N° d'ordre: 58/2016-C/ELN

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediene
Faculté d'Electronique et d'Informatique

THÈSE

Présentée pour l'obtention du **diplôme** de **DOCTORAT 3ème Cycle (LMD)**

En: ELECTRONIQUE

Spécialité: Télécommunications

Par: Mr KADI Lahcene Kamil

Thème

Discrimination des Voix Pathologiques par Analyse Paramétrique du Signal de Parole Large Bande

Soutenue publiquement le 19/12/2016, devant le jury composé de:

Mme.	L.FALEK	Professeur	à l'USTHB	Présidente
Mme.	M.BOUDRAA	Professeur	à l'USTHB	Directeur de thèse
Mr.	S.SELOUANI	Professeur	à l'U.Moncton Canada	Co-Directeur de thèse
Mme.	R.KHEDDAM	MCA	à l'USTHB	Examinatrice
Mr.	Z.BENSELAMA	MCA	à l'USTHB	Examinateur
Mme.	F.MERAZKA	MCA	à l'USTHB	Examinatrice

Table des matières

Liste des figures

Liste des tableaux

Remerciements

Pour réaliser ce document et le travail qu'il présente, j'ai largement bénéficié de l'aide de nombreuses personnes. Je tiens à les remercier très sincèrement.

En premier lieu, je remercie les membres du jury de ma thèse pour l'intérêt qu'ils ont porté à mon travail et pour leur disponibilité.

Je tiens également à exprimer ma profonde gratitude à Madame Malika Boudraa, Monsieur Bachir Boudraa et Monsieur Sid Ahmed Selouani qui ont encadré mes recherches et qui se sont toujours souciés de m'offrir, de tout point de vue, les meilleures conditions de travail possibles.

Travailler au sein du Laboratoire LCPTS de l'USTHB et du Laboratoire LARIHS de l'Université de Moncton a été très agréable et enrichissant et je tiens à remercier tous ceux et celles qui ont contribué à créer cette atmosphère.

Je remercie également très fortement pour leur soutien, leur patience et leur bienveillance mes proches : mon épouse, mes parents et ma sœur.

Résumé

Des millions d'enfants et d'adultes, dans le monde, souffrent d'un trouble neuro-moteur de la communication d'origine congénitale ou acquise qui peut affecter l'intelligibilité de la parole. La caractérisation automatique du trouble de la parole peut contribuer à mettre en place des outils pour offrir des possibilités de communication alternative aux malades, et ainsi, contribuer à l'amélioration de leur qualité de vie. Par ailleurs, la caractérisation automatique peut assister les experts dans le diagnostic, l'évaluation et la conception de traitements adaptés à chaque individu.

Dans ce travail de recherche, de nouvelles approches sont présentées pour améliorer l'analyse et la discrimination des voix pathologiques et plus particulièrement, de la parole produite par des personnes atteintes de dysarthrie, qui est l'une des maladies de trouble de la parole d'origine neurologique la plus commune. Les deux principaux objectifs de cette thèse sont de proposer, d'abord, un système de diagnostic et/ou évaluation automatique du niveau de gravité de la dysarthrie, puis, un système de reconnaissance automatique du locuteur spécialement adapté aux personnes dysarthriques. Deux des plus importantes bases de données de renommée mondiale sont exploitées dans ces travaux, la première étant la base de données Nemours de paroles dysarthriques, la seconde est la récente base de données Torgo contenant des données acoustiques et articulatoires de la parole dysarthrique.

En premier lieu, une analyse linéaire discriminante *(LDA)* est combinée avec deux approches de classification automatiques, le modèle de mélanges Gaussiens *(GMM)* et les machines à vecteurs de support *(SVM)*, pour évaluer automatiquement la parole dysarthrique. Le *front-end* utilisé contient un ensemble de onze caractéristiques prosodiques sélectionnées par *LDA* sur la base de leur aptitude de discrimination. En exploitant la base de données Nemours, le système *LDA-SVM* a atteint une performance

de 93% dans l'évaluation de quatre niveaux de sévérité, variant de non-affecté à gravement malade.

En second lieu, nous proposons un modèle simulant l'oreille externe, l'oreille moyenne et l'oreille profonde. Ce modèle auditif procure des indices pertinents qui sont combinés avec les coefficients Cepstraux conventionnels *MFCC*, pour représenter les énoncés vocaux atypiques. Les expériences sont réalisées en utilisant les deux bases de données, Nemours et Torgo. Par ailleurs, trois systèmes de classification automatique sont comparés, *GMM*, *SVM* et le système hybride *GMM-SVM*. Un taux de discrimination du niveau sévérité de la dysarthrie de 93.2% a été atteint, ce qui surpasse l'état de l'art dans le domaine.

En dernier lieu, dans un contexte où les personnes atteintes de troubles de la parole sont exclues des solutions biométriques utilisant la voix, un système de reconnaissance des locuteurs dysarthriques est proposé. Une consolidation des indices auditifs distinctifs et des *MFCCs* a permis d'atteindre le meilleur taux d'identification de 97.2%. Ce résultat peut être considéré comme prometteur vu l'état de l'art actuel.

Les méthodes proposées pourront être utiles pour l'accessibilité des locuteurs dysarthriques aux systèmes biométriques et pour être un outil d'aide aux cliniciens dans le cadre de l'évaluation et/ou diagnostic des patients.

Partie 1 :

Introduction générale

Partie 1 :

I. Introduction générale

La communication est un processus dynamique et multidimensionnel qui est nécessaire à l'expression des pensées, des émotions et des besoins, permettant une interaction entre les personnes et leur environnement. La cognition, l'ouïe, la parole et la coordination motrice interviennent dans le processus de la communication. L'affaiblissement de l'une de ces facultés engendrerait un trouble de ce processus (Melf, 2015). Le trouble de la communication impacte grandement la qualité de vie des personnes atteintes de ce type de maladie, cela peut affecter l'expression de besoins, d'opinions ou de souhaits. De plus, ces pathologies réduisent les capacités individuelles à exprimer la personnalité, à être autonome, et ont aussi un impact sur le relationnel et l'estime de soi (Roth, 2011). Il est donc important et nécessaire d'améliorer la qualité de communication des personnes atteintes de troubles de la communication verbale en leur offrant plus de possibilités pour interagir avec leur environnement.

Des millions d'enfants et d'adultes, dans le monde, souffrent d'un trouble neuro-moteur de la communication d'origine congénitale ou acquise qui peut affecter l'intelligibilité de la parole. Dans ce travail, nous nous intéressons à l'une des maladies de trouble de la parole d'origine neurologique la plus commune, la dysarthrie (Roth, 2011), (American Speech-Language-Hearing Association).

Les malades dysarthriques peuvent être atteints de faiblesse, de ralentissement et d'incoordination lors du processus de production de la parole affectant la voix dont les caractéristiques sont différentes de la normale (Roth, 2011). La caractérisation automatique de la parole produite par des patients atteints de dysarthrie peut contribuer à mettre en place des outils pour offrir des possibilités de communication alternative aux malades, et ainsi, contribuer à l'amélioration de leur qualité de vie. Par ailleurs, la caractérisation automatique peut assister les experts dans le diagnostic, l'évaluation et la conception de traitements adaptés à chaque

individu. Dans ce travail de recherche, de nouvelles approches sont présentées pour améliorer l'analyse et la discrimination des voix pathologiques et plus particulièrement, de la parole produite par des personnes atteintes de dysarthrie. Les deux principaux objectifs de cette thèse sont de proposer, d'abord, un système de diagnostic et/ou évaluation automatique du niveau de gravité de la dysarthrie, puis, un système de reconnaissance automatique du locuteur spécialement adapté aux personnes dysarthriques.

Deux des plus importantes bases de données de renommée mondiale sont exploitées dans ces travaux, la première étant la base de données Nemours de paroles dysarthriques, la seconde est la récente base de données Torgo contenant des données acoustiques et articulatoires de la parole dysarthrique.

Les principales contributions introduites dans cette thèse peuvent être résumées selon les points suivants :

- De nouvelles approches pour améliorer l'efficacité de l'évaluation automatique du niveau de gravité de la maladie chez les personnes dysarthriques, telles que la proposition d'un modèle de calcul pour l'extraction d'information sur la perception auditive, et l'utilisation des caractéristiques prosodiques de la voix, combinées à une analyse linéaire discriminante.
- La proposition d'une évaluation globale basée sur des scores de chacun des patients dysarthriques intervenant dans la base de données Torgo et se basant sur le récent protocole d'évaluation médicale, *Frenchay Dysarthria Assessment second edition (FDA-2).*
- Un système de reconnaissance des locuteurs dysarthriques dans un contexte où les personnes atteintes de troubles de la parole sont exclues des solutions biométriques utilisant la voix.

Partie 2 :

Background

Partie 2 :

II. Background

Ce chapitre résume les éléments conceptuels fondamentaux utilisés tout au long de cette thèse. La section II.A décrit l'anatomie et la physiologie de l'appareil phonatoire et ses caractéristiques. La section II.B définie la dysarthrie en tant que pathologie de la parole, et ses fondamentaux aspects médicaux. Dans la section II.C, l'analyse automatique est abordée, ainsi que les paramètres acoustiques exploités dans ce travail pour la discrimination des différents signaux de parole. Cette dernière section présente aussi des méthodes, parmi les plus utilisées, de reconnaissance automatique des formes.

A. Anatomie de la parole

L'étude de l'anatomie du système vocal est utile pour arriver à la compréhension de la production de la parole humaine, soit le mécanisme de production du signal de parole.

Tous les organes utilisés dans le système vocal ont évolué à des fins autres que la parole, comme la respiration ou l'alimentation. Comparativement, ces organes ont été adaptés à la parole que récemment, ce qui, dans un certain sens, les transforme en un mécanisme de communication sous-optimal. (O'Shaughnessy, 2001). Par ailleurs, comme l'anatomie subit des évolutions, des adaptations et des ajustements précis, le système vocal et le systeme auditif ont évolué pour travailler à l'unisson. Par conséquent, la compréhension du fonctionnement de l'appareil auditif humain peut aider à améliorer l'extraction des caractéristiques de la voix.

De plus, cette évolution touche l'habilité de reconnaissance de la perception auditive, comme la reconnaissance des voix des personnes et des mots du langage (Beigi, 2011).

Les organes de la parole peuvent être subdivisés en trois groupes:

- Les poumons
- Le larynx
- Le conduit vocal supérieur constitué de la mâchoire, des lèvres, de la langue et des parois de la bouche.

Les poumons procurent le flux d'air nécessaire qui est transformé au niveau du reste du conduit vocal en ondes de pression d'air non-stationnaire qui constituent la parole. La pression produite durant le processus de production de la parole est entre 10 cm H2O et 20 cm H2O comparée à 1 ou 2 cm H2O nécessaire pour une respiration normale, sachant que cette dernière est quasi-inaudible vu que le flux d'air produit par les poumons n'est pas obstrué par le conduit vocal (O'Shaughnessy, 2001). Cependant, beaucoup d'animaux peuvent créer un bruit vocal en obstruant sinusoïdalement la pression d'air par le biais du larynx et à l'aide de la thyroïde, de la cricoïde, de l'aryténoïde et des cartilages épi glottiques (voir figure 1).

Deux muscles laryngales sont principalement responsables de la variation de la fréquence fondamentale (pitch), le muscle cricothyroïdienne dans les cordes vocales et le muscle vocalis qui peuvent augmenter la fréquence fondamentale en contractant et étendant les cordes vocales jusqu'à 4mm. La fréquence fondamentale peut aussi être réduite par une contraction des muscles sterno et thyro-aryténoïdien (Lofqvist, McGarr, et Honda, 1984), (Titze, 1994). Si ces muscles sont mal contrôlés, les cordes vocales peuvent être étroitement rapprochées et ne pas vibrer de façon normale résultant un pitch irrégulier et une voix cassante (Schneiderman et Potter, 2002).

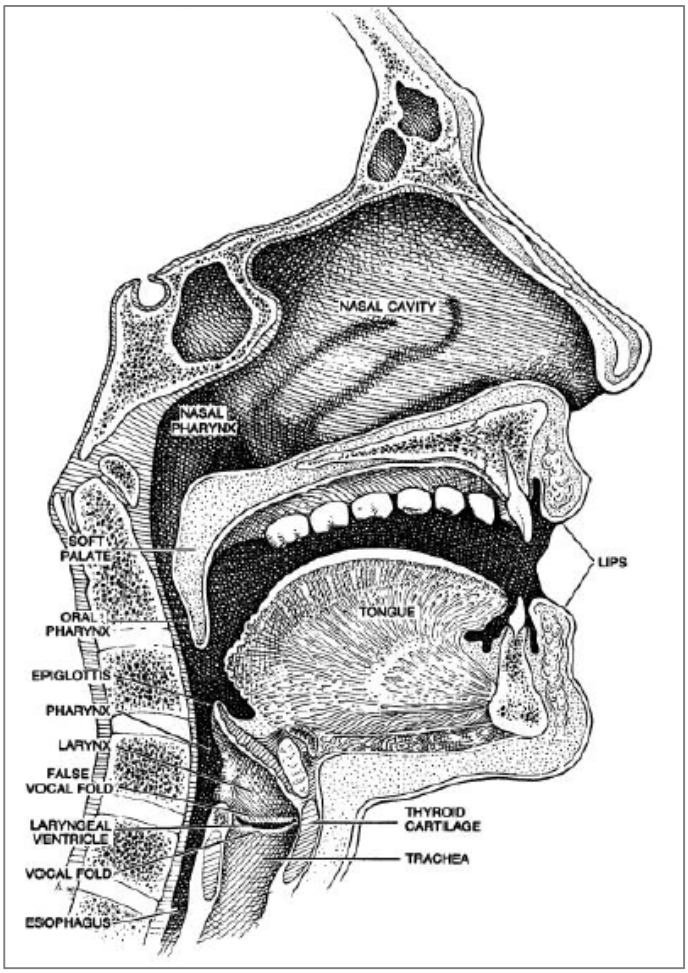

Figure 1| Les organes vocaux (Sunderg, 1977)

La langue est composée de 12 paires de muscles et de tissus et est probablement l'articulateur le plus complexe et le plus important dans la production de la parole (O'Shaughnessy, 2001). A l'exception du voile du palais, qui abaisse et soulève l'arrière de la cavité buccale pour laisser l'air passer dans la cavité nasale, la langue fournit presque tous les mouvements dans la bouche. Dans des conditions normales, il n'y a pratiquement pas de déplacement latéral de la langue, encore que la langue soit très agile et peut être réajustée entre les positions pertinentes en moins de 50ms. La pointe et la face dorsale de la langue sont deux éléments importants qui permettent des constrictions rapides à diverses positions le long du conduit vocal. (Stevens, 1998).

Le système de production de la parole peut être modélisé comme une conjonction d'au moins deux parties, une source qui génère des ondes sonores, et un filtre qui façonne ces ondes. La glotte représente la source de ce modèle source-filtre, dont le taux de vibration fournit des harmoniques à des multiples plus élevées que la fréquence d'onde d'origine (fréquence fondamentale). Les emplacements de ces harmoniques sont déterminés essentiellement par des changements soudains dans la largeur de la cavité buccale et l'interaction des ondes sonores avec les parois de cette cavité. Ces changements brusques sont dus quasi exclusivement à la configuration de la langue, qui est le principal agent causal du filtre (Rudzicz, 2011).

Outre le mécanisme de production de la parole et les organes sensoriels, le système nerveux joue un rôle majeur, il est donc responsable de la production et du déchiffrement des signaux de parole. Dans les deux hémisphères du cortex auditif, se déroule le traitement auditif de bas niveau. Certaines parties de l'hémisphère gauche du cerveau se spécialisent dans la production et la compréhension des langages, parallèlement, les miroirs de ces parties qui sont situés dans l'hémisphère droit sont sollicités pour la production et la compréhension de la fréquence fondamentale, le tempo, le rythme et la voix propre du locuteur, qui sont considérés comme des caractéristiques musicales (Beigi, 2011).

La partie nerveuse et la partie mécanique ont été schématisées et délimitées dans la figure 2 qui représente le mécanisme de production et perception de la parole.

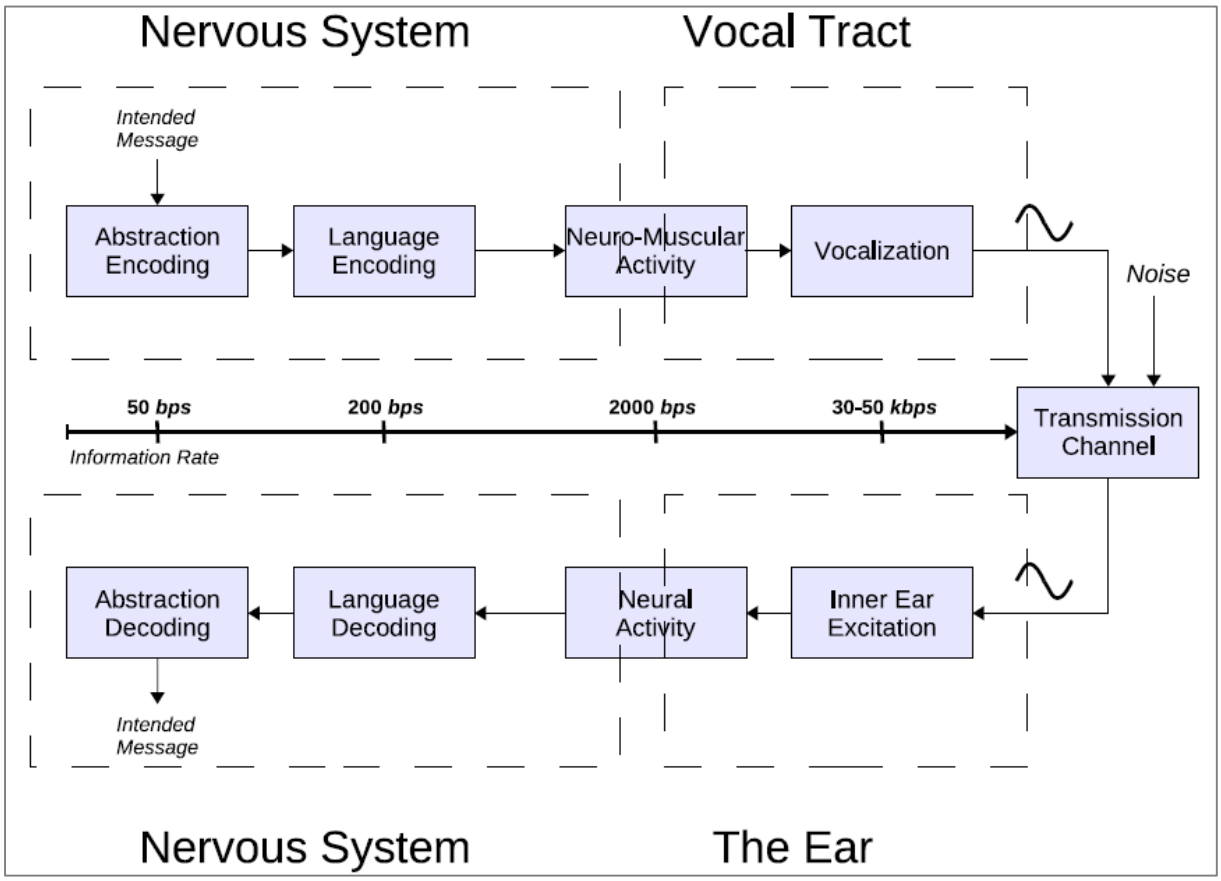

Figure 2| Production et perception de la parole (Beigi, 2011)

Le cerveau contrôle l'ensemble des muscles impliqués dans le processus de production de la parole où les mouvements volontaires sont initiés par le cortex moteur. Cependant, les messages produits sont transmis par les nerfs crâniens (*CN*) hautement spécialisés qui émergent à travers les fissures au niveau du cerveau inferieur, autour du cervelet et des noyaux gris centraux. Ces nerfs crâniens portent les impulsions qui commandent les muscles de l'appareil phonatoire mais aussi communiquent les données sensorielles dans le sens inverse, c'est-à-dire vers le cerveau (Rudzicz, 2011).

