Approche Statistique pour l'Analyse Objective et la Caractérisation de la Voix Dysphonique

Gilles Pouchoulin

ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THÈSE

présentée à l'Université d'Avignon et des Pays de Vaucluse
pour obtenir le grade de Docteur

SPÉCIALITÉ : Informatique

École Doctorale 166 I2S « Mathématiques et Informatique »
Laboratoire d'Informatique (EA 4128)

Approche Statistique pour l'Analyse Objective et la Caractérisation de la Voix Dysphonique

par

Gilles Pouchoulin

Soutenue publiquement le 19 décembre 2008 devant un jury composé de :

M.	Henry Méloni	PR, LIA, Avignon, France	Président du jury
M^{me}	Martine Adda-Decker	CR-HDR, LIMSI-CNRS, Orsay, France	Rapporteur
M.	Francis Grenez	PR, ULB, Bruxelle, Belgique	Rapporteur
M^{me}	Lise Crevier-Buchman	CR, CNRS/Sorbonne-Nouvelle, Paris, France	Examinateur
M.	Antoine Giovanni	PR, CHU Timone, Marseille, France	Examinateur
M.	Alain Ghio	IR, LPL, Aix-en-Provence, France	Examinateur
M.	Jean-François BONASTRE	PR, LIA, Avignon, France	Directeur de thèse
M^{lle}	Corinne FREDOUILLE	MCF, LIA, Avignon, France	Co-Encadrante de thèse

---

Laboratoire d'Informatique d'Avignon

Remerciements au LAPEC et au LPL

En premier lieu, je souhaite exprimer toute ma gratitude au LAPEC et au LPL pour leur précieuse collaboration et leurs conseils avisés durant ces trois années de thèse.

Je commencerai donc par remercier vivement A. Giovanni, D. Robert et J. Révis pour :

- avoir mis à disposition le corpus de voix dysphoniques sans lequel cette thèse n'aurait pu être possible
- avoir organisé des séances d'évaluation perceptive de voix dysphoniques
- leur disponibilité, enthousiasme et gentillesse dont ils ont témoigné
- leur grande connaissance et compétence dans leur domaine respectif qu'ils n'hésitent pas à partager
- avoir permis une collaboration entre le LIA et des stagiaires en orthophonie (Marion, Ondine, Audrey), montrant ainsi leur confiance et reconnaissance envers notre laboratoire.

De plus, j'exprime ici toute mon amitié à M. Ouaknine que j'ai eu plaisir à retrouver par le hasard de la vie et avec qui j'ai usé les bancs des amphis du CMI à Marseille, il y a quelques années déjà.

Concernant le LPL, je commencerai par remercier son directeur P. Blache pour son accueil si cordial.

Un grand merci à B. Teston pour son enthousiasme à partager ses connaissances et expériences avec une grande facilité de narration et en toute simplicité. De même, pour A. Ghio pour avoir toujours répondu à mes questions avec précision, pour sa grande disponibilité et spontanéité. Et bien sûr, un grand merci à Thierry L., Serge P., Muriel L., Alain M., Daniel H., Danielle D., ...

Merci aussi à tous ceux que je ne peux citer ici pour leur sympathie et leur bonne humeur. Je m'excuse auprès d'eux de ne pas tous les citer.

Remerciements Personnels

Ces trois années de travail de thèse passées au LIA n'aurait pu aboutir sans la collaboration et le soutien de nombreuses personnes. Ces quelques lignes essayent d'exprimer ma gratitude et reconnaissance envers ceux qui m'ont permis de le mener à bien.

Tout d'abord, mes premiers remerciements sont destinés à l'ensemble des membres de mon jury : Henry Méloni pour m'avoir fait l'honneur d'en être le président, Martine Adda-Decker et Francis Grenez pour avoir accepté d'en être les rapporteurs en y consacrant une partie de leur temps précieux, ainsi que Lise Crevier-Buchman, Antoine Giovanni et Alain Ghio pour leur participation en tant qu'examinateurs.

Je tiens à exprimer ma plus profonde reconnaissance à mes deux co-directeurs de thèse : Jean-François bonastre qui, malgré son emploi du temps très chargé, a toujours su prendre le temps de répondre à mes questions et surtout de m'avoir donné les moyens de finir cette thèse. Je lui en serais toujours reconnaissant. Et bien sûr à Corinne Fredouille sans qui rien n'aurait pu être possible. Un GRAND MERCI Corinne pour ton encadrement, ta disponibilité, tes compétences et le soutien que tu m'as accordé durant ces années. Tout comme pour Jef, je t'en serais toujours reconnaissant.

Plus localement, je remercie mes deux acolytes de bureau, Christophe L. et Anthony L., pour m'avoir supporter et pour la bonne ambiance qu'ils ont su entretenir.

Plus généralement, je remercie tous les membres du LIA et de l'IUP (thésards, enseignants, chercheurs, personnel administratif, ...) de ces deux institutions, les jeunes docteurs qui sont retournés dans leur pays lointain, ..., ainsi que tous ceux que j'ai oubliés et qui sauront me le pardonner.

Pour finir, je souhaiterais exprimer ma plus grande affection à toute ma famille pour leurs encouragements et soutien qu'ils m'ont témoigné.

Le dernier mot sera pour mon *père* ...

Résumé

Dans notre société où la communication verbale est essentielle, l'évaluation de la qualité de la voix pathologique et des causes de sa dégradation occupent une place de plus en plus importante pour le corps médical. Une personne atteinte d'un trouble vocal momentané ou durable comme une dysphonie peut subir de graves conséquences dans ses relations sociales, tant sur le plan professionnel que personnel : manque d'efficacité dans la communication, arrêt ou perte du travail, exclusion sociale, voire perte identitaire. Cependant, face aux limites du jugement à l'oreille (analyse perceptive) du dysfonctionnement vocal, les thérapeutes de la voix ressentent le besoin de plus en plus pressant d'une méthode d'évaluation objective de la qualité de la voix pathologique, complémentaire à l'analyse perceptive.

Cette thèse s'inscrit dans ce cadre et plus précisément dans l'adaptation des techniques de Reconnaissance Automatique du Locuteur (RAL) à la classification automatique des voix dysphoniques suivant le grade global de l'échelle perceptive GRBAS. Toutes les études sont conduites sur un corpus de 80 voix dysphoniques (dont 20 voix de contrôle) fourni par le département ENT du Centre Hospitalier et Universitaire de La Timone (Marseille). L'objectif des travaux présentés ici est d'acquérir une meilleure compréhension des phénomènes acoustiques liés à la dysphonie. L'originalité d'une telle approche réside dans l'utilisation d'un système de classification automatique comme outil de caractérisation des phénomènes pathologiques dans le signal de parole afin d'apporter aux experts humains de nouvelles connaissances sur les altérations de la voix. En ce sens, les spécialistes de la voix comme les phonéticiens pourront valider et/ou enrichir ces nouvelles connaissances en les approfondissant. Le cas échéant, les experts pourront en retour suggérer des indications/directives permettant au système automatique d'explorer de nouvelles pistes d'investigation. Cette démarche se distingue des méthodologies proposées dans la littérature qui visent davantage à améliorer les performances du système pour la tâche visée.

Les travaux réalisés dans cette thèse se subdivisent en deux volets : un premier volet qui décrit le système automatique adapté au contexte pathologique et un deuxième volet qui s'intéresse à la recherche de l'information pertinente. Dans cette optique, trois axes de recherche sont proposés.

Le premier axe est consacré à l'étude de différentes représentations paramétriques du signal de parole utilisées classiquement en RAL et appliquées ici dans un contexte pathologique. Les analyses spectrale, cepstrale et prédictive sont comparées ainsi que la complémentarité des coefficients statiques et dynamiques (Δ, $\Delta\Delta$, $\Delta\Delta\Delta$) avérée utile en RAL. Cette étude a montré l'intérêt de l'analyse spectrale dans notre contexte expérimental, ainsi que celui des informations dynamiques.

Le deuxième axe de recherche étudie la manière dont les caractéristiques acoustiques de la dysphonie sont dispersées sur l'ensemble de l'espace fréquentiel. Cette étude a montré la pertinence de la bande de fréquences [0-3000]Hz. Partant de ce résultat, une évaluation perceptive des signaux de parole du corpus des voix dysphoniques filtrés en [0-3000]Hz a été proposée ainsi qu'un parallèle avec la bande téléphonique.

Dans le dernier axe de cette thèse, les manifestations de la dysphonie sont étudiées en observant le comportement du système de classification par phonème ou classe de phonèmes. La principale observation concerne la pertinence de la classe des consonnes sur les deux bandes fréquentielles [0-8000]Hz et [0-3000]Hz. Le comportement « peu attendu » des consonnes et plus particulièrement des consonnes sourdes (vis à vis du type de pathologie étudié) permet ici au système automatique de remplir pleinement son rôle d'outil caractérisant les phénomènes pathologiques. En effet, l'analyse du comportement du système a permis de mettre en évidence des phénomènes (comme par exemple le VOT) qui nécessitent à présent une expertise phonétique et clinique approfondie.

