevée
F_{lo}	Fréquence fondamentale la plus basse
STD	Ecart type de la fréquence fondamentale.
Jit_a	Jitter absolue (s)
jit_t	Jitter (%) .
$Shim_a$	Shimmer absolue.
$Shim_t$	Shimmer (%).
$Shim_{db}$	Shimmer (en décibels).
APQ	Quotient de perturbation d'amplitude.
sAPQ	Quotient de perturbation d'amplitude lissé.
HNR	Rapport harmonique/ bruit.
NUV	Nombre des segments non voisé.
DUV	Degré de la non voisement. (%)

Quelques auteurs qui utilisent la base de données MEEI (Massachusetts Eye & Ear Infirmary, base de données commerçiale développée en 1994 au sein du laboratoire Kaypentax) utilisent les paramètres du MDVP fournis avec chaque enregistrements [Dibazar 2002a], [I.G.Llorente 1999], et [Chen 2007]. D'autres utilisent ce logiciel dans la paramétrisation de leur propre base de données [Wang 2006], [Llorente 2008].

Praat est un logiciel gratuit disponible sur le site web [2] développé pour l'analyse acoustique du signal de la parole. Il contient plusieurs taches telles que :

- Analyse générale (wavforme, intensité, durée).
- Analyse de la fréquence fondamentale.

2. http ://www.fon.hum.uva.nl/praat

- Analyse spectrographique.
- Analyse harmonique.

2.3 Etat de l'art : travaux selon l'approche de classification

Les classificateurs peuvent se diviser en deux grandes catégories fondamentalement différentes. La première englobe les méthodes dites "génératives" telles que les modèles de mélanges gaussiennes et modèles de Markov cachés et la deuxième inclue les méthodes dites "discriminantes" telles que les k-voisins plus proches, les réseaux de neurones et les séparateurs à vaste marge.

2.3.1 Modèle à mélange de gaussiennes (GMM)

Les GMM se caractérisent par leur capacités de représentation des données reconnues par les capacités génératives. Ce classificateur est le plus répandu dans le domaine de la reconnaissance du locuteur [Reynolds 2000]. Les GMM présentent des bonnes performances lorsque ils sont appliqués pour la discrimination entre la voix normale et la voix pathologique. Les auteurs [Wang 2006] proposent les paramètres calculés en utilisant le logiciel MDVP pour entrainer les GMM. Chaque classe est représentée par un modèle, le nombre de gaussiennes est choisi empiriquement, le nombre optimal est celui qui présente les meilleurs performances. Le taux de reconnaissance atteint les 92.9% pour la voix normale et les 98.6% pour la classe pathologique. Dans l'étude [Godino-Llorente 2006] les GMM sont utilisés pour examiner l'efficacité des coefficients cepstraux à court terme dans la discrimination des pathologies des cordes vocales. les meilleurs résultats sont obtenus en utilisant 24 coefficients MFCC modélisés par 6 gaussiennes.

La plupart des travaux cités auparavant sont appliqués sur la base de données MEEI. Les GMM sont appliqués pour la première fois sur la base de données SVD (Saarbrcken Voice Database)[3] par les auteurs David et al.[M.David 2012]. Ils ont

3. http ://www.stimmdatenbank.coli.uni-saarland.de.

proposé les coefficients MFCC et les trois mesures de bruit (HNR, GNE, NNE) comme paramètres. Les expériences sont effectuées en utilisant différentes voyelles tenues /a/,/i/et /u/ dans différents types d'intonations (normale, basse, haute et basse-haute-basse). Dans le premier groupe des expérimentations, les taux sont calibrés pour améliorer les performances. Dans le deuxième groupe toutes les voyelles dans différentes intonations sont fusionnées. les meilleurs performances sont obtenues en utilisant la technique de fusion.

Les GMM sont utilisés aussi dans le contexte de la prédiction automatique du grade perceptive de la dysphonie. La première expérience est réalisée par Corine Fredouille en 2005 [Fredouille 2005], chaque paramètre de l'échelle de GRBAS est représenté par modèle. (Les GMM sont détaillés dans le chapitre 3.)

2.3.2 Séparateurs à vaste marge (SVM)

Les séparateurs à vaste marge, comme son nom l'indique, consistent à séparer les données dites d'apprentissage tout en définissant un hyperplan optimal qui sert à maximiser la marge entre les données de différentes classes. Ils sont basés sur une théorie mathématique solide développée par Vapnik [Vapnik. 1998]. Ce classificateur a prouvé son efficacité dans plusieurs domaines de recherche et notamment pour la reconnaissance de la parole/ locuteur .

Cette technique a une large utilisation dans la classification de la voix pathologique. La première application est réalisée par [Godino-Llorente 2005], cette étude est réalisée sur la base de données MEEI (53 voix normales et 77 voix pathologiques). Ils ont proposé les paramètres MFCC et les trois mesures de bruit (HNR, GNE, NNE) pour être modélisés avec les SVM, les paramètres du noyaux RBF (C, sigma) sont choisis pour être optimaux. Le taux de reconnaissance atteints 95.12 %. L'objectif de cette étude est de tester l'efficacité des méthodes discriminatives dans le contexte de la classification de la voix pathologique. Le classificateur SVM a donné les meilleurs résultats.

Une deuxième étude proposée par Chen et al. [Chen 2007] consiste à calculer les 25 paramètres en utilisant le logiciel MDVP. Ces paramètres sont réduits en utilisant l'analyse à composante principale (ACP) pour garder uniquement deux com-

posantes, puis les SVM sont entrainés via différents types de noyaux. Les auteurs notent que l'influence des paramètres des noyaux n'est pas significatif devant l'importance des composantes choisis.

Dans l'étude [Arias 2009], les auteurs ont fixé comme objectif l'évaluation de l'efficacité des mesures de complexité, qui traduisent le comportement non linéaire du signal de la parole, à travers les SVM.

Pour évaluer l'impact des coefficients MFCC et les paramètres de modulation dans la classification de la voix pathologique, les auteurs Markaki et al.[Markaki 2010] ont effectué trois groupes d'expériences. La première expérience consiste à entrainer SVM uniquement avec les coefficients MFCC . Dans la deuxième expérience les paramètres de modulation sont utilisés pour entrainer le même classificateur. Dans la dernière expérience, les deux paramètres sont combinés. La base de données MEEI est utilisée pour l'apprentissage, d'autre base de données sont utilisées dans le test et vice versa. La combinaison des paramètres présente de meilleurs performances. (les SVM sont détaillés dans le chapitre 3).

2.3.3 Modèle de markov caché (HMM)

Le modèle HMM est également une modélisation statistique postulant que le système à modéliser est un processus de Markov (i.e. une suite d'états où l'état courant dépend du ou des états précédents) dont certains états sont " cachés ". Ce type de modèle est particulièrement adapté pour la représentation et l'identification de séquences temporelles et a été largement utilisé pour la modélisation de la parole.

Dans sa première utilisation [Wester 1998], ce classificateur a été entrainé par les paramètres : HNR calculés dans différentes bandes fréquentielles, la fréquence fondamentale et l'intensité. L'objectif c'était l'évaluation du grade global de la dysphonie, le degré de la raucité et du souffle. La comparaison avec les modèles de regression montre que les HMM sont plus performants. L'efficacité du HMM est testée dans le contexte de la discrimination entre voix normale et voix pathologique [Dibazar 2002b]. Les auteurs proposent comme paramètres la fréquence fondamentale, les coefficients MFCC et leur dérivés. Les expériences sont réalisées en utilisant des voyelles tenues et des phrases, Les taux de reconnaissance atteignent les 98,3 %

et 97.75% respectivement. Les résultats sont comparés avec ceux obtenus en utilisant le MDVP.

2.3.4 Les réseaux de neurones artificiels (ANN)

Les réseaux de neurones artificiels est un modèle discriminant formé d'un grand nombre de cellules élémentaires (neurones) fortement interconnectées, dont la sortie de chaque neurone est fonction de ses entrées. Un réseau de neurones est défini par plusieurs paramètres : la topologie des connexions entre les neurones, les fonctions d'agrégation des entrées, et l'algorithme d'apprentissage utilisé. En reconnaissance automatique de la voix pathologique, l'ensemble des patients qui partagent la même dysphonie sont modélisés, par réseaux de neurones, à partir des paramètres qui caractérisent cet ensemble. Différents types de réseaux de neurones sont utilisés dans la tache de discrimination de la voix pathologique.
La première utilisation de ce classificateur dans la discrimination normale pathologique était proposée dans [I.G.Llorente 1999]. Un réseau de neurones multi couches (MLP) est entrainé par 26 paramètres calculés via le logiciel MDVP.
Les auteurs Alonso et al.[Alonso 2001] proposent les paramètres classiques (jitter et shimmer) pour entrainer le réseaux de neurones multi couches (MLP). L'algorithme proposé est testé sur une base de données Espagnole qui contient 5 voyelles. Ils atteignent un taux de 94% d'exactitude.
Leur deuxième étude [Alonso 2005] s'inscrit dans le même cadre dont ils ont testé l'efficacité des paramètres non linéaires en utilisant la même base de données et le même classificateur. Ils ont obtenu 91,77 % d'exactitude. La combinaison des nouveaux paramètres avec ceux utilisés dans l'étude [Alonso 2001] présente meilleurs résultats.

2.3.5 Autres classificateurs

Dans les sections précédentes nous avons présenté les classificateurs les plus utilisés dans le domaine de la classification automatique de la voix pathologique. Il existent d'autres classificateurs qui sont peu utilisés, mais ils présentent des ré-

sultats qui encouragent leur utilisation. Parmi ces classificateurs nous trouvons : le classificateur K- plus proches voisins (KNN). Ce classificateur est défini comme suit : *on dispose d'une base de données d'apprentissage constituée de N couples " entrée-sortie ". Pour estimer la sortie associée à une nouvelle entrée x, la méthode des k plus proches voisins consiste à prendre en compte (de façon identique) les k échantillons d'apprentissage dont l'entrée est la plus proche de la nouvelle entrée x, selon une distance à définir*[4]. Ce classificateur est utilisé dans l'étude [Goddard 2007] afin de tester l'efficacité de différentes méthodes de transformation, y compris l'analyse à composante principale, dans la réduction de dimensions des paramètres. Les paramètres projetés sont : le jitter, le shimmer et HNR. La méthode proposée est testée en utilisant la base de données MEEI. Le taux de classification atteint les 79.3%. Les auteurs [Mujumdar] utilisent KNN comme classificateur dans le but de concevoir un système de discrimination entre différents types de dysarthrie.

L'algorithme k-moyenne est utilisé généralement pour le regroupement des données en plusieurs sous ensembles ou chaque ensemble est représenté par un centre dit moyenne. Cet algorithme est utilisé dans la classification des voix pathologiques [Shama 2007]. Les paramètres utilisés sont les HNR calculés dans 4 bandes fréquentielles et l'énergie dans le domaine spectrale. Un taux de 94,28% est obtenu lorsque l'algorithme est entrainé en utilisant le premier groupe des paramètres. Le deuxième groupe des paramètres donne 92,38%.

2.3.6 Les systèmes hybrides

Les systèmes hybrides consistent à combiner généralement deux classificateurs afin d'améliorer les performances d'un système de détection ou de classification. Les classificateurs combinés sont complémentaires.

L'hybridation GMM-SVM est la plus répandue. Ce système vient pour exploiter les capacités des deux classificateurs. Les GMM sont connus par leurs capacités de représentation des données, appelées communément les capacités génératives. Les SVM possèdent des capacités importantes dans la décision, appelés communément les capacités discriminatives. En d'autres mots, les SVM vient pour combler l'insuf-

4. Wikipédia

fisance des GMM dans la discrimination. Ce système a prouvé son efficacité dans le domaine de la reconnaissance automatique du locuteur [W.M.Campbell 2006].

La discrimination entre voix pathologiques et normales ne fait pas exception, nous trouvons dans la littérature des travaux très intéressants qui encouragent l'utilisation des systèmes hybrides notamment le système cité auparavant. Cette hybridation existe en deux formes : parallèle et séquentielle. Dans la première catégorie, la fusion entre les deux classificateurs se fera au niveau de l'étape de décision. i.e les SVM sont appris sur des nouveaux paramètres en fonction des scores de vraisemblance GMM, on fait un post traitement des scores GMM par des SVM. Un exemple sur cette méthode est présenté dans l'étude de [Markaki 2010]. Les auteurs proposent les coefficients MFCC pour être modélisés par les GMM. Ils ont proposé les paramètres du domaine spectral comme entrée du classificateur SVM.

Pour la deuxième catégorie incluant l'approche séquentielle, les GMM ne sont pas utilisés dans la tache de classification mais ils sont utilisés dans la modélisation des paramètres. Puis les GMM seront l'entrée du SVM. Autrement dit les capacités discriminatives du SVM sont au profit de la séparation entre modèles. La première utilisation du GMM- SVM était proposée par [W.Xiang 2011] Dans ce travail, les auteurs proposent 36 coefficients MFCC dans l'étape de paramétrisation (18 coefficients + leur premières dérivées). Les paramètres sont modélisés par les GMM. Puis la SVM est utilisé dans l'étape de classification, la séparation entre GMM est assurée par le noyau du SVM. Dans la plupart des cas le noyau le plus populaire est le noyaux RBF. Les auteurs confirment l'efficacité du système hybride par rapport au GMM dans la classification de quelques voix pathologiques notamment les pathologies des cordes vocales. Un taux d'exactitude de 96,1% est atteint.

La fusion séquentielle des deux classificateurs est peut être réalisée sous une autre forme. Tout d'abord, les centres des gaussiennes du GMM de chaque classe (normale ou pathologique) sont concaténés dans un vecteur appelé communément GMM super vecteurs. Ce vecteur sera l'entrée du SVM. Les auteurs McLaren et al. proposent cette technique pour la tache de vérification du locuteur. Cette approche est réalisée dans le chapitre 4.

Un nouveau système GMM-SVM est proposé récemment par les auteurs Evaldas

Vaiciukynas et al.[V.Evaldas 2012]. L'idée de base dans cette étude est exploitée la fonction de similarité du noyau RBF pour séparer les modèles GMM. Dans l'approche citée auparavant nous séparons uniquement des centres des gaussiennes. Dans cette approche nous bénéficions de toute l'information statistique définissant le modèle. C'est-à-dire nous bénéficions de l'information inclue dans les poids, les centres des gaussiennes (moyennes) et la matrice de covariance. Après avoir les modèles GMM, leur séparation est assurée dans l'espace de redescription au moyen du noyau RBF. Ce noyau est connu par sa capacité de mesurer la similarité entre les objets, cette caractéristique vient du fait que ce noyau possède ce qu'on appelle la fonction de similarité. Cette fonction ce n'est qu'une distance euclidienne. Les auteurs remplacent cette distance par une distance qui permet de mesurer la similarité entre distributions. Cette étape est donnée dans l'équation suivante :

$$k_{i,j} = \exp(-\frac{D_{i,j}}{2\sigma^2}) \tag{2.5}$$

D est la matrice de distance calculée en utilisant la distance proposée, σ est la largeur du RBF

Les auteurs proposent l'approximation de la distance de Kullback-Leibler par la méthode de monte carlo (KL-MON) et la disantce Kullback-Leibler combinée avec la distance EMD (Earth mover distance).

2.4 Conclusion

Dans ce chapitre nous avons montré que le détection et la classification de la voix pathologique à partir du signal de la parole a attiré l'attention du chercheurs le long des deux dernières décennies. Malgré qu'il y a un grand nombre de travaux et qu'il y a une variété dans les méthodes l'objectif était l'amélioration des performances du système de détection.

L'état de l'art est divisé en deux grandes parties, la première partie concerne l'étape de paramétrisation dans laquelle les chercheurs s'intéressent à utiliser les techniques de traitement du signal dans le but de trouver les indicateurs les plus pertinents

permettant d'établir la différence entre différents types de pathologies. La deuxième étape concerne la classification, les auteurs s'intéressent à exploiter le potentiel des classificateurs tels que les capacités de modélisation, les capacités de discrimination. Nous avons remarqué que la plupart des travaux portent des contributions dans la première étape c'est-à-dire les auteurs pensent à trouver un nouveau paramètre qui n'est pas connu dans les travaux précédents ou bien proposer des nouveaux algorithmes afin d'améliorer l'estimation d'un paramètre bien déterminé. Par rapport à la première étape, les travaux qui portent des contributions au niveau du classificateur sont peu. L'hybridation GMM-SVM, est trés recommandée, dans le cadre de cette thèse nous nous interessons à améliorer ce système.

Chapitre 3

Système de détection automatique de la voix pathologique.

Contents

3.1 Introduction

La détection et la classification automatique de la voix pathologique peuvent être traitées comme un problème de reconnaissance de forme. C'est à dire nous pouvons exploiter les méthodes et les algorithmes utilisés dans ce domaine afin d'atteindre cette objectif. Aprés la création des modèles, le système admet en entrée un signal de la parole et produit en sortie la catégorie de la pathologie véhiculée par la voix du patient. Dans ce chapitre, nous présentons les différentes étapes à suivre pour la création de ce système.