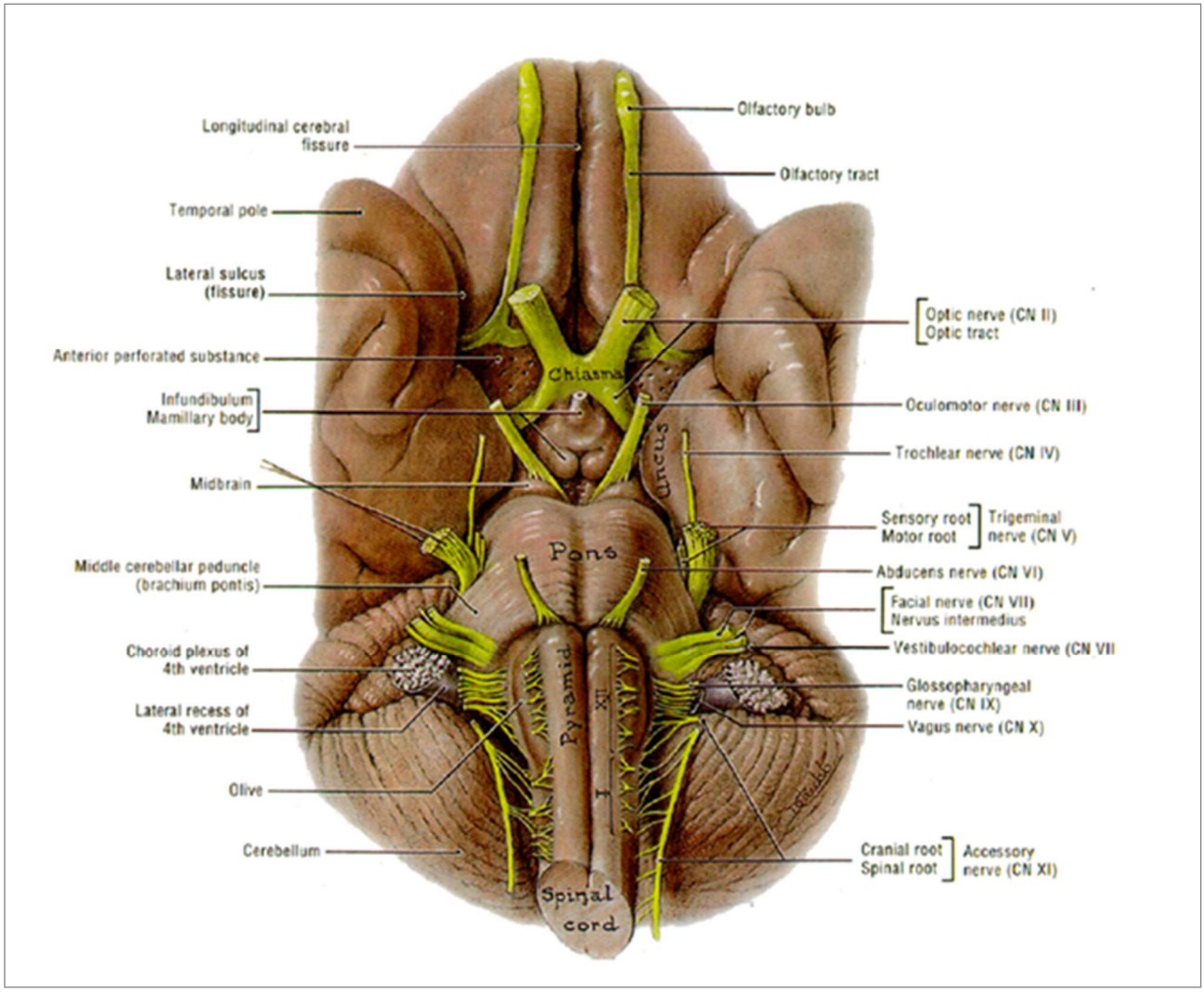

Figure 3|: Les nerfs crâniens mis en évidence dans le cerveau (Moore et Dalley, 2005).

La figure 3 est une illustration des nerfs crâniens humains. Ce qui pourrait être considéré comme le nerf le plus important est le nerf hypoglosse *CN XII* qui contrôle pratiquement tous les muscles intrinsèques et extrinsèques de la langue. Toute la musculature faciale est innervée par le nerf facial primaire *CN VII* tandis que le nerf intermédiaire facial commande les mouvements sublinguaux et submandibulaires. Lorsque ces nerfs crâniens sont atteints de trouble ou rendus inopérants, le signal contenant l'information n'est plus transmis à cause d'une

paralysie partielle de la musculature respective. Par ailleurs, si ces nerfs sont actionnés involontairement, les muscles peuvent réagir de façon relativement imprévisible.

B. Pathologie de la parole

1. Généralités

La communication est un processus dynamique et multidimensionnel qui est nécessaire à l'expression des pensées, des émotions et des besoins, permettant une interaction entre les personnes et leur environnement. La cognition, l'ouïe, la parole et la coordination motrice interviennent dans le processus de la communication. L'affaiblissement de l'une de ces facultés engendrerait un trouble de ce processus (Melf, 2015).

La parole est une fonction très complexe qui nécessite une contraction appropriée et synchronisée d'un grand nombre de groupes musculaires, cette contraction doit être associée à la respiration, à la fonction laryngale, au flux d'air et à l'articulation. Ce processus peut être altéré de différentes manières, ce qui pourrait être une indication sur la nature de la pathologie de la parole mise en cause (Enderby, 2013).

Des millions d'enfants et d'adultes, dans le monde, souffrent d'un trouble neuro-moteur de la communication d'origine congénitale ou acquise qui peut affecter l'intelligibilité de la parole. Dans ce travail, nous nous intéressons à l'une des maladies de trouble de la parole d'origine neurologique la plus commune, la dysarthrie (Roth, 2011), (American Speech-Language-Hearing Association). Les malades dysarthriques peuvent être atteints de faiblesse, de ralentissement et d'incoordination lors du processus de production de la parole affectant la voix dont les caractéristiques sont différentes de la normale (Roth, 2011).

2. La dysarthrie

La production de la parole est construite à partir de séquences respiratoires, laryngales, et de mouvements oro-faciaux ; ces séquences sont largement automatisées et sont coordonnées, pouvant être considérées comme un ensemble de gestes phonétiques élémentaires qui sont planifiés ou programmés (Ackerman, Hertrich et Ziegler, 2010).

La dysarthrie est un trouble de la parole d'origine motrice qui est due à un dysfonctionnement des muscles contrôlant l'appareil phonatoire, suite à une détérioration au niveau du système nerveux. Ces troubles affectent la respiration, la phonation, la résonance, l'articulation et la prosodie. Les personnes atteintes de dysarthrie peuvent voir leur intelligibilité de la parole considérablement réduite et avoir des anomalies au niveau des caractéristiques vocales. Quelques causes de la dysarthrie peuvent être citées comme: la maladie de Parkinson, les accidents vasculaires cérébraux, les traumatismes crâniens, les tumeurs, les dystrophies musculaires et les paralysies cérébrales (American Speech-Language-Hearing Association), (Enderby, 2013).

Le tableau 1 représente les dimensions et les impacts de la dysarthrie selon la Classification Internationale du Fonctionnement, handicape et santé (*ICF, World Health Organisation, 2001*).

Dimension *ICF*	**Impacte**
Altération	Altération du tonus musculaire affectant la puissance et la précision ; ainsi que la gamme des mouvements qui affectent l'orale, le vocal, et les mouvements de respiration. Incoordination de la musculature impliquée dans la production de la parole résultant une anormalité dans les caractéristiques de la parole. Par exemple : - Une mauvaise articulation des phonèmes - Une voix altérée au niveau de la qualité, du ton et du volume - Une résonance altérée, - Une émission nasale - Un manque d'appui au niveau de la respiration.
Activité	Réduction de l'intelligibilité de la parole à travers une voix basse. Réduction de l'habilité à communiquer. Le poids de l'activité communicative peut reposer sur le partenaire de communication.
Engendrement	La réduction des capacités de communication peut affecter l'estime de soi, le relationnel, la scolarité et l'emploi. Préjudices et restrictions pour la participation sociale et l'interaction.

Tableau 1| Dimensions et Impactes de la Dysarthrie *(World Health Organisation, International Classification of Functioning) (Enderby, 2013).*

a) **Types de dysarthries**

Les personnes atteintes de dysarthrie ont des discours qui semblent anormaux et ont une intelligibilité réduite, ce qui peut rendre la communication verbale laborieuse. Par ailleurs, si la nature du trouble de la parole est bien identifiée, cela peut être un apport important pour le diagnostic des différentes pathologies sous-jacentes. Le type de la dysarthrie est reflété par le caractère du trouble de la parole comme suit (tableau 2) (Enderby, 2013) :

Type de dysarthrie	Caractéristiques	La partie impliquée du système nerveux
Flaccide	Isolation des fonctions en fonction des régions de neurones moteurs affectées. Certains aspects de la parole peuvent être normaux.	Neurones moteurs inférieures.
Spastique	Voix roque, hyper nasalité, lenteur, articulation imprécise. Fréquemment accompagné de difficultés pour saliver et avaler.	Neurones moteurs supérieures.
Hypokinétique	Causant un bruit audible de la respiration, voix monotone avec une intensité réduite. L'articulation tend à être accélérée et réduite.	Tractus extrapyramidal, substantia nigra.
Hyperkinétique	Enrouement tendu et arrêts vocaux.	Tractus extrapyramidal, noyaux gris centraux.

Ataxique	Intensité excessive, tremblement et pauses articulatoires irrégulières. Le pitch et l'intonation sont généralement affectés. Difficulté à alterner les mouvements de la langue.	Cérébelleux.
Mixte	Symptômes similaires à ceux de la dysarthrie spastique plus un son humide de la voix, mauvais mouvements du larynx et de la langue, avec un mauvais contrôle des lèvres.	Neurones moteurs inférieures et supérieures.

Tableau 2| Caractéristiques des types de dysarthrie et la partie du système nerveux impliquée

b) **Evaluation et traitement**

L'intelligibilité quantifie le degré auquel la parole émise par un individu est perceptible aux auditeurs humains, généralement en mesurant la précision moyenne des transcriptions au niveau des mots des énoncés sur des groupes d'auditeurs naïfs (Kent et al., 2004).

Si les échantillons de parole sont phonétiquement équilibrés, les erreurs les plus répandues peuvent être classées automatiquement selon des caractéristiques phonétiques discrètes, comment le conduit vocal restreint la circulation de l'air, où se situe la constriction le long du conduit, et si les cordes vocales vibrent pendant la production. En résumé, la façon, la position et le voisement.

D'autres procédures de mesure de l'intelligibilité peuvent être citées, comme, la mesure de l'intelligibilité de la parole des enfants qui inclut les statistiques du développement et l'évaluation de Yorkston Beukelman Traynor (Hammen, Yorkston et Minifie, 1994) qui a été

automatisé incluant des facteurs tels que la vitesse l'élocution et l'intelligibilité. Les scores d'intelligibilités sont aussi accompagnés des résultats de l'évaluation Frenchay (Enderby, 1983) qui évalue individuellement le score des différents aspects comme la respiration, le réflexe et la vitesse d'élocution (Menendez-Pidal et al., 1996).

Etant donné que la dysarthrie ne peut pas encore être guérie avec des médicaments ou encore de la chirurgie, la rééducation comportementale est souvent utilisée pour renforcer les muscles articulatoires ou élaborer des stratégies de prononciations alternatives afin d'améliorer l'intelligibilité (Kent, 2000). Ces interventions comportementales peuvent impliquer le traitement automatisé à l'aide d'ordinateurs, cela peut être utile en procédant à une rééducation et avoir un retour généré à l'aide de la reconnaissance automatiquement de la parole (Thomas-Stonell et al., 1998).

C. Analyse et discrimination automatique

1. Analyse acoustique

Une des étapes les plus importantes dans la reconnaissance des formes est l'extraction de descripteurs efficaces qui, dans le cas du traitement de la parole, représenteront la forme d'onde de parole prononcée. L'objectif est d'extraire de l'information pertinente contenue dans le signal de parole en excluant la partie non-informative. Sachant que l'inclusion de la partie non-informative dans les systèmes de reconnaissance des formes, peut non seulement engendré une surcharge en termes d'espace mémoire mais peut aussi détériorer les performances de reconnaissance et de discrimination automatiques.

Plusieurs caractéristiques acoustiques peuvent être utilisées comme paramètres d'entrée dans les systèmes de reconnaissance et de la caractérisation des troubles liés à la parole. Parmi ces caractéristiques, nous pouvons citer les *LPC (Linear Predictive Coding)* et les coefficients suivants : les très connus *MFCC* (*Mel-Frequency Cepstral Coefficients), Short-time Spectral Envelope, Short-time Energy* et *Zero Crossing Rate.* Il est important de noter qu'un nombre important d'études ont démontré que l'utilisation des propriétés du processus auditif humain

peut permettre d'avoir une représentation *front-end* de la parole pertinente (Selouani et al., 2015)

Il existe différents niveaux d'indices perceptuels, allant des niveaux bas comme l'acoustique et la phonétique de la parole, aux niveaux supérieurs comme la prononciation, la prosodie et la sémantique (figure 4) (Liss et al., 2009).

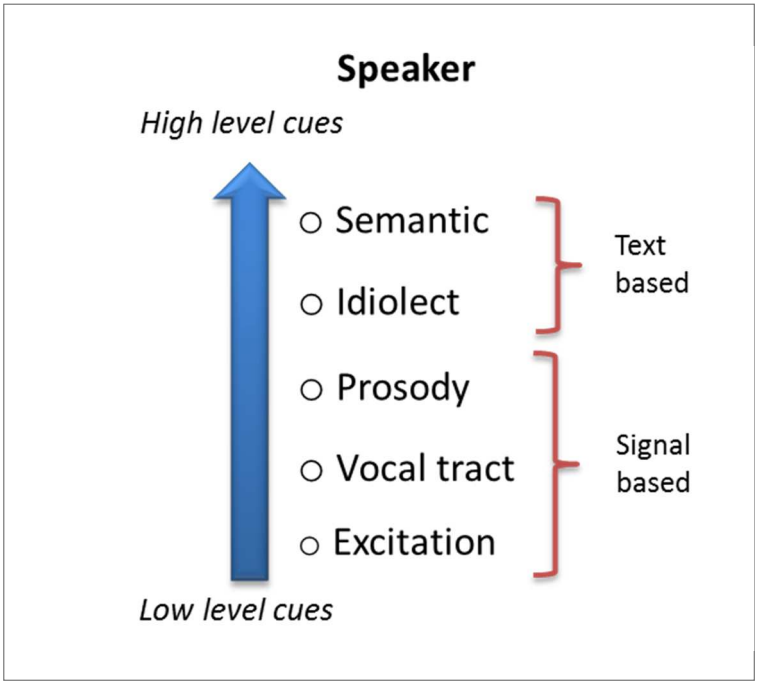

Figure 4| Diffèrent niveaux d'indices

Ainsi, des indices de niveau supérieur relatifs à la prosodie sont utilisés dans la première série d'expériences comme *front-end* du système d'évaluation automatique de la dysarthrie. Au cours de la deuxième série d'expériences, les indices de niveau plus bas suivants sont exploités: les paramètres bien connus *MFCCs* et les indices présentés dans ce travail pour l'extraction d'information sur la perception auditive '*Ear model*'.

a) **Prosodie**

La parole est essentiellement destinée à transmettre un message à travers une séquence d'unités linguistiques sonores. La prosodie est définie comme une branche de la linguistique qui se spécialise dans la description et la représentation des composants de la parole. Les paramètres prosodiques incluent des entités perceptuelles complexes qui sont : l'intonation, le stress et le rythme. Elle est fondamentalement exprimée par 3 paramètres acoustiques : le pitch, la durée et l'énergie (Mary, 2012). Le stress, le timing et l'intonation dans la parole sont étroitement liés à la prosodie de la parole, ce qui rehausse l'intelligibilité du message transmit, permettant à la personne qui écoute de segmenter facilement la parole continue en mots et en phrase.

Selon (Duffy, 2000), « la parole est la plus complexe des aptitudes motrices que l'humain acquière d'une façon innée, une activité caractérisée chez l'adulte normal par la production d'environ quatorze sons distincts par seconde à travers l'action coordonnée d'environ cent muscles innervés par des nerfs crâniens et spinaux multiples ». Dans la dysarthrie, certains dommages neurologiques affectent typiquement les nerfs qui contrôlent le système musculaire articulatoire impliqué dans la parole, causant une faiblesse, un ralentissement et une incoordination. Selon la gravité de la dysarthrie ce trouble peu affecter la prosodie de manières différentes.

Un ensemble de onze paramètres prosodiques a été sélectionné par le biais d'une analyse discriminante LDA *(Linear Discriminant Analysis)*, plus précisément en utilisant la mesure Wilks' lambda (cette méthode sera présentée dans la prochaine partie de ce chapitre). Les indices suivants ont donc été déterminés comme étant les plus adaptés à notre étude : Jitter, Shimmer, Pitch médian, Ecart-type du Pitch, Nombre de Périodes, Ecart-type de la Période, Proportion de la durée Vocalique (%V), Ratio Harmoniques sur Bruit (dB), Ratio Bruit sur Harmoniques (%), Vitesse d'Articulation et Degré des Pauses de la voix (Kadi et al., 2013).

Ci-dessous la description des principaux paramètres prosodiques exploités dans ce travail.

Mean Pitch :

Le corrélat physique du pitch est la fréquence fondamentale (F0) estimée par le taux de vibration des cordes vocales pendant la phonation de sons voisés (Shriberg, Stolcke et Hakkani, 2000). L'ensemble des variations du pitch au cours d'une énonciation est définie comme l'intonation (Hart, Collier et Cohen, 1990). La gamme typique d'un locuteur masculin est de 80 à 200 Hz (pour un discours conventionnel), dépendamment de la masse et de la longueur des cordes vocales (Mary, 2012). Au cours de notre étude, le pitch moyen est calculé en faisant la moyenne de la fréquence fondamentale à travers une phrase, en utilisant la méthode de l'autocorrélation. La valeur du pitch moyen d'une phrase prononcée par un locuteur dysarthrique peut aider à détecter une anomalie du signal glottique.

Jitter :

Le Jitter représente la variation de la fréquence fondamentale tout au long de l'évolution temporelle de l'énoncé. Il indique la variabilité ou la perturbation de la période de temps (T0) à travers plusieurs cycles d'oscillation. Le Jitter est principalement affecté par une insuffisance dans le contrôle de la vibration des cordes vocales (Westzner et al., 2005). Le seuil de comparaison pour un Jitter normal/pathologique est de 1,04 %, selon le *Multi-Dimensional Voice Processing program* (MDVP), ce programme a été conçu par la société *Kay Elemetrics* (MDVP, online). Le Jitter brute et le Jitter normalisé sont définis respectivement comme :

$$Jitter(seconds) = \sum_{i=1}^{N-1} |T_i - T_{i+1}| \Big/ N - 1 \qquad (1)$$

$$Jitter(\%) = Jitter(seconds) \Big/ \frac{1}{N} \sum_{i=1}^{N} T_i \qquad (2)$$

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

Shimmer :

Le Shimmer indique la variabilité ou la perturbation de l'amplitude du signal de parole. Il est lié aux variations dans l'intensité de l'émission vocale et est partiellement affecté par la réduction de la résistance glottique (Westzner et al., 2005). Le MDVP donne une valeur de 3,81% comme seuil pour la pathologie. Le Shimmer est estimé de façon similaire au Jitter, en utilisant l'amplitude en tant que paramètre.