Table des matières

Introduction Générale

Dans notre société, la communication est omniprésente et occupe une place prépondérante dans nos relations sociales et professionnelles. Dans ces conditions, un trouble vocal peut avoir un impact sur la vie quotidienne pouvant engendrer des comportements d'exclusion, de repli sur soi voire de marginalisation. En effet, face à des situations d'échanges sociaux, une personne atteinte d'une perturbation vocale peut soit les éviter ou les refuser, soit les affronter et se sentir « diminuée » par manque d'efficacité dans la communication. Dans le cas d'une dysphonie, cas particulier de troubles de la voix, la plainte du patient peut être d'ordre fonctionnel (forçage vocal, douleurs laryngées, ...), esthétique (voix éraillée, enrouée, irrégulière, ...) mais aussi relationnel (difficultés à communiquer, ...). Souvent, le mal-être sous-jacent est l'élément déclencheur qui incite les sujets dysphoniques à consulter un spécialiste de la voix.

Le jugement à l'oreille, connu également sous la terminologie d'analyse ou de jugement perceptif, est l'une des méthodes d'analyse et d'évaluation de la voix pathologique la plus utilisée en milieu clinique. Elle permet d'évaluer si une voix est normale ou pathologique et de mesurer le dysfonctionnement vocal. Néanmoins, le jugement perceptif est reconnu comme intrinsèquement subjectif et variable. Il est donc légitime de s'interroger sur sa pertinence pour évaluer la qualité de la voix pathologique. Dans ce sens, de nombreuses études [Dejonckere et al., 2001; Revis et al., 2006] ont été consacrées à la définition de protocoles standardisés d'analyse perceptive de la voix afin d'en améliorer la fiabilité, portant notamment sur l'élaboration d'échelles d'évaluation, sur la constitution et la formation de jury d'écoute, sur le choix du matériau phonétique, ... A ce jour, malgré le caractère individuel et subjectif de l'évaluation perceptive et son manque de fiabilité persistant, elle reste la référence et conserve sa légitimité dans le cadre des travaux scientifiques décrivant la voix et sa pathologie.

Face à ces limites, les spécialistes de la voix expriment, depuis plusieurs années, le besoin de plus en plus pressant d'une méthode d'évaluation objective de la qualité de la voix pathologique, complémentaire à l'analyse perceptive. Idéalement, ces méthodes objectives devraient s'intégrer à l'examen clinique comme une aide en terme de diagnostic, de thérapie rééducative, de suivi thérapeutique, de prévention (milieu scolaire et médecine du travail), de connaissance (pour la formation, l'enseignement et l'apprentissage des professionnels de la voix tels les orthophonistes, les phoniatres, les médecins ORL, ...), ... De nombreux travaux [Teston & Galindo, 1995; Wuyts et al., 2000;

Hernandez-Espinosa et al., 2000; Saenz-Lechon et al., 2006; Lee et al., 2007] ont porté sur l'élaboration d'outils d'évaluation objective standardisés et applicables aux voix dysphoniques. Ces derniers reposent sur l'acquisition de mesures acoustiques, aérodynamiques et/ou physiologiques par le biais de capteurs. Ces techniques dites instrumentales constituent une première catégorie de méthodes objectives. Néanmoins, elles comportent un certain nombre de limites et de contraintes comme le choix du matériau phonétique utilisé pour les enregistrements (voyelles tenues) souvent trés éloigné de la parole continue, la nécessité d'appareillage médical coûteux, la pratique de méthodes invasives non dénuées de risque pour le patient, des résultats statistiques dépendants fortement de la population de patients observés, ... Tout comme pour l'évaluation perceptive, qualifier et quantifier objectivement le dysfonctionnement de la voix pathologique n'est pas une tâche triviale.

Une deuxième catégorie de méthodes objectives s'est alors intéressée aux techniques utilisées en Traitement Automatique de la Parole (TAP) et à leur adaptation dans le cadre des voix pathologiques [Dibazar et al., 2002; Wang & Jo, 2006; Yi & Loizou, 2008]. Comparées aux autres méthodes instrumentales analytiques, l'avantage et l'originalité de ces approches reposent sur :

1. la capacité à analyser de la parole continue proche de l'élocution naturelle ;
2. la capacité à traiter de grands corpus, permettant de mener des études à grande échelle et d'obtenir des informations statistiques significatives ;
3. une analyse acoustique, simple et automatique, permettant une utilisation clinique facile à caractère non invasif et à faible coût humain.

Cet intérêt a été motivé par l'essor du Traitement Automatique de la Parole au cours des quinze dernières années. En effet, les technologies liées à la Reconnaissance Automatique de la Parole (ou RAP qui consiste en l'étude du contenu linguistique d'un énoncé observé [Haton et al., 2006]), à la Reconnaissance Automatique du Locuteur (ou RAL qui consiste à reconnaître l'identité d'une personne par analyse de sa voix [Bimbot et al., 2004]) et à l'Identification Automatique de la Langue (ou IAL qui consiste à déterminer la langue parlée [Lamel & Gauvain, 1994] ou les accents régionaux [Ferragne & Pellegrino, 2006] à partir d'un échantillon de parole) ont prouvé leur pertinence dans l'extraction des informations, linguistiques et extra-linguistiques, véhiculées par la parole et la voix. Ainsi, hormis le message linguistique porté par un signal de parole, d'autres informations sur les spécificités d'un individu peuvent en être extraites telles que son identité, son émotivité (colère, joie, ...), son état pathologique (rhume, rhinolalie, ...) ou ses particularités régionales. Si l'on considére les phénomènes liés à la dysphonie comme une classe d'information extra-linguistique au même titre que celles citées ci-dessus, il est possible d'admettre qu'elles puissent être extraites et traitées par les techniques utilisées pour la reconnaissance automatique de la parole.

Ce travail s'inscrit dans ce contexte, et plus particulièrement, dans l'adaptation des techniques de RAL à la tâche de classification des voix dysphoniques suivant leur degré de sévérité. Contrairement aux méthodes objectives citées ci-dessus, son objectif n'est pas d'améliorer les performances du système sur la tâche visée, mais plutôt, de mieux

appréhender et comprendre les phénomènes acoustiques liés à la dysphonie dans le signal de parole. L'originalité de ce travail repose, par conséquent, sur l'utilisation d'un système de classification automatique comme outil pour caractériser ces phénomènes et apporter aux experts humains de nouvelles connaissances sur la dysphonie et ses répercussions sur le signal de parole. Ces nouvelles connaissances pourront par la suite faire l'objet d'un approfondissement par l'expert humain (phonéticiens) en vue de les enrichir et/ou de les valider. Les approches proposées dans cette thèse reposent sur les hypothèses suivantes :
◦ la dysphonie n'est pas considérée comme un phénomène homogène dans le signal de parole mais plutôt comme un phénomène intermittent et superposé aux caractéristiques phonétiques et linguistiques ;
◦ les techniques de reconnaissance automatique sont capables d'extraire et de modéliser les phénomènes irréguliers et extra-linguistiques liés à la dysphonie contenus dans les échantillons de parole.

Dans ces conditions, deux volets sont décrits dans ce document :

1. le système de classification et ses performances ;
2. la recherche des informations pertinentes.

Le premier volet est consacré à la description du système de RAL adapté au contexte pathologique (Partie II).

Dans ce cadre, le chapitre 2 présente le contexte expérimental dans lequel une description détaillée des deux corpus utilisés est fournie ; le premier est le corpus des voix dysphoniques sur lequel va porter la caractérisation des phénomènes de la dysphonie, le deuxième est un corpus de voix françaises nécessaire à l'application des techniques de RAL. Le système RAL du LIA (Laboratoire Informatique d'Avignon) basé sur une modélisation statistique GMM (Gaussian Mixture Model) et adapté au contexte pathologique est décrit au chapitre 3. Ce dernier débute par une présentation des différentes étapes de la paramétrisation acoustique, détaillant les approches utilisées dans ce travail de thèse. La phase d'apprentissage est ensuite abordée, mettant en avant la nécessité des techniques de RAL pour pallier le manque de données pathologiques disponibles. Enfin, la phase de décision est décrite dans le cadre de la voix pathologique. Finalement, ce premier volet se termine par le chapitre 4 dans lequel différentes tâches de classification des voix dysphoniques sont détaillées. Ici, les tâches classiques de décision binaire et en quatre classes suivant l'échelle d'évaluation GRBAS [Hirano, 1981] sont définies. L'évaluation du système sur ces deux tâches ont conduit à l'élaboration d'une troisième tâche de classification originale, basée sur une échelle plus fine de sept classes. Cette première évaluation permettra d'ouvrir une discussion sur les stratégies de décision les mieux adaptées au contexte pathologique.

Le deuxième volet de cette thèse est consacré à la recherche des informations pertinentes pour caractériser la voix dysphonique (Partie III).