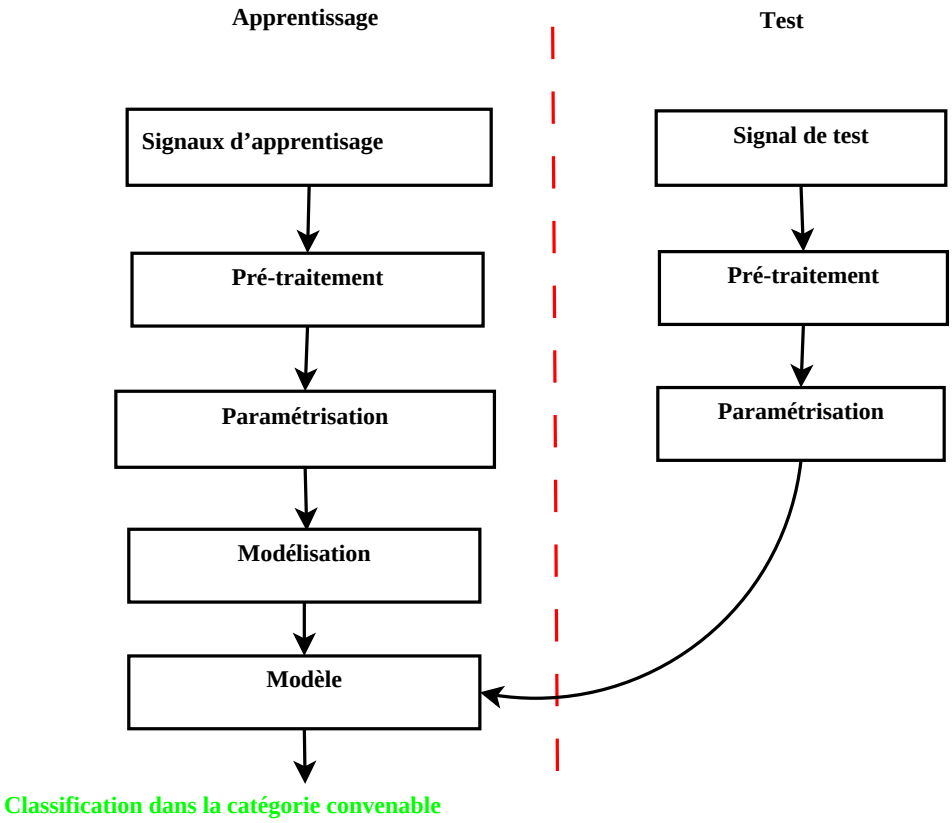

FIGURE 3.1 – Différentes étapes de la conception d'un système de classification automatique

3.2 Architecture générale d'un système de détection

La création d'un système de détection se réalise dans deux étapes essentielles, la première étape concerne l'apprentissage et la deuxième étape est le test. Dans la première phase, chaque classe (pathologique ou normale) est représentée par un modèle où les frontières de la classe sont définies. Le système sera capable de donner une décision pour une nouvelle entrée, La figure 3.1 résume les différentes étapes. L'apprentissage est assuré dans les étapes suivantes :

- Choix de la base de données.
- Prétraitement.
- Extraction des paramètres.
- Modélisation.

Dans ce chapitre, nous aborderons avec plus de détails quelques aspects théoriques sur les méthodes utilisées au cours de la phase d'extraction des caractéristiques et de la conception du classificateur.

3.2.1 Base de données

Dans la conception d'un système de reconnaissance de la voix pathologique le signal de la parole présente la matière première à analyser. La collection des signaux de la parole, dite base de données, doit être performante en terme matériel et environnement d'enregistrement. Ces deux facteurs jouent un rôle très important car une telle application nécessite l'utilisation des signaux de bonne qualité.
D'autre part, l'utilisation d'une base de données standard est un facteur essentiel dans le progrès reconnu dans le traitement du signal de la parole et ces applications y compris la reconnaissance du locuteur (identification et vérification) et la classification de la voix pathologique [N.sanez 2006]. L'utilisation des données standards, nous permet de comparer les résultats obtenus et de juger l'efficacité de la méthode proposée par rapport à l'état de l'art.

3.2.2 Pré-traitement

Il est nécessaire, avant de faire toute analyse, de procéder à certaines opérations sur les fichiers audio appelées prétraitement. L'objectif est l'amélioration de la qualité des paramètres.
Tout d'abord, les sons doivent être sous-échantillonnés . Ensuite, un filtre dit de préaccentuation est appliqué sur tous les fichiers dont la fonction de transfert est donnée par :

$$H(z) = 1 - aZ^{-1} \tag{3.1}$$

Le rôle de ce filtre est de réduire l'effet du microphone en amplifiant les amplitudes des hautes fréquences.

3.2.3 Extraction des paramètres

La première étape dans tous les systèmes de reconnaissance automatique de la voix pathologique est l'extraction des paramètres. Ces paramètres ont le rôle de caractériser la qualité vocale et de quantifier le degrés de la dysphonie.
Vue la nature analogique du signal de la parole et sa complexité traduit par la multitude d'informations et la redondance, son exploitation directe est difficile voire même impossible. L'objective principale de la paramétrisation est de réduire sa complexité en proposant une représentation plus simple. Cette étape consiste à segmenter le signal en un ensemble de portions dites "trames" en utilisant des fenêtres de pondération de longueur fixe durant toute l'analyse. La durée de la fenêtre est entre 20 ms et 32 ms du fait que le signal de la parole est considéré comme stationnaire pendant ces durées. La deuxième trame est obtenue en glissant la fenêtre de telle façon que les fenêtres sont chevauchées, le chauvechement atteint 50%. Pour chaque trame nous appliquons une analyse destinée à un paramètre bien déterminé pour obtenir un vecteur appelé vecteur acoustique. Le principe est illustré dans la figure 3.2

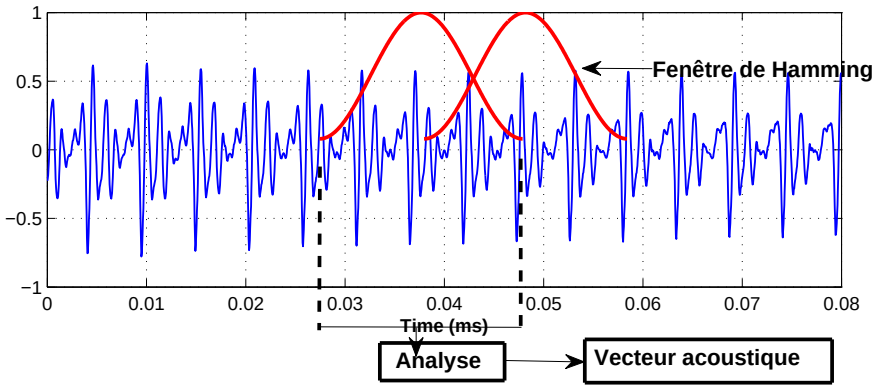

FIGURE 3.2 – Exemple de fenêtrage

Comme nous avons déjà mentionné, les maladies sur lesquelles nous avons travaillé provoquent le disfonctionnement des cordes vocales. La possibilité de parler devient difficile et douloureuse. Ce disfonctionnement influe directement sur la qualité de la voix, le signal produit parrait interrompu et rauque. Ces changements acoustiques

sont audibles par l'oreille nue, et surtout par l'oreille clinicienne. Les coefficients MFCC apparaissent très efficace pour caractériser ces changements. Dans ce qui suit nous présentons la théorie de base derrière ces coefficients. Cette partie est inspirée de la thèse [Y.Attabi].

3.2.3.1 Coefficients MFCC

Les coefficients MFCC ont été intensivement utilisés dans le domaine de la reconnaissance automatique de la parole et de locuteur. Leur rôle principal est la séparation des caractéristiques du conduit vocal des caractéristiques générées par l'excitation.

Les MFCC sont une représentation définie comme étant la transformée cosinus inverse du logarithme du spectre de l'énergie du segment de la parole. L'énergie spectrale est calculée en appliquant un banc de filtres uniformément espacés sur une échelle fréquentielle modifiée, appelée échelle Mel. L'échelle Mel redistribue les fréquences selon une échelle non linéaire qui simule la perception humaine des sons. Dans ce qui suit nous présentons les différentes étapes d'extraction. Figure 3.3

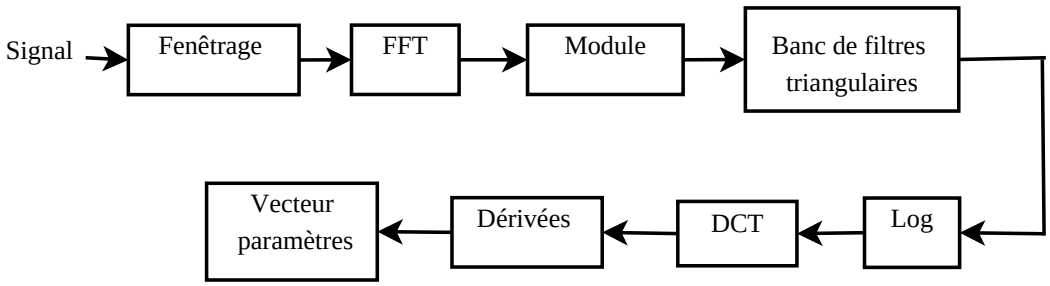

FIGURE 3.3 – Différentes étapes d'extraction des coefficients MFCC

- Fenêtrage :

  Dans cette étape nous utilisons la fenêtre dite de pondération. Généralement c'est la fenêtre de hamming w(n), elle est donnée par la formule suivante :

$$w(n) = 0.54 - 0.46cos(\frac{2\pi n}{N-1})0 \leq n \leq N-1 \tag{3.2}$$

  D'où N est le nombre de points de la fenêtre.

Le résultat de multiplication d'un signal x(t) par la fenêtre w(n) est donné par :

$$x_a(n) = w(n)x(n) \tag{3.3}$$

Cette opération permet d'assurer la continuité aux bords.

- Calcul de la TFD :

  Dans cette étape la transformée de Fourier discrète (TFD) de chaque trame est calculée en utilisant l'algorithme FFT (Transformée de Fourier Rapide). Cette opération permet de transformer chacune des trames du domaine temporel au domaine fréquentiel.

$$X[k] = \sum_{n=0}^{N-1} x_a[n] \exp(\frac{j2\pi nk}{k}) \tag{3.4}$$

  Ensuite, l'énergie est calculée (Prendre le module de la transformée).

- Filtrage :

  Un banc de filtre triangulaires (M filtres) répartis sur l'échelle de mel, l'échelle de mel est considérée comme linéaire sur les fréquences au-dessus de 1khz et logarithmique au de la de cette fréquence. Cette échelle simule la perception humaine.

- Calcul du logarithme :

  Nous calculons le logarithme de l'énergie de chaque filtre en utilisant la formule suivante :

$$S[m] = ln[\sum_{k=0}^{N-1} X_a[k]H_m[k]] \quad 0 < m \leq M \tag{3.5}$$

- Calcul du cepstre :

  Le cepstre est obtenu par le calcul de la transformée en cosinus discrète, qui joue le rôle d'une TFD inverse, du logarithme de la sortie de M filtres.

$$c[n] = \sum S[m]cos(\frac{\pi n(m-\frac{1}{2})}{M}) \quad 0 \leq n < M \tag{3.6}$$

  Seuls les premiers coefficients sont conservés. Dans le domaine de la reconnaissance de la parole les 12 premiers coefficients sont les plus significatifs.

Le premier coefficient , c0 , représente l'énergie moyenne de la trame de de la parole.

3.2.3.2 Calcul des caractéristiques dynamique :

Les changements temporels dans le cepstre (c) jouent un rôle important dans la perception humaine. Ces changements peuvent être obtenus par le calcul des dérivées des coefficients. Δc, sont les dérivées premières et $\Delta\Delta c$ sont les dérivées deuxièmes. Une matrice des coefficients MFCC contient alors :

$$\begin{pmatrix} c \\ \Delta c \\ \Delta\Delta c \end{pmatrix} \tag{3.7}$$

3.2.4 Modélisation

Cette section est réservée pour présenter le fondement de base des deux classificateurs utilisés dans cette thèse :

Modèle de mélange Gaussiennes :

La modélisation par GMM (Gaussian mixture model) est une représentation statistique des données. Ce modèle a montré son efficacité dans plusieurs domaines de recherche, la compression des images, La reconnaissance des émotions [Y.Attabi]et la reconnaissance du locuteur[W.M.Campbell 2006]. La reconnaissance de la voix pathologique ne fait pas exception à la règle, comme nous avons déjà mentionné dans l'état de l'art, les GMM sont intensivement utilisés dans la détection et la classification de la voix pathologique.

- Principe :

  Une densité de probabilité d'un ensemble de données $X = \{x_1, x_2 \dots x_d\}$ peut-être approximée par une somme pondérée de M composantes de densités

$$p(x/\lambda) = \sum_{i=1}^{M} w_i b_i(x) \tag{3.8}$$

  x est un vecteur de dimension d, λ est le modèle, w_i sont les poids de différentes

composantes où

$$\sum_{i=1}^{M} w_i = 1 \text{ Les poids sont positifs} \tag{3.9}$$

$b_i(x)$ est la I ième densité, elle peut s'écrire sous la forme :

$$b_i(x) = \frac{1}{(2\pi)^d |\Sigma_i|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu_i)^T \sum_{i}^{-1}(x-\mu_i)) \tag{3.10}$$

μ_i et Σ_i représente la moyenne et la matrice de covariance de la I ième gaussiennes respectivement.

Le modèle λ est défini par les trois paramètres (w_i, μ_i, Σ)

La phase d'apprentissage consiste à définir les paramètres du modèle pour un ensemble de données (paramètres de la classe). La première étape, connue sous le nom initialisation, à le rôle de déterminer une structure initiale, déterminer les centres des gaussiennes, à partir desquels nous commençons l'ajustement jusqu'à ce que le critère de vraisemblance est atteint (maximum likelihood (ML)). L'algorithme EM (Expectation-maximisation) est l'algorithme le plus utilisé pour assurer ce critère.

- Maximum de vraisemblance (ML) :

  Le maximum de vraisemblance est une méthode qui sert à trouver les paramètres qui maximisent la probabilité d'appartenance des vecteurs paramètres au modèle λ. Ce critère est donné par :

$$p(X/\lambda) = \prod_{n=1}^{N} p(x_n/\lambda) \tag{3.11}$$

  La maximisation de cette fonction est assurée par l'algorithme EM (Espérance-Maximisation).

- Algorithme Espérance-Maximisation (EM) :

  L'algorithme EM est un algorithme itératif utilisé dans l'apprentissage statistique. Il permet de définir les paramètres du modèle λ selon le critère de vraisemblance. Lorsque les seules données dont on dispose ne permettent

pas l'estimation des paramètres, et/ou que l'expression de la vraisemblance est analytiquement impossible à maximiser, l'algorithme EM peut être une solution. De manière grossière et vague, il vise à fournir un estimateur lorsque cette impossibilité provient de la présence de données cachées ou manquantes ou plutôt, lorsque la connaissance de ces données rendrait possible l'estimation des paramètres.

L'algorithme EM tire son nom du fait qu'à chaque itération il opère deux étapes distinctes :

- La phase " Expectation ", souvent désignée comme " l'étape E ", procède comme son nom le laisse supposer à l'estimation des données inconnues, sachant les données observées et la valeur des paramètres déterminée à l'itération précédente.

- La phase " Maximisation ", ou " étape M ", procède donc à la maximisation de la vraisemblance, rendue désormais possible en utilisant l'estimation des données inconnues effectuée à l'étape précédente, et met à jour la valeur du ou des paramètre(s) pour la prochaine itération.

  En bref, l'algorithme EM procède selon un mécanisme extrêmement naturel : s'il existe un obstacle pour appliquer la méthode ML, on fait simplement sauter cet obstacle puis on applique effectivement cette méthode.

  Dans la pratique, il est facile de maximiser le log-vraisemblance que maximiser la vraisemblance elle-même.

Séparateur à vaste marge :

La théorie d'apprentissage statistique étudie les propriétés mathématiques des machines d'apprentissage. Ces propriétés représentent les propriétés de la classe de fonctions ou modèles que peut implémenter la machine. L'apprentissage statistique utilise un nombre limité d'entrées (appelées exemples) d'un système avec les valeurs de leurs sorties pour apprendre une fonction qui décrit la relation fonctionnelle exis-

tante, mais non connue, entre les entrées et les sorties du système.
On suppose premièrement que les exemples d'apprentissage, appelés aussi exemples d'entrainement, sont générés selon une certaine probabilité inconnue (mais fixe) c'est-à dire indépendants et identiquement distribués. C'est une supposition standard dans la théorie d'apprentissage. Les exemples sont de dimension m ($m = R^m$)et dans le cas d'apprentissage supervisé, accompagnés d'étiquettes caractérisant leurs types ou classes d'appartenance. Si ces étiquettes sont dénombrables, on parle de classification sinon on parle de régression. Dans le cas d'une classification binaire cette étiquette est soit +1 et-1. L'ensemble des exemples et leurs étiquettes correspondantes est appelés ensemble d'apprentissage.
Une machine efficace d'apprentissage est une machine qui apprend de l'ensemble d'entrainement une fonction qui minimise les erreurs de classification sur l'ensemble lui-même.

3.2.4.1 Cas des données linéairement séparables

Etant donnée une base d'exemples $A_n = \{(x_1, y_1)(x_2, y_2) \dots (x_n, y_n)\}$. y_i présente l'étiquette qui correspond à la donnée x_i, $y_i = \{+1, -1\}$. Les données étiquetées positivement sont les éléments de la première classe et les données étiquetées négativement appartiennent à la deuxième classe. A partir de cette ensemble A_n,dite l'ensemble d'apprentissage, l'hyperplan est peut être déterminé par l'équation suivante :

$$\begin{aligned} f : R^N &\to R \\ x &\mapsto f(x) = w^T x + b \end{aligned} \tag{3.12}$$

Si les données sont linéairement séparables (marge dure), il y a une infinité d'hyperplans qui séparent ces données figure 3.4. Un hyperplan optimal est celui qui sépare les données de la base d'exemple A_n sans erreur. En d'autre mots, c'est celui qui se situe à une distance maximale par rapport au vecteur x_i les plus proches (appelées communément support vector) parmi les éléments des A_n.
La classification d'un nouvel exemple x est basée sur le signe de la fonction $f(x)$. x appartient à la première classe si la fonction $f(x)$ est positive et à la deuxième

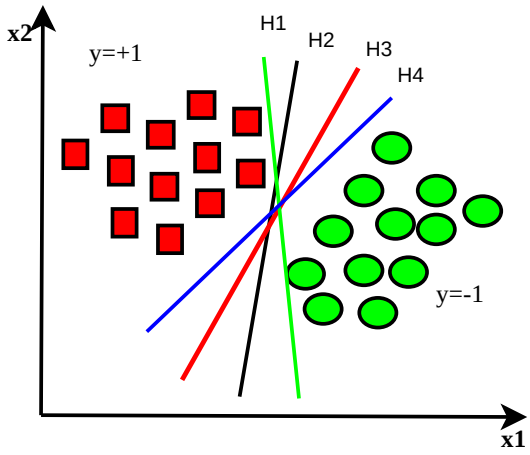

FIGURE 3.4 – Hyperplans séparateurs

classe si $f(x)$ est négative.

$$\begin{cases} w^T x + b \geq 0 & y_i = 1 \\ w^T x + b \leq 0 & y_i = -1 \end{cases} \tag{3.13}$$

A partir des deux équations nous pouvons tirer l'inégalité suivante :

$$y_i(w^T x + b) - 1 \geq 0 \tag{3.14}$$

L'hyperplan $w^T x + b = 0$ est l'hyperplan séparateur, l'hyperplan $w^T x + b = +1$ est celui qui passe par l'exemple le plus proche de la première classe. l'hyperplan $w^T x + b = -1$ est celui qui passe par l'exemple le plus proche de la deuxième classe. La région entre les deux hyperplans s'appelle marge figure 3.5. Plus cette marge est large plus le classificateur est efficace.