L'intensité et le pitch peuvent être plus difficiles à contrôler si la pression de l'air qui afflue vers les cordes vocales est très variable ou extrêmement faible (Baghai-Ravary and Beet, 2013).

Articulation rate :

Le débit d'articulation est le nombre de syllabes prononcées par seconde, en excluant les pauses (Liss et al., 2009). Au cours de nos expériences, nous avons constaté que, généralement, plus le niveau de gravité de la dysarthrie est important plus le débit d'articulation est faible.

Proportion of vocalic duration :

La durée vocalique (DVOC) est représentée par la séparation entre le relâchement (R) et la constriction (C) encadrant la cible d'une voyelle (voir la figure 5) (Calliope, 1989). La proportion de la durée vocalique (%V) est la fraction de la durée de l'énonciation qui se compose des intervalles vocaliques (Liss et al., 2009). La difficulté à maintenir la voix durant une voyelle soutenue peut être considérée comme un signe de pathologie (Boersma and Weenink, 2001).

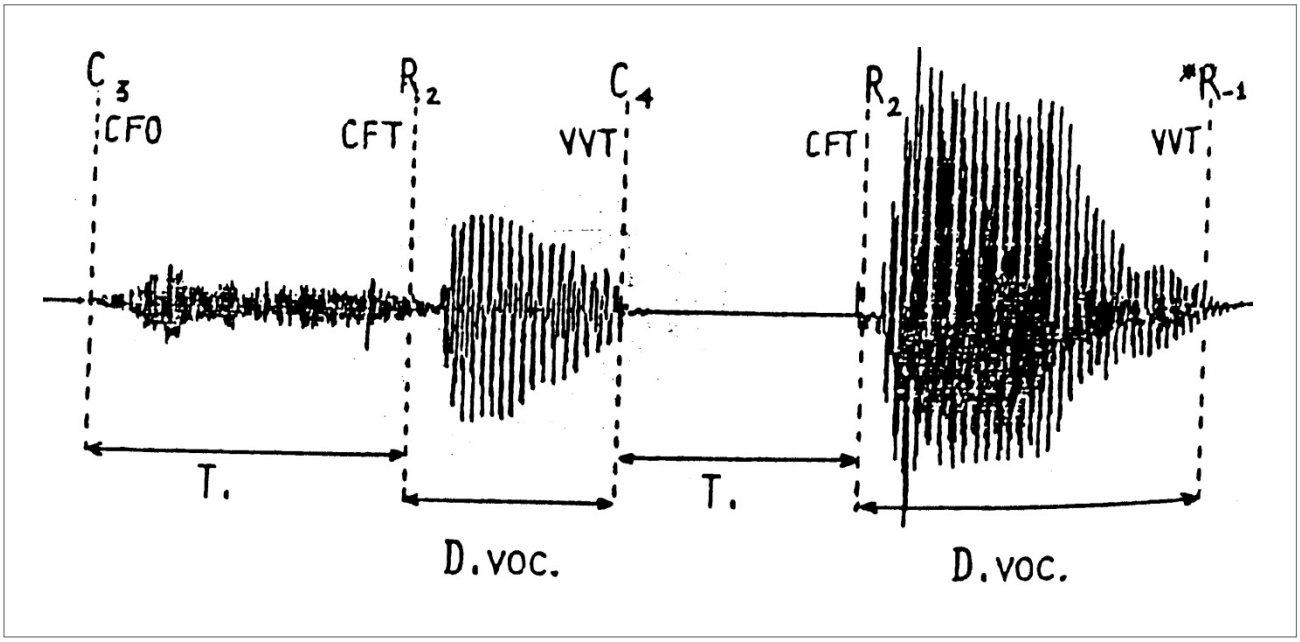

Figure 5| Illustration des coordinations majeurs D.voc= durée vocalique et T = tenue consonantique (Calliope, 1989)

Harmonics to noise ratio :

Le ratio harmonique sur bruit (HNR) représente le degré de périodicité acoustique. L'harmonicité est mesurée en dB, calculée par le ratio entre l'énergie de la partie périodique et l'énergie du bruit. Le HNR peut être utilisé comme une mesure de la qualité de la voix. Par exemple, un locuteur saint peut produire un 'a' soutenu avec un HNR d'environ 20 dB (Boersma and Weenink, 2001). Le HNR est défini comme suit:

$$HNR(dB) = 10log\left(\frac{E_p}{E_n}\right) \qquad (3)$$

où E_P est l'énergie de la partie périodique et E_n est l'énergie du bruit.

Degree of voice breaks :

Le degré de pauses de la voix est calculé par la durée totale des pauses à travers le signal de parole, divisée par la durée totale, en excluant le silence au début et à la fin de la phrase (Liss et al., 2009). Une pause de la voix peut se produire à cause d'un arrêt soudain du flux d'air dû à une carence passagère dans le contrôle du mécanisme de phonation (Guerra et Lovey 2003).

b) Paramètres acoustiques Cepstraux

Les coefficients MFCC à court terme sont souvent utilisés dans le traitement du signal appliqué aux troubles de la parole, comme la reconnaissance de la parole dysarthrique ou encore la classification de la maladie. L'échelle Mel introduite pas Davis et Mermelstein est une correspondance entre une échelle de fréquence linéaire et non-linéaire, celle-ci étant basée sur la perception auditive humaine. L'échelle Mel est représentée par l'approximation suivante :

$$Mel(f) = 2595 log_{10}\left(1 + \frac{f}{700}\right) \qquad (4)$$

où f représente l'échelle de fréquence linéaire.

Pour calculer les MFCCs, une transformée en cosinus discrète est appliquée aux sorties de *M* filtres passe-bande critiques. Ces filtres sont triangulaires et sont portés sur l'échelle de Mel qui est linéaire pour les fréquences inférieures à 1 kHz et logarithmique au-delà. 15 à 24 filtres peuvent être utilisés ; dans ce travail, environ 2 filtres par octave ont été utilisés. Les MFCCs sont définis comme suit :

$$MFCC_n = \sum_{m=1}^{M} X_m \cos\left(\frac{\pi\, n}{M}\,(m - 0.5)\right), \qquad n = 1, 2, \dots, N \qquad (5)$$

où *M* est l'ordre de l'analyse, *N* représente le nombre des coefficients Cepstraux et X_m (m = 1, 2,..., M) est la sortie du $M^{ième}$ filtre appliqué au log du spectre de magnitude du signal.

c) **Modélisation Auditive**

Dans ce travail, un modèle de calcul pour l'extraction d'information sur la perception auditive est proposé dans le but de caractériser le signal de parole dysarthrique. Tout au long de l'évolution du traitement de la parole, de nombreuses tentatives ont été réalisées pour l'extraction de l'information la plus pertinente du phénomène de la parole. Ce phénomène peut être considéré comme imprécis, imprédictible et versatile. Les méthodes usuelles d'extraction de caractéristiques acoustiques restent inefficaces pour assurer un bon niveau de robustesse, tandis que les humains peuvent facilement gérer l'incertitude des environnements acoustiques et des conditions défavorables pour la compréhension de la parole. Cela a conduit de nombreux chercheurs à étudier le système auditif humain afin de mieux comprendre cette capacité extraordinaire de perception. La capacité du système auditif à traiter efficacement et à interpréter la parole, même dans de mauvaises conditions, comme l'inintelligibilité de la parole dysarthrique, fait de la modélisation auditive une approche intéressante et prometteuse.

Habituellement, la caractérisation basée sur la modélisation auditive vise à examiner la réponse de la membrane basilaire et du nerf auditif à différents sons. Un traitement avancé simulant le cortex auditif peut également être effectué. Un modèle de calcul a été proposé par Flanagan (1960) pour évaluer le mouvement de la membrane basilaire. Ce modèle s'est avéré utile pour avoir renseigné sur le comportement auditif subjectif et le processus acoustique-mécanique de l'oreille (Flanagan, 1960). Autres modèles bien connus, comme le *Cochlea Model* (Lyon, 1982), la représentation de *Mean Synchrony Auditory* (Seneff, 1988), ainsi que le traitement *Ensemble Interval Histogram* (Ghitza, 1994), ont été la base de nombreuses approches contemporaines. Dans tous ces modèles, un banc de filtres passe-bande est utilisé pour simuler le filtrage cochléaire.

Ces dernières années, il y a eu de nouveaux intérêts dans l'amélioration du traitement *front-end* pour le calcul de caractéristiques robustes inspirées par la modélisation auditive (Stern

et al., 2012). Dans le domaine de la reconnaissance de la parole, le traitement basé sur la psycho-acoustique et la physiologie auditive devient le principal composant des méthodes d'extraction de caractéristiques robustes comme les *Gammatone Features* (GFCC par Schluter et al., 2007) et les *Power-Normalized Cepstral Coefficients* (PNCC par Kim et Stern, 2012).

Le modèle auditif utilisé tout au long de nos expériences simule les parties externe, moyenne et interne de l'oreille. Ce modèle auditif fut d'abord proposé par Caelen (1985) et par la suite adapté pour être utilisé comme un module *front-end* dans les systèmes de reconnaissance de la parole par Selouani (2011). Pour modéliser les divers mouvements adaptatifs des osselets au niveau de l'oreille externe et moyenne, un filtre passe-bande a été utilisé. Un banc de filtres non-linéaires stimule la membrane basilaire (BM) qui agit sur l'oreille interne. La transduction électro-mécanique des cellules ciliées et des fibres afférentes, d'où le signal encodé est généré (dans les terminaisons synaptiques) est également considérée par le modèle BM. Le long des multiples organes impliqués dans la perception et l'audition, différentes régions sont sensibles aux sons avec des propriétés spectrales différentes, en raison de la différence de l'anatomie et la physiologie. Ainsi, chaque partie le long de la BM a une fréquence de résonance donnée pour un certain son d'entrée (Caelen, 1985). La figure 6 représente le modèle auditif utilisé dans cette étude.

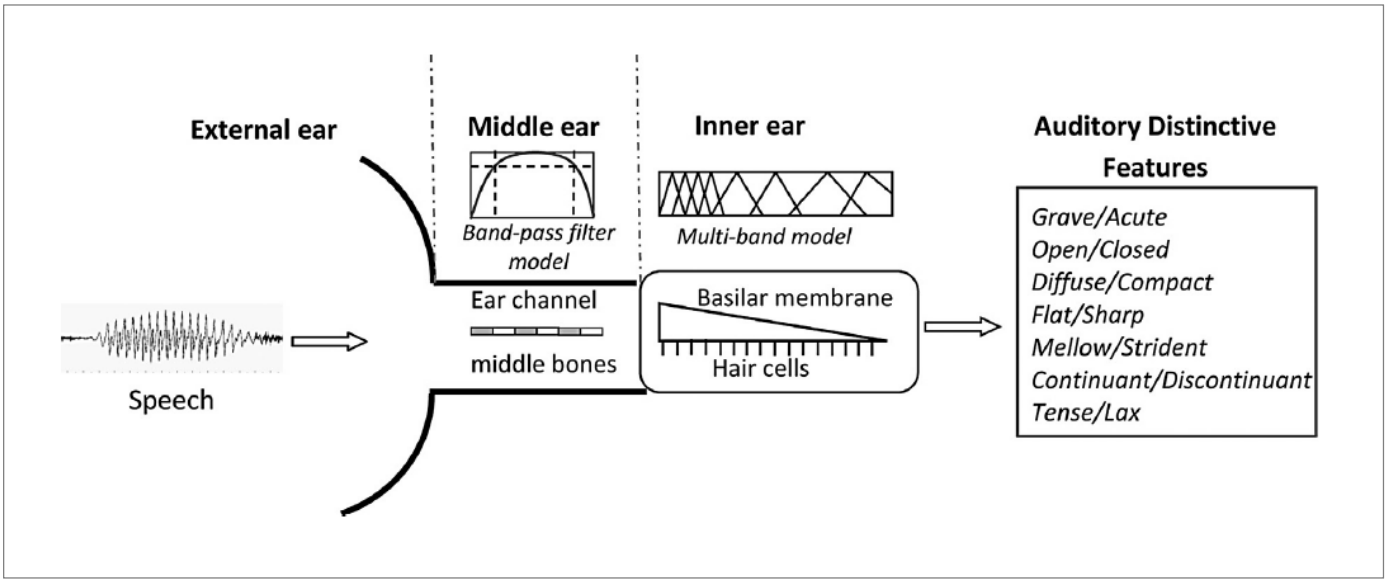

Figure 6| Diagramme du model auditif utilisé pour représenter l'oreille externe, moyenne et interne

Mid-external ear model :

Un filtre passe-bande simule les parties externe et moyenne de l'oreille, cela peut être défini comme suit :

$$s'(k) = s(k) - s(k-1) + \propto_1 s'(k-1) - \propto_2 s'(k-2), \qquad k = 1, \dots K \qquad (6)$$

Où $s(k)$ est l'onde du signal de parole, $s'(k)$ est le signal de sortie filtré, K est le nombre d'échantillons de la fenêtre, $\propto_1$ et $\propto_2$ sont des coefficients qui dépendent de la fréquence centrale du filtre, de son facteur Q et de la fréquence d'échantillonnage Fs. La valeur de 1500 Hz et 1,5 sont respectivement utilisés comme fréquence centrale et facteur Q (Selouani et al., 2007).

Basilar membrane model :

Le BM est modélisé par 24 filtres chevauchés qui représentent le banc de filtres cochléaire. La vibration d'une partie spécifique de la BM est simulée par la réponse fréquentielle d'un certain filtre pour une stimulation auditive au niveau de l'oreille externe (Caelen, 1985). Pour chaque filtre du banc, la sortie est définie comme suit:

$$y_i(k) = \beta_{1,i} y_i(k-1) - \beta_{2,i} y_i(k-2) + G_i[s'(k) - s'(k-2)] \qquad (7)$$

La fonction de transfère est:

$$H_i(z) = \frac{G_i[1 - z^{-2}]}{1 - \beta_{1,i}\, z^{-1} + \beta_{2,i}\, z^{-2}} \tag{8}$$

où $y_i(k)$ est la réponse de la BM au signal *mid-external* $s'(k)$. Cela constitue l'amplitude de vibrations dans la position x_i de la BM. Les paramètres du filtre i sont les coefficients $\beta_{1,i}$, $\beta_{2,i}$ et le gain G_i . Le nombre de filtres cochléaires qui se chevauchent (ou canaux) est fixé à 24 et représenté par N_c . Chaque filtre couvre environ Δx = 1,46 mm de la BM. A des fins de simplification, seulement les effets de couplage de la transduction électromécanique dans les cellules ciliées et les fibres ont été considérés. Les paramètres de couplage C_i , E_i et A_i représentent le comportement des fibres et des cellules ciliées. L'algorithme 1 représente un algorithme détaillé qui donne $y'_i(k)$, le stimulus de sortie après le passage à travers toutes les constituantes du modèle d'oreille. L'énergie de chaque canal est ensuite calculée à partir de $y'_i(k)$.

Definitions

F_s: sampling frequency (16000 Hz)

Δ_x: basic length unit of the basilar membrane (1.46 mm)

x: total length of the basilar membrane (35 mm)

N_c: total number of channels (24)

i: channel index

k: time index

K: number of total samples by frame

F_i: central frequency of channel i

$H_i, r_{i,j}, v\ and\ u$: temporary calculation functions

E_i: direct coupling function

A_i: inverse coupling function

C_i: coupling parameter

Q_i: *Q*-factor

G_i: gain of the filter

$\beta_{1,i}\ and\ \beta_{2,i}$: filter coefficients

$s'(k)$: filtered speech signal that passed through the mid-external ear

$y_i'(k)$: resulting stimulus

Initialize $f_x = (F_s\,\Delta_x)^2, H_0 \quad = 0, r_{i,j} = 0, E_0 = 0.$

For $i = 1\ to\ N_c$ ***Do***

$$x_i = i\Delta x; v = e^{(-106.5x_i)};\ F_i = 7100v - 100;\ C_i = \frac{(27v)^2}{f_x};$$

$$Q_i = (-8300x + 176.3)x_i + 4;\ G_i = e^{(-80x_i)}; u = e^{-\frac{\pi F_i}{F_s Q_i}};$$

$$\beta_{1,i} = 2u\ cos\left(\frac{2\pi F_i}{F_s}\right);\ \beta_{2,i} = u^2;$$

$$E_i = \frac{1}{1 + (2 - E_{i-1})C_i};\ A_i = E_i C_i;$$

EndDo

For $k = 1\ to\ K$ ***Do***

For $i = 1\ to\ N_c$ ***Do***

$$H_i = \left(G_i\left(s'(k) - s'(k-2)\right) + \beta_{1,i}\, r_{i,2} - \beta_{1,i}\, r_{i,1}\right)E_i + H_{i-1}\, A_i$$

EndDo

For $i = N_c\ to\ 1$ ***Do***

$$r_{i,2} = A_i\, r_{i+1,3} + H_i, and\ y'_i(k) = r_{i,3}$$

EndDo

For $i = 1\ to\ N_c$ ***Do***

For $j = 1\ to\ 2$ ***Do***

$$r_{i,j} = r_{i,j+1}$$

EndDo

EndDo

EndDo

Algorithme 1| Algorithme simulant l'oreille interne pour objectif de calculer le stimulus $y'_i(k)$ échantillon par échantillon

Auditory distinctive features :

Pour la sortie de chaque canal, l'énergie absolue est définie comme :

$$W'_i(T) = 20 \log \sum_{k=1}^{K} \left| y'_i(k) \right|, \quad i = 1, 2, \dots, N_c \qquad (9)$$

T représente l'index de la fenêtre du signal de parole, i indique le numéro du canal exploité et N_c est le nombre total de canaux qui est 24. K est la longueur de la fenêtre en terme d'échantillon et k réfère à l'échantillon traité. Pour réduire la fluctuation de l'énergie, une fonction de lissage est appliquée :

$$W_I(T) = c_0 W_i(T-1) + c_1 W'_i(T) \qquad (10)$$

où $W_I(T)$ représente l'énergie des sortie de la fonction de lissage. La somme des coefficients c_0 et c_1 est égale à 1

Les paramètres auditifs sont calculés par des combinaisons linéaires des énergies de sorties des canaux. D'après les études précédentes (Selouani et Caelen, 1999), nous utilisons sept indices dérivés du modèle de l'oreille comme caractéristiques distinctifs : *Grave/Acute (G/A), Open/Closed (O/C), Diffuse/Compact (D/C), Flat/Sharp (F/S), Mellow/Strident (M/S), Continuant/Discontinuant (C/D)* et *Tense/Lax (T/L).*

Le calcul des sept paramètres à travers chaque fenêtre est détaillé dans le Tableau 3. Basé sur des connaissances phonétique et acoustique, le calcul de l'indice (M/S) peut être par exemple expliqué. En effet, les phonèmes stridents sont ca ractérisés par la présence de bruit

dû à une turbulence au niveau du point d'articulation. Par conséquent, un phonème est considéré comme strident si la bande de fréquences entre 3800 Hz et 5300 Hz contient plus d'énergie que la bande allant de 1900 Hz à 2900 Hz. Cette approche a été testée avec succès dans le contexte de la reconnaissance de la parole dans des conditions défavorables (Selouani *et all*, 2003).