Dans cette partie expérimentale, le chapitre 5 est dédié à l'étude de différentes représentations paramétriques du signal acoustique éprouvées en TAP dans le contexte pathologique. Dans un premier temps, l'étude comparative s'applique uniquement sur les coefficients statiques afin de relever la technique paramétrique la plus pertinente. Trois grandes classes de paramètres sont étudiées : analyse prédictive, analyse spectrale et analyse cepstrale. Dans un deuxième temps, l'apport de l'information dynamique est évalué sur les différentes paramétrisations par ajout aux coefficients statiques des dérivées 1re, 2^{e} et 3^{e}. Une étude sur la longueur de la fenêtre temporelle, utilisée dans l'estimation des paramètres dynamiques, est également menée. Le chapitre 6 est consacré à la caractérisation de la dysphonie dans le domaine fréquentiel. Ici, une approche en sous-bandes de fréquences est élaborée afin d'évaluer les plages fréquentielles les plus pertinentes pour la reconnaissance du degré de sévérité de la dysphonie. Cette approche fera l'objet d'une comparaison entre une évaluation perceptive et une évaluation objective, toutes les deux appliquées sur la plage de fréquences apparaissant comme la plus pertinente de cette étude. Finalement, au regard des résultats obtenus précédemment, l'étude du domaine fréquentiel a été étendue à une analyse phonétique présentée au chapitre 7. Cette étude se propose d'analyser les manifestations de la dysphonie suivant différentes bandes de fréquences en fonction de la nature des segments phonétiques observés. Ces travaux conduiront finalement à une discussion sur les effets de la bande téléphonique sur la qualité de la voix dysphonique.

Enfin, ce travail de thèse se clôturera par un ensemble de conclusions et de perspectives (Partie IV).

Première partie

Etat de l'Art «Multidisciplinaire»

Chapitre 1

La voix pathologique et son évaluation

Sommaire

Résumé

Dans ce chapitre consacré à la voix pathologique, un rappel sur l'anatomie et la physiologie de l'appareil vocal sera présenté avant d'aborder succintement les troubles de la voix, de la parole et du langage. Les dysphonies, pathologies sur lesquelles cette thèse porte, seront abordées selon leur origine fonctionnelle ou organique. Pour conclure, l'évaluation de la voix dysphonique sera décrite avec ses différentes méthodes subjectives et objectives.

Support privilégié de la communication parlée, la voix a comme rôle primordial de véhiculer nos pensées, de transmettre nos émotions et nos sentiments. Le message véhiculé peut être verbal ou non verbal. Elle nous permet d'exprimer nos émotions tout en donnant des indications sur notre personnalité. C'est pourquoi elle est perçue comme le reflet de notre personnalité, le propre de notre identité. Elle est aussi une fenêtre sur notre santé, physique ou psychique, et peut aussi être porteuse d'informations sur notre situation sociale et culturelle. De part sa dimension esthétique, c'est souvent lorsqu'une altération ou une lésion de la voix apparaît que l'on prend réellement conscience de son importance dans notre vie quotidienne et que nous consultons des spécialistes tels que les ORL, les phoniatres et les orthophonistes.

Après une observation clinique du larynx et une évaluation à l'oreille de la qualité de la voix du patient, les spécialistes posent un diagnostic. Etant donné la complexité de la tâche et le caractère subjectif d'une écoute individuelle, la recherche s'est orientée vers le développement d'appareillages informatisés afin d'apporter un outil d'aide aux praticiens. Il existe à l'heure actuelle deux types d'évaluation du trouble vocal : l'évaluation perceptive et l'analyse instrumentale.

Après un rappel sur l'anatomie et la physiologie de l'appareil phonatoire, nous présenterons dans ce chapitre les troubles de la voix, de la parole et du langage. Nous insisterons plus particulièrement sur la qualité de la voix dysphonique avec les différentes techniques, perceptives et instrumentales, permettant son évaluation clinique.

1.1 L'appareil phonatoire

Comme le montre la figure 1.1, l'appareil phonatoire repose sur la partie haute de l'appareil respiratoire. Il est composé de trois systèmes :

1. « la soufflerie » : l'énergie aérodynamique est fournie par la pression de l'air trachéal *i.e.* les poumons fournissent l'air à travers la trachée artère jusqu'aux CV sous l'action des muscles du thorax et de l'abdomen ;
2. « le vibrateur » : le larynx transforme l'énergie aérodynamique en énergie acoustique *i.e.* la vibration sonore apparaît au niveau des CV lorsqu'elles sont rapprochées sur la ligne médiane ;
3. « l'articulation » : le long du canal vocal, le son laryngé est modifié par des éléments fixes et mobiles *i.e.* rapprochement d'un organe vers un lieu d'articulation désignant l'endroit où s'effectue l'obstruction au passage de l'air. Ainsi, les articulateurs sont à l'origine des cavités qui permettent la mise en résonance du son. Ce dernier est alors amplifié et filtré par les cavités aériennes situées au dessus des CV jusqu'au niveau des lèvres (« les résonateurs bucco-laryngés »).

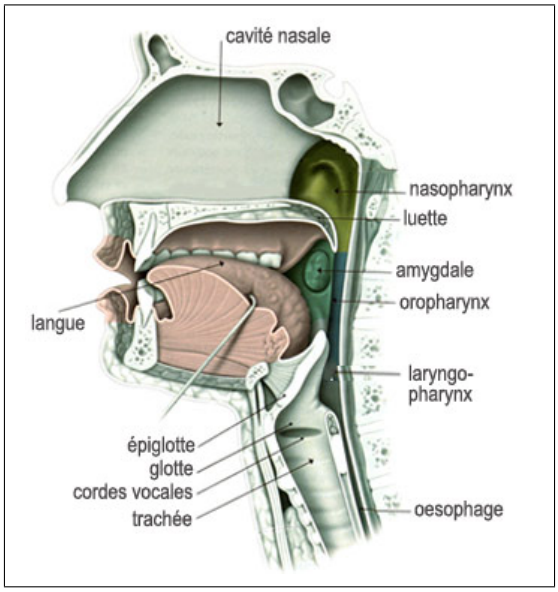


FIG. *1.1* – *L'appareil phonatoire (Source http ://lecerveau.mcgill.ca/).*

L'appareil phonatoire comprend donc le larynx avec ses dépendances ainsi que l'appareil respiratoire coordonné avec la posture du corps. L'ensemble de ces organes ne sauraient être isolés du reste de l'organisme et encore moins du cerveau.

La voix est donc un souffle le plus souvent expiratoire sonorisé dont la qualité dépend de la façon de respirer, de la physiologie de l'appareil phonatoire (en particulier des CV), de la façon d'articuler et de la résonance. Chez l'Homme, le larynx a un rôle fondamental dans la physiologie de la phonation, de la déglutition, de la respiration et de la toux.

Concernant la phonation, elle décrit l'ensemble des processus physiologiques et physiques correspondant à l'apparition d'une vibration sonore au niveau des cordes vocales (CV), permettant la production de la voix.

1.1.1 La fonction du larynx

Comme décrit sur les figures 1.2 et 1.3, le larynx est un conduit « fibro-musculo-cartilagineux » situé au carrefour des voies aérodigestives. Il peut être divisé en trois étages :

1. « supra-glottique » : cavités situées au dessus des CV jusqu'à l'hypopharynx ;
2. « glottique » : partie centrale où se trouvent les deux CV ;
3. « sous-glottique » : région au-dessous des CV jusqu'au bord trachéal supérieur.

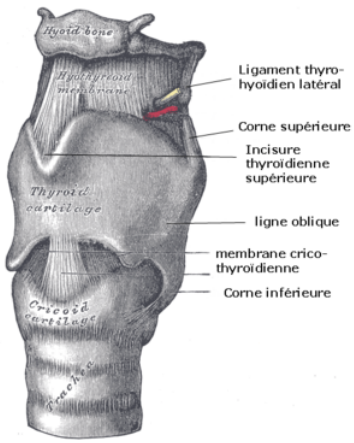


FIG. *1.2* – *Larynx (vue antérieure).*

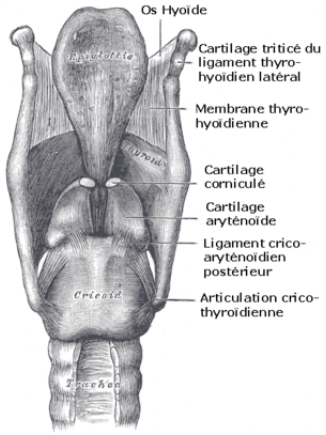


FIG. *1.3* – *Larynx (vue postérieure).*

Source http ://fr.wikipedia.org/wiki/Larynx

Organe principal de la production vocale, le larynx n'y est pourtant pas uniquement dédié. Il a pour fonction première celle de sphincter permettant l'obturation de la trachée, rôle vital de protection des voies respiratoires lors de la déglutition. De plus, faisant partie intégrante des voies aériennes supérieures, le larynx assume une fonction respiratoire.

Le vibrateur laryngé est constitué par les deux CV[1] (« vocal fold » sur la figure 1.4) qui apparaissent comme deux lèvres horizontales placées à l'extrémité supérieure de la trachée. Elles sont attachées horizontalement entre le cartilage thyroïde situé à l'avant et les cartilages arythéroïdes situés à l'arrière.

[1] également appelées plis vocaux

L'espace compris entre les CV est appelé la glotte. C'est un espace variable divisé en deux parties : la glotte phonatoire (partie antérieure musculo-ligamentaire) et la glotte respiratoire (partie postérieure cartilagineuse). La glotte s'ouvre lors de l'inspiration (abduction) et se referme lors de la phonation (adduction) permettant aux CV de vibrer grâce à l'action du souffle pulmonaire.