Calculer cette marge vient de calculer la distance euclidienne entre x et hyperplans, la distance est donnée par :

$$a = -\frac{w^T x + b}{\|w\|} \tag{3.15}$$

A partir des deux équations nous pouvons déduire que :

$$a = \frac{1}{\|w\|} \tag{3.16}$$

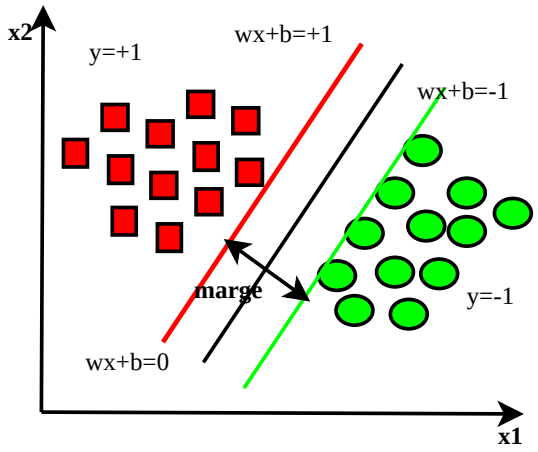

FIGURE 3.5 – Hyperplan optimal

Maximiser cette quantité vient de minimiser $\frac{1}{2}w^2$. L'hyperplan optimal peut être obtenu après résolution du système suivant :

$$\begin{cases} Minimiser \frac{1}{2}w^2 \\ \text{sous les contraintes} \\ y_i(w^T x + b) \geq 1 \end{cases} \tag{3.17}$$

Le système est un problème de programmation quadratique, vue le nombre important des variables, ce problème est convertit en un problème dual qui fait intervenir les multiplicateurs de Lagrange sous la forme :

$$Q(w, b, \alpha) = \frac{1}{2}w^T w - \sum_{i=1}^{n} \alpha_i \{y_i(w^T x + b) - 1\} \tag{3.18}$$

$\alpha = (\alpha_1 \alpha_2 \ldots \alpha_n)$ sont les multiplicateurs non négatifs de Lagrange. L'optimum de la fonction objective Q peut être obtenu en la minimisant par rapport à w et b et en la maximisant par rapport à α. Les conditions d'annulation des dérivées partielles

du Lagrangien mènent directement aux relations vérifiées par l'hyperplan optimal.

$$\begin{cases} \frac{\delta Q(w,b,\alpha)}{\delta w} = 0 \\ \frac{\delta Q(w,b,\alpha)}{\delta b} = 0 \\ \alpha_i\{y_i(w^T x + b) - 1\} = 0 \\ \alpha_i \geq 0 \end{cases} \tag{3.19}$$

Nous pouvons déduire que :

$$\begin{cases} w = \sum_{i=1}^{n} \alpha_i y_i x_i \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases} \tag{3.20}$$

Nous remplaçons dans l'equation :

$$\begin{cases} Minimiser \quad Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \alpha_i \geq 0 \end{cases} \tag{3.21}$$

l'hyperplan optimal est donné par :

$$f(x) = \sum_{i=1}^{m} \alpha_i^* y_i < x, x_i > + w_0^* \tag{3.22}$$

3.2.4.2 Cas des données non linéairement séparables

Dans la plupart des temps, il n'existe pas un hyperplan optimum permettant de séparer les données linéairement. L'utilisation d'un hyperplan à marge dure augmente le nombre des exemples mal classés et que par conséquent diminuer la capacité de généralisation du classificateur. Afin de surmonter les inconvénients d'un séparateur à marge dure deux solutions sont envisageables :

- **Solution 1** :Marge souple :

  Il convient de reformuler le problème résolu auparavant en introduisant des coefficients qui servent à assouplir la marge. afin de prendre en compte tous les exemples d'une classe. On introduit alors les contraintes des variables ξ_i

dites ressort. l'équation de nouvel hyperplan devient :

$$y_i(w^T x + b) \geq 1 - \xi_i, i = 1 \ldots n \tag{3.23}$$

si ξ_i est supérieur à 1, l'observation x_i est mal classée dans la structure. Les erreurs commises sur tous les éléments sont peut être caractérisées par la fonction cout $\sum_{i=1}^{n} \xi_i$. Le problème d'optimisation défini auparavant doit être changer d'une façon à trouver hyperplan le moins tolérant aux erreurs. Le problème primal est peut être exprimé par :

$$\begin{cases} min \frac{1}{2}\|w^2\| + C \sum_{i=1}^{n} \xi_i \\ \text{sous les contarintes} \\ y_i(w^T x + w_0) \geq 1 - \xi_i, i = 1 \ldots M \\ \xi_i \geq 0, i = 1 \ldots M \end{cases} \tag{3.24}$$

$C > 0$ est un constant fixé préalablement, Ce paramètre permet de contrôler le compromis entre la maximisation de la marge et la minimisation des erreurs commises.

Le nouveau problème dual est donné par :

$$\begin{cases} max \left\{ \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{M} \alpha_i \alpha_j y_i y_j < x_i, x_j > \right\} \\ \textit{sous les contraintes} \\ 0 \leq \alpha_i \leq C \quad i = 1 \ldots M \\ \sum_{i=1}^{M} \alpha_i y_i = 0 \end{cases} \tag{3.25}$$

La seule différence par rapport au premier problème (marge dure) et la borne supérieur à C pour les multiplicateur de Lagrange α.

- **Solution 2** : Astuce noyau :

  La projection des données d'un espace de dimension faible à un espace plus dimensionnel vient pour permettre aux SVM de répondre aux problèmes complexes de classification. C'est à dire, transformer les données qui ne sont pas linéairement séparables dans l'espace initiale en des données linéairement sé-

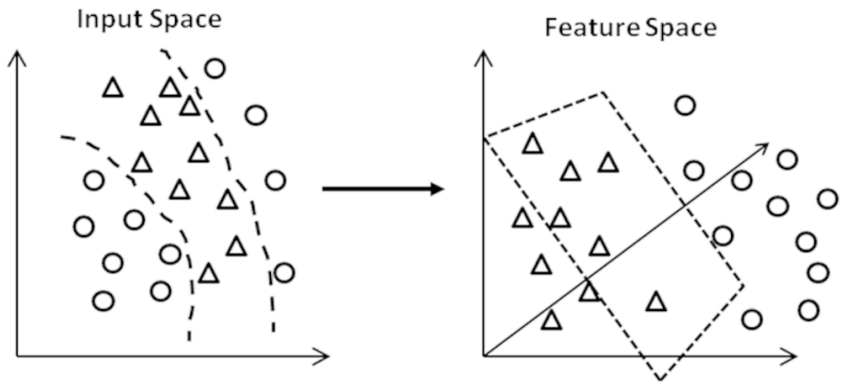

FIGURE 3.6 – Exemple de projection rendant les exemples linéairement séparable.

parable dans l'espace dit espace de redescription. Un exemple de projection est montré dans la figure 3.5.

Exemples de noyaux

- Noyaux polynomiaux :

  Le noyau polynomial est défini comme le produit scalaire des observations élevé à la puissance naturelle q.

$$k(x_1, x_2) = (1 + < x_1, x_2 >)^q \tag{3.26}$$

  si $q = 1$, le noyau devient linéaire.

- Noyaux RBF (Radial basis function)

  Le noyau RBF est le noyau le plus utilisé dans le domaine de détection et la classification est donné par :

$$k(x_1, x_2) = \exp(\frac{-\|x_1 - x_2\|^2}{\sigma^2}) \tag{3.27}$$

  Le mapping ϕ induit par ce type de noyau est un peu spécial. En fait, un exemple va être mappé sur une fonction gaussienne représentant la similarité de l'exemple avec tous les points de X.

  L'espace d'arrivé F (l'espace de représentation) de la fonction ϕ est de dimension infinie étant donné que ϕ fait correspondre une fonction continue à chaque exemple, le kernel RBF permet donc de calculer des

similarités dans l'espace de dimension infinie. Avec le noyau RBF tous les données sont placées sur une sphère de rayon 1 :

$$k(x,x) = \exp(\frac{-\|x-x\|^2}{\sigma^2}) \tag{3.28}$$

Le parmètre σ permet de régler la largeur de la gaussienne. En prenant un σ grand, la similarité d'un exemple par rapport à ceux qui l'entoure sera assez élevée, alors qu'en prenant un σ tendant vers zéro, l'exemple ne sera similaire à aucun autre.

- noyau sigmoidaux

  Le noyau sigmoïdal est défini comme étant :

$$k(x_1,x_2) = \tanh(\alpha_0 < x_1, x_2 > +\beta_0) \tag{3.29}$$

  L'utilisation d'un tel noyau est équivalente à celle d'un réseau de neurones à une couche cachée. Ce noyau dépend de deux paramètres α_0 et β_0 ce qui peut poser des problèmes lors de sa mise en oeuvre.

3.3 Critères d'évaluation

Il existe plusieurs critères pour évaluer les performances d'un système de détection. La matrice de confusion est souvent utilisée par les chercheurs. Elle permet de visualiser les vraies classifications versus les classifications perdîtes. La matrice de confusion d'un problème à deux classes (dans notre cas, normale versus pathologique) est présentée dans le Tableau 3.1.

Décision du système	Décision actuel	
	Pathologique	Normale
Pathologique	VP	FP
Normale	FN	VN

TABLE 3.1 – Matrice de confusion

Les vraies positives (VP) sont les voix pathologiques classés correctement.
Les fausses négatives (FN) sont les voix pathologiques mal classés.
Les fausses positives (FP) sont les voix normales détectées comme pathologiques.
Les vraies négatives (VN) sont les voix normales classés correctement.
A partir de la matrice de confusion nous pouvons calculer différents critères tels que : l'exactitude, précision, rappel. L'exactitude représente le nombre de fichiers correctement classées par rapport au nombre totale des fichiers. Elle reflète la justesse du classificateur en général.

$$exactitude = \frac{VP + VN}{VP + FP + FN + VN} * 100‘\% \tag{3.30}$$

La sensitivité représente le nombre des fichiers pathologiques correctement classés par rapport au nombre totale des fichiers pathologiques.

$$sensitivité = \frac{VP}{VP + FN} * 100‘\% \tag{3.31}$$

La spécificité représente le nombre des fichiers normaux correctement classés par rapport au nombre totale des fichiers normaux.

$$spécificité = \frac{VN}{VN + FP} * 100‘\% \tag{3.32}$$

3.4 Conclusion

Dans ce chaptire nous avons présenté les étapes essentielles pour le developpement d'un classificateur automatique (architecture générale). En cour de cette présentation nous avons profité l'occasion pour présenter la théorie et le fondement de base des outils utilisés dans ce travail y compris les coefficients MFCC et les deux classificateurs GMM ,SVM. Ces étpaes seront respecteés dans les deux chapitres suivants.

Chapitre 4

Méthodologie et expérimentation du classificateurs GMM et SVM.

Contents

4.1 Introduction

Ce chapitre concerne la mise en oeuvre d'un système de détection automatique de la voix pathologique à partir du signal de la parole. Notre système suivra les différentes étapes présentées dans le chapitre 3 (architecture générale).
En effet, deux exemples sont développés, le premier est basé sur les GMM, le deuxième est basé sur les SVM. L'objectif principale n'est pas de concevoir un système dont les performances seront compétitives à celles qui existent dans l'état de l'art, mais nous nous intéressons à tester l'efficacité des deux classificateurs dans le cadre de notre étude ou il s'agit d'une base de données spécifique et des cas pathologiques bien déterminées. L'exploitation des capacités génératives et discriminatives respectivement des GMM et des SVM est l'objectif de ces expérimentations.

Dans ce qui suit, nous décrirons la méthodologie suivie et nous présenterons les expérimentations et les résultats obtenus.

4.2 Méthodologie

L'objectif de cette partie de thèse est de concevoir et d'expérimenter un système de détection en utilisant l'énoncé (une phrase) comme unité d'analyse, les coefficients MFCC comme paramètres, les GMM et les SVM comme classificateurs. A notre connaissance, un tel système n'a pas été encore expérimenté sur la base de données SVD alors qu'il représente l'état de l'art des systèmes de détection automatique de la voix pathologique.
A travers ce système , nous visons à reproduire les résultats obtenus dans l'état de l'art et éventuellement améliorer les taux de détection en explorant deux points différents, à savoir le domaine de l'extraction des paramètres acoustiques, celui de la modélisation en choisissant les paramètres optimaux du classificateur. Dans ce qui suit, nous détaillons chacun de ces deux points :

- Pour l'extraction des paramètres acoustiques, nous envisageons l'utilisation d'une plus large gamme d'information spectrale du signal acoustique permettant de mieux caractériser la classe pathologique, en utilisant un vecteur des 13 premiers coefficients, augmentés par les dérivées (Δ) premières et deuxièmes ($\Delta\Delta$) représentant la vélocité et l'accélération respectivement. Sachant que, les treize premiers coefficients MFCC, leurs dérivées, fournissent toute l'information et aucun gain dans les performances n'est enregistré par l'augmentation du nombre des coefficients (Huang, Acero et Hon, 2001).
- Pour l'optimisation des paramètres du classificateur, nous allons chercher une meilleure modélisation de chaque classe.
  Pour les GMM : le nombre de gaussiennes joue un rôle très important dans le choix du modèle. La définition de nombre de gaussiennes n'obéit pas à une règle notamment dans le cas où les données d'apprentissage ne sont pas visualisable ; comme le cas des coefficients MFCC. Le choix de ce paramètre est empirique c'est-à dire nous varions le nombre de mélanges de gaussiennes du

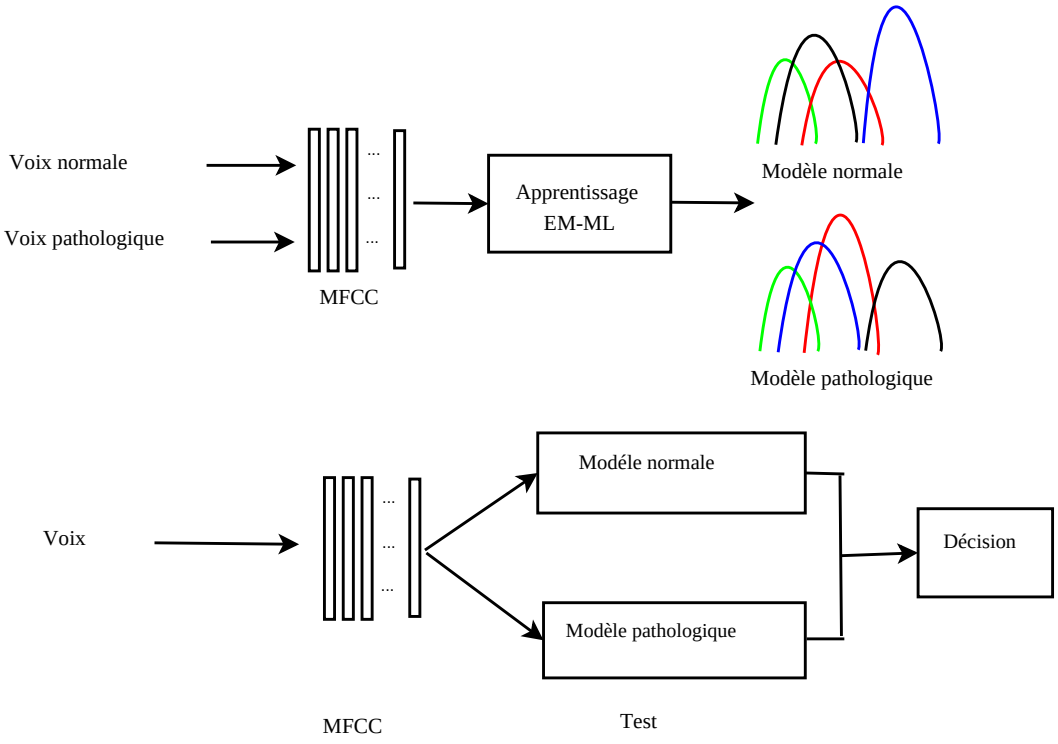

FIGURE 4.1 – Diagramme block d'un système de reconnaissance automatique basé sur les GMM

modèle afin de trouver le nombre optimale. Les valeurs possibles du nombre de mélanges de gaussiennes expérimentées appartiennent à l'ensemble 2, 4, 8, 16, 32, 64. Car un modèle GMM optimale est assuré en utilisant l'algorithme itérative EM-ML, l'initialisation et le nombre d'itération sont pris en considération.

Pour les SVM, le choix du noyau et ces paramètres est le facteur le plus important pour avoir un détecteur performant. En effet, le noyau RBF est le noyau le plus utilisé dans ce domaine. Ce noyau se décrit par deux parmètres : C est définit par le facteur de pénalité et *sigma* est définit par la largeur du noyau. Ces étapes sont illustrées dans la figure 4.1.

Pour la phase de test, nous calculons le logarithme de la vraisemblance de la nouvelle voix (matrice de paramètre) pour chacun des deux modèles (pathologique et normal). Nous disons que la voix appartient à un modèle P lorsque la valeur du logarithme de

la vraisemblance est maximale. Le logarithme de la vraisemblance pour le modèle d'une classe est calculé selon la formule (4.1) :

$$\log(X\backslash\lambda) = \sum_{n=1}^{N} \log(x_n\backslash\lambda) \tag{4.1}$$

$$P = argmax \log(X\backslash\lambda_i) \tag{4.2}$$

4.2.1 Corpus de données

Notre travail est effectué sur une base de données allemande, SVD (Saarbrcken Voice Database), qui présente certains avantages par rapport à MEEI . MEEI présente quelques inconvénients cités par [N.sanez 2006] :

- Plusieurs enregistrements sont affectés à un seul patient.
- Les enregistrements sont échantillonnés dans différentes fréquences.
- La base de données est créée dans diffèrents lieu (différents conditions).
- La durée d'enregistrement est différente surtout dans le cas des patients qui ne peuvent pas parler pour une longue durée.
- Chaque patient possède un seul enregistrement alors qu'il est commode d'avoir plusieurs enregistrements.