Indice	Description
(**G**/**A**)	Mesure la différence d'énergie entre les basses fréquences (50-400 Hz) et les hautes fréquences (3800-6000 Hz)
(**O**/**C**)	Un phénomène est considéré fermé si l'énergie des basses fréquences (230-350 Hz) est plus grande que dans les moyennes fréquences (600-800 Hz)
(**D**/**C**)	La compacité reflète la proéminence de la région du formant central (800-1050 Hz) par rapport aux régions autour (300-700 Hz) et (1450-2550 Hz)
(**F**/**S**)	Un phénomène est considéré comme aigu si l'énergie dans (2200-3300 Hz) est plus importante que dans (1900-2900 Hz)
(**M**/**S**)	Les phénomènes stridents sont caractérisés par la présence de bruit due à une turbulence au niveau du point d'articulation, ce qui provoque plus d'énergie dans la bande (3800-5300 Hz) par rapport à la bande (1900-2900 Hz)
(**C**/**D**)	Quantifie la variation du spectre de magnitude en comparant l'énergie de la fenêtre courante et précédente
(**T**/**L**)	Mesure la différence d'énergie entre les fréquences moyennes (900-2000 Hz) et les fréquences relativement hautes (2650-5000 Hz)

Tableau 3| Description des Indices Auditifs

2. Discrimination

a) Fonction d'Analyse Discriminante

La sélection des caractéristiques pertinentes peut apporter une grande contribution au processus de classification. En effet, les paramètres sélectionnés peuvent éviter certaines mauvaises prédictions automatiques et améliorer le taux de classification. Dans ce travail, nous procédons à une analyse discriminante en utilisant *Wilks' lambda* comme outil de mesure et de sélection des caractéristiques les plus efficaces parmi un grand nombre de paramètres de la parole calculés.

L'analyse discriminante est utilisée pour modéliser une variable catégorielle dépendante basée sur sa relation avec une ou plusieurs variables prédictives. A partir d'un ensemble de variables indépendantes, l'analyse discriminante détermine les combinaisons linéaires de ces variables qui discriminent le mieux les classes. Ces combinaisons sont appelées 'fonctions discriminantes' et sont définies comme suit (Kadi et al., 2013) :

$$d_{ik} = b_{0k} + b_{1k}\, x_{i1} + \ldots + b_{pk}\, x_{ip} \qquad (11)$$

où d_{ik} est la valeur de la $k^{ième}$ fonction discriminante de la $i^{ième}$ classe, p est le nombre de prédicteurs (qui représentent les variables indépendantes), b_{jk} est la valeur du $j^{ième}$ coefficient de la $k^{ième}$ fonction et x_{ij} est la valeur de la $i^{ième}$ classe du prédicteur j .

Le nombre de fonctions discriminantes est égal à *min* (nombre de classes-1, nombre de prédicteurs).

La procédure choisit automatiquement une première fonction qui séparera les classes autant que possible. Par la suite, une seconde fonction est calculée, cette dernière est à la fois non corrélée avec la première fonction et fournit autant de discrimination supplémentaire possible. La procédure poursuit l'ajout de fonctions par la même méthode jusqu'à atteindre le nombre maximum de fonctions, tel que déterminé par le nombre des prédicteurs et le nombre

de catégories des variables dépendantes. Pour sélectionner les variables les plus pertinentes pour le modèle, la méthode pas à pas peut être utilisée.

La méthode de sélection de variables 'lambda de Wilks' pour l'analyse discriminante pas à pas sélectionne les variables en fonction de leur capacité à réduire au minimum le lambda de Wilks. A chaque étape, la variable qui réduit le lambda de Wilks global est repêché. Cette méthode exploite certaine mesures de capacité de discrimination.

Pour mesurer la capacité discriminatoire de chaque variable X_p , nous utilisons analyse de la variance ANOVA (*univariate*). Sa formule de décomposition est :

$$\underbrace{\sum_{i=1}^{I}\sum_{n=1}^{N_i}\left(X_{jin}-\overline{X_j}\right)^2}_{\text{Total covariance}} = \underbrace{\sum_{i=1}^{I} N_i\left(\overline{X_{ji}}-\overline{X_j}\right)^2}_{\text{Separate – groups covariance}} + \underbrace{\sum_{i=1}^{I}\sum_{n=1}^{N_i}\left(X_{jin}-\overline{X_{ji}}\right)^2}_{\text{Within – groups covariance}} \qquad (12)$$

Nous considérons un ensemble de N observations constitué des variables X_1 , …, X_p . Ces observations sont partitionnées par une variable qualitative en I classes. Ces classes contiennent de N_1 à N_I observations, respectivement.

X_{jin} est la valeur X_j de la $n^{ième}$ observation appartenant à la classe i

$\overline{X_{jl}}$ est la moyenne de X_j à travers la classe i

$\overline{X_j}$ est la moyenne de X_j .

Pour chaque variable X_p , le lambda de Wilks est calculé comme le ratio entre la covariance intra-groupes et la covariance totale. Plus la valeur lambda est petite plus cela indique une plus grande capacité de discrimination.

$$\Lambda_p = \frac{within\ group\ sum\ of\ squares}{total\ sum\ of\ squares} \qquad (13)$$

Dans ce travail, nous établissons un modèle qui classe les locuteurs dysarthriques dans l'un des quatre groupes prédéfinis de 'niveaux de gravité de dysarthrie'. Ce modèle utilise onze caractéristiques prosodiques qui ont été sélectionnées avec la méthode 'lambda de Wilks' en utilisant une analyse discriminante. Pour déterminer la relation entre une variable dépendante catégorielle (niveau de gravité) et les variables indépendantes (onze caractéristiques), nous utilisons une procédure de régression linéaire (Kadi et al., 2014).

b) **Gaussian Mixture Models**

Gaussian Mixture Models (GMM) est un modèle statistique bien connu dans le domaine de la classification des données. Ce modèle utilise une technique générative pour l'estimation de la fonction de densité de probabilité (PDF) à partir d'un ensemble de données.

Les deux étapes principales pour la modélisation par GMM sont :

Apprentissage : calcul des caractéristiques de la distribution des variables aléatoires en les modélisant comme une somme d'un ensemble de gaussiennes. Les paramètres calculés sont donc la moyenne, la variance (*w*) et le poids (). Dans ce travail, nous utilisons l'algorithme E*xpectation-Maximization* (EM) (Campbell et Karam, 2009) pour optimiser itérativement les paramètres gaussiens selon un critère de maximum de vraisemblance, dans le but d'obtenir un modèle qui approche le plus possible la distribution concernée.

Reconnaissance : une fois l'estimation des modèles effectuée, il s'agit d'attribuer à chaque observation la catégorie (classe) à laquelle il appartient avec un maximum de probabilité. Pour obtenir cette probabilité, la fonction PDF *(Probability Density Function*) est utilisée.

c) **Support Vector Machines**

La méthode *Support Vector Machine* (SVM) est une méthode de classification discriminative. Proposée par Vapnik comme une approche de *machine learning* originale, en introduisant la fonction noyau (Abd El-Samie, 2011). L'approche des SVM permet une projection des données sur un nouvel espace de dimension supérieure, puis l'établissement d'un hyperplan séparateur qui maximise la marge entre 2 groupes de données.

La fonction noyau est le constituant essentiel du classifieur SVM. La capacité d'apprentissage et la capacité de généralisation dépendent du choix de cette fonction. Dans nos expériences, la fonction noyau utilisée est *Radial Basis Function* (RBF). Cette dernière est formulée comme suit

$$k(x,y) = exp\left(-\frac{\|x-y\|^2}{2\,\sigma^2}\right) \qquad (14)$$

où le paramètre σ représente la largeur de la gaussienne, sachant que σ est adapté et perfectionné à travers les expériences. Dans ce travail, une validation croisée est utilisé pour l'obtention des paramètres optimaux : le paramètre C (*box constraint*) et le parametre $-\sigma$.

d) **Systèmes hybrides**

Pour bénéficier des deux, la capacité de description des données de la GMM et la performance élevée de classification de la SVM, nous combinons les deux systèmes décrits ci-dessus afin d'utiliser les distributions gaussiennes comme base paramétrique des classifieurs SVMs.

Pour atteindre une meilleure efficacité, seuls les paramètres représentant les moyennes des modèles GMM sont impliqués dans le traitement frontal SVM (Liu et al., 2006).

Partie 3 :

Travaux connexes et bases de données

Partie 3 :

III. Travaux connexes et bases de données :

A. Travaux connexes :

Plusieurs outils et méthodes ont été développés pour aider les locuteurs atteints de dysarthrie. En effet, de remarquables travaux ont été réalisés dans le domaine de la reconnaissance de la parole dysarthrique, l'amélioration de l'intelligibilité de la parole et l'évaluation automatique de la maladie.

Parmi les travaux les plus récents, significatifs et qui impactent le plus la recherche scientifique, dans le domaine du traitement automatisé de la parole dédié aux locuteurs atteins de trouble de la parole, plus précisément la dysarthrie, nous pouvons citer Rudzicz qui a mis au point un système de reconnaissance de la parole dysarthrique et a proposé des méthodes de rehaussement de l'intelligibilité du discours dysarthrique, en s'appuyant sur la base de donnée Torgo qui a été récemment développée par (Rudzicz, Namasivayam et Wolff, 2012).

Bien que, il convient de noter que les efforts de recherche n'ont pas exploré d'autres fonctionnalités qui pourraient être utilisées aux patients atteints de trouble de la parole, tel que la reconnaissance automatique du locuteur qui est de plus en plus utilisée dans divers systèmes de gestion d'identité et de systèmes de sécurité basés sur la biométrie.

Il est également important de noter que le diagnostic et l'évaluation automatisés, qui peuvent aider les cliniciens dans la prise en charge des patients souffrant de troubles de la parole, n'ont pas reçu suffisamment d'attention et que les travaux dans ce champ de recherche sont sporadiques.

Parmi les méthodes d'évaluation automatique de la dysarthrie, nous pouvons citer les travaux développés les plus significatifs pour notre domaine de recherche, traitant précisément un des plus importants troubles de la parole qui est la dysarthrie :

- Dans (Rudzicz, 2012), un système d'évaluation automatique de l'intelligibilité effectuant une classification binaire a été proposé. Cette approche se base sur l'extraction de variations atypiques dans la parole dysarthrique et l'utilisation de deux types de classifieurs, SVM et LDA. Les expériences ont été conduites sur la récente base de données Torgo.

- Une classification basée sur la méthode de distance de Mahalanobis et l'analyse discriminante a été développée pour une classification du niveau de sévérité de la dysarthrie. Une performance de 95% de bonne évaluation a été atteinte sur une échelle de deux niveaux de sévérité, et ce en exploitant des caractéristiques acoustiques (O'Shaughnessy, 2001).

- Un système hybride utilisant la méthode statistique GMM et les réseaux de neurones ANNs a été présenté dans (Shahamiri et Salim, 2014) pour faire la discrimination entre quatre niveaux de sévérité. Prenant les paramètres MFCCs comme front-end, une précision de 86% a été atteinte.

- Dans (Selouani et al., 2012) des réseaux de neurones de type *'feed-forward'*, des SVMs ainsi que des paramètres phonologiques ont été utilisés pour concevoir des modèles discriminants pour la parole pathologique.

B. Bases de données

1. Bases de données existantes

La disponibilité de données est une question cruciale dans le domaine du traitement et d'analyse de la parole pour les pathologies de communication verbale. En effet, il n'est pas aisé de concevoir une base de données contenant des évaluations médicales et plusieurs

enregistrements vocaux pour chaque patient, les contraintes sont liées, d'une part, à la disponibilité et l'approbation des malades et des organismes médicaux et, d'autre part, aux conditions délicates d'enregistrement. De telles données existent seulement pour quelques langues, majoritairement en anglais, en espagnol, en coréen et en allemand (Baghai-Ravary et Beet, 2013).

Étant donné que les locuteurs dysarthriques sont en minorité et sont sensibles à la fatigue, la collecte de données de cette population peut être particulièrement difficile. Collecter des données avec la participation de locuteurs dysarthriques implique généralement moins de 5 participants (Hasegawa-Johnson et al., 2006), produisant souvent seulement 25 enregistrements chacun (Jayaram et Abdelhamied, 1995).

- La base de données Nemours de l'institut *A.I. duPont* est une source populaire de données acoustiques étiquetées au niveau phonémique, Nemours inclut les enregistrements de onze locuteurs dysarthriques masculins qui ont des degrés d'intelligibilité différents, et des enregistrements d'un locuteur témoin (non dysarthrique).

- La base de données MOCHA de l'Université d'Édimbourg (Ecosse) se compose de 460 phrases dérivées de TIMIT (Zue, Seneff et Glass, 1989), prononcées par un locuteur et une locutrice tous deux Britanniques et ne sont pas atteints de dysarthrie (Wrench, 1999). Toutes les données acoustiques sont alignées temporellement par une articulographie électromagnétique, laryngographie et électropalatographie.

- Des données de microfaisceaux de rayons x ont été recueillies pour quinze personnes atteintes de la maladie de Parkinson et de *Amyotrophic lateral sclerosis* (ALS) (Yunusova et al. (2008). Cette base de données comprend dix locutions par patient, ce qui est peu pour des systèmes de reconnaissance automatique.

- A l'Université de l'Illinois (USA) a été enregistrée la base de données *Universal Access (UA-Speech).* Elle est composés de 17 participants atteints de paralysie cérébrale (Kim et al., 2008). Chaque participant dans cette base de données a

prononcé 765 mots isolés provenant de sources telles que des chiffres, l'alphabet radio, des mots communs de dictionnaire. *UA-Speech* ne contient pas d'enregistrements de mots connectés constituant des phrases.

- Plus récemment, la base de données Torgo a été développée par l'Université de Torronto (Canada) en collaboration avec l'Hopital Holland–Bloorview de réhabilitation pour enfants à Toronto. Torgo contient des enregistrements de huit patients dysarthriques et sept locuteurs témoins. Chaque locuteur a produit des enregistrements audio de différents types : non-mots, mots courts, phrases prédéfinies et discours libres (Rudzicz et al., 2012).

2. Les données exploitées

Les deux bases de données, Torgo et Nemours ont été utilisées dans ce travail pour l'élaboration et l'évaluation des méthodes proposées, ce qui assure la disponibilité d'une large quantité de données et une diversité suffisante dans les enregistrements.

a) **Nemours**

Nemours est l'une des rares bases de données de parole dysarthrique, contenant des enregistrements de 11 malades dysarthriques américains qui sont atteints à des degrés de sévérité différents. Chaque locuteur dysarthrique prononce 74 phrases courtes et non-sens, plus deux paragraphes communément utilisés, « *Grandfather* » et « *Rainbow* ». Un ensemble de 74 noms monosyllabes et un autre ensemble de 37 verbes dissyllabes sont utilisés pour former les phrases. La structure de celles-ci est « *The X is Ying the Z* », *X* et *Y* sont sélectionnés aléatoirement à partir des deux ensembles de mots, respectivement. A la fin, chaque locuteur aurait prononcé deux fois chaque nom et chaque verbe. Toutes les données sont étiquetées au niveau des mots, ainsi qu'au niveau des phonèmes pour les phrases de 10 patients. Par la suite,

le pathologiste de la parole qui a conduit les enregistrements a prononcé le corpus tout entier, pour une considération comme témoin sain.

Un pathologiste spécialisé dans les troubles de la parole a effectué une évaluation des fonctions motrices pour chaque patient, suivant le protocole *Frenchay Dyrathria Assessment (FDA, Enderby, 1983).* L'évaluation contient 8 différentes sections : reflexe, respiration, lèvres, palais, larynx, langue et intelligibilité, ainsi que des sections de facteurs influents comme la sensation et la vitesse.

Pour la conception de systèmes automatiques d'évaluation de la dysarthrie, un score global qui représente l'évaluation FDA de chaque patient doit être utilisé. En ce qui concerne la base de données Nemours, les scores présentés dans les travaux (Menendez et al., 1996) seront exploités.

b) **Torgo**

La base de données Torgo contient, d'une part, des données de 8 patients dysarthriques anglophones (3 femmes et 5 hommes) dont les niveaux de sévérité de la maladie sont différents et, d'autre part, des enregistrements de 7 locuteurs non-dysarthriques, considérés comme groupe de contrôle. Torgo a été développée par l'Université de Toronto, au département d'Informatique et d'Orthophonie, en collaboration avec l'Hopital Holland–Bloorview de réhabilitation pour enfants à Toronto.

Ainsi, 23 heures de parole dysarthrique ont été collectées entre 2008 et 2010 utilisant deux différents microphones, un microphone matriciel et un microphone à casque. Les fréquences d'échantillonnages utilisées sont respectivement de 44.1 kHz et de 22.1 kHz, donnant suite à un sous-échantillonnage à 16kHz a été effectué sur tous les enregistrements.

Par ailleurs, la base de données Torgo contient des enregistrements 3D des mouvements articulatoires qui peuvent permettre une étude détaillée sur l'activité de la parole dysarthrique. Ces données ont été récoltées par un articulographe 3D électromagnétique (Rudzicz, 2012).

Chaque locuteur a produit des enregistrements audio structurés par différents types : non-mots, mots courts, phrases prédéfinies et discours libres. Pour les besoins des traitements automatiques de la parole, nous avons réorganisé la base de données en créant de nouveaux répertoires et en renouvelant le nom de chaque fichier audio par une nomination normalisée. Torgo contient aussi des fichiers de l'évaluation FDA détaillée pour chaque patient dysarthrique, ces fichiers contiennent un score sur une échelle de 9 points pour 28 dimensions perceptuelles. Etant donné que des scores globaux représentant l'évaluation FDA ne sont pas disponibles, une nouvelle évaluation FDA globale est proposée, dans ce travail, pour chacun des 8 patients dysarthriques de la base de données Torgo. Cette proposition, qui sera détaillée dans le prochain chapitre, utilise les scores des dimensions perceptuelles et est fondée sur la base du protocole *Frenchay Dyrathria Assessment second edition* (FDA-2, Enderby et Palmer, 2008).

c) **Les locuteurs**

Les locuteurs de la base de données Nemours sont 11 jeunes hommes adultes affectés par un type différent de dysarthrie qui fait suite à une paralysie cérébrale *(CP)* ou à un traumatisme crânien *(HT).* Parmi les participants, un homme adulte représente le locuteur contrôle (sain). Sept des onze patients sont atteints de paralysie cérébrale, parmi eux, deux ont une paralysie cérébrale athétoïde (un tétraplégique), trois ont une paralysie cérébrale spastique avec tétraplégie, et deux ont une combinaison d'athétoïde et spastique avec une quadriplégie. Un code de deux lettres est attribué à chaque locuteur dysarthrique: BB, BK, BV, FB, JF, KS, LL, MH, RK, RL et SC. Sachant que les données perceptuelles et l'évaluation de la parole de deux locuteurs n'ont pas été prises en compte, ces deux locuteurs représentent les cas dysarthriques, le moins atteint et le plus gravement atteint, correspondant respectivement aux patients FB et KS (Menendez et al., 1996).