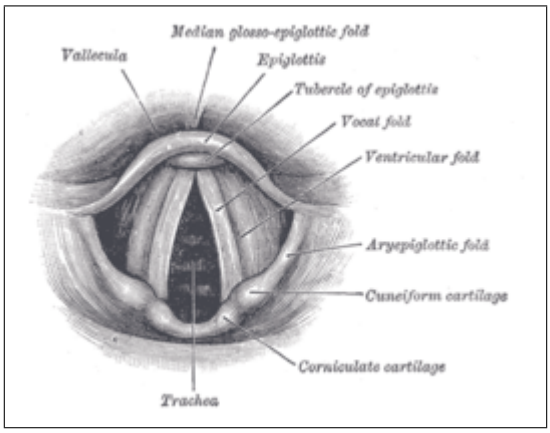


FIG. *1.4* – *Vue supérieure du larynx (Source http ://fr.wikipedia.org/wiki/Corde_vocale).*

Située dans la partie supérieure du larynx, l'épiglotte (« epiglottis » sur la figure 1.3) est un clapet qui se rabat en arrière au moment de la déglutition pour permettre le passage des aliments de l'oesophage vers l'estomac en évitant qu'ils ne passent dans les voies respiratoires.

1.1.2 Le squelette du larynx

Au niveau de la gorge, le larynx est situé entre le pharynx et la trachée. Il forme une saillie au niveau du cartilage thyroïde, visible de l'extérieur du cou, communément appelée « pomme d'Adam ».

Le squelette laryngé (figure 1.5) est constitué principalement :
- des éléments de soutien : les cartilages cricoïde [6] et thyroïde [3] ;
- des éléments de mobilité : les deux cartilages aryténoïdes [5], l'épiglotte [1] et les muscles endolaryngés [11] [12] (muscles adducteurs et abducteurs) ;
- de l'élément de stabilité : l'os hyoïde [8].

La fonction des éléments de soutien est de maintenir ouverte la voie aérienne, alors que celle des éléments de mobilité est de fermer le larynx.

Le larynx est composé de cinq structures cartilagineuses :

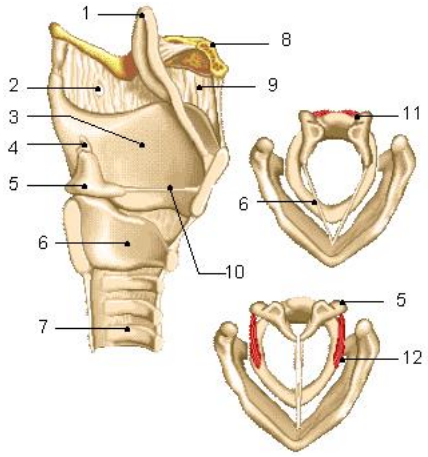
• Le **cartilage cricoïde** en forme d'anneau renforce la partie supérieure de la trachée afin de maintenir l'ouverture des voies aériennes.

• Les **deux cartilages aryténoïdes** constituent les seuls cartilages mobiles du larynx sur lesquels s'insèrent les muscles de la phonation ; ils assurent l'essentiel de la fonction laryngée.

• Le **cartilage épiglottique**, semblable à un rabat, permet l'occlusion des voies aériennes pendant la déglutition ; il s'abaisse contre le larynx qui s'élève pour empêcher les aliments de pénétrer dans la trachée.

• Le **cartilage thyroïde** réalise la partie principale de la structure du larynx ; il fixe l'épiglotte par les bandes ventriculaires et attache également les CV aux apophyses vocales du cartilage aryténoïde de la glotte.

Le larynx permet la vocalisation en faisant vibrer les CV à hauteur désirée lorsque l'air le traverse. La hauteur de la voix dépend étroitement de l'élasticité et de la tension des CV.



[1] Cartilage épiglottique ou l'épiglotte
[2] Orifice laryngien
[3] Cartilage thyroïde
[4] Cartilage corniculé
[5] Cartilage aryténoïde
[6] Cartilage cricoïde
[7] Trachée
[8] Os hyoïde
[9] Membrane hyo-thyroïdienne
[10] Cordes vocales
[11] Muscles crico-aryténoïdiens postérieurs
[12] Muscles crico-thyroïdiens obliques

FIG. *1.5* – *Structure du larynx.*

De plus, on distinguera :

• **les muscles intrinsèques** qui assurent l'abduction (écartement), l'adduction (rapprochement) et le réglage de la tension des CV.

• **les muscles extrinsèques** qui fixent les cartilages du larynx aux structures anatomiques environnantes et facilitent l'action des muscles endolaryngés intrinsèques.

1.2 Les troubles de la voix, de la parole et du langage

En communication parlée, la voix, la parole et le langage entretiennent des relations très étroites justifiant dans ce document une étude simultanée. La parole utilise comme support acoustique la voix pour exprimer le langage. Il en est de même dans certaines affections pour lesquelles il est difficile de séparer les troubles de la voix, de la parole et du langage qui apparaissent souvent étroitement intriqués. L'impossibilité d'émettre un son articulé et modulé, empêche l'enchaînement des syllabes, et par conséquent la compréhension des mots. Un trouble de la parole peut donc génèrer un trouble du langage tout en provoquant des dégradations de la voix.

Bien que cette thèse soit dédiée à la dysphonie, les troubles liés à la voix, à la parole et au langage sont présentés dans cette section en distinguant trois grandes classes de pathologie :

• **Les dysphonies** sont des anomalies de la qualité de la voix. Par exemple, la raucité peut provenir d'une paralysie, d'un surmenage vocal ou d'une laryngite.

• **Les dysarthries** sont des troubles de l'élocution liés à une paralysie ou à une incoordination des muscles entrant en jeu dans l'articulation. Elles sont d'origine neurologique.

• **Les aphasies** sont des troubles du langage dus à une lésion du cortex cérébral. Le patient ne comprend plus le sens des mots ou ne peut plus s'exprimer.

1.2.1 Les dysphonies

Une définition de la dysphonie (ou enrouement) désigne une altération du son laryngé impliquant une ou plusieurs des caractéristiques acoustiques de la voix telles que le timbre, l'intensité et la hauteur tonale. Elle peut être liée à une atteinte organique ou fonctionnelle du larynx. Ces dysfonctionnements vocaux peuvent être la conséquence audible de changements anatomiques du larynx ou d'atteintes du système nerveux et cela à titre momentané ou durable.

Il faut tout de même préciser que l'altération de la fonction vocale peut être purement fonctionnelle *i.e.* absence de lésion organique du larynx à l'origine de la dysphonie, et qui est entretenue principalement par un comportement vocal défectueux. Elle peut apparaître à la suite de traumatismes laryngés (intubation , hurlement, ...), de chocs psychologiques, chez certaines professions (enseignants, chanteurs, ...), en raison de pollutions aériennes et sonores, ... Elle est dite « simple » tant que la muqueuse des cordes vocales ne s'altère pas et « compliquée » dès l'apparition de complications laryngées provoquant des lésions essentiellement de la muqueuse du pli vocal.

Comme décrites dans [Teston, 2004], les dysphonies peuvent se définir en deux grandes familles suivant leur origine morphologique ou neurologique[2].

Les dysphonies d'origines morphologiques

Elles englobent les changements anatomiques de la glotte provoqués par les lésions des cordes vocales. Le tableau 1.1 énumère les facteurs provoquants ainsi que les conséquences des lésions organiques du larynx sur la voix.

facteurs provoquants	**conséquences sur la voix**
forçage vocale avec apparition de nodules, polypes et kystes (lésions bénignes des cordes vocales)	voix plus grave, rauque, soufflée timbre voilé, sourd, éraillé
laryngite (inflammations de l'ensemble des cordes vocales occasionnées par des infections)	voix plus grave, rauque, difficultés dans les aigus voix peu timbrée, extinction de voix possible
traumatismes chirurgicaux (suite à l'ablation d'un cancer cordal par exemple)	voix très dégradée (désonorisations), grave, faible intensité timbre très rauque, granuleux et soufflé

TAB. *1.1* – *Les dysphonies d'origines morphologiques*

Les dysphonies d'origines neurologiques

Elles sont provoquées par l'état neuromoteur du patient. Le tableau 1.2 présente les facteurs à l'origine de la dysphonie et leurs impacts sur la qualité de la voix.

facteurs provoquants	**conséquences sur la voix**
hypotonie	intensité faible, abaissement de la Fo
hypertonie	hésitation au démarrage du voisement émissions vocales discontinues augmentation de la Fo timbre sourd par manque d'harmonique et voilé
tremblements	modulation de la hauteur, de l'intensité et du timbre chevrotements (instabilité de la Fo en voix tenue)
dystonie laryngée	changements brutaux de la hauteur de la voix qui peut s'interrompre, repartir, glisser et chevroter timbre désagréable (voix inintelligible)
paralysie laryngée	monotone avec perte de mélodie, nombreuses désonorisations voix soufflée et rauque avec fuite d'air importante essoufflement en fin de phrase

TAB. *1.2* – *Les dysphonies d'origines neurologiques*

[2] en section 1.3, elles seront définies plus en détails suivant leur origine fonctionnelle ou organique

Notons qu'aux termes de « hypokinésie/hyperkinésie », il est préféré ceux de « hypotonie/hypertonie » car pour certains auteurs, il s'agit plus d'un problème de tension des cordes vocales que d'un défaut de mouvement.