SVD est développée par Manfred putzer [1] au sein du laboratoire de la phonétique, université de saarland. Elle contient des fichiers audio pour des cas sains et pour des cas pathologiques comme suit :

- Des voyelles tenus /a/,/i/et /u/ prononcées dans différentes intonations (normale, basse, haute et basse-haute-basse) .
- Une phrase prononcée en allemand "Guten Morgen, wie geht es Ihnen ? " et qui veut dire " Bon jour, comment allez- vous ? ".
- Des signaux EGG (Electroglottographie).

Tous les fichiers sont aux format wav, échantillonnés à 50 Khz.

A cause de sa nouveauté, SVD n'est pas très reconnue par les chercheurs dans

1. HTTP//www.stimmdatenbank.coli.uni-saarland.de.

Genre	Nature	Apprentissage	Test	Age	Unité d'analyse
Homme	Normale	75	25	20-60	Phrase
	Pathologique	75	25		
Femme	Normale	75	25		
	Pathologique	75	25		

TABLE 4.1 – Corpus de données

le domaine de la détection et la classification de la voix pathologique. L'étude de [M.David 2012] a montré que SVD est mieux performante que MEEI.
Dans notre travail nous avons choisis deux classes : une classe normale et une classe pathologique. Pour la classe pathologique nous avons utilisé des patients qui souffrent de la dysphonie spasmodique, la laryngite et la dysodie. Ces maladies sont des maladies qui touchent le larynx. Toutes les données sont divisées en : données d'apprentissage et données de test. Chaque locuteur prononce une phrase comme le montre le tableau 4.1.

4.2.2 Pré-traitement

Il est nécessaire avant de faire toute analyse pour obtenir les caractéristiques acoustiques et les paramètres du signal de parole, de procéder à certaines opérations sur les fichiers audio. Tous les sons sous-échantillonnés à 16 Khz .
Un filtre dit de préaccentuation est appliqué sur tous les fichiers dont la fonction de transfert est donnée par :

$$H(z) = 1 - 0.97Z^{-1} \tag{4.3}$$

Le rôle de ce filtre est de réduire l'effet du microphone en amplifiant les amplitudes des hautes fréquences.

4.2.3 Extraction des paramètres

L'analyse des signaux est effectuée à l'aide d'une fenêtre Hamming de 20 ms, ces fenêtres glissantes d'une façon régulière avec un chevauchement de 50 %. Les coefficients MFCC sont calculés à partir d'un banc de 24 filtres. Le vecteur de

paramètres se compose de 39 coefficients incluant 12 premiers coefficients MFCC , l'énergie, leur dérivées premières et deuxièmes Ces paramètres ont été extraits en utilisant la fonction melcepst (voicebox).

Une illustration des coefficients MFCC est montrée dans les figures :

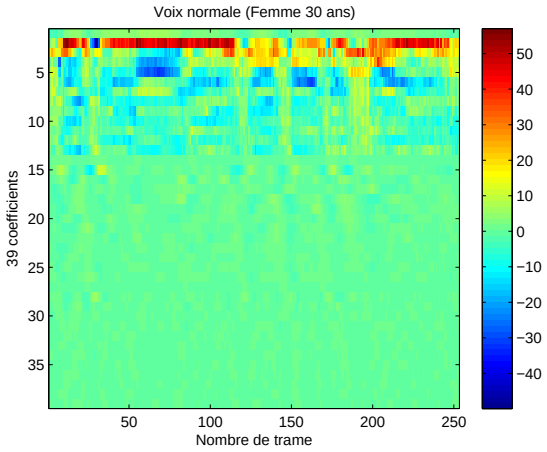

FIGURE 4.2 – Exemple de coefficients MFCC et leurs dérivées (39 coefficients), voix normale (Femme 30 ans)

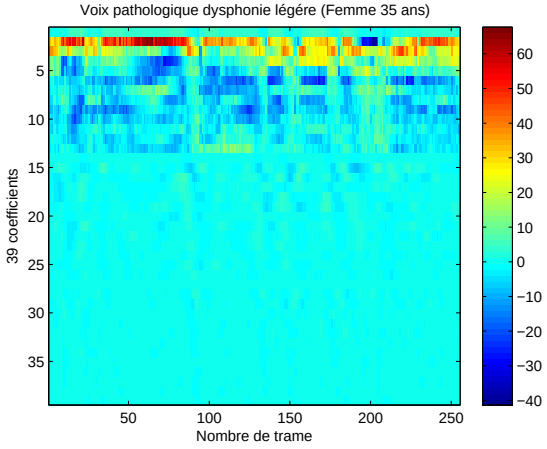

FIGURE 4.3 – Exemple de coefficients MFCC et leurs dérivées (39 coefficients), dysphonie légère (Femme 30 ans)

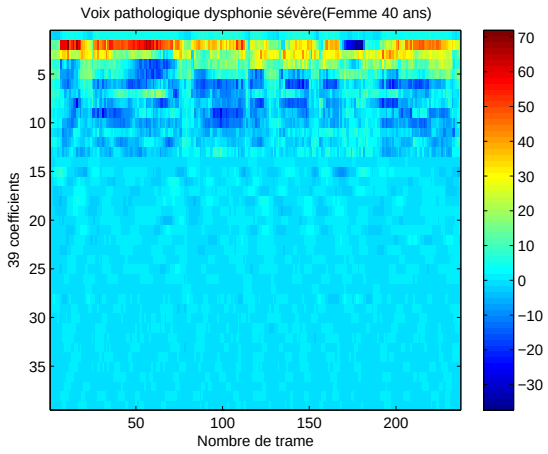

FIGURE 4.4 – Exemple de coefficients MFCC et leurs dérivées (39 coefficients), dysphonie sévère (Femme 30 ans)

4.3 Résultats et discussion

Les résultats sont présentés en deux groupes :

Groupe 1 : les résultats sont relatives à l'utilisation des GMM.

Dans ce groupe nous nous intéressons à tester l'effet de nombre des coefficients MFCC ainsi l'effet de nombre de gaussiennes définissant le modèle.

Les résultats sont représentés par les matrices de confusion données dans le tableau 4.2 et 4.3. Les figures 4.5, 4.6, 4.7 et 4.8 représetent respectivement le taux de spécificité et le taux de sensitivité pour hommes et femmes .

		Nombre de gaussiennes							
		8		**16**		**32**		**64**	
		P	**N**	**P**	**N**	**P**	**N**	**P**	**N**
13 coef	**P**	11	9	14	7	14	7	15	6
	N	14	16	11	18	11	18	10	19
26 coef	**P**	16	5	17	5	16	4	17	3
	N	9	20	8	20	9	21	8	22
39 coef	**P**	20	3	20	3	21	2	21	2
	N	5	22	5	22	4	23	4	23

TABLE 4.2 – Matrices de confusion : système MFCC-GMM (voix des hommes)

– Pour homme : A partir des deux figures 4.5 et 4.6 nous pouvons remarquer clairement que les deux perfomances sensitivité et spécificité s'améliorent lorsque

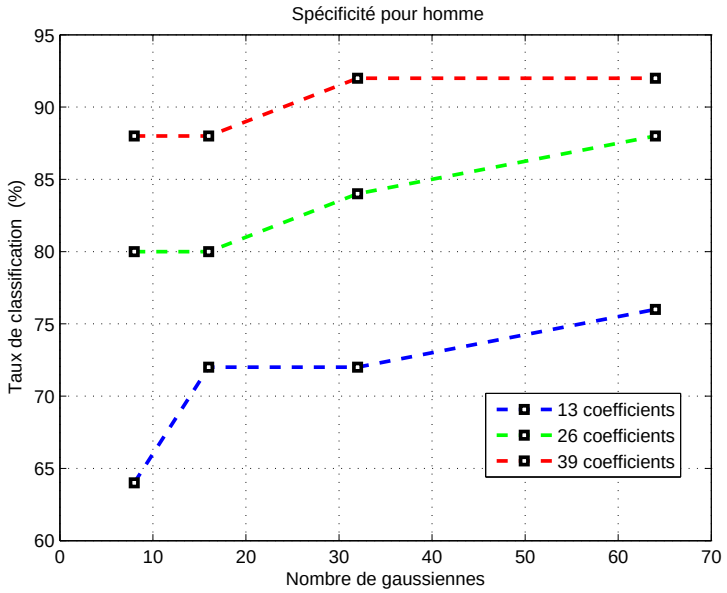

FIGURE 4.5 – Taux de reconnaissance (spécificité) en utilisant le système MFCC-GMM, voix des hommes

nous augmentons le nombre de gaussiennes deffinissant le modèle GMM et nombre de coefficients MFCC. Meilleur taux de reconnaissance est obtenu en utilisant 39 coefficients MFCC. La spécificité atteind les 92% et la sensitivité atteind les 84%.

Ces résultats nous permettent de dire que les coefficients dynamiques portent des informations supplémentaires sur la qualité de la voix, ces informations permettent d'établir la différence entre une voix normale et une voix pathologique.

Le deuxiéme facteur qui est le nombre de gaussiennes joue aussi un rôle trés important pour l'amélioration des performances du système. l'augmentation du nombre de gaussiennes (de 8 à 64) augmente le taux de reconaissance [Fezari 013]. Meilleures performances sont obtenues en utilisant 64 gausssiennes. cela veut dire qu'il fallait choisir un nombre de gaussiennes suffisant pour modéliser toutes les variations des coefficients MFCC qui font la différence entre les deux classes.

Selon ce que nous avons obtenu comme résultats nous constatons que le sys-

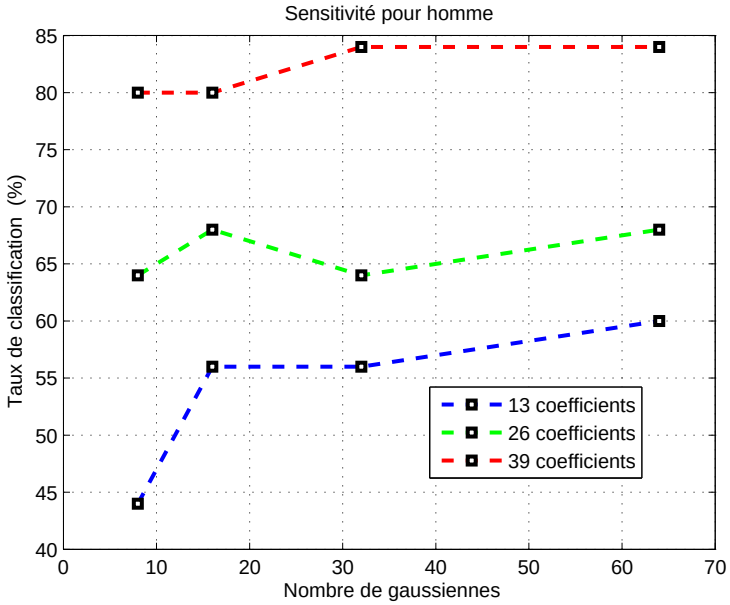

FIGURE 4.6 – Taux de reconnaissance (sensitivité) en utilisant le système MFCC-GMM, voix des hommes

tème MFCC-GMM est plus performant avec la classe normale qu'avec la classe pathologique dont il y a 8 % de différence (différence entre la spécificité et la sensitivité). C'est à-dire qu'il y'a des voix pathologiques considérées comme voix normales plus que des voix normales considérés comme pathologiques. Cela revient au degré de la dysphonie, quand la voix est légèrement altérée elle peut se détecter comme normale.

		Nombre de gaussiennes							
		8		**16**		**32**		**64**	
		P	N	P	N	P	N	P	N
13 coefficients	P	12	16	12	14	14	14	14	13
	N	13	9	13	11	11	11	11	12
26 coefficients	P	18	9	19	6	19	5	20	5
	N	7	16	6	19	6	20	5	20
39 coefficients	P	19	5	19	5	20	4	21	3
	N	6	20	6	20	5	21	4	22

TABLE 4.3 – Matrices de confusion : sysstème MFCC-GMM (voix de femmes)

– Pour femmes : Les résultats obtenus en utilisant les voix des femmes sont

trés similaires aux résultats obtenus avec les voix des hommes. Les deux performances sensitivité et spécificité s'améliorent lorsque nous augmentons le nombre de coefficients MFCC et le nombre de gaussiennes. Meilleurs taux sont obtenus en utilisant 39 coefficients modélisés par 64 gaussiennes.

Nous remarquons que les résultats obtenus en utilisant 13 coefficients sont trés faibles, ils atteignent 44% et 56% ; lorsque nous utilisons respectivement 64 gaussiennes, pour la spécificité et la sensitivité .

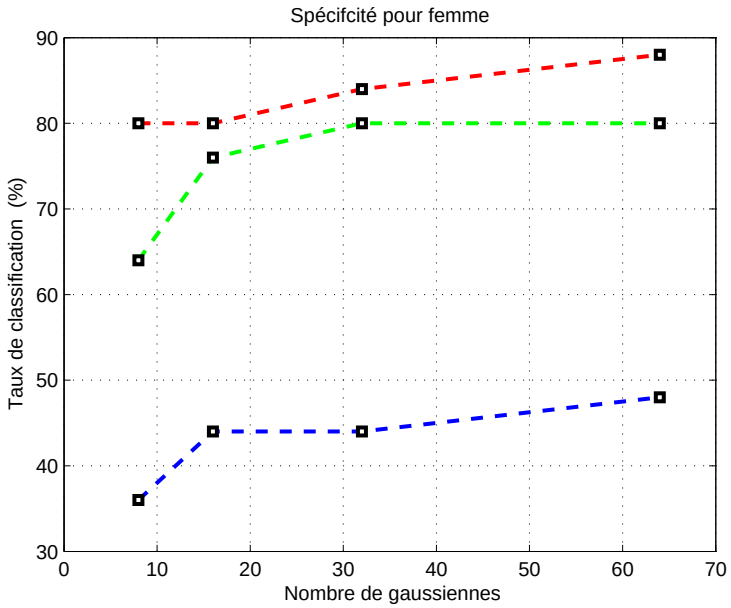

FIGURE 4.7 – Taux de reconnaissance(spécificité) en utilisant le système MFCC-GMM (voix de femmes)

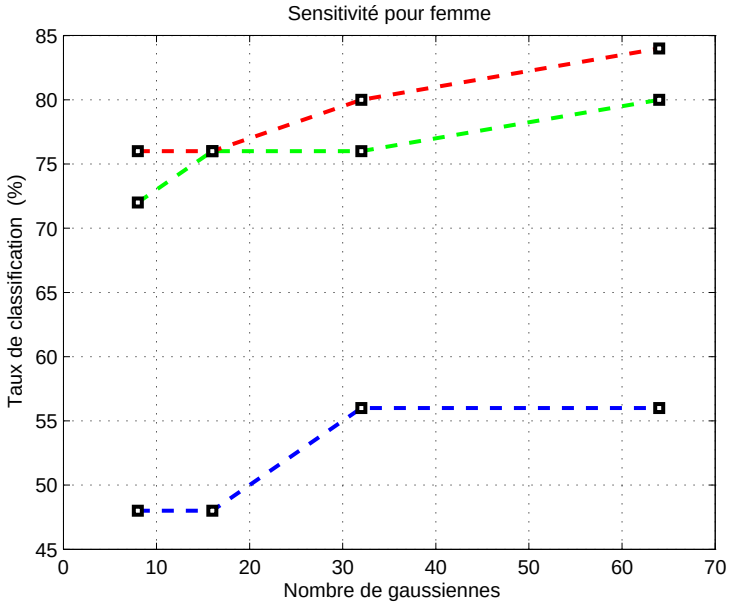

FIGURE 4.8 – Taux de reconaissance (sensitivité) en utilisant le système MFCC-GMM (voix de femmes)

Groupe 2 : les résultats sont relatives à l'utilisation du SVM. Dans ce groupe des expérimentations nous nous intéressons à tester d'une part l'effet des coefficients MFCC sur le taux de reconnaissance, d'autre part nous voulons trouver le noyau le plus performant. Le couple (C, σ) du noyau RBF est choisi en utilisant la méthode à recherche en grille (grid search).
Les résultats sont présentés par les matrices de confusion donnés par les tableaux 4.4 et 4.5. La sensitivité et la spécificité sont données dans les figures 4.9, 4.10, 4.11 et 4.12 pour les voix d'homme et les voix des femmes respectivement.