La base de données Torgo d'articulation dysarthrique est constituée de données acoustiques alignées, ces données sont récoltées à partir de locuteurs atteints d'une paralysie cérébrale *(CP)* ou sclérose latérale amyotrophique *(SLA)* qui sont deux des causes les plus récurrentes de trouble de la parole. Chacun des huit locuteurs dysarthriques a enregistré environ

500 élocutions, tandis que tous les locuteurs du groupe de contrôle ont enregistré environ 1200 élocutions chacun. Un code a été attribué à tous les participants, ce code commence par 'F' pour les participantes féminines et par 'M' ' pour les locuteurs masculins. La lettre 'C' est ajoutée avant le code qui spécifie le genre pour les locuteurs du groupe de contrôle. Les deux derniers chiffres désignent l'ordre auquel les participants ont été inscrits. Ainsi, les 8 locuteurs dysarthriques, de la base de données Torgo, impliqués dans nos expériences sont : F01, F03, F04, M01, M02, M03, M04 et M05 (Rudzicz, 2011).

3. Combinaison de données de sources différentes

Il est important de contrôler les différences entre les bases de données afin de minimiser les effets inattendus de la combinaison des données sur les expériences.

Concernant la base de données Nemours, les sessions d'enregistrement ont été menées dans une petite pièce avec absorbation sonore, alors que pour la base de données Torgo, une réduction du bruit acoustique a été effectuée. Pour les deux bases de données, la méthode d'échantillonnage Pulse Code Modulation (PCM) a été utilisée pour le codage de la parole, et le signal de parole a été échantillonné à une fréquence de 16 kHz avec une résolution d'échantillon de 16 bits. Par ailleurs, le format de fichier RIFF *(Resource Interchange File Format)* a été utilisé pour formater les fichiers audio.

De plus, nous avons effectué une normalisation et une suppression de silences à l'étape de prétraitement sur tous les enregistrements que comportent les deux bases de données, Nemours et Torgo.

L'évaluation de la dysarthrie est basée sur le protocole *Frenchay Dyrathria Assessment* (*FDA)* (Enderby, 1983). Les deux bases de données Nemours et Torgo contiennent des fichiers d'évaluation FDA détaillés pour chaque locuteur dysarthrique. Cette partie sera présentée plus en détail dans le prochain chapitre.

Partie 4 :

Evaluation-diagnostique automatique de la parole pathologique

Partie 4 :

IV. Evaluation-diagnostique automatique de la parole pathologique

L'intelligibilité de la parole dysarthrique peut être classée sur une échelle entre 'proche de la normale' et 'inintelligible', selon la gravité de la maladie. Habituellement, pour évaluer la sévérité du trouble de la parole ou pour estimer le progrès d'une rééducation, une large batterie de tests est nécessaire pour l'évaluation du niveau d'intelligibilité. Des solutions de diagnostic et évaluation des pathologies de la parole pourraient aider les cliniciens dans la prise en charge des patients et le suivi de l'évolution de la maladie (Kadi et al.,2014).

A. Evaluation du niveau de gravité de la dysarthrie (FDA)

Pour les deux bases de données Nemours et Torgo, une évaluation des fonctions motrices de chaque locuteur dysarthrique a été effectué en se basant sur le protocole d'évaluation standardisée de dysarthrie 'Frenchay Dysarthria Assessment' (FDA, Enderby 1983), ces évaluations ont été conduites par un orthophoniste. Le test est divisé en 8 sections : reflexe, respiration, lèvres, mâchoire, palais, larynx, langue et intelligibilité. Par ailleurs, l'évaluation FDA contient une section sur l'influence de facteurs tels que la sensation et la vitesse d'élocution. À travers les huit sections, un ensemble global de 28 dimensions perceptuelles sont évaluées en utilisant un score sur une échelle de 9 points.

Afin de concevoir notre système automatique qui sert à évaluer la gravité de la dysarthrie, une seule valeur numérique est nécessaire pour représenter le score global FDA de

chaque patient. En se basant sur (Menendez et al., 1996), les niveaux de gravité des locuteurs dysarthriques de Nemours sont représentés dans le tableau 4 :

Patients	KS	SC	BV	BK	RK	RL	JF	LL	BB	MH	FB
Severity (%)	-	49.5	42.5	41.8	32.4	26.7	21.5	15.6	10.3	7.9	7.1

Tableau 4| Niveau de sévérité des locuteurs dysarthriques de la base de données Nemours

Nous devons préciser que ces scores ne sont pas disponibles pour les locuteurs dysarthriques de la base de données Torgo. Seul un fichier contenant les taux sur l'échelle de 9 point des 28 dimensions perceptuelles sont fournis pour chaque patient. Par conséquent, un nouveau score global FDA est proposé, dans ce travail, pour chacun des huit participants dysarthriques de la base de données Torgo. Cette estimation utilise les scores des dimensions perceptuels fournis avec la base de données, en se basant sur le récent protocole FDA-2 qui est une amélioration de la version précédente (Enderby, 2008).

Section	Perceptual dimension
Reflex	Cough
	Swallow
	Dribble/Drool
Resp.	At Rest
	In Speech
Lips	At Rest
	Spread
	Seal
	Alternate
	In Speech
Jaw	At Rest
	In Speech
Palate	Fluids
	Maintenance
	In Speech
Laryneal	Time
	Pitch
	Volume
	In Speech
Tongue	At Rest
	Protrusion
	Elevation
	Lateral
	Alternate
	In Speech
Intel.	Words
	Sentences
	Conversation

Tableau 5| Structure de l'échelle de notation du protocole FDA

Le tableau 5 présente le barème FDA utilisé par les cliniciens pour évaluer la capacité du locuteur dysarthrique, sur une gamme de comportements liés à chaque fonction de la parole évaluée, ces fonctions représentent les sections sur le tableau 5. La capacité de chaque paramètre est évaluée de fonction-normal à non-fonction utilisant l'échelle de 9 points, qui contient 5 descripteurs (a, b, c, d et e) + ½ point. La nouvelle méthode pour l'estimation des scores globaux FDA proposé dans cette étude est basée sur le processus qui contient les deux étapes suivantes :

- Dans la première étape, pour chaque section, un score sous-jacent est estimé en calculant la moyenne des taux des dimensions perceptuelles qui appartiennent à la section correspondante. Suivant le protocole FDA-2, nous avons supprimé les tests de la mâchoire de l'évaluation car les nouvelles avancées dans le domaine ont montré que les patients dysarthriques sont rarement affectés par un trouble de la mâchoire, ainsi l'information ne contribuait pas de manière positive à l'évaluation.

- Pour la deuxième étape, un taux global est calculé en pourcentage (%) en utilisant la moyenne des sept scores sous-jacents (de la première étape). Le protocole FDA-2 précise que la fonction laryngale et les mouvements des lèvres contribuent substantiellement au manque de l'intelligibilité. C'est pourquoi, il est approprié de se concentrer sur ces fonctions de la parole (Laryneal et Lips) plus que les autres aspects (Enderby et Palmer, 2008). Sur cette base, nous avons attribué plus de poids pour les sections Laryneal et Lips dans l'estimation du nouveau score global FDA.

Le tableau 6 contient les résultats finaux de la nouvelle méthode d'estimation des scores pour les patients de la base de données Torgo:

Patients	**M04**	**M01**	**M02**	**F01**	**M05**	**M03**	**F03**	**F04**
Score rate (%)	44.35	49.79	49.79	55.87	57.96	94.29	96.67	96.67

Tableau 6| La nouvelle proposition de scores FDA-2 pour les locuteurs de Torgo

Le score FDA exprime le niveau d'intelligibilité, et il est inversement correspondant au niveau de gravité de dysarthrie. Tous les locuteurs dysarthriques peuvent être classés en trois sous-groupes, selon leur évaluation. Les sous-groupes sont: 'L1 faiblement atteints' qui comprend les patients F04, F03, M03, FB, MH, BB et LL ; 'L2 sévère' contenant les participants M05, F01, JF, RL, RK, BK et BV ; et le sous-groupe des plus gravement atteints 'L3 sévère", qui comprend les locuteurs M01, M02, M04, SC et le KS.

B. Front-end

1. Paramètres prosodiques

Un ensemble de 11 paramètres prosodiques a été sélectionné par le biais d'une analyse discriminante LDA, plus précisément en utilisant la mesure Wilks' lambda. Les indices suivants ont donc été déterminés comme étant les plus adaptés à la classification des niveaux de sévérité de la dysarthrie : Jitter, Shimmer, Pitch médian, Ecart-type du Pitch, Nombre de Périodes, Ecart-type de la Période, Proportion de la durée Vocalique (%V), Ratio Harmoniques sur Bruit (dB), Ratio Bruit sur Harmoniques (%), Vitesse d'Articulation et Degré des Pauses de la voix. Les valeurs de Wilks' lambda varient entre 0 et 1. Plus la mesure est petite, plus l'aptitude de discrimination du paramètre concerné est grande (tableau 7) (Kadi et al., 2014).

D'autre part, nous utilisons une consolidation des paramètres présentés précédemment, à savoir les MFCCs conventionnels et les indices acoustiques dérivés d'études sur le phénomène de l'audition. Le choix de consolider deux sources d'information du signal de parole a pour but de minimiser la partie d'apport d'information qui peut manquer en utilisant une seul source.

Caractéristique	Wilks' lambda
Articulation rate	0.565
Number of period	0.595
Mean pitch	0.701
Voice breaks	0.835
%V	0.861
HNR	0.864
Jitter	0.925
Shimmer	0.962
Std Pitch	0.979
Std Period	0.984
NHR	0.989

Tableau 7| Wilks' lambda des paramètres prosodiques

2. Indices du modèle d'oreille

Le modèle auditif, exploité dans nos expérimentations, simule l'oreille externe, l'oreille moyenne et l'oreille profonde. Cette modélisation a été proposée par Caelen et adaptée pour être exploitée comme un module front-end dans les systèmes de reconnaissance de la parole par Selouani. Les indices auditifs sont calculés à partir d'une combinaison linéaire des énergies de sorties des canaux correspondant à 24 filtres chevauchés.

Les 7 indices, dérivés du modèle d'oreille, utilisés comme des caractéristiques distinctives sont : Grave/Acute (G/A), Open/Closed (O/C), Diffuse/Compact (D/C), Flat/Sharp (F/S), Mellow/Strident (M/S), Continuant/Discontinuant (C/D) et Tense/Lax (T/L). Pour rappel, la théorie relative à ces indices auditifs est présentée et détaillée au niveau de la 2eme partie de cette thèse (Partie 2 : Backround).

Les figures 7 et 8 donnent un exemple graphique des caractéristiques basées sur le modèle d'oreille présenté. Ces caractéristiques ont été calculées à partir d'une phrase prononcée par un patient dysarthrique et par un locuteur témoin, la phrase est « The sin is sitting the who » extraite de du corpus Nemours.

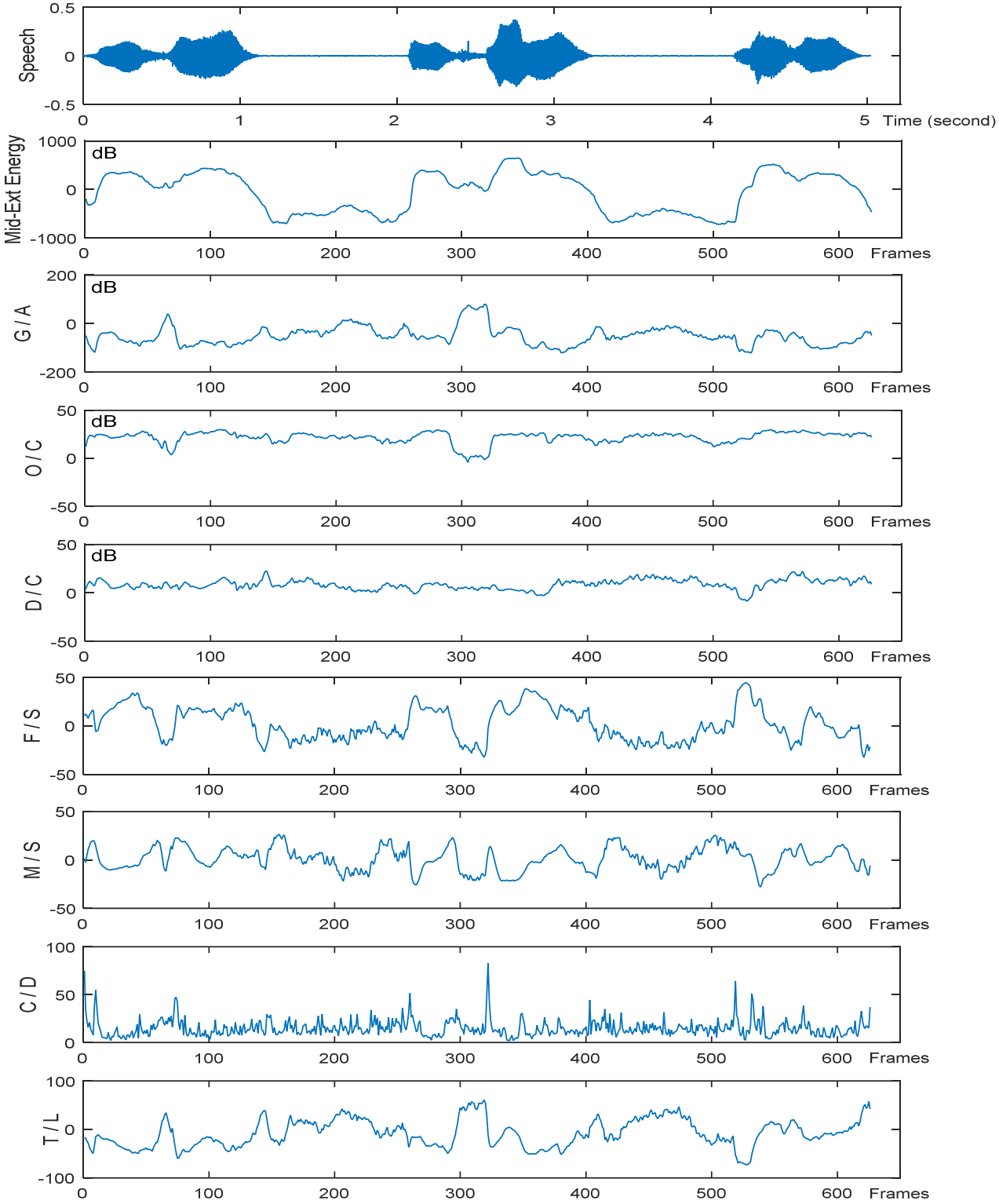

Figure 7| Indice des caractéristiques de la fonction auditive, calculés à partir de la phrase: "The sin is sitting the who" prononcés par un locuteur dysarthrique, G/A: Grave/Acute, O/C: Open/Closed, D/C: Diffuse/Compact, F/S: Flat/Sharp, M/S: Mellow/Strident, C/D: Continuant/Discontinuant, T/L: Tense/Lax

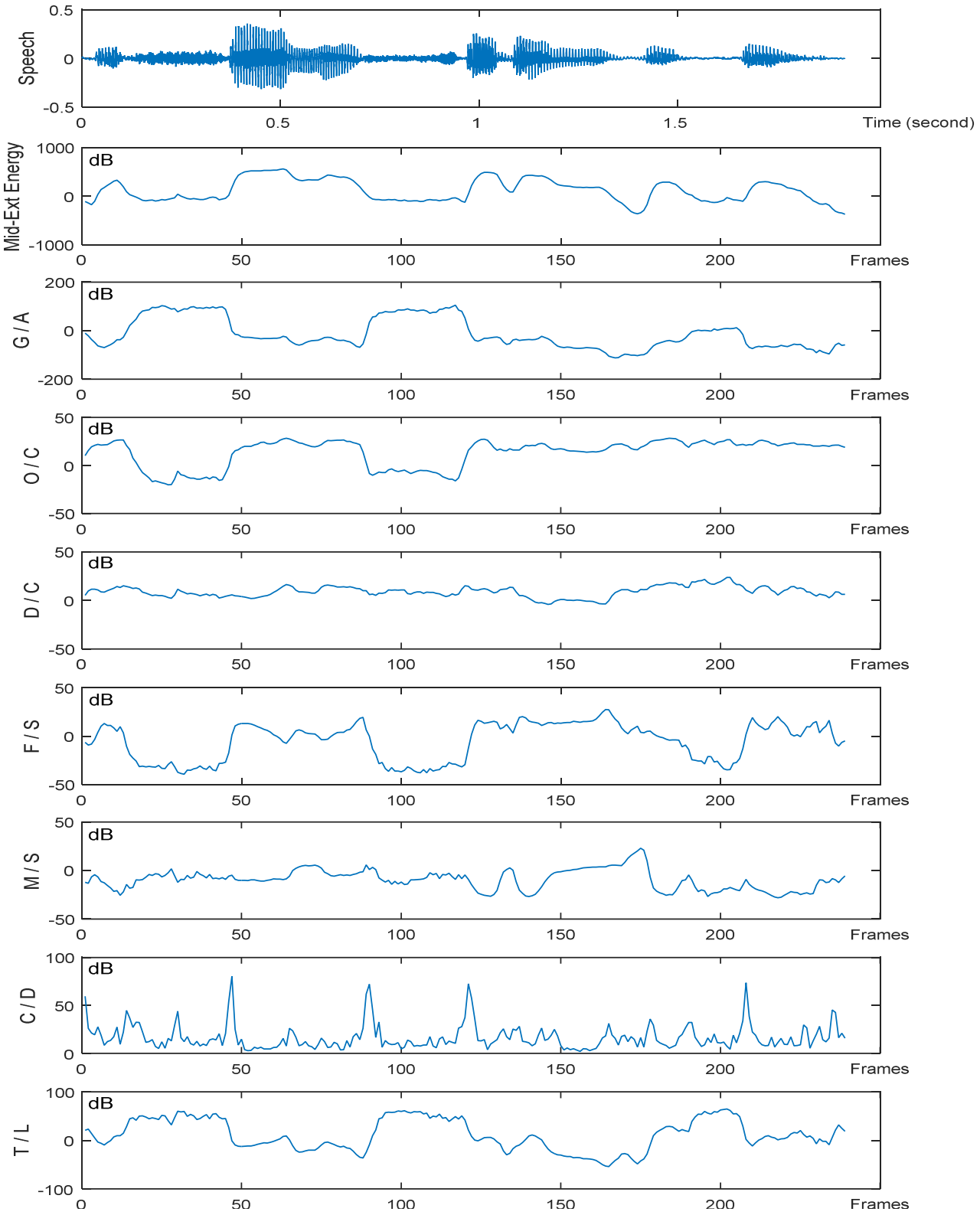

Figure 8| Indice des caractéristiques de la fonction auditive, calculés à partir de la phrase: "The sin is sitting the who" prononcés par un locuteur témoin G/A: Grave/Acute, O/C: Open/Closed, D/C: Diffuse/Compact, F/S: Flat/Sharp, M/S: Mellow/Strident, C/D: Continuant/Discontinuant, T/L: Tense/Lax.