De plus, la dystonie laryngée, appelée aussi dysphonie spasmodique, correspond à un stade très avancé de la dysphonie hypertonique et se manifeste par des spasmes laryngés et/ou respiratoires. Sa composante psychologique est très importante.

Concernant les paralysies laryngées (unilatérales ou bilatérales) dont l'origine la plus fréquente est une lésion du nerf laryngé inférieur (NLI), elles correspondent à des troubles de la mobilité des cordes vocales avec défaut d'accolement entre elles.

1.2.2 Les dysarthries

D'après Darley et al. (1975), la dysarthrie est définie comme un trouble de la réalisation motrice de la parole, secondaire à des lésions du système nerveux central ou périphérique. Actuellement, ce terme englobe les troubles moteurs de la parole d'origines neurologiques à l'exception de l'apraxie[3] de la parole. Il concerne les troubles d'origines neurogènes ayant des répercussions sur la respiration, la phonation, l'articulation, la résonance et la prosodie.

Les dysarthries se distinguent selon la localisation des lésions neurologiques, la nature de l'affection en cause (AVC, TC, ...) ou la réaction du sujet à son propre trouble (notamment pour le traitement rééducatif). Historiquement, la classification des dysarthries a été envisagée selon différentes considérations telles que neurologiques (étiologiques ou topographiques), physiopathologiques ou cliniques.

n°	Catégories de dysarthries	Exemples de pathologies
1	flaccide	paralysies bulbaires
2	spastique	paralysies pseudo-bulbaires
3	ataxique	lésions cérébelleuses
4	hypokinétique	syndrome parkinsonien
5	hyperkinétique rapide	chorée
6	hyperkinétique lente	athétose, dystonie
7	mixte (flaccido-spastique)	Sclérose Latérale Amyotrophique (SLA)

TAB. *1.3* – *Classification des dysarthries de Darley, Aronson et Brown (1969)*

Depuis les travaux de Darley et al. (1969a,b), les dysarthries sont classées fréquemment en sept catégories comme présentées dans le tableau 1.3. La classification est basée sur la description perceptive des anomalies regroupées pour construire des hypothèses physiopathologiques.

[3] décrit une atteinte de la capacité de planification ou de programmation des mouvements de la parole

S'écartant de celle de Darley, la classification dans Le Huche & Allali (2001b) propose six groupes de dysarthries constitués suivant la symptomatologie et l'étiologie. Le tableau 1.4 présente cette nouvelle classification en montrant la correspondance avec celle de Darley (1969).

n°	Classification de	
	Le Huche et Allali	**Darley, Aronson et Brown (1969)**
1	paralytique	(1) flaccide (2) spastique (7) mixte
2	akinétique	(4) hypokinétique
3	dyskinétique	(5) hyperkinétique rapide (6) hyperkinétique lente
4	ataxique	(3) ataxique
5	apraxique	
6	dystonique	(6) hyperkinétique lente

TAB. *1.4* – *Classification des dysarthries : ⟨ Le Huche et Allali ⟩ versus ⟨ Darley, Aronson et Brown (1969) ⟩*

Plusieurs observations peuvent être faites au regard du tableau 1.4 :

1. l'ensemble des dysarthries paralytiques est regroupé sous une même catégorie sans distinction sur l'origine de la lésion qu'elle soit centrale, périphérique ou mixte ;
2. le terme « akinétique » est préféré à « hypokinétique » car le trouble de la parole du malade parkinsonien correspond plus à la perte de l'initiative motrice entraînant une diminution du mouvement corporel, transparaissant également sur son visage par un faciès inexpressif ;
3. les dysarthries hyperkinétiques (lentes et rapides) sont regroupées sous la rubrique « dyskinétique », et différenciées des dystonies déplaçées dans une catégorie qui leur est plus spécifique ;
4. une nouvelle classe est dédiée aux dysarthries apraxiques, dénotant ainsi une attention toute particulière.

Même si, pour chacune des catégories de dysarthries, des profils perceptifs peuvent être définis, caractérisant les troubles de la voix et de la parole, comme par exemple
- *pour les dysarthries flaccides : (1) l'hypernasalité (2) l'imprécision des consonnes (3) une voix voilée (4) une voix monotone (5) une déperdition nasale (6) ... ;*
  ou
- *pour les dysarthries akinétiques : (1) la monotonie (2) la réduction de l'accentuation (3) l'absence de variation de l'intensité de la voix (4) ... ;*

ces derniers ne peuvent être utilisés pour diagnostiquer le type de dysarthrie ; la connaissance du diagnostic (examen) médical adapté reste prioritaire face à de tels profils purement symptomatiques et perceptifs qui risquent d'aboutir à des propositions de rééducation complètement inadaptées, voire à des conséquences fâcheuses pour les patients.

Un descriptif des différentes dysarthries est présenté en annexe V pour plus de détails.

1.2.3 Les aphasies

Pathologie du système nerveux central, l'aphasie désigne un trouble du langage altérant aussi bien l'expression que la compréhension du langage parlé ou écrit. Ce trouble apparaît consécutivement à une lésion le plus souvent des centres du langage dans le cortex cérébral[4]. Il survient donc en dehors de tout déficit sensoriel, voire même de dysfonctionnement de l'appareil phonatoire.

De manifestations très diverses, ce trouble peut n'être que « léger » pour certaines personnes aphasiques comme par exemple, une difficulté à trouver leurs mots ; tandis que pour d'autres, cela peut être beaucoup plus grave allant jusqu'à la perte de la faculté de s'exprimer et de comprendre. Une personne aphasique peut également éprouver certaines difficultés à lire et à écrire.

La difficulté dans le choix, l'utilisation et la compréhension des mots font que l'aphasie est considérée comme un trouble du langage avant tout et non de la parole. Celui-ci est qualifié d'« acquis » dans le sens où il survient chez une personne ayant un langage normal avant l'apparition du trouble. Il se distingue donc des problèmes tels que, une dyslexie développementale ou un bégaiement, pouvant apparaître durant le développement du langage chez l'enfant.

Voici énumérées les principales causes de l'aphasie :
- l'accident vasculaire cérébral (AVC) pouvant aussi provoquer certains problèmes d'élocution, en raison de la paralysie de la bouche et des muscles de la phonation (dysarthrie) ;
- un traumatisme cranio-cérébral à la suite d'une blessure à la tête ;
- une tumeur cérébrale dont la croissance peut entraîner une compression progressive du centre du langage ;
- un syndrome dégénératif tel que la maladie d'Alzheimer où les cellules cérébrales responsables du langage sont progressivement détruites ;
- un syndrome infectieux ou inflammatoire comme l'encéphalite.

Selon la localisation des atteintes dans les régions cérébrales, on distinguera différents types d'aphasie. Cependant, comme la majorité des aphasies ont leur origine lésionnelle dans les deux principales aires du langage, Broca et Wernicke, il est classiquement décrit deux grandes formes d'aphasies :

• **l'aphasie de Broca (aphasie motrice)** : incapacité de formuler oralement ses idées, nombreuses transformations phonétiques, manque de mots, troubles syntaxiques et prosodiques, ainsi que de la compréhension ; une personne aphasique de Broca apparaît comme ralentie, monotone et souvent triste ;

[4] trouble de l'hémisphère dominant où se situe le centre du langage, responsable de l'expression et de la compréhension du langage

• **l'aphasie de Wernicke (aphasie sensorielle)** : dont les symptômes pathologiques s'opposent quasiment à ceux de l'aphasie de Broca avec un débit très rapide voire incontrôlé, vocabulaire riche, pas de problème articulatoire ; par contre, le discours est incompréhensible et dénué de sens ; une personne aphasique de Wernicke apparaît tonique, parfois joyeuse, car ne se rendant pas compte de son état.

Cependant, ces deux formes d'aphasies sont souvent intriquées car il n'existe pas d'aphasie qui soit purement « motrice » ou « sensorielle ».

1.3 Compléments sur les dysphonies

En général, la dysphonie est assimilée à une voix ayant un timbre altéré. Plus précisément, le terme de dysphonie est associé à toute voix perçue comme pathologique c'est à dire dont l'altération vocale est perçue sur un ou plusieurs de ses caractères acoustiques comme le timbre, l'intensité et la hauteur tonale.

Cependant, il peut exister des voix « dysphoniques » sans perturbation acoustique et des voix « altérées » non pathologiques. C'est la raison pour laquelle Le Huche & Allali (2001a) la définissent comme « *un trouble momentané ou durable de la fonction vocale ressenti comme tel par le sujet lui-même ou son entourage.* »

Nous conviendrons que la dysphonie est un trouble de la voix qui résulte d'une dysfonction de production et/ou d'une lésion organique, principalement liée aux cordes vocales. Dans les prochaines sections, nous présenterons les dysphonies d'origines fonctionnelles puis les dysphonies d'origines organiques.