– Pour homme : La figure 4.9 nous montre la sensitivité de notre système basée sur les SVM, c-à-d les taux de reconnaissance de la voix normale. Nous remarquons clairement que le noyau RBF est le noyau le plus performant par rapport aux deux autres noyaux, et que le noyau plynomial est mieux que le noyau linéaire. Le meilleur taux est obtenu en utilisant 39 coefficients MFCC modélisés par le noyau RBF ($(c, \sigma) = (10^3, 10^{-3})$), le taux atteint 72%. Nous

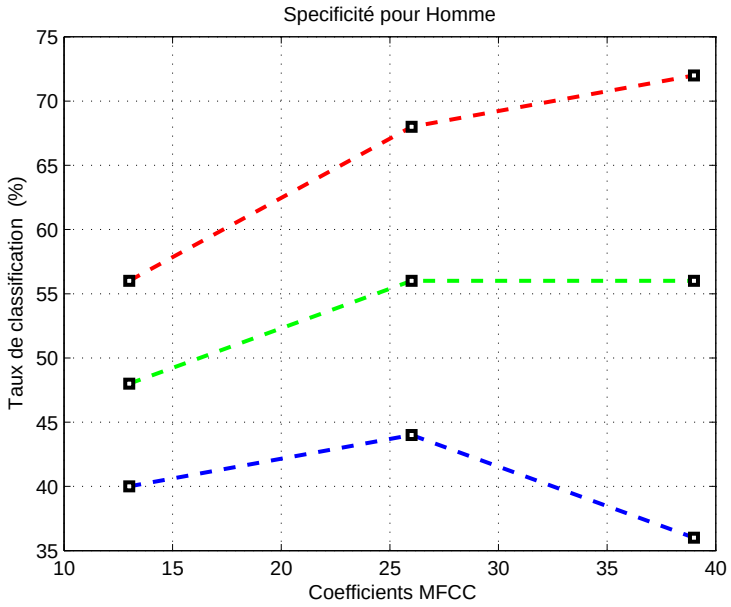

FIGURE 4.9 – Taux de reconnaisance (spécificité) en utilisant le système MFCC-SVM (voix d'hommes)

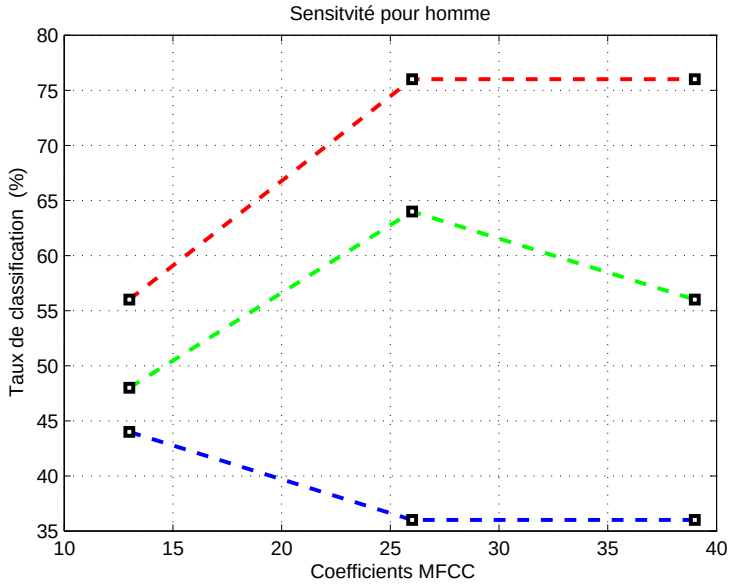

FIGURE 4.10 – Taux de reconnaissance (sensitivité) en utilisant le systèmeMFCC-SVM voix d'hommes

		Noyaux					
		Linéaire		Polynomial		RBF	
		P	N	P	N	P	N
13 coef	P	11	15	12	14	14	16
	N	14	10	13	11	11	9
26 coef	P	9	13	16	11	19	11
	N	16	12	9	14	6	14
39 coef	P	9	12	14	8	19	7
	N	16	13	11	17	6	18

TABLE 4.4 – Matrices de confusion : MFCC-SVM (voix des hommes)

remarquons qu'il y a toujours une amélioration lorsque nous augmentons le nombre de coefficients sauf dans le cas du noyau linéaire ou le taux s'abaisse lorsque nous utilisons 39 coefficients. Cette dimunition revient à la nature des données qui ne sont pas linéairement séparables surtout dans l'ordre de 39 coefficients.

Presque les mêmes remarques sont données à la spécificité, pour les deux noyaux linéaire et polynomial le taux s'abaisse lorsque nous augmentons l'ordre des coefficients. Meilleur taux est atteint lorsque nous utilisons le système 39 MFCC et le noyau RBF, nous avons obtenu 76%.

		Noyaux					
		Linéaire		Polynomial		RBF	
		P	N	P	N	P	N
13 coef	P	12	13	11	12	9	12
	N	13	12	14	13	16	13
26 coef	P	12	13	9	10	9	9
	N	13	12	16	15	16	16
39 coef	P	18	6	19	5	21	6
	N	7	19	9	20	4	19

TABLE 4.5 – Matrices de confusion pour le système MFCC-SVM (voix de femmes)

– Pour femmes : Ce que nous avons remarqué pour les voix des hommes est valable pour les voix des femmes. Meilleur taux de spécificité atteint 80% lorsqu'on utilise 26 coefficients avec le système à noyau RBF ($(c, \sigma) = (10^3, 2 * 10^{-3})$), par contre le taux de sensitivité atteint 84% en utilisant 39 coefficients .

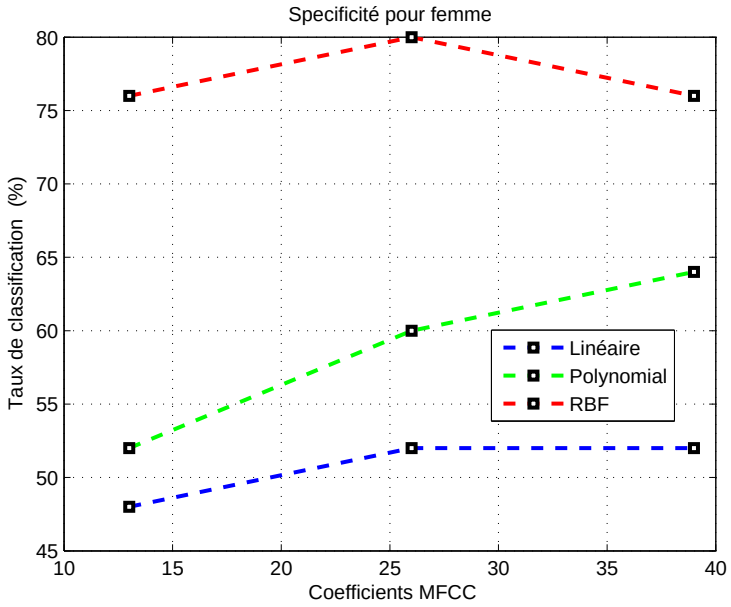

FIGURE 4.11 – Résultats de classification (spécificité) en utilisant MFCC-SVM voix de femmes

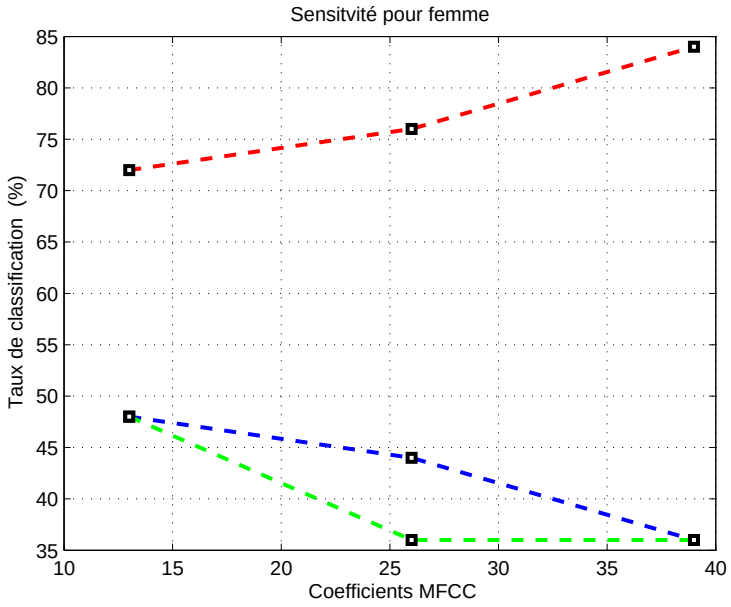

FIGURE 4.12 – Résultats de classification (sensitivité) en utilisant MFCC-SVM voix de femmes

A partir des performances obtenues nous pouvons dire que les caracteristiques dynamiques portent des informations trés critiques sur la qualité de la voix. Ces informations permettent d'établir la différence entre la voix normale est la voix pathologique [Emary].

Ces paramètres ne sont pas linéairement séparables, la chose qui explique les performances onbtenues lorsque nous utilisons le noyau RBF.

4.4 Conclusion

Dans ce chapitre nous avons présenté la méthodologie suivie pour développer les deux systèmes de reconnaissance automatique ; MFCC-GMM et MFCC-SVM. L'objectif principale est de trouver les meilleurs systèmes donnant les meilleures performances (sensitivité et spécificité). En d'autres mots, nous exploitons les capacités géneratives et les capacités discriminatives des GMM et des SVM respectivement. Les paramètres mis en jeu : nombre de coefficients MFCC est nombre de gaussiennes déffinissant le modèle GMM pour le premier système. Pour le deuxième système, nous cherchons le noyau qui présente le meilleur taux de reconaissance.
Nous avons étudié les performances des deux systèmes les plus reconnus dans le domaine de la classification de la voix pathologique avec la base de données SVD. Les résultats obtenus sont trés acceptables par rapport a ce qui existe dans l'état de l'art. Notre stratégie est de combiner les deux classificateurs afin d'exploiter les deux capacités. L'hybridation est le sujet du cinquième chapitre.

Chapitre 5

Méthodologie et expérimentation du système hybride GMM-SVM.

Contents

5.1 Introduction

Ce chapitre concerne la mise en oeuvre d'un système hybride pour la détection automatique de la voix pathologique à partir du signal de la parole. Cette hybridation consiste à combiner les deux classificateurs utilisés dans le chapitre 4 (GMM-SVM). Les GMM sont utilisés dans l'étape de modélisation (représentation statistique) par contre les SVM sont utilisés comme module de décision. Comme nous avons déja mentionner dans l'état de l'art, la majorité des travaux réalisés jusqu'a présent, les chercheurs s'interessent à séparer les centres des GMM appelés GMM supervecteurs en utilisant les noyaux de produit de probabilité tel que le noyau de bahattacharyya. Ce genre d'hybridation néglige l'importance des poids et de la matrice de covariance définissants le modèle GMM. Ces deux paramètres portent des informations cruciales permettant de mesurer la similarité entre deux modèles.

Dans ce chapitre, on aspire à bénificier de toute l'information statistique, des divergences permettant de mesurer la distance entre modèle GMM sont utilisées. Ces distances sont embarquées dans le noyau RBF (noyau à base de divergence) c-à-d la fonction de similarité est remplacée par ces divergences. Dans un premier temps, nous avons utilisé la divergence de Kullback-leibler et de Bahattacharyya addaptées avec les GMM. Sachant que ces divergences ne sont pas des distances métriques car elles ne respectent pas tous les axiomes déffinissants un métrique notamment l'inégalité triangulaire. Car il n'est pas evident de montrer l'impact de la violance de l'inégalité triangulaire dans le contexte de la classiffication, nous avons utilisé la version modifiée des deux divergences et qui verifient tous les axiomes afin de tester expérimentalement l'impact de cette inégalité.

Ce chapitre est divisé en deux parties essentielles une partie théorique et une partie expérimentale. Dans la première partie nous avons inroduit les notions de base sur les différents noyaux permettant de séparer les GMM et surtout les noyaux à partir de divergences (utilisés dans ce mémoire). La deuxième partie est résevée au développement des deux approches hybrides.

5.2 Fusion GMM-SVM

Dans La plupart des travaux les auteurs utilisent des approches ou bien génératives, ou bien discriminatives, si la combinaison est prise en considération c'est pour des raisons de comparaison et ce n'est pas pour améliorer les performances de ce système. Cette dernière est notre objectif dans ce travail.
Le système GMM-SVM existe sous deux formes : fusion parallèle et fusion séquentielle. dans ce mémoire nous nous intéressons à la deuxième méthode dont les GMM ne sont pas utilisés dans la phase de décision, ils sont utilisés pour la représentation statistique des parametres MFCC. Puis le noyau pour la SVM est considéré pour exploiter les informations contenues dans les modèles.
Comme nous avons parlé des noyaux, la séparation des GMM est assurée dans l'espace de redescription (feature space).
Dans ce qui suit nous introduisons les différents types de noyaux (la section 5.3 et 5.4 sont inspirée de la thèse [Louradour 2007]) :

5.3 Noyaux entre vecteurs

Les noyaux projectifs sont ceux qui peuvent s'exprimer en formulant toutes les opérations qui impliquent les variables d'entrée x et y sous forme du produit scalaire $x^T y$. Les noyaux radiaux sont ceux qui s'expriment comme une fonction monotone des distances entre vecteurs $\|x - y\|$. Nous rappelons que la distance euclidienne associée au produit scalaire est donnée par :

$$\|x - y\|^2 = x^T x - 2x^T y + y^T y \tag{5.1}$$

La condition pour qu'un noyau radial soit défini positif est donnée par le théorème de [Schoenberg, 1938], qui a été étendu par [Micchelli, 1986b] avec la formulation :
Théorème :
Une fonction $f : R^+ \mapsto R$ est dite absolument décroissante si elle est indéfiniment dérivable,

$$etsi \forall (n,t) \in N * R^{+}, \begin{cases} f^{n}(t) > 0 & \text{si n est pair} \\ f^{n}(t) < 0 & \text{si n est impair.} \end{cases} \tag{5.2}$$

si f est absolument décroissante, alors $k(x,y) = f(\|x-y\|^2)$ est un noyau de mercer. Par essence, les noyaux projectifs et les noyaux radiaux sont invariants aux rotations. Par contre, seuls les noyaux radiaux sont invariants aux translations, et aucun de ces noyaux n'est invariant aux homothéties. Enfin, notons que tous les noyaux cités dans cette partie vérifient les conditions de Mercer.

5.3.1 Noyaux linéaire

Le produit scalaire est le noyau le plus basique. Il peut être généralisé en incorporant une matrice de transformation linéaire des données dans l'espace de départ. On rencontre dans la littérature le terme de noyau linéaire généralisé. L'expression mathématique est de la forme :

$$k(x,y) = x^{T} R y \tag{5.3}$$

Où Rest une matrice définie positive, comme par exemple l'inverse d'une matrice de covariance. Une telle normalisation réduit les corrélations entre variables d'entrée et rend les variances unitaires (pour les nouvelles variables normalisées $R^{\frac{1}{2}}x$). Elle permet d'introduire une invariance aux transformations linéaires, pour une meilleure stabilité des méthodes à noyaux.

5.3.2 Noyaux Radiaux

Les noyaux radiaux correspondent à des fonctions représentantes $k(x_0, .)$ qui sont des fonctions centrées en x_0 avec un maximal local en ce centre. La plupart des noyaux radiaux sont bornés dans [0, 1], avec une valeur nulle ou pratiquement nulle atteinte lorsque deux points sont suffisamment éloignés (dans ce cas les noyaux radiaux non bornés tendent vers $-\infty$). La distance marquant la zone d'influence de chaque vecteur x_0 est paramétrée par un facteur positif ρ (homogéne à une distance entre vecteurs). Le noyau Gaussien est le noyau radial de loin le plus populaire. Il

est désigné par abus de language par le sigle RBF (Radial Basis Function) et est souvent paramétré dans la littérature par $\gamma = \frac{1}{2\rho^2}$.
Les noyaux radiaux correspondent à un espace de redescription de dimension $D = +\infty$,et permettent aux SVMs de construire des frontières particulièrement complexes. Notons aussi que ces noyaux ne nécessitent pas de normalisation sphérique étant donné qu'ils vérifient déjà $k(x, y) = cte$. Sachant que le noyau RBF est donné par :

$$k(x, y) = \exp(\frac{-\|x_1 - y_2\|^2}{\sigma^2}) \tag{5.4}$$

Ce noyau permet de mesurer la similarité entre les données, cette propriété est assurée au moyen de la fonction de similarité qui est en réalité une distance. Comme il est mentionné dans l'équation, cette fonction est une distance Euclidienne $\|x_1 - x_2\|^2$ Cette distance est peut être remplacée par une autre distance afin d'améliorer la capacité de séparation du fait qu'il y a des distances plus appropriées avec des données.

5.4 Noyaux entre densité de probabilité

Les noyaux entre densités de probabilité permettent d'étendre les algorithmes comme les SVMs pour traiter d'autres types d'entrées que les vecteurs de taille fixe. Ces fonctions s'appliquent à des distributions de données et permettent en pratique de manipuler des ensembles de tailles variables, pour divers types d'applications.

5.4.1 Noyaux de produit de probabilité

Les noyaux de produits de probabilités (Probability Product Kernel) sont définis selon [Jebara et Kondor, 2003] par la forme :

$$k_P^p(P_1, P_2) = \int_X p_1(x)^p p_2(x)^p dx \tag{5.5}$$

Où p est un degré strictement positif. Parmi ces noyaux, les plus communément utilisés sont :

- Le noyau de corrélation (P=1)

$$k_P^1(P_1, P_2) = \int_X p_1(x)p_2(x)dx \tag{5.6}$$

- Le noyau de Bhattacharyya (pour $p = \frac{1}{2}$)

$$k_P^B(P_1, P_2) = \int_X \sqrt{p_1(x)p_2(x)dx} \tag{5.7}$$

Ce noyau est considéré comme le produit scalaire de référence pour les fonctions positives, parmi lesquelles les densités de probabilités sont toutes de norme unitaire.

5.4.2 Noyaux à partir de divergence

Un grand nombre de mesures sont disponibles pour mesurer l'écart entre deux distributions. La table 5.1 liste les principales. A partir de ces mesures de divergences, il est possible de concevoir des noyaux, de la même manière que les noyaux radiaux qui sont construits à partir d'une distance. Typiquement, l'analogue d'un noyau vectoriel Gaussien avec l'astuce de *l'Exponential Embedding* :

$$D(p_1, p_2) \mapsto k(p_1, p_2) = \exp^{-\gamma D(p_1,p_2)^2} \tag{5.8}$$

Dans le cas général, le calcul des distances probabilistes n'est pas trivial, et requiert un algorithme d'approximation d'intégrale comme l'algorithme de Monte Carlo. Toutefois elles ont des expressions analytiques simples pour les familles de distributions exponentielles. Les expressions des principales distances entre deux modèles Gaussiens sont listées dans le tableau 5.2. Dans le cas particulier où $p_1$et p_2 sont des distributions Gaussiennes de même covariance Σ alors la divergence de Kullback symétrique (et la distance de Bhattacharyya modulo un facteur multiplicatif) se réduit à une distance de Mahalanobis au carré entre les vecteurs moyennes [Mahalanobis, 1936] c'est-à-dire :

$$\widetilde{D}_{KL}(N_1, N_2) = 8D^B(p_1, p_2) = (\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2) \tag{5.18}$$

Divergence de kullback Leibler	$D^{KL}(p_1 \parallel p_2) = \int_X p_1(x) \log \frac{p_1(x)}{p_2(x)} dx$	(5.9)
Divergence de Jensen-Shannon	$D^{JS}(p_1 \parallel p_2) = H[\alpha_1 p_1 + \alpha_2 p_2] - \alpha_1 H[p_1] - \alpha_2 H[p_2]$ avec $H[p] = \int_X p(x) \log(p(x)) dx$	(5.10)
Divergence de Rényi	$D_p^R(p_1 \parallel p_2) = \frac{1}{p-1} \log \int_X p_1^p(x) p_2(x)^{1-p} dx$	(5.11)
Distance de Bhattacharyya	$D^B(p_1 \parallel p_2) = -\log \int_X \sqrt{p_1(x) p_2(x)} dx$	(5.12)
Distance de Chernoff	$D^c(p_1 \parallel p_2) = -\log \int_X [p_1]^{\alpha_1} [p_2]^{\alpha_2} dx$	(5.13)
Distance de Hellinger	$D^H(p_1 \parallel p_2) = \sqrt{\int_X [\sqrt{p_1(x)} - \sqrt{p_2(x)}]^2 dx}$	(5.14)
Distance de Patrick-Fisher	$D^{PF}(p_1 \parallel p_2) = \sqrt{\int_X [\pi_1 p_1(x) - \pi_2 p_2(x)]^2 dx}$	(5.15)
Distance de Kolmogorov	$D^K(p_1 \parallel p_2) = \int_X \lvert \pi_1 p_1(x) - \pi_2 p_2(x)]^2 \rvert dx$	(5.16)
Distance de Lissack-Fu	$D^{LF}(p_1, p_2) = \int_X \lvert \Pi_1 p_1(x) - \Pi_2 p_2(x) \rvert^{\alpha_1} [\Pi_1 p_1(x) + \Pi_2 p_2(x)]^{\alpha_2} dx$	(5.17)

TABLE 5.1 – Quelques divergences.