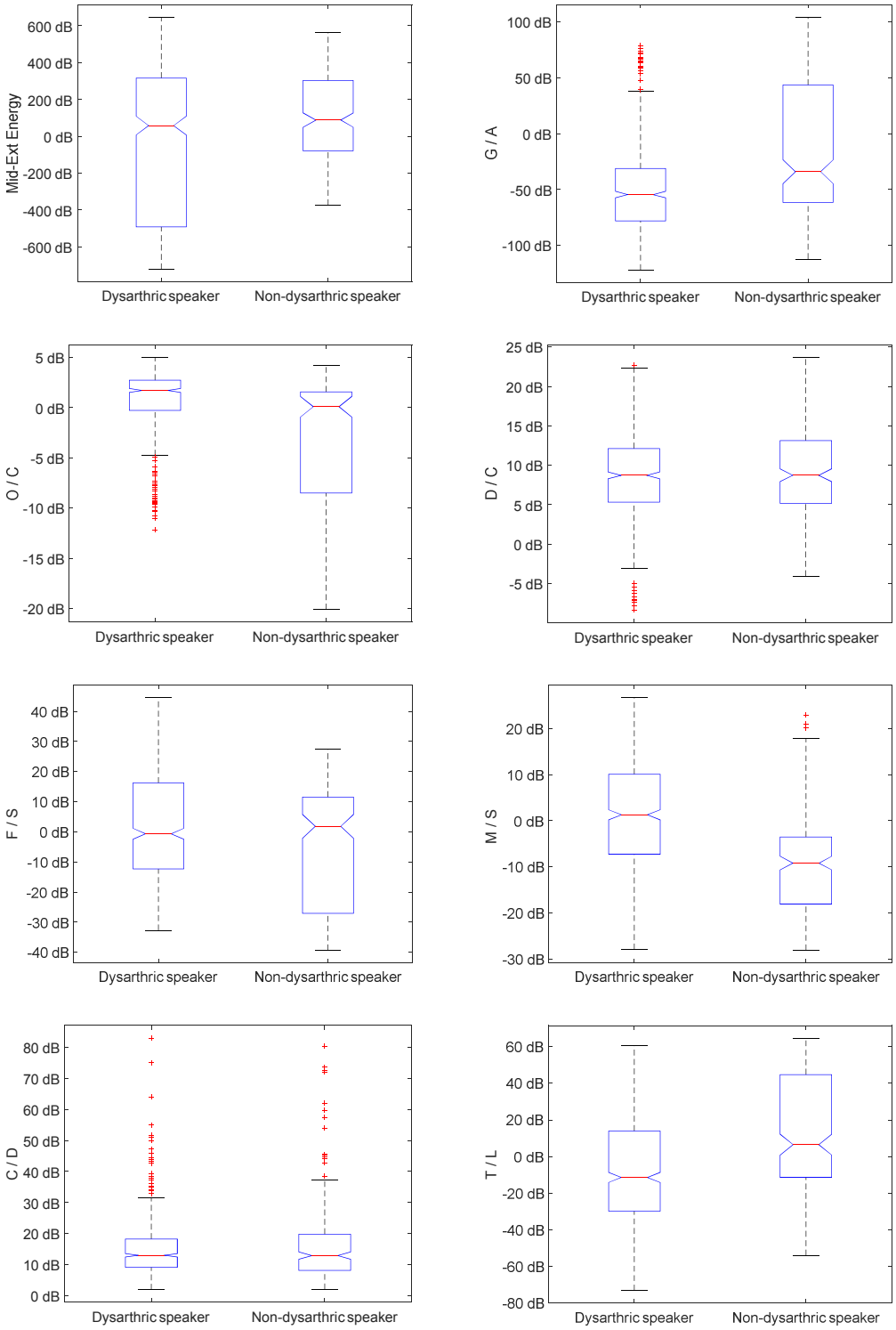

Figure 9| Boxplot de chaque aractéristique auditif (locuteur dysarthrique / locuteur non-dysarthrique).

Les différences visibles peuvent être observées sur les figures représentant les indices auditifs d'un locuteur dysarthrique et d'un autre locuteur non-dysarthrique. Ces différences sont statistiquement analysées à l'aide d'un diagramme *boxplot*, ce dernier montre la distribution des indices auditifs basé sur leur minimum, premier quartile, médiane, troisième quartile et maximum. La figure 9 représente la distribution des caractéristiques auditives représentées dans les figures précédentes.

La médiane est représentée par la marque centrale sur chaque boîte, les marges sont les 25ème et 75ème centiles, et les moustaches s'étirent jusqu'au point extrême des indices auditifs. Tel qu'illustré sur la figure 9, la discrimination peut facilement être faite entre les phrases prononcées par le locuteur dysarthrique et le locuteur non-dysarthrique pour les indices Grave/Acute, Open/Closed, Flat/Sharp, Mellow/Strident et Tense/Lax. Cependant, la différence n'est pas évidente pour les indices Continuant/Discontinuant et Diffuse/Compact.

En plus de la représentation par boxplot, une analyse de variance à un facteur (one-way ANOVA) a été réalisée pour chacun des huit indices en les considérant comme les variables dépendantes. Les résultats dans le tableau 8 montrent que le facteur dysarthrique/non-dysarthrique a un nombre important d'effets significatifs sur les caractéristiques auditives proposées. Cela inclut 7 paramètres sur 8 : Mid-Ext Energy, G/A, O/C, F/S, M/S, C/D, T/L. Ce qui correspond aux observations qui peuvent être faites sur les figures 7 et 8 et sur les diagrammes boxplot.

Features	**Mid-Ext Energy**	**G/A**	**O/C**	**D/C**	**F/S**	**M/S**	**C/D**	**T/L**
p-value	2.02e^(-9)	6.08e^(-25)	5.32e^(-31)	0.410	9.81e^(-8)	1.23e^(-30)	0.047	3.21e^(-18)

Tableau 8| Significance statistique (valeurs du p) basée sur la méthode one-way ANOVAs, avec locuteur dysarthrique versus locuteur non-dysarthrique comme variable indépendante. La significance est atteinte lorsque p < 0.05 (en gras).

C. Expérience série 1

La première série d'expériences concerne l'évaluation automatique de la dysarthrie en utilisant les 11 paramètres prosodiques sélectionnés par LDA comme front-end et deux méthodes d'apprentissage automatique, GMM et SVM. La base de données Nemours est exploitée dans le cadre de cette série. En effet, les patients de Nemours peuvent être divisés en 3 groupes selon leurs scores FDA (voir tableau 4) ; le niveau 'L1' regroupe les patients les moins atteints : FB, BB, MH et LL ; le niveau 'L2' inclut les patients atteints assez sévèrement : RK, RL et JF ; le groupe 'L3' représente un niveau très sévère de la maladie incluant les locuteurs : KS, SC, BV et BK. Le niveau 'L0' est réservé au groupe de contrôle (non-dysarthrique).

Les deux méthodes, GMM et SVM, effectuent un apprentissage et une classification. Nous divisons l'ensemble des phrases du corpus en deux sous-ensembles : un sous-ensemble pour l'apprentissage qui contient 70 % des phrases avec des niveaux de gravité de dysarthrie différents et un sous-ensemble de test qui contient 30 % des phrases. Le sous-ensemble d'apprentissage comprend 459 phrases de parole dysarthrique et 153 phrases de parole non-dysarthrie (HC) ; le sous-ensemble de test contient 207 phrases de parole dysarthrique et 69 phrases de parole non-dysarthrie (HC).

1. LDA

Dans cette section, les résultats de l'analyse discriminante sont présentés. Cette analyse utilise la méthode pas à pas, une procédure de régression linéaire et le lambda de Wilks'.

La méthode pas à pas a été choisie dans la mesure où nous avons plusieurs caractéristiques prosodiques calculées, et nous n'avons pas de prédictions réelles sur l'importance ou non de chaque caractéristique. Néanmoins, le risque de cette méthode est de sélectionner une variable sur une base statistique seulement, mais qu'en pratique cette variable n'ait pas de pertinence pour le cas étudié.

La représentation graphique résultante de l'analyse discriminante linéaire (LDA) montre la discrimination des 4 niveaux L0, L1, L2 et L3 par les indices prosodiques utilisés (figure 10). Cette analyse utilise la méthode pas à pas, une procédure de régression linéaire et la mesure Wilks' lambda. Au cours de l'analyse, trois fonctions discriminantes sont générées (nombre de groupes -1) dont les deux premières sont les plus significatives.

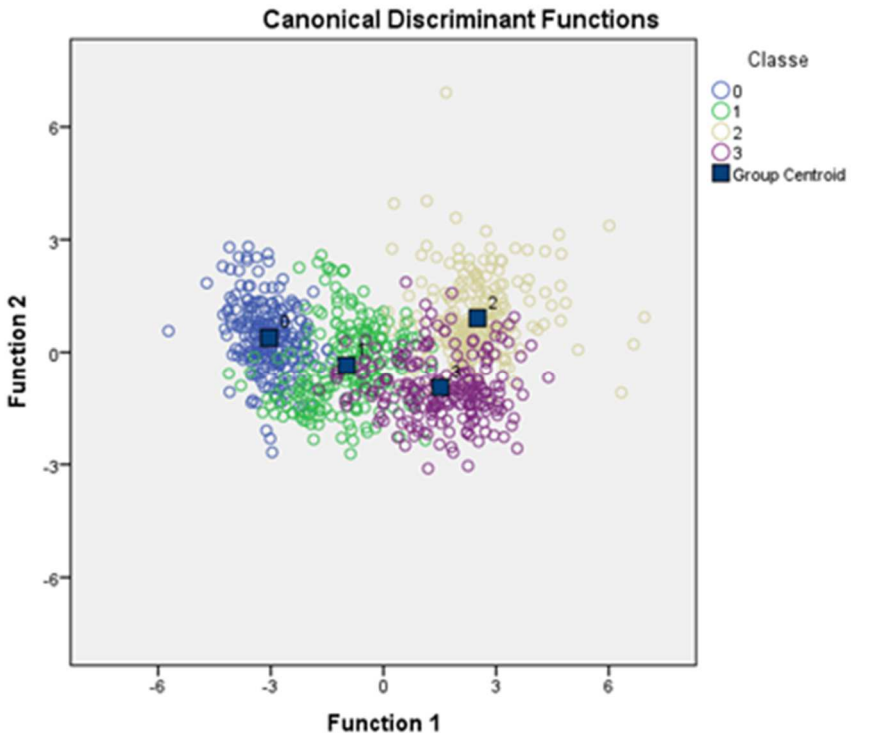

Figure 10| Représentation de la discrimination des 4 classes par les deux premières fonctions LDA (fonction1 et fonction 2)

Dans ce travail, nous établissons un modèle qui classe les locuteurs dysarthriques dans l'un des quatre groupes prédéfinis de 'niveaux de gravité de dysarthrie'. Ce modèle utilise onze caractéristiques prosodiques qui ont été sélectionnées avec la méthode 'lambda de Wilks' en utilisant une analyse discriminante. Pour déterminer la relation entre une variable dépendante catégorielle (niveau de gravité) et les variables indépendantes (onze caractéristiques), nous utilisons une procédure de régression linéaire (Ka di et al., 2014).

Le sommaire de l'analyse dirimante linéaire LDA avec les taux de bonne classification de la dysarthrie sont présentés dans le tableau 9.

La taille des quatre groupes représentant chaque catégorie L0, L1, L2 et L3 est la même. Aussi, nous considérons les mêmes probabilités à priori de l'appartenance à un certain groupe, pour toutes les classes.

	Class	Predicted Group Membership				Total
		0	1	2	3	
Count	0	215	7	0	0	222
	1	27	185	1	9	222
	2	0	1	193	28	222
	3	0	31	29	162	222
%	0	96,8	3,2	0,0	0,0	100
	1	12,2	83,3	0,5	4,1	100
	2	0,0	0,5	86,9	12,6	100
	3	0,0	14,0	13,1	73,0	100

Tableau 9| Résultats de la LDA

2. LDA-GMM

La classification automatique de vecteurs observés en I nombre de classes peut être effectuée à l'aide de la méthode de modélisation GMM. Les deux principales étapes de cette modélisation sont :

Etape 1 : Apprentissage

Pour chaque classe Ci du corpus, l'apprentissage est initié pour obtenir un model contenant les caractéristiques de chaque distribution de gaussienne m de cette classe. Ces caractéristiques sont : le vecteur des moyennes $\mu_{i,m}$, la matrice de covariance $\sum_{i,m}$ et les poids gaussiens $w_{i,m}$. Ces paramètres sont calculés un certain nombre d'itérations de l'algorithme EM (expectation-maximization) (Dempster, Laird et Rubin, 1977). Finalement, un modèle GMM sera généré pour chaque niv eau de sévérité de la dysarthrie.

Etape 2 : Reconnaissance

Chaque signal *X* est représenté par un vecteur acoustique x contenant p composants qui sont les paramètres prosodiques. La taille du vecteur acoustique est le nombre des paramètres acoustique extrais du signal de parole, qui est onze dans le cas présent. La probabilité de chaque vecteur acoustique d'appartenir à une certaine classe *Ci* est calculée comme suit (Reynolds et Rose, 1995):

$$p(x\backslash C_i) = \sum_{m=1}^{M} w_{i,m} \cdot \frac{1}{\sqrt{(2\pi)^d \left|\sum_{i,m}\right|}} \cdot e^{A_{i,m}}$$

$$A_{i,m} = \left(-\frac{1}{2}\left(x - \mu_{i,m}\right)^T \cdot \frac{1}{\sum_{i,m}} \cdot \left(x - \mu_{i,m}\right) \right) \qquad (15)$$

où M est le nombre de Gaussiennes.

Pour cette série d'expériences, chaque phrase est représentée par un vecteur acoustique contenant onze caractéristiques prosodiques, non pas par un vecteur pour chaque fenêtre d'analyse. La probabilité du signal est notée comme *p(x\Ci)* et l'algorithme estime que le signal *X* va appartenir au groupe *Ci* selon la plus grande probabilité *p(x\Ci)* .

La figure 11 représente le système LDA-GMM. En utilisant l'algorithme EM, un modèle de mélange de gaussiennes est créé pour chaque niveau de gravité de la dysarthrie L1, L2 et L3 ainsi que pour L0. La meilleure performance du système a été atteinte avec huit Gaussiennes (M=8), elle est de 88.89% de classification correcte. Sachant que la métrique d'évaluation de performance du système d'évaluation de la maladie est obtenue comme suit :

$$Performance\ rate = \left(\frac{number\ of\ correct\ severity\ classifications}{total\ number\ of\ severity\ classification\ trials}\right) \times 100 \qquad (16)$$

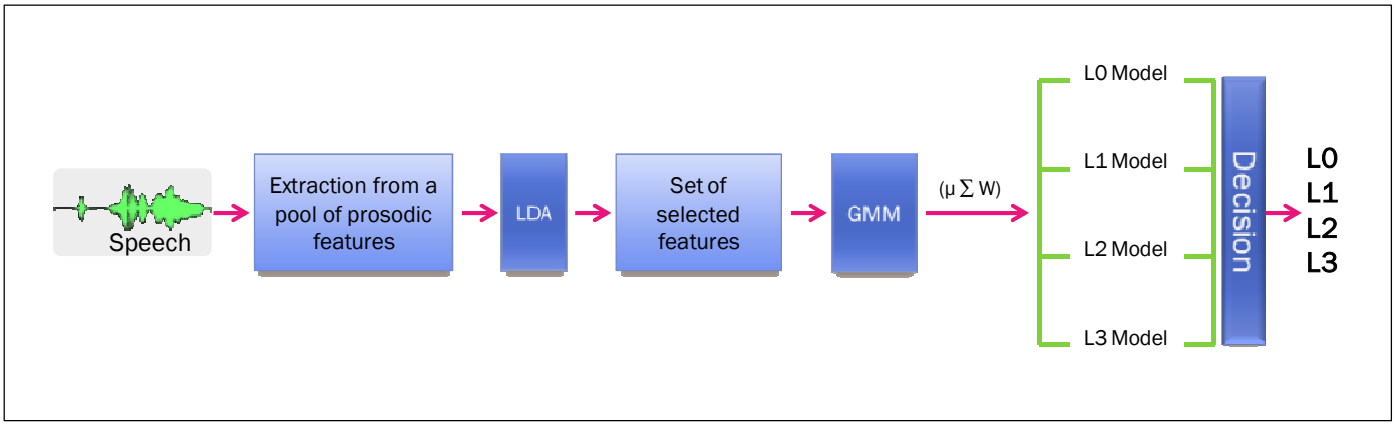

Figure 11| le système LDA-GMM

3. LDA-SVM

La méthode SVM est une méthode de classification binaire qui utilise un hyperplan séparateur le plus optimal possible pour faire la discrimination en deux classes distinctes.

L'approche SVM a été proposée par Vapnik comme une nouvelle méthode de *machine learning* à travers l'introduction d'une fonction noyau (Vapnik, 1999). La fonction noyau appelée Kernel fonction projette les données, non-séparables linéairement, vers un nouvel espace de grande dimension dans lequel une séparation linéaire serait possible. Les SVMs déterminent l'hyperplan séparateur linéaire qui maximise la marge entre les 2 classes de données.

La deuxième méthode de classification automatique LDA-SVM est illustrée dans la figure 12. Cette méthode utilise un front-end similaire à celui utilisé pour l'approche LDA-GMM, c'est-à-dire, les 11 caractéristiques prosodiques sélectionnées par la LDA. Par ailleurs le classifieur automatique est conçu en utilisant un système multi-classe-SVM de type 'one-againts-one' avec la fonction noyau RBF (Radial Basis Fonction). Ce SVM multi-classes a été développé pour classer automatiquement les quatre niveaux de sévérité de la dysarthrie. Par conséquent, plusieurs SVM binaires sont développés pour différencier les classes Ci et Cj , sachant que $0 < i \leq I$ et $0 < j < i$, où *I* est le nombre de classes (Fleury, Vacher et Noury, 2010). Le nombre nécessaire de SVM binaires pour classer un nombre I classes est $(I\ (I-1)) / 2$.

Le multi-classe-SVM comprend donc six SVM binaires. Se basant sur le résultat de ces six classifieurs, la stratégie de décision finale est basée sur la méthode du vote majoritaire (best candidate). Pour chacun des six SVMs, une validation croisée a été effectuée sur 4 sous-groupes du corpus, afin de déterminer la paire paramétrique (C, σ) optimale de la fonction RBF. Cette approche LDA-SVM a atteint une correcte évaluation de la dysarthrie dans 93% des cas.

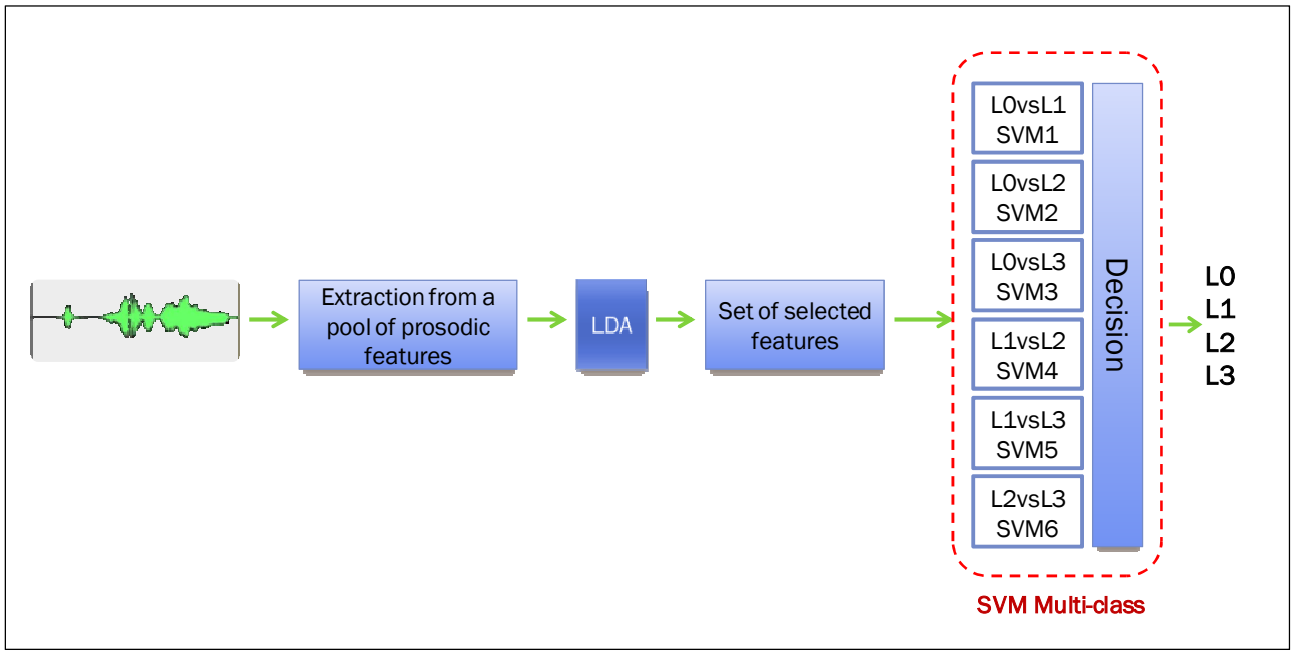

Figure 12| le système LDA-SVM

Un résumé et une comparaison des différentes méthodes présentées dans cette série 1 d'expériences sont mentionnés dans le tableau 10 suivant :

Système	LDA	LDA-GMM	LDA-SVM
Taux de classification correcte (%)	84.2	88.9	93

Tableau 10| Comparaison des performances des méthodes proposées

D. Expérience série 2

L'approche de la deuxième série d'expériences consiste à consolider deux des types de paramètres présentés précédemment, à savoir les MFCCs conventionnels et les indices acoustiques dérivés d'études sur le phénomène de l'audition, comme front-end pour l'évaluation de la dysarthrie en niveaux de sévérité. Les méthodes de reconnaissances expérimentées sont, GMM, SVM et le système hybride GMM-SVM. Le choix de consolider deux sources d'information du signal de parole a pour but de minimiser la partie qui peut manquer en utilisant une seule source.