1.3.1 Les dysphonies dysfonctionnelles

On définit la dysphonie dysfonctionnelle comme une altération du geste vocal en l'absence de perturbation organique permanente à l'origine de cette dysphonie.

Le terme « dysfonctionnelle » est préféré à celui de « fonctionnelle », sachant que parler de « trouble fonctionnel » suggère communément « absence de lésion organique ». Or il apparaît que le dysfonctionnel et l'organique sont souvent entremêlés :
« *une dysphonie d'origine fonctionnelle peut se compliquer de lésions organiques du larynx (nodule du pli vocal par exemple) provoquées par le forçage vocal ou se constituer à l'occasion d'une altération organique transitoire du larynx (laryngite aiguë par exemple).* »

Le Huche & Allali (2001a) définissent la dysphonie dysfonctionnelle comme « *une altération de la fonction vocale entretenue essentiellement par une perturbation du geste vocal* », et considérent trois notions clés présentes dans le mécanisme d'installation et d'entretien de la dysphonie dysfonctionnelle : le forçage vocal, les facteurs déclenchants et les facteurs favorisants.

Le forçage vocal :

Souvent de façon inconsciente, le forçage est utilisé afin de compenser la baisse de rendement vocal. En réalité, il existe un cercle vicieux de forçage vocal : l'effort vocal devient proportionnel à l'inefficacité de production de la voix ; ainsi, plus le patient force sur sa voix, plus celle-ci se dégrade et devient de moins en moins efficace. Sous l'effet de facteurs favorisants, son usage continu peut entraîner une altération de la muqueuse laryngée pouvant détériorer de plus en plus la fonction vocale pour aboutir à une dysphonie aggravée (voire à une aphonie).

Le forçage vocal correspond à une augmentation des tensions péri-laryngées au cours de la phonation, dues à un travail musculaire inapproprié. Il se manifeste le plus souvent par une posture caractéristique (projection du visage vers l'avant, tensions musculaires cervicales, ...).

Les facteurs déclenchants :

On peut distinguer deux classes :

• certaines affections ORL : laryngite aiguë, traumatismes laryngés (intubation, hurlements), angine, ... Ce sont des altérations ponctuelles de l'organe vocal (ou des organes voisins) susceptibles de perturber le comportement vocal et de déclencher le forçage vocal.

• des facteurs psychologiques : contrariétés ou chocs psychologiques à l'occasion d'événements professionnels, familiaux ou sentimentaux. Dans la vie d'un individu, il peut s'agir de périodes de surmenage, de maladie durant lesquelles les capacités de résistance physique sont affaiblies, et donc propices au déclenchement du processus de la dysphonie.

Les facteurs favorisants :

Ils offrent un terrain propice aux facteurs déclenchants. Cela concerne des professions où l'usage de la voix (parler ou chanter) est l'outil principal comme les enseignants, les commerciaux, les démonstrateurs, les chanteurs, ...

De même, le tempérament nerveux ainsi que certains facteurs psychologiques comme la tendance à l'anxiété et au perfectionnisme, prédisposeraient à la constitution de la dysphonie dysfonctionnelle.

Il faut aussi inclure l'intoxication alcoolique et tabagique qui sont nocives pour la muqueuse des plis vocaux, la pollution sonore (élévation de la voix pour se faire entendre dans un environnement bruyant), les polluants aériens (atmosphère chargée de poussières, odeurs, fumées, nettoyant chimique, ...), l'air sec et conditionné, ...

Nous distinguerons la « dysphonie dysfonctionnelle simple » où l'on n'observe pas de lésion spécifique du larynx, de la « dysphonie dysfonctionnelle compliquée » présentant des complications laryngées.

Les dysphonies dysfonctionnelles simples

Elles ne présentent pas de complication laryngée (pas de lésion organique du larynx). Les difficultés vocales éprouvées par les sujets atteints de dysphonie dysfonctionnelle simple ne présentent pas de description type de la façon dont les troubles apparaissent. On observe une grande variabilité des facteurs déclenchants et favorisants, ainsi que des moments de la journée où se manifestent les troubles.

Voici les principales plaintes émises par les patients au sujet de l'appréciation sur leurs possibilités vocales et sur les impressions ressenties dans la région de l'organe vocal :

IMPRESSIONS SUBJECTIVES SUR LES POTENTIALITÉS PHONATOIRES
- enrouement
- irrégularité du timbre
- inefficacité de la voix d'appel
- baisse des performances de la voix après un temps d'usage

IMPRESSIONS SUBJECTIVES RESSENTIES DANS LA RÉGION LARYNGÉE
- gêne dans la gorge avec picotements
- hemmage (raclement de la gorge)
- sensations d'irritation laryngée et de brûlure
- fatigue ou douleurs à la phonation prolongée

En raison de la diversité des différentes altérations vocales observées, il n'est pas possible de définir des catégories précises dans lesquelles on pourrait classer les troubles.

Les dysphonies dysfonctionnelles compliquées

Comme cela a été précisé en 1.3.1, l'existence d'un comportement vocal défectueux (forçage vocal) dans le cas d'une dysphonie dysfonctionnelle peut entraîner des complications laryngées. On parle alors de « laryngopathies dysfonctionnelles ». Elles correspondent à des lésions essentiellement de la muqueuse du pli vocal produites ou entretenues par un geste vocal défectueux.

La prise en charge médicamenteuse ou chirurgicale de la laryngopathie dysfonctionnelle n'est pas une condition suffisante pour rétablir la fonction vocale. En effet, il ne faut pas perdre de vue qu'elle correspond à une lésion organique venant s'ajouter à une perturbation fonctionnelle. Une rééducation vocale de la dysphonie originelle reste malgré tout le plus souvent indispensable.

Voici présentées les principales laryngopathies dysfonctionnelles.

NODULE DU PLI VOCAL

Définition	*épaississement localisé de la muqueuse, siégeant sur le bord libre d'un pli vocal (ou des deux), à l'union du tiers antérieur et du tiers moyen de celui-ci*
Plaintes	picotements, douleurs, fatigue à la phonation prolongée
Voix parlée	timbre fréquemment éraillé, parfois des désonorisations plus rarement un petit sifflement se rajoutant à la voix
Evolution	peut disparaître complètement lorsque cesse le forçage vocal
Traitement	rééducation vocale pour éliminer le forçage vocal, chirurgie possible

OEDÈME CHRONIQUE DES PLIS VOCAUX *ou oedème de Reinke ou pseudo-myxome*

Définition	*transformation oedémateuse du chorion de la muqueuse du pli vocal intéressant l'espace de Reinke et déformant la face supérieure et le bord libre de ce pli*
Plaintes	fatigue à la phonation prolongée, manque de portée de la voix perte du registre aigu, toux d'irritation, tic d'hemmage
Voix parlée	aggravation de sa tonalité, souvent rauque, éraillée et fatigable
Evolution	son développement peut provoquer des difficultés respiratoires l'arrêt du tabac permet une stabilisation ou même une relative réduction
Traitement	chirurgie et rééducation vocale

POLYPE DU LARYNX

Définition	*pseudo-tumeur bénigne du pli vocal*
Plaintes	irrégularité de la production vocale, la voix se dérobe fatigue vocale, sensation d'irritation laryngée, hemmage
Voix parlée	abaissement de la tonalité de la voix timbre sourd, éraillé ou graillonnant, plus rarement rauque
Evolution	n'est pas susceptible de régresser spontanément augmentation du volume du polype avec un risque respiratoire
Traitement	ablation chirurgicale et rééducation vocale

KYSTE MUQUEUX PAR RÉTENTION

Définition	*tuméfaction apparaissant au niveau du pli vocal qui résulte d'une accumulation de sécrétion mucoïde due à l'obstruction du canal excréteur d'une glande muqueuse*
Plaintes	timbre vocal assourdi, parfois éraillé, par moment bitonal diminution de l'intensité vocale, brefs moments de désonorisation baisse de qualité du timbre vocal dès que la voix est plus sollicitée
Evolution	sans traitement, peut rester stationnaire pendant des années tendance à augmenter de volume par poussées successives
Traitement	microchirurgie suivie d'une rééducation vocale

Les dysphonies dysfonctionnelles particulières

Les dysphonies dysfonctionnelles peuvent prendre parfois certaines formes particulières telles que :

• **La raucité vocale infantile** apparaît en général entre 6 et 8 ans. Classiquement, elle peut être déclenchée par une laryngite banale durant laquelle la modération vocale n'a pas été respectée ou une amygdalectomie. Elle évolue de façon irrégulière pour devenir constante. Malgré la douleur, l'enfant parle trop fort afin de compenser l'altération laryngée. Du point de vue acoustique, on observe une tonalité aggravée, souvent un timbre rauque très marqué, des désonorisations sur les finales, une alternance entre des périodes de parole de forte intensité et d'aphonie.

• **Le trouble de la mue** chez le garçon apparaît au moment de la puberté, période durant laquelle l'accroissement du larynx s'effectue sur 6 mois environ. Un allongement « trop rapide » des cordes vocales est à l'origine de ce trouble. La voix alterne entre l'aigu (voix de tête) et le grave (voix de poitrine).