Divergence de kullback Leibler	$\widetilde{D}_{KL}(N_{\mu_1,\Sigma_1}, N_{\mu_2,\Sigma_2}) = \frac{1}{2}(\mu_1-\mu_2)^T(\Sigma_1^{-1}+\Sigma_2^{-1})(\mu_1-\mu_2) + \frac{1}{2}tr(\Sigma_1\Sigma_2^{-1}+\Sigma_2\Sigma_1^{-1}-2I_d)$	(5.19)
Distance de Bhattacharyya	$D_B(N_{\mu_1,\Sigma_1}, N_{\mu_2,\Sigma_2}) = \frac{1}{4}(\mu_1-\mu_2)^T(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2) + \frac{1}{2}\log\frac{det(\Sigma_1+\Sigma_2)}{2\sqrt{(det\Sigma_1\Sigma_2)}}$	(5.20)
Distance de Chernoff	$D_C(N_{\mu_1,\Sigma_1}, N_{\mu_2,\Sigma_2}) = \frac{\alpha_1\alpha_2}{2}(\mu_1-\mu_2)^T(\alpha_1\Sigma_1+\alpha_2\Sigma_2)^{-1}(\mu_1-\mu_2) + \frac{1}{2}\log\frac{det(\alpha_1\Sigma_1+\alpha_2\Sigma_2)}{(det\Sigma_1)^{\alpha_1}(det\Sigma_2)^{\alpha_2}}$	(5.21)
Distance de Patrick-Fisher	$D_{PF}(N_{\mu_1,\Sigma_1}, N_{\mu_2,\Sigma_2}) = \frac{1}{2\sqrt{((2\pi)^d)}}((det^{\frac{1}{2}})+(det^{\frac{-1}{2}})) - \frac{2}{\sqrt{((2\pi)^d)det(\Sigma_1+\Sigma_2)}}\exp^{-(\mu_1-\mu_2)^T(\Sigma_1)^{-1}(\mu_1-\mu_2)}$	(5.22)

TABLE 5.2 – Version analytique de quelques distances

5.4.3 Noyaux dérivées de métriques Hilbertiennes

Hein et Bousquet et al. étendent les métriques Hilbertiennes semi-homogènes aux densités de probabilités. Ces métriques Hilbertiennes sont basées sur des distances définies négatives pour les réels positifs, paramétrées par $\alpha \in [1, +\infty]$ et $\beta \in [-\infty, -1] \bigcup [\frac{1}{2}, \alpha]$ selon la formule :

$$\forall\{x,y\} \in (R^+), d^2_{\alpha,\beta}(x,y) = \frac{2^{\frac{1}{\beta}}(X^\alpha + Y^\alpha)^{\frac{1}{\alpha}} - 2^{\frac{1}{\alpha}}(X^\beta + Y^\beta)^{\frac{1}{\beta}}}{2^{\frac{1}{\alpha}} - 2^{\frac{1}{\beta}}} \tag{5.23}$$

Pour les densités de probabilités, à valeurs réelles positives, cette notion est généralisée pour donner la distance au carré entre densités :

$$D^2_{\alpha,\beta} = \int_X d^2_{\alpha,\beta}(p_1(x), p_2(y))dx \tag{5.24}$$

On peut alors utiliser cette nouvelle mesure de distance avec un Exponential Embedding comme dans la sous-section précédente. On peut aussi concevoir des noyaux de type "projectif", en remarquant que pour certains choix de α et β, on peut déterminer une expression analytique du produit scalaire correspondant à la distance Hilbertienne. On peut y voir que le noyau de corrélation est un cas particulier pour $\alpha = \frac{1}{2}$ et $\beta = 1$.

5.5 Distances utilisées

5.5.1 Kullback-Leibler

En théorie des probabilités et en théorie de l'information, la divergence de K-L, est une mesure de dissimilarité entre deux distributions de probabilités P et Q. Elle doit son nom à Solomon Kullback (en) et Richard Leibler.
Pour deux distributions de probabilités discrètes P et Q la divergence de Kullback-Leibler de Q par rapport à P est définie par :

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{5.25}$$

Pour des distributions P et Q continues on utilise une intégrale :

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)dx} \tag{5.26}$$

Bien que perçue souvent comme une distance, elle n'en remplit pas les conditions : elle n'est pas symétrique et ne respecte pas l'inégalité triangulaire. Il est recommandé d'utiliser la version symétrique donnée par :

$$D_{KL}(P \parallel Q) = \left| \frac{1}{2} \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} dx + \frac{1}{2} \int_{-\infty}^{+\infty} q(x) \log \frac{q(x)}{p(x)} dx \right| \tag{5.27}$$

Nous utilisons l'approximation de monte- carlo (MCS) donnée par :

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} dx \approx \frac{1}{N} \sum_{i=1}^{N} log \frac{p(x_i)}{q(x_i)} \tag{5.28}$$

La forme finale de la version symétrique est donnée par :

$$D_{KL}(P \parallel Q) = \left| \frac{1}{2N} \sum_{x \mapsto p}^{N} log(p(x)) - \frac{1}{2N} \sum_{x \mapsto p}^{N} log(q(x)) + \frac{1}{2N} \sum_{x \mapsto q}^{N} log(q(x)) - \frac{1}{2N} \sum_{x \mapsto q}^{N} log(p(x)) \right| \tag{5.29}$$

5.5.2 Bhattacharyya

En statistiques, la distance de Bhattacharyya est une mesure de la similarité de deux distributions de probabilités discrètes. Elle est reliée au coefficient de Bhattacharyya, qui est une mesure statistique du recouvrement de deux ensembles d'échantillons. Cette mesure est la plus utilisée pour la mise en correspondance entre deux observations basées sur l'histogramme de couleur. Cette mesure est régulièrement utilisée dans des problèmes de classification, en particulier dans le domaine de la vision par ordinateur.

Le nom de la distance et du coefficient proviennent du statisticien indien A. Bhattacharya , qui travaillait dans les années 1930 à l'Institut indien de statistiques. Le coefficient peut être utilisé pour déterminer la proximité relative des deux ensembles considérés. Il est utilisé pour mesurer la séparabilité de classes en classification automatique.

Pour deux distributions de probabilités discrètes p et q définies sur le même espace de probabilité, la distance de Bhattacharyya est calculée par :

$$D_B(p, q) = -ln(BC(p, q)) \tag{5.30}$$

BC est le coefficient de Bhattacharyya

Pour des distributions de probabilités continues, le coefficient est défini par :

$$BC(p, q) = \int sqrt(p(x)q(x))dx \tag{5.31}$$

Dans les deux cas, $0 < BC \leq 1$ et $0 \leq DB \leq 1$. La distance de Bhattacharyya n'obéit pas à linégalité triangulaire.

Pour deux distributions gaussiennes :

$$D_B = \frac{1}{8}(\mu_1 - \mu_2)^T P^{-1}(\mu_1 - \mu_2) + \frac{1}{2}\ln(\frac{detP}{\sqrt{(detPdetQ)}}) \tag{5.32}$$

où m_i et p_i sont les moyennes et les covariances des distributions, et $P = \frac{p_1+p_2}{2}$
Cette écriture montre que dans le cas gaussien, le premier terme de la distance de Bhattacharyya est relié à la distance de Mahalanobis.

5.5.3 Impact de l'inégalité triangulaire

Une distance est appelée un métrique lorqu' elle vérifie les axiomes suivants :

- Séparation $d(x, y) \geq 0$.
- Coincidence $d(x, y) = 0$ si et seulement si $x = y$.
- Symétrique $d(x, y) = d(y, x)$.
- Inégalité traingulaire $d(x, z) \leq d(x, y) + d(x, z)$.

Les trois premiers axiomes séparation, coincidence et symétrique sont vérifiés par la plupart des distances et divergenges. Cepandant, la propriété de l'inégalité triangulaire est rarement verifiée bien qu'elle joue un rôle cruçiale dans le domaine de la reconnaissance automatique.
Lorsqu' une distance qui n'obéit pas à l'inégalité triangulaire est considérée pour mesurer la similarité, les objets peuvent être vus comme proches et qui sont en réalité loin et vice verça.
Notre hypothése est la suivante : ce que nous pouvons dire pour une distance peut se dire pour le mapping en utilisant le noyau RBF. Cest-à-dire, si la distance embarquée dans le noyau n'obéit pas à cet axiome le noyau résultant aussi n'obéit pas à cet axiome. La chose qui peut etre en cause pour dégrader les performances du système.

5.5.4 Nouvelles versions

Comme nous avons mentionné, la divergence de KL et de Bhattacharyya n'obéit pas à l'inégalité triangulaire. Dans l'étude de Karim.T [T.Karim 2012] ces distances sont modifiées afin de vérifier cette condition. Cette modification est présentée briè-

vement :

La divergence de KL et la distance de Bhattacharyya sont exprimées dans le tableau 5.2 sous la forme analytique. Les deux équations sont divisées en deux termes, le premier terme mesure la différence entre les moyennes pondérées avec la matrice de covariance. Tandis que le deuxième terme mesure la différence entre les matrices de covarriance.

Il est clair que : si $\Sigma_1 = \Sigma_2$ les deux équations s'expriment comme suit :

$$\begin{aligned} KL(P \parallel Q) &= (\mu_1 - \mu_2)^T(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2) \\ D_B &= \tfrac{1}{8}(\mu_1 - \mu_2)^T(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2) \end{aligned} \tag{5.33}$$

si $\mu_1 = \mu_2$ Les deux équations sont réduites à :

$$\begin{aligned} D_{KL} &= \tfrac{1}{2}tr(\Sigma_1\Sigma_2^{-1} + \Sigma_2\Sigma_1^{-1} - 2I_d) \\ D_B &= \tfrac{1}{2}\log\frac{det(\Sigma_1+\Sigma_2)}{2\sqrt{(det\Sigma_1\Sigma_2)}} \end{aligned} \tag{5.34}$$

Cette séparation facilite la modification, pour le premier terme les auteurs prennent la racine carré, le deuxième terme est remplacé par la distance de Reimman donnée par :

$$d_R(\Sigma_1, \Sigma_2) = \left(\sum_{j=1}^{p} log\lambda_j\right)^{\frac{1}{2}} \tag{5.35}$$

λ est la valeur propre.

Les deux nouvelles distances s'expriment comme suit :

$$\begin{aligned} D_{KL}(P,Q) &= \tfrac{1}{2}(\mu_1 - \mu_2)^T(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2) + d_R(\Sigma_1, \Sigma_2) \\ D_B(P,Q) &= \tfrac{1}{8}(\mu_1 - \mu_2)^T(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2) + d_R(\Sigma_1, \Sigma_2) \end{aligned} \tag{5.36}$$

Le coefficient β est utilisé pour la pondération des deux termes :

$$\begin{aligned} D_{KL}(P,Q) &= \beta\tfrac{1}{2}(\mu_1 - \mu_2)^T(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2) + (1-\beta)d_R(\Sigma_1, \Sigma_2) \\ D_B(P,Q) &= \beta\tfrac{1}{8}(\mu_1 - \mu_2)^T(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2) + (1-\beta)d_R(\Sigma_1, \Sigma_2) \end{aligned} \tag{5.37}$$

Où $\beta \in [0, 1]$

5.5.5 Adaptation avec les GMM

Les distances proposées mesurent la similarité entre les distributions gaussiennes. Dans l'étude de G.Sfikas [G.Sfikas 2005] ces distances sont adaptées avec des modèles GMM.

Pour la distance de Kullback-Leibler :

$$D_{KL}(GMM1, GMM2) = \sum_{i=1}^{n} \sum_{j=1}^{m} \prod_{i} \prod_{j} D_{KL}(P, Q) \tag{5.38}$$

Pour la distance de Bhattacharyya :

$$D_{B}(GMM1, GMM2) = \sum_{i=1}^{n} \sum_{j=1}^{m} \prod_{i} \prod_{j} D_{B}(P, Q) \tag{5.39}$$

Remarque : les deux distances KL et bhattacharyya et leurs versions modifiées sont utilisées dans la deuxième approche.

5.6 Approches hybrides GMM-SVM proposées :

Dans cette section nous allons décrire les différents systèmes que nous avons développé. Ces systèmes sont des systèmes hybrides GMM-SVM.

5.6.1 Protocole expérimental :

Dans ce chapitre, nous avons suivi le même protocole expérimental dont nous avons utilisé le même corpus de données, les même paramétres (13,26,39 coefficients MFCC).

5.6.2 Premiére approche

Dans cette approche nous nous intéressons à séparer les GMM supervecteurs en utilisant des SVM.[F.AMARA 2013] Une fois nous avons obtenu les modèles GMM nous passons à la construction des vecteurs d'entrée des SVM.

Construction des GMM supervecteurs :

Les GMM supervecteurs ce ne sont que les centres des gaussiennes, aprés ces extractions nous les concaténons dans un vecteur. La taille de ce vecteur est de M*D, M est le nombre de gaussienne déffinissant le modèle GMM. D représente la dimension de l'espace des paramétres MFCC. Supposons que nous utilisons 39 coefficients MFCC (12 premiers coefficients + energie + $\Delta + \Delta\Delta$) modélisés par 64 gaussiennes, la taille du vecteur sera donc de 64*39.

Modélisation :

Chaque classe (normale ou pathologique) est représentée par un vecteur. Puis nous passons à la modélisation en utilisant la SVM. Le noyau RBF est le noyau le plus utilisé, ce noyau est donné par $k(x, y) = \exp(\frac{-\|x_1-y_2\|^2}{\sigma^2})$. La fonction $-\|x_1 - y_2\|^2$ dite fonction de similarité est en réalité une distance euclidienne. Cette distance est l'une des plus importantes distances utilisées dans le domaine de la reconnaissance automatique de la parole. La projection dans un espace de plus grande dimension et la possibilité de rendre les données linéairement séparables est basée sur l'éfficacité de cette distance à mesurer la similarité entre ces données spécifiques. Afin

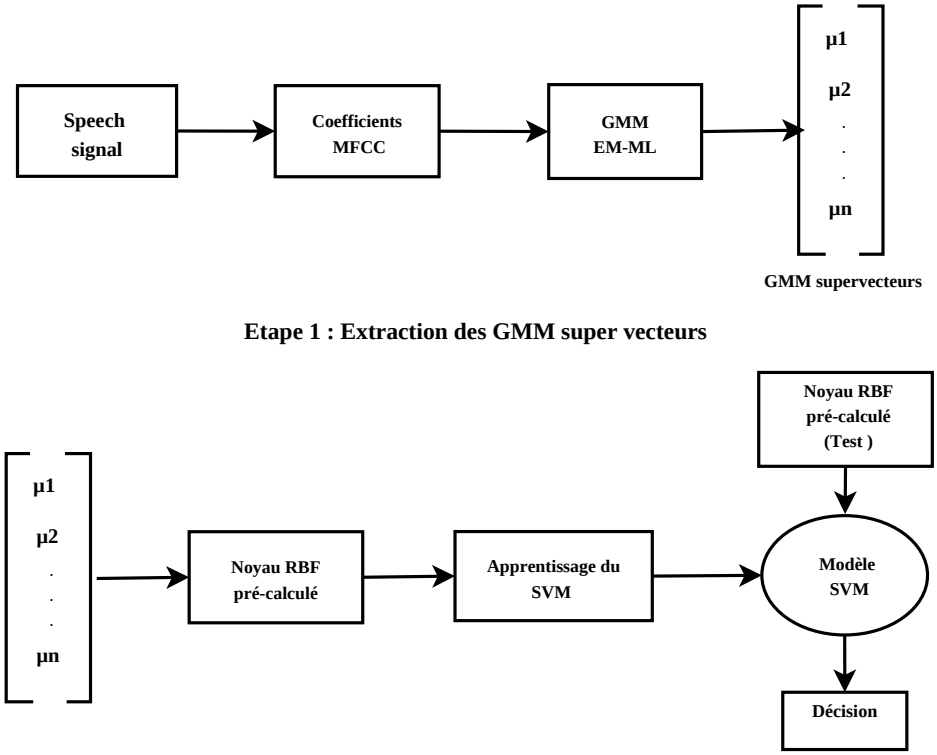

FIGURE 5.1 – Diagramme block du systéme hybride GMM-SVM.

de tester son éfficacité, cette fontion est exploitée en utilisant différentes distances (Euclidienne, Euclidienne standarisée et Mahalanobis). Pour atteindre cette objectif, nous avons calculé le noyau préalablement (precomputed kernel). Nous commençons par le calcule de la matrice distance (distance entre GMM superveteur) puis nous remplaçons cette matrice dans le noyau. Ce noyau sera l'entrée du SVM. Le couple (c, σ) est choisi par la méthode de recherche en grille.

Pour la validation croisée les experimentations sont répétées 10 fois. Les taux présentés dans les tableaux sont les moyennes du 10 répétitions.

Test :

Dans la phase de test, chaque locuteur (sain ou malade) est modélisé séparément. De la même façon, nous calculons le noyau de test. Les différentes étapes sont résumées dans la figure 5.1.