L'évaluation des niveaux de sévérité de la dysarthrie repose sur l'évaluation Frenchay Dysarthria Assessment (FDA). Des fichiers contenant l'évaluation FDA détaillée des locuteurs dysarthriques sont disponibles pour les deux bases de données Nemours et Torgo. Le protocole FDA a été publié en 1983, suite à des recherches qui ont permis d'identifier le modèle et les caractéristiques de la production de la parole, et les mouvements oro-moteur impliqués à des maladies neurologiques évidentes. Le protocole de test est censé aider au diagnostic, guider le traitement et avoir une grande fiabilité. La deuxième édition de l'évaluation FDA nommée FDA-2 a été modifiée pour intégrer de récentes connaissances concernant les troubles moteurs de la parole et leur contribution au diagnostic neurologique. Pour la deuxième édition (2008), certains éléments inclus dans la première édition (1983) ont été omis car ils étaient considérés comme inadéquats ou redondants pour le diagnostic et la prise en charge (FDA-2, 2008). Pour déterminer le score global de l'évaluation de chaque locuteur dysarthrique, nous nous sommes appuyés sur les recherches dans (Menendez et al., 1996) pour la base de données Nemours. Par

ailleurs, pour la base de données Torgo qui est plus récente que Nemours, nous avons pris en compte les amendements établies pour le protocole FDA-2, tout en exploitant les résultats détaillés de l'évaluation FDA-1 qui sont inclus pour chacun des huit locuteurs dysarthriques.

Les locuteurs de Nemours sont regroupés selon les niveaux L1, L2 et L3 déjà utilisés dans la série d'expériences 1. De même, les patients de la base de données Torgo peuvent être regroupés en se basant sur leurs évaluations (voir tableau 6) en niveau faible de la maladie, niveau sévère et niveau très sévère. Ainsi la combinaison de Torgo et Nemours donne les groupes de locuteurs suivants :

- Groupe L1 : F04, F03, M03, FB, MH, BB et LL.
- Groupe L2 : M05, F01, JF, RL, RK, BK et BV.
- Groupe L3 : M02, M01, M04, SC et KS.

La tâche consiste à classer trois niveaux de sévérité de la dysarthrie en utilisant la base de données de parole dysarthriques 'Nemours' et la base de données acoustique et articulatoire de la parole 'Torgo' (Figure 13).

Le sous-ensemble d'apprentissage est composé de 70% de l'ensemble des enregistrements et il est utilisé pour entrainer les différents classifieurs à la distinction des niveaux de gravité L1, L2 et L3. Par la suite, l'évaluation de la capacité des systèmes à classer les phrases dans les niveaux de gravité correspondants est estimée en utilisant les 30% restants de l'ensemble des enregistrements. Ce tiers des données représente le sous-ensemble de test.

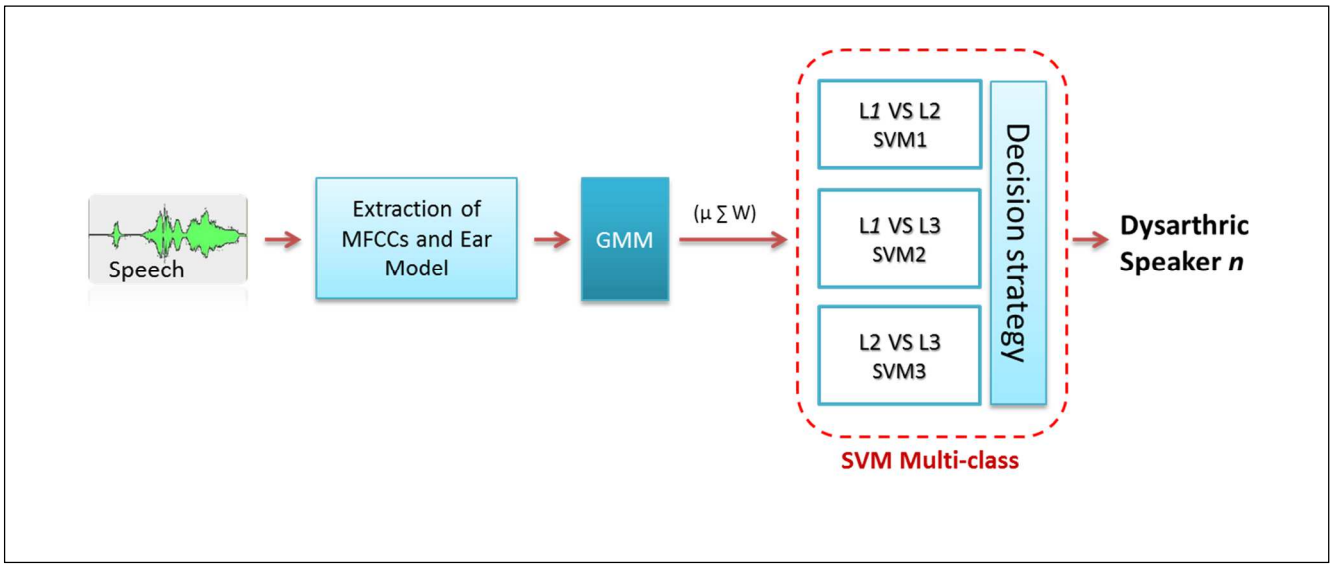

Figure 13| Diagram simplifié du processus d' »valuation de la parole patholohgique utilisant le GMM/SVM

1. Discrimination par GMM :

Pour représenter chacun des 3 niveaux de sévérités, un modèle de distribution de Gaussiennes est généré. Chaque fenêtre du signal *X* est représentée par un vecteur acoustique x contenant p composants. Sachant que ce vecteur acoustique peut contenir trois variantes de caractéristiques extraites du signal de parole, comme suit :

- Vecteur de 12 MFCCs conventionnels.
- Vecteur de 8 paramètres du modèle d'oreille proposé (7 indices auditifs + l'énergie)
- Consolidation des coefficients MFCCs et des paramètres du modèle d'oreille, ce qui résulte un vecteur acoustique à 20 composants.

Le tableau 11 contient les performances d'évaluation automatique selon de l'ordre du modèle et du front-end utilisé.

Ordre du Modèle	*MFCCs*	*Indices Auditifs*	*MFCCs+Indices Auditifs*
M=4	70,4	71,3	68,4
M=8	76,8	65,7	72,6
M=16	83,9	75,2	83,6
M=32	88,9	83,6	87,1
M=64	90,5	86,5	90,7
M=128	**92,3**	87,1	**93,2**
M=256	92.2	**88.6**	92.9

Tableau 11| Performances d'évaluation pour différents paramètres acoustiques et plusieurs ordres du modèle GMM

La méthode GMM utilisant les MFCCs et les indices auditifs atteint 93.2% de taux de bonne classification à travers 3 niveaux de dysarthrie, L1, L2 et L3.

2. Discrimination par SVM :

L'approche des SVMs permet une projection des données sur un nouvel espace de dimension supérieure, puis l'établissement d'un hyperplan séparateur qui maximise la marge entre 2 groupes de données. Dans cette partie, un classifieur multi-classes SVM est conçu en utilisant la méthode 'one-against-one' pour discriminer les 3 niveaux de sévérité de la dysarthrie (L1, L2 et L3). Il est composé donc de 3 SVM binaires comme suit:

- $SVM_1 : L1\ vs\ L2$
- $SVM_2 : L1\ vs\ L3$
- $SVM_3 : L2\ vs\ L3$

Il est à noter que le front-end exploité est similaire à celui déjà calculé pour la méthode précédente (GMM). Le tableau 12 présente les résultats de classification en utilisant un système SVM multi-classes, constitué de trois SVM binaire.

	MFCCs	***Indices Auditifs***	***MFCCs+Indices Auditifs***
multiclass-SVM	76.6	66.4	75.3

Tableau 12| les performances du One-against-one SVM par paramètre acoustique

La meilleure performance, 76.6%, est atteinte par le système SVM utilisant les MFCCs.

3. Discrimination par GMM-SVM

Afin de tirer bénéfice des points forts des deux méthodes GMM et SVM, c'est-à-dire, l'aptitude de description des GMMs et la haute capacité de classification des SVMs, nous avons combiné les deux systèmes décrits précédemment en utilisant la distribution de Gaussiennes comme base paramétrique du classifieur SVM. Ainsi, un système hybride GMM/SVM a été testé avec différents ordres de modèles GMM et un SVM multi-classes de type 'one-against-one'. Les résultats sont présentés dans le tableau 13.

Ordre du modèle	***MFCCs***	***Indices Auditifs***	***MFCCs+Indices Auditifs***
M=8	76.6	62.9	**78.8**
M=16	**75.5**	61.8	74.2
M=32	73.2	**63.1**	74.2
M=64	74.5	62.4	72.6

Tableau 13| Performances du système hybride GMM-SVM pour différents paramètres acoustiques et ordres du modèle GMM

Le système hybride atteint une performance de 78.8% avec un modèle d'ordre 8 et en combinant les MFCCs et les indices auditifs.

PARTIE 5 :

Vers l'accessibilité aux systèmes biométriques, des personnes ayant des besoins spéciaux

Partie 5 :

V. Vers l'accessibilité aux systèmes biométriques, des personnes ayant des besoins spéciaux

De nombreux outils et méthodes ont été développés pour aider les locuteurs dysarthriques. En effet, des travaux importants ont été réalisés dans le domaine de la reconnaissance de la parole, l'évaluation automatisée et le rehaussement de l'intelligibilité. Au cours des dernières années, Rudzicz a développé un système de reconnaissance de la parole dysarthrique et proposé des méthodes pour l'amélioration de l'intelligibilité du discours dysarthrique, en exploitant la récente et importante base de données Torgo (Rudzicz et al., 2011) (Rudzicz et al., 2013). Il est à noter que les efforts de recherche n'ont pas exploré d'autres fonctionnalités comme la reconnaissance du locuteur qui est de plus en plus utilisée dans différents systèmes de sécurité ou de gestion de l'identité basés sur la biométrie.

A. Biométrie

La biométrie est une technique qui permet d'associer à une identité une personne voulant procéder à une action, grâce à la reconnaissance automatique d'une ou de plusieurs caractéristiques physiques et/ou comportementales de cette personne préalablement enregistrées (Encyclopédie Larousse, online).

En effet, deux catégories de la biométrie peuvent être citées, comportementale et physiologique. Néanmoins, la frontière entre les deux catégories n'est pas si claire, le traitement spécifique de certaines méthodes biométriques peut les classer dans l'une des catégories ou comme une combinaison des deux. Par exemple, prenons l'objet de ce chapitre qui est la reconnaissance du locuteur, l'approche peut être considérée comme une biométrie comportementale si le système de reconnaissance automatique se focalise sur le traitement des transitions audio et sur les traits représentant la manière de parler. En revanche, l'approche peut être considérée comme biométrie physiologique si les caractéristiques physiologiques de l'appareil phonatoire (MFCC, indices du Modèle d'oreille…) sont utilisées comme base paramétrique du système. Généralement, les méthodes de reconnaissance du locuteur qui sont indépendantes du texte, exploitent la biométrie physiologique. Tandis que, les approches dépendantes du texte ont tendance à utiliser certaines données comportementales en plus des caractéristiques physiologiques (Beigi, 2011).

L'information sur l'identité d'une personne est intégrée dans sa voix, celle-ci peut être détectée par les systèmes de reconnaissance automatique du locuteur. Dans cette partie, nous décrivons un nouveau système conçu pour inclure les individus qui souffrent de trouble de la parole dans les systèmes biométriques basés sur la voix. Cette approche peut être utilisée dans plusieurs applications comme par exemple le contrôle à distance, la correspondance de la voix pour la criminalistique ou pour l'identification des patients dans les outils médicaux. La reconnaissance par la voix peut aussi avoir un grand apport dans les cas de combinaison de modalités biométriques, comme par exemple la combinaison de la voix avec l'empreinte, la reconnaissance faciale ou l'identification de l'iris.

La différence inter locuteurs provient essentiellement de la variation des caractéristiques physiologiques du système phonatoire et des réflexes de prononciation acquis ou appris. En ce sens, le front-end du système d'identification du locuteur proposé, qui est indépendant du texte, est basé sur les caractéristiques du signal bas niveau précédemment présentées. Le but étant de déterminer l'identité d'un locuteur dysarthrique inconnu en comparant sa voix avec N modèles (1:N) ; l'empreinte de la voix est ainsi représentée par les Indices Auditifs et les MFCCs tandis que les méthodes GMM, SVM et l'hybride GMM-SVM sont appliquées pour la modélisation et la classification automatique (Kadi et al., 2015).

B. Reconnaissance automatique des locuteurs atteints de trouble de la parole

Le but est de développer des systèmes d'identification indépendants du texte adaptés aux locuteurs souffrant de dysarthrie, en expérimentant les méthodes de reconnaissance automatique, GMM, SVM et GMM-SVM sur l'ensemble des locuteurs ayant participé aux bases de données Torgo et Nemours réunis.

Ainsi, des caractéristiques de la parole de niveau bas sont utilisées comme front-end du système d'identification de locuteurs dysarthriques proposé (Figure 14 et Figure 15). La tâche vise à déterminer l'identité d'un locuteur inconnu en comparant le modèle de sa voix contre N modèles (1: n). L'empreinte vocale est représentée par les paramètres MFCCs et les Indices Auditifs tandis que les GMM, SVM et l'hybride GMM/SVM comme méthodes sont appliquées à la modélisation et la classification.

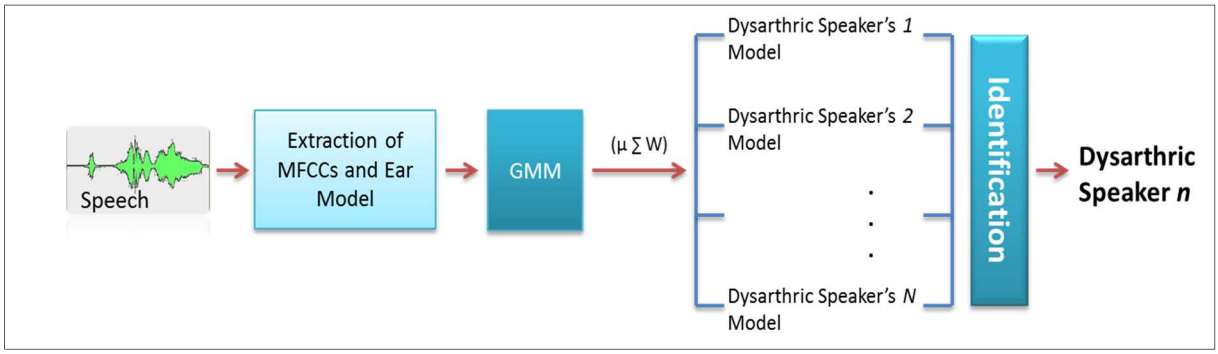

Figure 14| Diagramme simplifié du processus d'identification du locuteur dysarthrique utilisant un système GMM

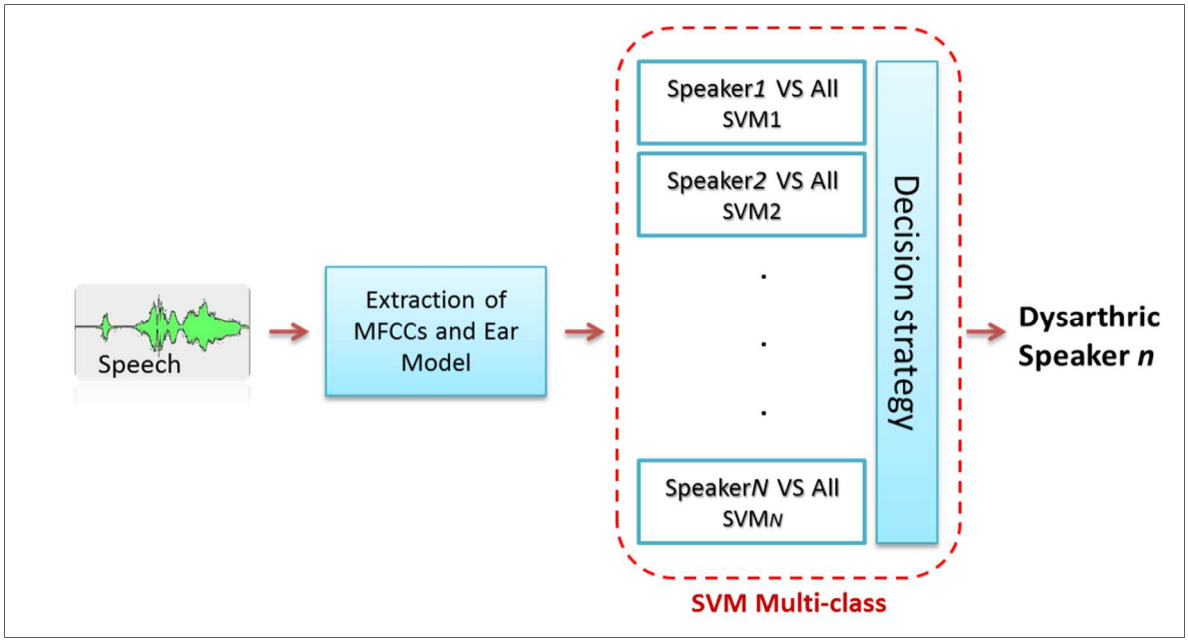

Figure 15| Diagramme simplifié du processus d'identification du locuteur dysarthrique utilisant un SVM Multi class

C. Expérience série 3

L'ensemble des données est divisé en deux sous-ensembles : un sous-ensemble d'apprentissage *(training)* qui contient 70 % des phrases enregistrées par chaque patient et un sous-ensemble de test qui inclut les 30 % restants des phrases. Le sous-ensemble de formation d'apprentissage sert à construire et à entrainer les classifieurs dans le but de reconnaitre chacun des 19 locuteurs dysarthriques des deux bases de données réunies. Le sous-ensemble de test est ensuite utilisé pour évaluer la capacité des systèmes pour l'identification des patients. Par ailleurs, les paramètres front-end des systèmes sont semblables aux paramètres décrits au niveau de la série 2 d'expériences, soit trois variantes :

- Vecteur de 12 MFCCs conventionnels.
- Vecteur de 8 paramètres du modèle d'oreille proposé (7 indices auditifs + l'énergie)
- Consolidation des coefficients MFCCs et des paramètres du modèle d'oreille, ce qui résulte un vecteur acoustique à 20 composants.