• **La dysphonie chez les chanteurs ou dysodie** est une altération de la voix qui survient brutalement et peu de temps avant le début de la représentation ou du spectacle. Les causes peuvent être variées telles que infectieux, un état de fatigue générale entraînant le malmenage vocal, des facteurs psychologiques (stress, appréhension, inquiétude, ...), ...

• **La dysphonie spasmodique** constitue un trouble de la fonction vocale qui se manifeste par des spasmes respiratoires et/ou laryngés. Touchant des personnes très anxieuses et angoissées, elle perturbe considérablement leur vie sociale en nécessitant des efforts importants pour parler. Elle provoque un étranglement bref de la voix dont l'intensité est limitée et le timbre étouffé. La dysphonie spasmodique est un trouble rare, apparentée aux dystonies (se référer aux dysarthries dystoniques en annexe V pour plus de précisions).

• **Les aphonies et dysphonies psychogènes** sont caractérisées par la disparition de la voix ou par une altération vocale en rapport avec un processus d'inhibition psychologique. Touchant en général les femmes, la malade fait beaucoup d'efforts pour obtenir une voix chuchotée.

1.3.2 Les dysphonies d'origine organique

Dans les dysphonies d'origine organique, la présence d'une lésion laryngée est principalement responsable des troubles de la fonction vocale.

Comme cela a déjà été mentionné en 1.3.1, le dysfonctionnel et l'organique sont souvent entremêlés. Ainsi en réaction à un déficit organique, un sujet peut mettre en place un comportement vocal tel que le forçage ou la retenue. Cette nouvelle composante venant s'ajouter à la dimension fonctionnelle est à prendre en compte lors de l'établissement d'une thérapeutique afin que celle-ci soit efficace et adaptée au trouble du souffrant.

Les principales dysphonies d'origine organique sont présentées en annexe V selon leur étiologie :

- la laryngite aiguë
- la laryngite chronique
- les laryngites spécifiques
- les traumatismes laryngés
- les paralysies laryngées
- les anomalies laryngées congénitales

1.4 L'examen clinique du patient dysphonique

L'examen clinique d'une personne atteinte de troubles de la voix ne se limite pas uniquement à l'examen endoscopique du larynx. La connaissance du comportement phonatoire du sujet dysphonique (par exemple, en situation de forçage vocal), l'importance de la voix dans son activité professionnelle et sociale, sa propre perception de la qualité de sa voix, la manière dont il vit sa phonation (sa souffrance, sa gêne, ...) ou dont son entourage le perçoit, constituent un ensemble primordial et complémentaire de l'examen médical pour déterminer correctement le caractère pathologique ou non pathologique d'une voix.

Ainsi pour Le Huche & Allali (2001a) « *En matière de pathologie vocale, les signes les plus importants à considérer ne relèvent pas directement de l'acoustique de la voix. C'est la gêne du sujet ou de son entourage qui est le fait primordial : sa gêne ou sa souffrance. C'est elle qu'il conviendra d'évaluer avec le plus de soin.* ».

Dans cette section, nous présentons trois méthodes permettant d'établir un bilan clinique d'un patient dysphonique : l'interrogatoire, l'examen physique et l'analyse du comportement vocal.

Sortant du cadre de cette thèse, nous ne développerons pas ici d'autres méthodes complémentaires telles que l'étude du comportement postural du patient [Giovanni et al., 2006], le profil psychologique et l'étude comportementale [Roy et al., 2000] et l'autoévaluation [Woisard et al., 2004].

1.4.1 L'interrogatoire avec le patient

L'examen clinique d'un sujet dysphonique commencera par un interrogatoire durant lequel le patient expose l'historique de sa plainte actuelle, comment et depuis combien de temps les troubles sur la voix sont apparus, les différents traitements entrepris ou explorations déjà faites (en terme de rééducation, de chirurgie, de cure thermale, ...), les résultats observés, ... L'ensemble de ces éléments vont permettre de constituer l'anamnèse c'est-à-dire « l'histoire de la maladie », d'établir un diagnostic complet pemettant d'interpréter différemment les résultats d'examens complémentaires.

Il est important aussi que le patient dysphonique parle de son ressenti à propos de sa propre phonation, ce qu'il pense de sa voix (d'un point de vue esthétique, agréabilité, ...) et de ses possibilités vocales (en terme de puissance, de fatigue, en milieu bruité ou téléphonique, ...), sur ses capacités de contrôle, à chanter, ... Il en est de même sur les sensations qu'il peut éprouver dans la région laryngée (comme picotement, irritation, douleur, ...) ou respiratoire. Ainsi, même si ces éléments constituent des « signes subjectifs », ils pourront permettre au patient de prendre conscience de sa guérison dès leur disparition.

D'autres éléments pourront être reccueillis par le clinicien afin de mieux comprendre la dysphonie du patient tels que sa santé générale, son genre de vie, son tempérament, l'importance de la voix dans sa profession, sa consommation d'alcool et de tabac, son exposition à la pollution aérienne et sonore, au surmenage, ...

1.4.2 L'examen physique du patient

A l'interrogatoire du patient, l'examen du larynx reste indispensable et nécessaire au médecin pour établir un diagnostic complet sur la pathologie. De façon générale, il comporte systématiquement la laryngoscopie et la laryngostroboscopie auxquelles est associé un examen ORL, permettant ainsi d'orienter le diagnostic étiologique. L'examen laryngoscopique dit « indirect » est réalisé à l'aide d'un miroir tel que celui présenté sur la figure 1.6. Il permet d'une part, de vérifier la présence ou non d'une anomalie organique laryngée, d'inspecter l'aspect de la muqueuse laryngée et de contrôler l'existence d'un processus tumoral. Il permet d'autre part, d'évaluer, durant la phonation, le dysfonctionnement de la cinétique laryngée et d'étudier la mobilité des cordes vocales et des aryténoïdes durant les passages alternés entre le mode de phonation (glotte fermée) et respiratoire (glotte ouverte).

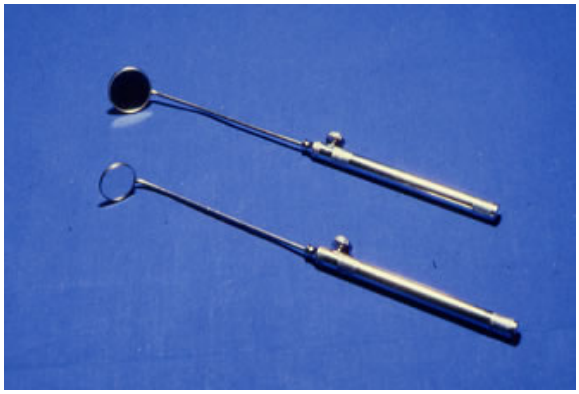


FIG. 1.6 – *Miroir utilisé en laryngoscopie « indirecte ».*

FIG. 1.7 – *Laryngoscopie « directe ».*

Source http ://www-sante.ujf-grenoble.fr

Depuis l'apparition de la fibre optique dans les années 1970, la fibroscopie par voie nasale (ou nasofibroscopie) peut être utilisée en remplacement de l'examen laryngé pratiqué classiquement au miroir. Elle permet l'observation du larynx durant la phonation et le chant. La laryngoscopie au tube rigide (ou épipharyngoscopie) consiste à introduire l'endoscope dans la bouche afin de visualiser l'ensemble du pharyngo-larynx avec une très grande précision. Offrant une meilleure définition de l'image laryngée, elle permet surtout l'examen en lumière stroboscopique. La laryngostroboscopie [Crevier-Buchman et al., 1993b] consiste à examiner les cordes vocales en phonation en utilisant la lumière stroboscopique. Appliquant le principe du stroboscope *i.e.* usage d'une lumière obtenue par flashs successifs, l'examen stroboscopique permet d'obtenir une image au ralenti de la vibration des cordes vocales. Devenu indispensable à l'étude

de la fonction laryngée, il permet d'observer la qualité des vibrations glottiques, de révéler la présence éventuelle d'une lésion intra-cordale et d'explorer l'ondulation de la muqueuse des plis vocaux. En cas de dysphonie durable sans pathologie organique apparente, cet examen est obligatoire.

Il est à préciser que l'examen en fibroscopie ou au tube en lumière normale ne permet pas l'observation de l'ondulation glottique mais seulement des mouvements d'adduction et d'abduction des plis vocaux. Comme décrit par Hirano (1974), l'organisation tissulaire des plis vocaux permet l'ondulation cordale dont la qualité est déterminé par les propriétés viscoélastiques de l'espace de Reinke et par là-même la qualité du timbre vocal selon que la vibration ou ondulation est symétrique, régulière, ...

Connectée au fibroscope ou à un endoscope droit, la caméra à haute vitesse[5] permet au médecin d'observer le phénomène extrêmement rapide de vibrations des plis vocaux, de visualiser la vidéo après examen du patient afin de compléter le dossier médical avec des photographies extraites de la vidéo. Il est à noter que les possibilités d'imagerie du larynx se sont considérablement améliorées grâce à l'apport du numérique.