Distances	Paramètres	Spécificité (%)	Sensitivité(%)	Exactitude(%)
Euclidienne	13 coef	68	64.8	66.40
	26 coef	72.40	76.40	74.40
	39 coef	83.2	84	83.6
Euclidienne standarisée	13 coef	65.2	72	68.6
	26 coef	76.2	72.8	74.5
	39 coef	82	80.4	81.5
Mahalanobis	13 coef	72	80	76
	26 coef	76	78	74
	39 coef	89.6	87.2	**88.4**

TABLE 5.3 – Taux de detection du système GMM-SVM (voix d'hommes)

Résultats et interprétation :

Pour les hommes :

La première lecture du tableau 5.3 montre que le système GMM-SVM présente les meilleures performances lorsque nous utilisons la distance Mahalanobis pour séparer les GMM supervecteurs. L'exactitude atteint 88.4%. La distance Euclidienne et la distance Euclidienne standarisée donnent presque les mêmes résultats dont l'exactitude atteint respectivement 83.6% et 81.5%. Nous pouvons remarquer aussi que l'augmentation du nombre des coefficients améliore le taux de detection.
La distance de Mahalanobis ce n'est que la distance Euclidienne pondérée par la matrice de covariance, la chose qui explique l'efficacité de cette distance dans ce contexte.

Pour les femmes :

Le tableau 5.4 montre les résultats du système hybride GMM-SVM en utilisant les voix de femmes. Meilleure exactitude est obtenue lorsque nous utilisons la distance de Mahalanobis , le taux avoisine les 85%. Cette distance nous permet d'avoir une amélioration de 2.6 % et 4.7% par rapport à la distance Euclidienne et Euclidienne standarisée respectivement. Les deux premières distances présentent pratiquement les mêmes résultats.
Si nous comparons les performances du système en utilisant des voix d'hommes et

Distances	Paramètres	Spécificité (%)	Sensitivité(%)	Exactitude(%)
Euclidienne	13 coef	69.6	67.2	68.4
	26 coef	72	75.6	73.6
	39 coef	83.2	82	82.6
Euclidienne standarisée	13 coef	68	72.20	70.10
	26 coef	73.6	72	72.8
	39 coef	81.8	79.2	80.50
Mahalanobis	13 coef	80	77.2	78.6
	26 coef	80	83.6	81.8
	39 coef	85.6	84.8	**85.2**

TABLE 5.4 – Taux de detection du système GMM-SVM (voix de femmes)

des voix de femmes, nous remarquons que le système se comporte de la même façon avec les deux genres.

5.6.3 Deuxième approche

Une fois que nous avons obtenu les modèles GMM nous calculons la matrice distance. Cette matrice contient les distances entre modèles GMM, dans la phase d'apprentissage cette matrice est calculée entre les modèles d'apprentissage (training VS training). La taille de cette matrice et de nxn dont n et le nombre de modèles (chaque patient est représenté par un modèle). Par contre dans la phase de test la matrice est calculée entre les modèles de test et les modèles d'apprentissage (testing VS training). La taille de cette matrice et de nxm, m est le nombre des modèles de test.

Comme nous l'avons déja mentionné cette partie est assurée en utilisant les divergences de Kullback-leibler et bnattacharyya ainsi que leurs versions modifiées.

La deuxième étape concerne le calcule du noyau RBF (precomputed kernel), le noyau préalablement calculé nous permet d'embarquer la distance voulue. Le noyau s'exprime comme suit :$k(x, y) = \exp(\frac{-D^2}{\sigma^2})$. D représente la matrice distance. Nous obtenons une matrice qui a la même taille que la matrice distance. La figure 5.2 résume toutes les étapes.

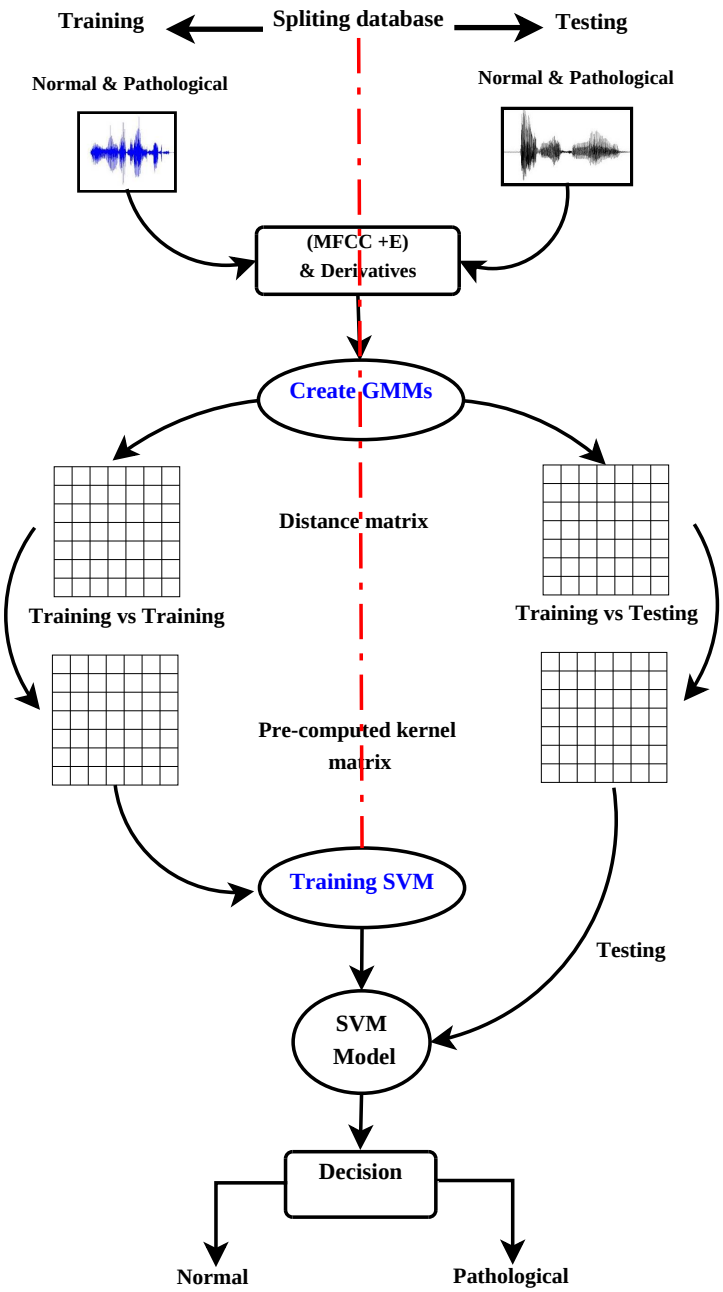

FIGURE 5.2 – Diagramme block du système hybride GMM-SVM basé sur les distances de Kullback-Leibler et Bhattacharyya.

Les paramètres du noyau RBF c, σ sont toujours calculés de la même façon (recherche en grille). La distance de KL et de Bhattacharyya utilisées sont les versions pondérées données dans l'équation (137), le coefficient β prend ces valeurs dans l'interval [0 1]. Afin de tester l'efficacité du système, les experiences sont répétées 10 fois (validation croisée).

Résultats et discussion

Les résultats sont relatifs à l'utilisation des deux distances et leurs versions modifiées :

Distance de Kullback-Leibler :

Pour les hommes :

La première lecture du tableau 5.5 nous permet de remarquer que les performances du système s'améliorent lorsque nous utilisons la nouvelle version de la distance de KL et lorsque nous augmentons le nombre de coefficients MFCC. Pour le nombre de coefficients il n'y a pas de nouveauté car nous connaissons à partir du quatrième chapitre que les caractéristiques dynamiques portent des informations supplémentaires permettant d'améliorer la discrimination entre les voix pathologiques et les voix normales. Le meilleur taux est obtenu en utilisant KL modifié avec 39 coefficients, l'exactitude atteint 98% où il y'a une amélioration de 6.2% par rapport à l'ancienne version utilisée avec le même nombre de coefficients. Cela nous permet de dire que la projection des données en utilisant le noyau RBF s'influence par la propriété de l'inégalité triangulaire. Autrement dit, lorsque la distance embarquée respecte cette axiome, le noyau conséquent le respecte aussi.

La deuxième remarque nous conduit à dire que notre système est plus performant avec la voix normale qu'avec la voix pathologique dont nous remarquons les taux de spécificité sont plus élevés que les taux de sensitivité. Prenons l'exemple de KL modifiée avec 39 coefficients, la spécificité atteint 100% par contre la sensitivité n'atteint que 96%. Cela revient au degré de dysphonie, si les voix sont légérement altérées il devient difficile de les séparer des voix normales.

Si nous comparons ce système avec celui qui utilise uniquement les centres de gaussiennes, nous trouvons une amélioration importante, cela confirme que les poids et les matrices de covariance apportent des informations supplémentaires qui per-

Divergences	Paramètres	Spécificité (%)	Sensitivité(%)	Exactitude(%)
KL-MCS [V.Evaldas 2012]	13 coef	68,80	64	66.40
	26 coef	72.40	76.40	74.40
	39 coef	93.6	90	**91.8**
K-L proposée	13 coef	72.20	68.4	70.30
	26 coef	84	80	82
	39 coef	100	96	**98**

TABLE 5.5 – Taux de detection du système GMM-SVM en utilisant la divergence de Kullback-Leibler (voix d'hommes)

mettent d'améliorer les performances du système.

Les courbes roc affirment l'efficacité du système, l'aire sous la courbe (AUC) atteind 0.97 lorsque nous utilisons la version modifiée du KL.

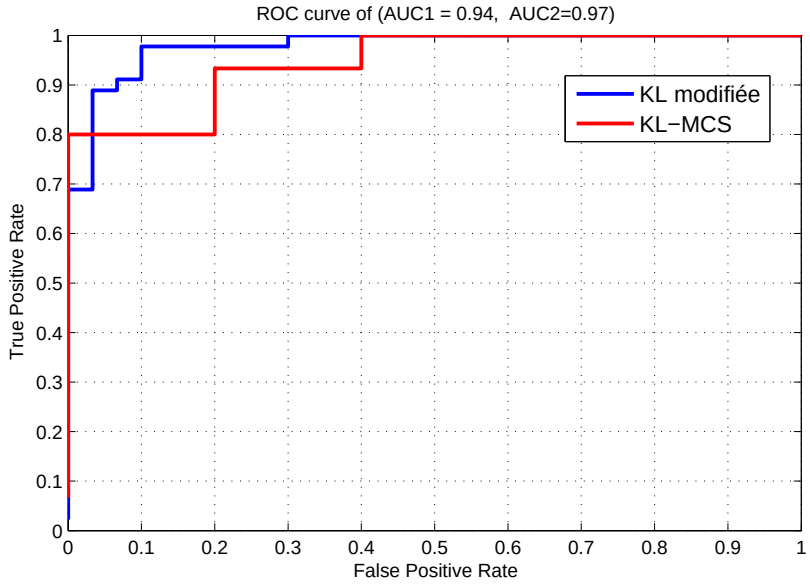

FIGURE 5.3 – Courbe ROC en utilisant la distance de KL (voix d'hommes)

Pour les femmes :

Ce que nous avons remarqué pour la voix des hommes est pratiquement le même par rapport à la voix des femmes. A partir du tableau 5.6 les meilleurs taux sont

Divergences	Paramètres	Spécificité (%)	Sensitivité(%)	Exactitude(%)
KL-MCS [V.Evaldas 2012]	13 coef	64.80	72	68.40
	26 coef	72.40	68	70.20
	39 coef	95.20	93.80	**94.50**
K-L proposée	13 coef	64	68	66
	26 coef	80	83.2	81.6
	39 coef	98	96	**97**

TABLE 5.6 – Taux de detection du système GMM-SVM en utilisant la divergence de Kullback-Leibler (voix de femmes)

obtenus en utilisant la nouvelle version du KL en augmentant le nombre de coefficients. Meilleure exactitude atteint 97% où il y a 2.5 % d'amélioration par rapport à l'anciennce version utilisée avec le même nombre de coefficients (39 coefficients). Nous remarquons aussi que notre système est plus performant avec la classe normale qu'avec la classe pathologique (taux de spécificité elevé par rapport aux taux de sensitivité). AUC atteint 0.92 et 0.97 respectivement pour KL et KL modifiées.

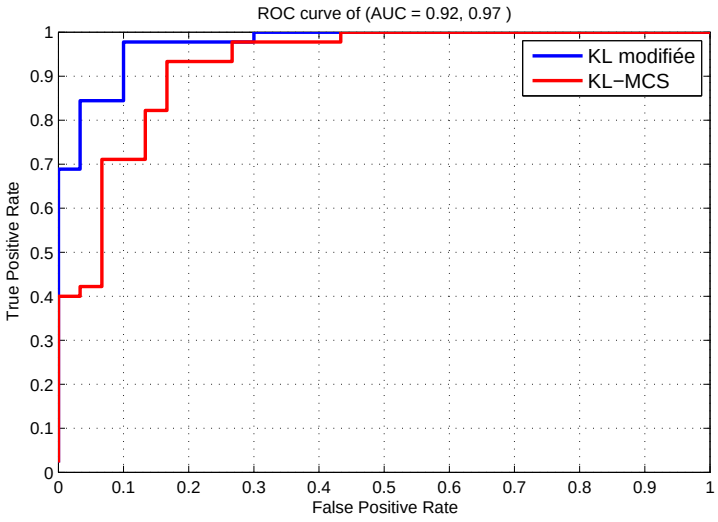

FIGURE 5.4 – Courbe ROC en utilisant la distance de KL (voix de femmes)

Distance de Bhattacharyya :

Pour hommes :

En utilisant la distance de Bhattacharyya, nous remarquons que la version modifiée respectant l'inégalité triangulaire présente les meilleures performances, le tableau 5.7 nous montre que l'exactitude atteint les 90 % et 97.50 % (avec 39 coefficients) respectivement pour Bhattacharyya et sa version modifiée.

Une autre remarque trés importante est tirée de ce tableau c'est que ce système est plus performant que le premier avec les coefficients MFCC à ordre faible surtout pour 13 coefficients. AUC atteint 0.98 en utilisant Bhattacharyya modifiée.

Distance	**Paramètres**	**Spécificité (%)**	**Sensitivité(%)**	**Exactitude(%)**
Bhattacharyya	13 coef	73.20	75.40	74.30
	26 coef	76	80	78
	39 coef	92	88	**90**
Bhattacharyya proposée	13 coef	78	78	78
	26 coef	84	82	83
	39 coef	99.2	95.8	**97.50**

TABLE 5.7 – Taux de detection du système GMM-SVM en utilisant la divergence de Bhattacharyya (voix d'hommes)

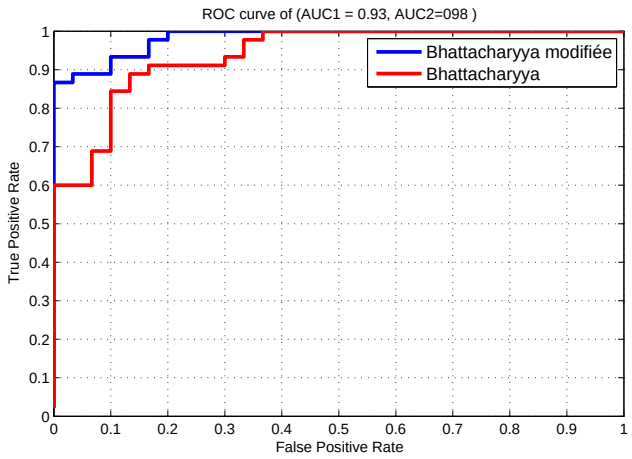

FIGURE 5.5 – Courbe ROC en utilisant la distance de Bhattacharyya (voix d'hommes)

Pour femmes :

Le tableau 5.8 présente les résultats obtenus en utilisant la distance de Bhattacharyya via le système hybride GMM-SVM. La distance modifiée donne les meilleures performances sauf que nous pouvons remarqué que ce système est moins performant par rapport aux voix des hommes. Nous avons obtenu AUC=0.93 et AUC=0.95 respectivement en utilisant Bhattacharyya et Bhattacharyya modifiée.

Distance	**Paramètres**	**Spécificité (%)**	**Sensitivité(%)**	**Exactitude(%)**
Bhattacharyya	13 coef	68	72	70
	26 coef	72	75.20	73.60
	39 coef	92	92	**92**
Bhattacharyya proposée	13 coef	72	68	70
	26 coef	85	86	85.5
	39 coef	95.20	94.80	**95**

TABLE 5.8 – Taux de detection du système GMM-SVM en utilisant la divergence de Bhattacharyya (voix de femmes)

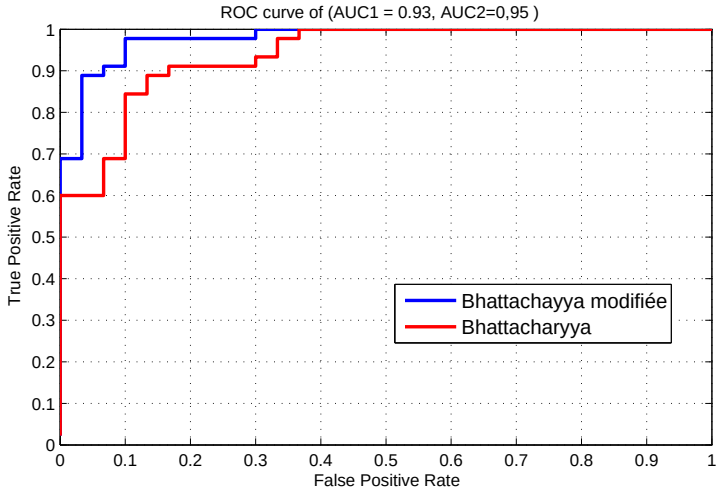

FIGURE 5.6 – Courbe ROC en utilisant la distance de Bhattacharyya (voix de femmes)

5.7 Conclusion

Dans ce chaptire nous avons présenté la méthodologie suivie pour développer le système hybride GMM-SVM. Les deux approches développées sont basées sur la capacité du SVM à séparer des GMM. Dans les deux approches nous étions penchés à trouver la meilleure distance à embarquer dans le noyau RBF afin d'améliorer les performances du système. Les expérimentations montrent qu'il y a la possibilité d'apporter d'autres améliorations dans les performances à travers le choix des paramètres et des distances, qui seront discutés dans la section futurs travaux.