La métrique d'évaluation des performances des différents systèmes est le taux de reconnaissance. Il est calculé par le rapport du nombre d'identifications correctes sur le nombre total des tentatives d'identification (Abd El-Samie, 2011).

1. Reconnaissance par GMM

L'efficacité des modèles de mélange de Gaussiennes dans la modélisation de l'identité d'un locuteur provient de leur aptitude à modéliser les densités arbitraires par la détection des formes de spectres propres au locuteur (Reynolds and Rose, 1995).

Apprentissage : pour chaque classe C_i qui représente un des locuteurs dysarthriques, l'apprentissage est initié pour obtenir un modèle dont les paramètres de chaque distribution de gaussiennes *m* de la classe sont : le vecteur des moyennes, la matrice de covariance et les poids gaussiens. Ces caractéristiques sont calculées après avoir effectué un nombre suffisant d'itérations pour assurer la convergence de l'algorithme EM. Un modèle GMM est donc généré pour représenter chaque identité des locuteurs dysarthriques.

Reconnaissance (test) : Chaque phrase du sous-ensemble de test est traitée et représentée par un vecteur d'entrée acoustique contenant l'une des trois variantes de caractéristiques de la parole présentées précédemment. La densité de probabilité (PDF) de chaque vecteur acoustique est alors calculée pour chaque classe donnée C_i , ce qui déterminera l'identité, la plus probable, du locuteur dysarthrique.

Le tableau 14 présente les taux de bonne identification des locuteurs dysarthriques pour des systèmes testés et les paramètres optimaux de la méthode.

Ordre du Modèle	*MFCCs*	*Indices Auditifs*	*MFCCs + Indices Auditifs*
M=4	93.4	81.9	91.3
M=8	93.1	89.3	93.4
M=16	94.8	92.7	95.1
M=32	95.3	93.7	95.3
M=64	**95.8**	93.9	**97.2**
M=128	95.6	**94**	97.1

Tableau 14| Performance d'identification pour différents paramètres acoustiques et plusieurs ordres du modèle GMM

Les résultats montrent que la combinaison des MFCCs et des paramètres Auditifs dans le *front-end* donne de meilleurs résultats comparativement au système de référence utilisant les MFCCs, et cela avec la plupart des ordres du modèle GMM. En effet, la méthode hybride MFCCs– Indices Auditifs atteint un taux de bonne identification de 97.2% avec un ordre de modèle de 64 gaussiennes.

2. Reconnaissance par SVM

Un système SVM multi classe a été mis en place pour distinguer chacun des locuteurs dysarthriques des deux bases de données, Torgo et Nemours, simultanément. Utilisant la méthode '*one-against-all* ',

Cette méthode consiste à construire un SVM binaire par identité, qui est formée pour discriminer les observations d'un patient des observations de tous les autres patients. Avec la méthode '*one-against-all* ' le nombre nécessaire de classifieurs binaires est égal au nombre de classes, il est donc de 19 SVM binaires.

Nous utilisons une validation croisée pour obtenir les paramètres optimaux, *C* et *Sigma*, du noyau RBF. Cette opération est effectuée pour les enregistrements de chacun des locuteurs dysarthriques et pour chacune des variantes du *front-end* utilisé :

- Vecteur de 12 MFCCs.
- Vecteur de 8 paramètres du modèle d'oreille (7 indices auditifs + l'énergie)
- Consolidation des coefficients MFCCs et des paramètres du modèle d'oreille (vecteur acoustique de 20 composants).

Les résultats d'identification utilisant différents *front-end* sont présentés dans le tableau 15.

	MFCCs	*Indices Auditifs*	*MFCCs + Indices Auditifs*
SVM multiclasse	**84.7**	70.1	84.2

Tableau 15| Performances d'identification du SVM one-against-all SVM pour différents paramètres acoustiques

Pour le système d'identification automatique du locuteur dysarthrique à base de SVM, le meilleur taux d'identification est atteint en utilisant les paramètres MFCCs.

3. Les effets de la normalisation temporelle sur les performances des classificateurs GMM et SVM

L'approche GMM réalise un taux élevé de bonne classification comparée à la méthode SVM. Cette différence pourrait s'expliquer par la capacité des GMM et des méthodes statistiques en général de s'adapter aux longueurs temporelles variables des données d'entrées. La méthode SVM ne peut pas traiter efficacement des signaux de parole qui ont des durées différentes, comme c'est le cas pour les données des deux bases de données Nemours et Torgo. Par conséquent, l'utilisation de SVM nécessite de définir la même durée pour tous les enregistrements traités. Ce processus de normalisation est inhérent à la préparation des données

d'entrée SVM pour assurer l'homogénéité du traitement *front-end*, ce qui peut induire une certaine perte d'information.

4. Reconnaissance par GMM-SVM

Pour bénéficier de la capacité de description des données de la méthode GMM et de la performance élevée de classification de la méthode SVM, nous combinons les deux systèmes décrits ci-dessus afin d'utiliser les distributions de gaussiennes comme base paramétrique des classifieurs SVMs.

Nous avons évalué l'approche hybride GMM-SVM dans cette tâche de reconnaissance des patients. Pour atteindre une meilleure efficacité, les moyennes des distributions de Gaussiennes sont exploitées dans le front-end du classifieur SVM (Liu et al., 2006). Le tableau 16contient les taux d'évaluation des performances du système GMM-SVM ainsi que les paramètres optimaux.

Ordre du Modèle	*MFCCs*	*Indices Auditifs*	*MFCCs + Indices Auditifs*
M=8	77.1	53.7	90
M=16	77.6	55.8	86.3
M=32	**82.6**	**57.1**	**91.1**
M=64	81	55.7	90.5

Tableau 16| Performances d'identification du système hybride GMM/SVM pour différents paramètres acoustiques et plusieurs ordres de modèle de Gaussiennes

La meilleure performance de reconnaissance atteinte avec la méthode hybride GMM-SVM est de 91.1% en utilisant la combinaison MFCCs-Indices Auditifs comme front-end.

Conclusion

Le premier accomplissement de nos études est la performance des méthodes de discrimination des niveaux de sévérité qui surpasse l'état de l'art dans le domaine de l'évaluation et de diagnostic automatique de la parole dysarthrique. En effet, l'utilisation des paramètres prosodiques pour caractériser la dysarthrie donne des résultats satisfaisants, particulièrement avec la méthode LDA-SVM. Par ailleurs, la combinaison de deux bases de données majeures a permis d'effectuer des expérimentations d'une plus grande ampleur. Ainsi, l'association des MFCCs conventionnels et des indices Auditifs proposés comme entrée du système GMM atteint une bonne classification de 93.2%.

Le deuxième accomplissement est le nouveau score-FDA globale proposé pour chaque participant dysarthrique de la base de données Torgo, basé sur les scores des dimensions perceptuelles, et en tenant compte du progrès introduit dans le nouveau protocole FDA-2. Cela permettra de développer et d'expérimenter d'autres méthodes d'évaluation automatisées sur Torgo.

Le troisième accomplissement est dans la discipline de la reconnaissance du locuteur où le système GMM basé sur un mélange de 64 Gaussiennes, utilisant la combinaison des MFCCs et des indices Auditifs distinctifs, a atteint le meilleur taux d'identification du locuteur dysarthrique de 97.2%. Ce résultat peut être considéré comme prometteur vu l'état de l'art actuel.

Les méthodes proposées sont escomptées être utiles pour l'accessibilité des locuteurs dysarthriques aux systèmes biométriques et pour être un outil d'aide aux cliniciens dans le cadre de l'évaluation et/ou diagnostic des patients.

Références:

Abd El-Samie F.E., Information Security for Automatic Speaker Identification, Springer Briefs in Speech Technology, Springer, 2011.

Ackermann Hermann, Hertrich Ingo and Ziegler Wolfram, Chapter 16 in: The Handbook of Language and Speech Disorders, A John Wiley & Sons, Ltd., Publication, 2010.

American Speech-Language-Hearing Association [Online], Available: http://www.asha.org/

Baghai-Ravary L., and Beet S.W., Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders, Springer Briefs in Electrical and Computer Engineering, 2013.

Beigi Homayoon, Fundamentals of Speaker Recognition, Springer, 2011.

Boersma P. and Weenink D., Praat, a system for doing phonetics by computer, Glot International, vol. 5, no. 9-10, pp. 341-345, 2001.

Caelen J., Space/time data-information in the ARIAL project ear model, Speech Communication, 1985, 4:457–467.

Calliope, La parole et son traitement automatique, Dunod, 1989.

Campbell W.M., Karam Z.N., A framework for discriminative SVM/GMM systems for language recognition, Interspeech, 2009, pp. 2195–2198.

Davis S., Mermelstein P., Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech and Signal Processing, 28(4), 1980, 357-366.

Dempster A.P., Laird N.M. and Rubin D.B., Maximum-likelihood from incomplete data via the EM algorithm, Journal of the Acoustical Society of America, vol. 39, no. 1, pp. 1-38, 1977.

Deng L. and D. O'Shaughnessy, Speech Processing: a dynamic and optimization-oriented approach, Marce Dekker, Inc., 2003.

Duffy J.R., Motor Speech Disorders: Clues to Neurologic Diagnosis, in Parkinson's disease and Movement Disorders, Springer, pp. 35–53, 2000.

Enderby P., Disorders of communication: dysarthria, in: Handbook of Clinical Neurology, Elsevier, 2013.

Enderby P.M., Frenchay Dysarthria Assessment, PRO-ED, 1983.

Enderby P.M. and Palmer R., Frenchay Dysarthria Assessment, Second Edition (FDA-2), PRO-ED, 2008.

Flanagan J.L., Models for approximating basilar membrane displacement, Bell System Technology Journal, 39:1163–1191, 1960.

Fleury A., Vacher M. and Noury N., SVM-Based multimodal classification of Activities of daily living in health smart homes: sensors, algorithms, and first experimental results, IEEE Trans, Information Technology in Biomedicine, vol. 14, no. 2, pp. 274-283, 2010.

Ghitza O., Auditory models and human performance in tasks related to speech coding and speech recognition, IEEE Trans, Speech Audio Proc. SAP, 1994, 2:115 – 132.

Giannakopoulos T., A. Pikrakis, Introduction to Audio Analysis, Academic Press, Elsevier, 2014.

Guerra C. E. and Lovey D. F., A modern approach to dysarthria classification, Engineering in Medicine and Biology Society, Proceedings of the 25th Annual International Conference of the IEEE, EMBS, 2003.

Hammen, Vicki , Yorkston, and Minifie. Effects of Temporal Alterations on Speech Intelligibility in Parkinsonian Dysarthria. Journal of Speech and Hearing Research, 37:244–253, 1994.

Hart J. T., Collier R. and Cohen A., A perceptual study of intonation, Cambridge University Press, 1990.

Hasegawa-Johnson, Mark, Jon Gunderson, Adrienne Perlman and Thomas Huang, HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria, In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006), volume 3, pages 1060–1063, 2006.

Jayaram, Gowtham and Kadry Abdelhamied. Experiments in dysarthric speech recognition using artificial neural networks, Journal of Rehabilitation Research and Development, 32(2):162–169, 1995.

Kadi K.L., Selouani S.-A., Boudraa B. and Boudraa M., Automated Diagnosis and Assessment of Dysarthric Speech Using Relevant Prosodic Features, in: Transactions on Engineering Technologies, Springer, 2014.

Kadi K.L., Selouani S.-A., Boudraa B. and Boudraa M., Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge, Biocybernetics and Biomedical Engineering Journal, 36:233-247, Elsevier, 2016.

Kadi K.L., Selouani S.-A., Boudraa B. and Boudraa M., Discriminative Prosodic Features to Assess the Dysarthria Severity Levels. Proceedings of the World Congress on Engineering 2013, 3–5 July London. Lecture Notes in Engineering and Computer Science, pp. 2201–2205, 2013.

Kent R.D and Rosen K., Motor control perspectives on motor speech disorders, chapter 12 in: Speech Motor Control in Normal and Disordered Speech, Oxford University Press, pp. 285-311, 2004.

Kent R.D, Research on speech motor control and its disorders: a review and prospective. Journal of Communication Disorders, 33(5):391–428, 2000.

Kim C., R. M. Stern, Power-normalized cepstral coefficients (pncc) for robust speech recognition, in: Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference, pp. 4101–4104.

Kim J., Kumar N., Tsiartas A., Li M. and Narayanan S., Automatic intelligibility classification of sentence-level pathological speech, Computer Speech and Language, Elsevier, vol.29, issue.1, pp.132-144, 2014.

Kim, Heejin, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas Huang, Kenneth Watkin and Simone Frame, Dysarthric speech database for universal access research, Interspeech, pages 1741–1744, Brisbane, Australia, 2008.

LDC, 2012, Catalog number LDC2012S02. Linguistic Data Consortium [Online]. Available: http://catalog.ldc.upenn.edu/docs/LDC2012S02/README.txt

Liss J.M., L. White, S.L. Mattys, K. Lansford, A.J. Lotto, S.M. Spitzer and J.N. Caviness, Quantifying speech rhythm abnormalities in the dysarthrias. J Speech, Lang. Hear. Res. 52,1334–1352, 2009.

liu M., Y. Xie, Z. Yao, B. Dai, A new hybrid GMM/SVM for speaker verification, the 18th International Conference on Pattern recognition, IEEE, 2006.

Lofqvist, Anders, Nancy S. McGarr and Kiyoshi Honda, Laryngeal muscles and articulatory control. The Journal of the Acoustical Society of America, 76(3):951–954, 1994.

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

Mary L., Extraction and representation of prosody for speaker, speech and language recognition, Springer Briefs in Speech Technology, ch.1, 2012.

Melf R.S., Communication Disorders [Online], Available: http://emedicine.medscape.com/article

Menendez-Pidal X., Polikoff J.B., Peters S.M., Leonzio J.E. and Bunnell H.T., The Nemours database of dysarthric speech, Fourth International Conference on Spoken Language, ICSLP, vol. 3, pp. 1962-1965, 1996.

Moore, Keith L. and Arthur F. Dalley. Clinically Oriented Anatomy, Fifth Edition. Lippincott, Williams and Wilkins, 2005.

O'Shaughnessy D., Speech communication: Human and machine, IEEE Press, 2001.

Paja M.S. and Falk T.H., Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech, Interspeech, 2012.

Polikoff J.B. and Bunnel H.T., The Nemours database of dysarthric speech: a perceptual analysis, 14th International Congress of Phonetic Sciences, ICPhS, pp. 783-786, 1999.

Reynolds D.A. and Rose R.C., Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Transactions on Speech and Audio Processing, 3(1): 72-83, 1995.

Roth C., Dysarthria, in: Encyclopedia of Clinical Neuropsychology, Springer, 2011

Rudzicz F., Namasivayam A.K. and Wolff T., The TORGO database of acoustic and articulatory speech from speakers with dysarthria, Language Resources and Evaluation, vol 46, issue 4, Springer, pp.523-541, 2012.

Rudzicz F., Adjusting dysarthric speech signals to be more intelligible, Journal of computer speech and language, Elsevier, 27:1163-1177, 2013.

Rudzicz F., Articulatory Knowledge in the Recognition of Dysarthric Speech, IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, No. 4, 2011.

Rudzicz F., Phonological features in discriminative classification of dysarthric speech, ICASSP, 2009.

Rudzicz F., Using articulatory likelihoods in the recognition of dysarthric speech, Journal of speech communication, Elsevier, 54:430-444, 2012.

Schluter R., L. Bezrukov, H. Wagner, H. Ney, Gammatone features and feature combination for large vocabulary speech recognition, Acoustics, Speech and Signal Processing, ICASSP, IEEE International Conference, vol 4, pp. 649–652, 2007.

Schneiderman, Carl R. and Robert E. Potter, Speech-language pathology: a simplified guide to structures, functions, and clinical implications. Academic Press, San Diego, CA, 2002.

Selouani S. A., Speech Processing and Soft Computing. Springer, 2011.

Selouani S.A., D. O'Shaughnessy and J. Caelen, Incorporating phonetic knowledge into an evolutionary subspace approach for robust speech recognition, International Journal of Computers and Applications, 29:143–154, 2007.

Selouani S.-A., H. Dahmani, R. Amami, and H. Hamam, Using speech rhythm knowledge to improve dysarthric speech recongnition, International Journal of Speech Technology, Vol. 15, no.1, pp.57-64, 2012.

Selouani S.A., H. Tolba, and D. O'Shaughnessy, Auditory-based acoustic distinctive features and spectral cues for robust automatic speech recognition in Low-SNR car environments, Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: HLT-NAACL - Vol. 2, pp. 91-93, Stroudsburg, PA, USA, 2003.

Selouani S.A., J. Caelen, Recognition of Arabic phonetic features using neural networks and knowledge-based system: A comparative study, International journal on artificial intelligence tools, 8(1), pp. 73-103, 1999.

Selouani S.-A., Y. Alotaibib, W. Cichockic, S. Gharsellaouia, et K. Kadi, Native and non-native class discrimination using speech rhythm- andauditory-based cues, Computer Speech and Language 31 :28–48, Elsevier, 2015.

Seneff S., A joint synchrony/mean-rate model of auditory speech processing, Journal of Phonetics, 1988, 16:55 – 76.

Shahamiri S.R., S.S.B. Salim, Artificial neural networks as speech recognizers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach, Journal of Advanced Engineering Informatics, Elsevier, vol28, issue1, pp.102–110, 2014.

Shriberg E., A. Stolcke, and D. Hakkani, Prosody-based automatic segmentation of speech into sentences and topics, Speech Communication. Special Issue on Accessing Information in Spoken Audio, vol.32, no.1-2, pp.127-154, 2000.

Stern R., N. Morgan, Hearing is believing: Biologically inspired methods for robust automatic speech recognition, IEEE Signal Process Mag, 29(6):34–43, 2012.

Stevens, S.S., J. Volkman and E.B. Newman.. A scale for the measurement of the psychological magnitude pitch. Journal of the Acoustical Society of America, 8(3):185–190, 1937.

Sundberg and Johan, The acoustics of the singing voice. Scientific American, 234:82–91, 1977.

Thomas-Stonell, Nancy, Ava-Lee Kotler, Herbert A. Leeper, and Philip C. Doyle. Computerized speech recognition: influence of intelligibility and perceptual consistency on recognition accuracy. Augmentative & Alternative Communication, 14(1):51–56, March, 1998

Vapnik V.N., An overview of statistical learning theory, IEEE Trans. Neural Networks, vol. 10, no. 5, pp. 988-999, 1999.

Westzner H. F., S. Schreiber, and L. Amaro, Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders, Brazilian Journal of Orthinolaryngology, vol. 71, no. 5, pp. 582-588, 2005.

Wrench, Alan. The MOCHA-TIMIT articulatory database, 1999.

Yunusova, Yana, GaryWeismer, John R.Westbury, and Mary J. Lindstrom, Articulatory movements during vowels in speakers with dysarthria and healthy controls. Journal of Speech, Language, and Hearing Research, 51:596–611, 2008.

Zue, Victor, Stephanie Seneff and James Glass. Speech Database Development: TIMIT and Beyond. In Proceedings of ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases, 1989.