Conclusion générale

Conclusions

Dans cette thèse, nous avons développé un système de détection automatique de la voix pathologique à partir du signal de la parole. Ce système joue le role d'un moyen de diagnostic complémentaire aux deux méthodes excitantes (l'évaluation perceptuelle et instrumentale).

Dans la première partie de cette thèse, nous avons développé deux systèmes bâtis autour des traits de type MFCC et leurs dérivés extraits au niveau d'une trame analysés à l'échelle d'un énoncé. Ensuite ces paramètres sont modélisés par un modèle de mélange de gaussiennes pour le premier système et par les SVM pour le deuxième système. Notre objectif était de tester l'efficacité des classificateurs qui se caractérisent par les capacités génératives dont nous avons choisi les GMM, le classificateur le plus répandu dans le domaine de la reconnaissance du locuteur et de la voix pathologique. L'autre objectif était de tester l'efficacité des capacités discriminatives du SVM. L'innovation marquante dans cette partie et de tester ces techniques avec des données spécifiques extraites d'une nouvelle base de données (SVD), n'est pas très utilisée à cause de sa nouveauté.
Les résultats des expériences réalisées ont montré qu'il pouvait exister dans les coefficients cepstraux supérieurs au-delà des treize premiers, une information utile pour la détection de la voix pathologique. Nous concluons aussi que le choix du nombre de gaussiennes est un facteur très critique surtout avec des données de grande dimension et qui ne sont pas visualisables tels que la MFCC. Pour les SVMs, ce classificateur est très puissant dans ce contexte, il suffit de bien choisir le noyau et de choisir les meilleurs paramètres d'apprentissage.

La combinaison des deux classificateurs est recommandée dans la plupart des travaux de recherche afin d'exploiter les deux capacités. Dans la deuxième partie de

ce travail nous avons développé le système hybride GMM-SVM.
Les GMM sont utilisés dans l' étape de modélisation et les SVM sont utilisés dans l'étape de discrimination. Deux approches sont réalisées : dans la première approche, nous avons utilisé le noyau RBF pour séparer les GMM supervecteur (vecteur contenant les centres des gaussiennes). Nous avons testé l'efficacité de la fonction de similarité dans la séparation des vecteurs. Les résultats sont très motivants par rapport aux résultats obtenus lorsque nous utilisons les deux classificateurs séparément.
Dans la deuxième approche, nous avons pensé à bénéficier de toute l'information statistique définissant le modèle GMM (poids, moyennes et matrice de covariance), par opposition à la première approche qui n'utilise que les moyennes.
Pour atteindre cette objectif nous avons embarqué dans le noyau RBF des distances qui permettent de mesurer le degré de similarité entre les distributions. L'utilisation de la distance de Bhattcharyya et de kullback-leibler permet d'améliorer les performances du système significativement et surtout les deux versions respectant l'inégalité triangulaire.

Malgré les résultats obtenus, il reste difficile à établir une classification de la voix pathologique à partir du signal de la parole et cela pour plusieurs considérations :

- Le degré de la dysphonie joue un rôle trés important, un patient qui à une légère altération est difficile à classer dans la catégorie exacte. De ce fait, il est commode d'utiliser une base de données ou les voix sont classées par grade de sévérité. (voix normale, dysphonie légére, modérée ou grave).
- Une voix normale caratérisée comme rauque, surtout pour les hommes, est peut etre classée comme pathologique. Il est important de combiner plusieurs parmètres qui font la différence.
- L'état émotionel du locuteur modifie sur la qualité de la voix, une voix normale peut etre vue comme pathologique.

Futurs travaux

Les futurs travaux seront la continuation de ce que nous avons réalisé jusqu'à présent. Nous nous intéressons à :

- Evaluer les performances de notre système sur d'autres bases de données notamment des bases où la voix est classée selon le grade de sévérité. Ce choix est motivé par le fait que le système est moins performant avec la classe pathologique qu'avec la classe normale.
- Aprés la classiffication binaire (normale/pathologique) nous voulons classer la voix pathologique selon la maladie.
- Utiliser des autres paramètres permettant d'enrichir les informations caractérisant la voix pathologique de la voix normale. Nous pensons à utiliser les paramètres de la source glottique.
- Utiliser d'autres distances à embarquer dans le noyau RBF qui permettent d'avoir le plus possible de données linéraiement séparables. Nous pensons aussi à utiliser d'autres techniques d'approximation de la distance de Kullback-Leibler et notamment le filtre de Kalman cubature.

Bibliographie

[A.Fort 1996] A.Fort, A. Ismaelli, C. Manfredi and P. Bruscaglioni. *Parametric and non-parametric estimation of speech formants : application to infant cry.* Medical Engineering and Physics, vol. 18, no. 8, pages 677–691, 1996. (Cited on page 28.)

[Alonso 2001] J. B. Alonso, J.Leon, I. Alonso and M. A. Ferrer. *Automatic Detection of Pathologies in the Voice by HOS-based Parameters.* EURASIP Journal on Advances in Signal Processing, vol. 20, no. 4, pages 275–284, 2001. (Cited on pages 26 and 34.)

[Alonso 2005] J. B. Alonso, F. D. de Maria, C. Travieso and M. A. Ferrer. *Using Non Linear Features for Voice Disorders Detection.* In Proceedings of the International Conference on Non- Linear Speech Processing (NOLISP' 05), 2005. (Cited on page 34.)

[Arias 2009] J. D. Arias, J. I. Godino Llorente and G. Castallanos Dominguez. *Short Time Analysis of Pathological Voices using Complexity Measures.* In Proceedings of the 3rd Advanced Voice Function Assessment AVFA 09 InternationalWorkshop, pages 93–96, 2009. (Cited on page 33.)

[Chen 2007] W. Chen, C. Peng, X. Zhu, B. Wan and D. Wei. *SVM-Based Identification of Pathological Voices.* In Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBS), no. 07, pages 3786–3789, 2007. (Cited on pages 30 and 32.)

[Childers 1992] D.G. Childers and K. Sung-Bae. *Detection of laryngeal function using speech and electroglottographic data.* IEEE Trans. Biomed. Eng., vol. 39, no. 1, pages 19–25, 1992. (Cited on page 28.)

[D.G.Silva 2009] D.G.Silva, L.Oliveira, and M.Andrea. *Jitter estimation algorithms for detection of pathological voices.* EURASIP Journal on Advances in Signal Processing, 2009. (Cited on page 26.)

[Dibazar 2002a] A. A. Dibazar and S. Narayanan. *A System for Automatic Detection of Pathological Speech.* In Proceedings of the 36th Asilomar Conference on Signals System and Computers, 2002. (Cited on page 30.)

[Dibazar 2002b] A.A. Dibazar, S. Narayanan and T.W. Berger. *Feature Analysis for Automatic Detection of Pathological Speech.* In Proceedings of the 2nd Joint Engineering in Medicine and Biology, 24th Annual Conference and the Annual Fall Meeting og the Biomedical Engineering Society (BMES/EMBS' 02), vol. 1, pages 182–183, 2002. (Cited on page 33.)

[D.Michaelis] D.Michaelis, T. Gramss and H. W. Strube. *Glottal-to-Noise Excitation Ratio : A New Measure for Describing Pathological Voices.* Acustica, vol. 83, no. 4, pages 700–706. (Cited on page 29.)

[Emary] M. El Emary, M. Fezari and F. Amara. Journal of Communications Technology and Electronics. (Cited on page 71.)

[F.AMARA 2013] F.AMARA, H.Bourouba and M.Fezari. *Pre-computed kernel for detection of spasmodic dysphonia from human voice.* ICSIP may 12-14,Guelma, Algeria., 2013. (Cited on page 86.)

[Fezari 013] M. Fezari and F.Amara. *Acoustic Voice Analysis, a Non-Invasive Tool for Detection of Voice Disorders Using Adaptive Features.* ICIST'13, March 22-24,Tangier, Morocco., 2013. (Cited on page 64.)

[Fredouille 2005] C. Fredouille, G. Pouchoulin, J.F. Bonastre, M. Azzarello, A. Giovanni and A. Ghio. *Application of Automatic Speaker Recognition Techniques to Pathologial Voice Assessment (Dysphonia).* In Proceedings of the 9th European Conference on Speech Communication and Technology, no. 05, pages 149–152, 2005. (Cited on pages 28 and 32.)

[Goddard 2007] J. Goddard, F. Martinez, G. Schlotthauer, M. E. Torres and H. L. Rufiner. *Visuallization of Normal and Pathological Speech Data.* In Proceedings of MAVEBA 07, pages 33–36, 2007. (Cited on page 35.)

[Godino-Llorente 2004] J. I. Godino-Llorente and P. Gomez-Vilda. *Automatic Detection of Voice Impairments by Means of Short-Term Cepstral Parameters*

and Neural Network Based Detectors. IEEE Transactions on Biomedical Engineering, vol. 51, no. 2, pages 380–384, 2004. (Cited on page 28.)

[Godino-Llorente 2005] J. I. Godino-Llorente, P. Gomez-Vilda, N. Saenz-Lechon, M. Blanco-Velasco, F. Cruz-Roldan and M. A. Ferrer. *Discriminative Methods for the Detection of Voice Disorders.* In Proceedings of the International Conference on Non-Linear Speech Processing (NOLISP' 05), pages 380–384, 2005. (Cited on pages 28 and 32.)

[Godino-Llorente 2006] J. I. Godino-Llorente, P. Gomez-Vilda and M. Blanco. *Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters.* IEEE Transactions on Biomedical Engineering, vol. 53, no. 10, pages 1943–1953, 2006. (Cited on pages 28 and 31.)

[G.Sfikas 2005] G.Sfikas, C.Constantinopoulos, A.Likas and N.P.Galatsanos. *An Analytic Distance Metric for Gaussian Mixture Models with Application in Image Retrieval.* ICANN, pages 835–840, 2005. (Cited on page 85.)

[H.Gray 1918] H.Gray. Anatomy of the human body. Lea and Febiger, Philadelphia, United States of America, 1918. (Cited on pages ix, 8, 9 and 10.)

[Hirano 1981] Hirano. *Clinical examination of voice.* Springer verlag, pages 83–84, 1981. (Cited on page 21.)

[H.Thomas 2009] H.Thomas. Reconstitution de la parole par imagerie ultrasonore et video de appareil vocal : vers une communication parlee silencieuse. 2009. (Cited on page 12.)

[I.G.Llorente 1999] I.G.Llorente, S. Aguilera-Navarro, C. Hernandez Espinosa, M. Fernandez Redondo and P. Gomez Vilda. *On the Selection of Meaningful Speech Parameters used by a Pathologic Non Pathologic Voice Register Classifier.* In Proceedings of the 6th European Conference on Speech Communication and Technology EUROSPEECH, 1999. (Cited on pages 30 and 34.)

[ins 2006] La voix, ses troubles chez les enseignants. Les éditions Inserm, 2006. (Cited on page 17.)

[Kacha 2006] A. Kacha, F. Grenez and J. Schoentgen. *Estimation des dyspériodicités dans la parole connectée dysphonique.* In Actes des XXVIemes Journées d'Etude sur la Parole, pages 363–366, 2006. (Cited on page 29.)

[Kasuya 1986] H. Kasuya, S. Ogawa and Y. Kikuchi. *Adaptative Comb Filtering Method as Applied to Acoustic Analysis of Pathological Voices.* In Proceedings of ICASSP, pages 669–672, 1986. (Cited on page 29.)

[Llorente 2008] J. I. Godino Llorente, V. Osma Ruiz, N. Saenz Lechon, I. Cobeta-Marco, R. Gonzalez-Herranz and C. Ramirez-Calvo. *Acoustic Analysis of Voice Using WPCVox : a Comparative Study with Multi Dimensional Voice Program.* European Archives of Otoringology, vol. 265, no. 04, pages 465–476, 2008. (Cited on page 30.)

[Louradour 2007] Jérome Louradour. Noyaux de s´equences pour la v´erification du locuteur par machines 'a vecteurs de support. 2007. (Cited on page 75.)

[Markaki 2010] M. Markaki, Y. Stylianou, J. D. Arias-Londono and J. I. Godino-Llorente. *Dysphonia Detection based on Modulation Spectral Features and Cepstral Coefficients.* In Proceedings of ICASSP 2010, pages 5162–5165, 2010. (Cited on pages 33 and 36.)

[M.David 2012] M.David, L.Eduardo, O.Alfonso, M.Antonio and V.Jesus. *Voice Pathology Detection on the Saarbrucken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit.* IberSPEECH, vol. 328, pages 99–109, 2012. (Cited on pages 31 and 61.)

[Michaelis 1998] D. Michaelis, M. Frohlich and H. W. Strube. *Selection and Combination of Acoustic Features for the Description of Pathologic Voices.* Journal of the Acoustical Society of America, vol. 103, no. 3, pages 1628–1639, 1998. (Cited on page 26.)

[Moran 2006] R. J. Moran, R. B. Reilly, P.Chazal and P. D. Lacy. *Telephony-Based Voice Pathology Assessment using Automated Speech Analysis.* IEEE Transactions on Biomedical Engineering, vol. 53, no. 3, pages 468–477, 2006. (Cited on page 26.)

[Mujumdar] M. V. Mujumdar and R. F. Kubichek. In Proceedings of ICASSP 2010. (Cited on page 35.)

[N.sanez 2006] N.sanez, I.Juan, O.Victor and G.Pedro. *Methodological issues in the development of automatic systems for voice pathology detection.* Biomedical Signal Processing and Control, vol. 1, pages 120–128, 2006. (Cited on pages 41 and 60.)

[O.C.Ai 2012] O.C.Ai, M. Hariharan, S. Yaacob and L.S. Chee. *Classification of speech dysfluencies with MFCC and LPCC features.* Expert Systems with Applications, vol. 39, no. 2, pages 2157–2165., 2012. (Cited on page 28.)

[Pouchoulin 2006] G. Pouchoulin, C. Fredouille, J. Bonastre, A. Ghio, M. Azzarello and A. Giovanni. *Modélisation statistique et informations pertinentes pour la caractérisation des voix pathologiques (dysphoniques).* In Proceedings of JEP 2006, 2006. (Cited on page 28.)

[R.Boite 2000] R.Boite, H. Bourlard, T.Dutoit, J. Hancq and H. Leich. *Traitement de la Parole.* Presses Polytechniques Universitaires Romandes, Lausanne., 2000. (Cited on page 26.)

[Reynolds 2000] D. A. Reynolds, T. Quatieri and R. B. Dunn. *Speaker Verification using Adapted Gaussian Mixture Models.* Digital Signal Processing : A Review Journal, vol. 10, no. 1, pages 19–41, 2000. (Cited on page 31.)

[Schafer. 2009] R. W. Schafer. *Homomorphic Systems and Cepstrum Analysis of Speech.* Springer handbook of speech processing, pages 161–180, 2009. (Cited on page 27.)

[Severin 2005] F. Severin, B. Bozkurt and T. Dutoit. *HNR Extraction in Voiced Speech Oriented Towards Voice Quality Analysis.* In Proceedings of EUSIPCO, no. 05, 2005. (Cited on page 29.)

[Shama 2007] K. Shama, A. Krishna and N. U. Cholayya. *Study of Harmonics-to-Noise Ratio and Critical-Band Energy Spectrum of Speech as Acoustic Indicators of Laryngeal and Voice Pathology.* EURASIP Journal on Advances in Signal Processing, vol. 2007, 2007. (Cited on page 35.)

[T.Drugman 2014] T.Drugman, P.Alku, A.Alwan and B.Yegnanarayana. *Glottal source processing : From analysis to applications.* Computer Speech and Language, vol. 28, no. 5, pages 1117–1138., 2014. (Cited on page 26.)

[T.Karim 2012] T.Karim and P.Frank. *A Note on Metric Properties for Some Divergence Measures :The Gaussian Case.* JMLR :Workshop and Conference Proceedings, vol. 25, 2012. (Cited on page 83.)

[Tokuda 2002] K. Tokuda, H. Zen and A. W. Black. *An HMM-Based Speech Synthesis System Applied To English.* IEEE Workshop on Speech Synthesis, pages 227–230, 2002. (Cited on page 28.)

[Vapnik. 1998] V.N. Vapnik. *Statistical Learning Theory.* 1998. (Cited on page 32.)

[Vasilakis 2007] M. Vasilakis and Y. Stylianou. *A mathematical model for accurate measurement of jitter.* In the 5th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications. Firenze University Press, 2007. (Cited on page 26.)

[V.Evaldas 2012] V.Evaldas, V.Antanas, G.Adas, B.Marija and U.Virgilijus. *Exploring similarity-based classification of larynx disorders from human voice.* Speech Communication science direct, pages 601–610, 2012. (Cited on pages 37, 92 and 93.)

[Wang 2006] J. Wang and C. Jo. *Performance of Gaussian Mixture Models as a Classifier for Pathological Voices.* In Proceedings of the 11th Australian International Conference on Speech Science and Technology, 2006. (Cited on pages 30 and 31.)

[Wester 1998] M. Wester. *Automatic Classification of Voice Quality : Comparing Regression Models and Hidden Markov Models.* In Proceedings of Symposium on Databases in Voice Quality Research and Education, 1998. (Cited on page 33.)

[W.M.Campbell 2006] W.M.Campbell, D.E.Sturim and D.E.Reynolds Senior. *Support Vector Machines Using GMM Supervectors for Speaker Verification.* IEEE signal processing letters, no. 13, 2006. (Cited on pages 36 and 45.)

[W.Xiang 2011] W.Xiang, Z.Jianping and Y.Yonghong. *Discrimination Between Pathological and Normal Voices Using GMMSVM Approach.* Journal of Voice, no. 25, pages 38–43, 2011. (Cited on page 36.)

[Y.Attabi] Y.Attabi. *Reconnaissance automatique des émotions à partir du signal acoustique.* (Cited on pages 28, 43 and 45.)

[Yumoto 1982] E. Yumoto and W. J. Gould. *Harmonics-to-Noise Ratio as An Index of the Degree of Hoarseness.* Journal of the Acoustical Society of America, vol. Journal of the Acoustical Society of America, no. 6, pages 1544–1550, 1982. (Cited on page 29.)