L'examen clinique du larynx peut éventuellement être complété par un examen audiométrique. Cependant, en cas d'atteinte organique, le seul moyen d'obtenir un diagnostic de certitude reste la laryngoscopie dite « directe » (figure 1.7) qui, réalisée au bloc opératoire sous neuro-analgésie, permet d'effectuer des prélèvements biopsiques. Il est à noter que l'examen ORL ne serait être complet sans la palpation de la glande thyroïde et des aires ganglionnaires.

1.4.3 L'examen du comportement vocal

L'appréciation de la voix durant l'examen interrogatoire du patient ne peut se limiter à la voix dite « conversationnelle ». L'observation du comportement vocal du patient dysphonique en diverses situations peut s'avérer d'un apport considérable pour approfondir la connaissance de la pathologie.

En effet, la dynamique de la voix dite « projetée », par exemple, se caractérise par un changement d'attitude et de posture (comme la verticalisation), par une modification du débit (ralentissement), de l'intensité et du mode respiratoire. Par rapport à la voix « conversationnelle », la dépense énergétique y est beaucoup plus importante. Ainsi en voix « conversationnelle », il consistera à relever les caractères acoustiques vocaux tels que l'intensité, la hauteur et le timbre. Il est aussi intéressant de noter la variabilité de la qualité vocale apparaissant dans le discours du patient (l'émotivité, les contextes favorisant les troubles pathologiques, ...), ainsi que la présence ou non d'un comportement de forçage vocal, de troubles d'articulation et de débit de parole.

[5] Caméra Ultra Rapide (CUR)

Cet examen étudie le comportement vocal dans les situations suivantes :
- épreuve de « comptage projeté » pour mettre en évidence le forçage vocal ;
- voix « d'appel » pour faire découvrir au patient ses possibilités vocales ;
- voix « chantée » pour étudier la tessiture, la justesse, les altérations du timbre ;
- épreuve de « Temps Maximal de Phonation »[6] pour objectiver l'efficacité du souffle phonatoire ;
- ...

De plus, faire écouter au patient sa propre voix enregistrée durant les différentes épreuves, peut lui permettre de mieux comprendre le mécanisme de ses difficultés et des troubles qu'il présente.

[6]TMP qui permet d'apprécier la durée de vibration des plis vocaux (donc l'équilibre pneumo-phonatoire) lors de l'émission d'une note soutenue le plus longtemps possible

1.5 L'évaluation de la voix dysphonique

Le bilan vocal de personnes dysphoniques ne saurait être exhaustif si l'examen clinique n'était pas complété par une évaluation de la qualité vocale. Cette évaluation reste encore aujourd'hui un sujet sensible au centre de nombreuses études dans des domaines multi-disciplinaires. Deux principales approches peuvent être considérées : l'analyse perceptive et l'analyse objective.

La première approche décrite dans cette section concerne le jugement « à l'oreille »[7] de la voix et de la parole qui est l'une des méthodes la plus utilisée en milieu clinique. Elle permet de déterminer si une voix est pathologique ou non et de mesurer le degré du dysfonctionnement vocal. Or, de part sa subjectivité intrinsèque et sa grande variabilité (intra-auditeur et inter-auditeurs), le jugement perceptif reste une méthode controversée. A travers cette section, l'évaluation perceptive est présentée, de la définition d'une terminologie adéquate à l'élaboration d'échelles d'évaluation. Ces deux étapes ont été nécessaires pour la mise en œuvre de protocoles standardisés d'analyse perceptive de la voix permettant d'atteindre une fiabilité raisonnable.

Afin de pallier les « faiblesses » du jugement auditif, l'analyse objective est proposée comme complémentaire à l'analyse perceptive. Ces techniques dites instrumentales se subdivisent en deux catégories. La première catégorie repose sur l'acquisition de mesures physiques par le biais de capteurs et applique des techniques statistiques (analyse discriminante, corrélation, régression, ...) sur les mesures extraites pour établir des résultats. La deuxième catégorie repose sur les techniques utilisées en Traitement Automatique de la Parole (TAP). Cette dernière a la particularité de s'appuyer uniquement sur une analyse acoustique du signal de parole permettant une facilité et simplicité d'utilisation. Ces deux techniques instrumentales, nommées respectivement « analytique » et « automatique », sont décrites dans cette section à la suite de l'analyse perceptive.

Enfin, la section se terminera sur la présentation de la méthode expérimentale « Phonetic Labeling ». Son principe repose sur une évaluation perceptive de chaque phonème d'un texte afin d'étudier l'influence des contraintes phonétiques et linguistiques dans le contexte de la voix dysphonique.

1.5.1 L'analyse perceptive et les méthodes d'évaluation

De part son essence perceptive, la voix est indissociable de l'oreille. Cela légitime le fait que la perception auditive soit l'outil premier pour l'évaluation du trouble vocal, de résultats post-opératoires ou rééducatifs. Dans la pratique quotidienne, les cliniciens utilisent « *l'écoute à l'oreille* » pour décrire la qualité de la voix à l'aide d'une terminologie établie sur une impression auditive. Même si des règles ont été instaurées (notamment

[7] appelée aussi analyse perceptive ou jugement perceptif

sur l'âge, le sexe ou la nature de voix), l'évaluation perceptive n'est pas sujet à consensus en ce qui concerne la définition des termes perceptifs décrivant la qualité de la voix. Selon les auditeurs, l'éraillement peut aussi bien désigner une voix rauque que pathologique. Il en est de même pour apprécier les nuances, parfois subtiles, entre deux termes très proches comme entre une voix *éraillée* et une voix *rocailleuse*.

Hammarberg et al. (1980) ont analysé les corrélations entre 28 variables sélectionnées parmi un ensemble de 50 termes les plus couramment utilisés à l'hôpital de Huddinge (Suède) pour d'écrire la qualité de la voix. Le but de l'étude était de proposer une terminologie structurée et précise de l'impression auditive des dysphonies. Une analyse en composantes principales a été appliquée sur les résultats issus d'une évaluation perceptive réalisée par un jury de 14 auditeurs jugeant 20 voix dysphoniques suivant les 28 paramètres retenus. Le tableau 1.5 affiche les résultats de l'analyse factorielle dans laquelle 85.3 % de la variance totale est expliquée par 5 facteurs bipolaires.

Facteurs bipolaires	**Stable / Instable**	**Soufflé / Serré**	**Hyperfonctionnel / Hypofonctionnel**	**Rude / Léger**	**Registre de tête / Registre de poitrine**
Variance expliquée	30.0 %	27.3 %	13.5 %	10.1 %	4.4 %

TAB. *1.5* – *Résultats de l'analyse factorielle [Hammarberg et al., 1980]*

Palliant le manque de terminologie structurée, précise et adéquate du domaine de la qualité vocale, ces premières études sont à l'origine du développement de protocoles standardisés d'analyse perceptive de la voix.

Les échelles d'évaluation perceptive

Le but de l'échelle d'évaluation perceptive est de décrire la qualité d'une voix suivant des critères qualitatifs et quantitatifs. L'échelle de mesure constitue la composante quantitative de la méthode d'analyse perceptive. Elle permet d'évaluer la proportion de paramètres perceptifs contenue dans la voix analysée. Ces derniers constituent la composante qualitative de l'échelle d'évaluation. Le choix de la mesure et des termes utilisés pour construire l'échelle est d'une grande importance et dépend des objectifs de l'évaluation de la qualité vocale que l'on souhaite atteindre.

Les trois principales échelles quantitatives sont :

- *l'échelle bipolaire* sémantique repose sur le principe du « oui/non » (ou encore du « présence/absence ») de deux paramètres de qualité de voix comme par exemple, les facteurs bipolaires affichés dans le tableau 1.5 ;
- *l'échelle de classe ou EAI (Equal-Appearing Interval)* propose plusieurs niveaux équidistants, habituellement de 4 ou 7 niveaux, pour quantifier séparément les paramètres qualitatifs. Pour Kreiman et al. (1993), une échelle à 7 niveaux semblerait plus appropriée pour mesurer la qualité pathologique qu'une échelle à 4 niveaux qui tendrait à réduire la sensibilité de discrimination des auditeurs ;

- *l'échelle analogique visuelle (EAV)* permet d'attribuer visuellement un degré de sévérité en traçant une croix sur une ligne, généralement de 10 cm de longueur ; la normalité se situant le plus à gauche de la ligne alors que la sévérité la plus forte se localisant le plus à droite. La distance mesurée en millimètres de la gauche à la croix traçée par l'auditeur correspond au degré de sévérité estimé.

Plusieurs méthodes d'évaluation perceptive [Osgood et al., 1957; Voiers, 1964; Holmgren, 1967; Isshiki et al., 1969; Hammarberg et al., 1980; Dejonckere et al., 1993] ont inspiré la plupart des échelles que l'on peut rencontrer dans la littérature. Ces dernières se proposent d'évaluer la qualité de la voix suivant des critères qualitatifs et quantitatifs.

Les échelles perceptives du tableau 1.6 sont présentées dans cette section.

	VPAS	**GRBAS**	**Hammarberg Scheme**	**BVP**
Auteurs	Laver (1980)	Hirano (1981)	Hammarberg (1986)	Wilson (1987)
Critères qualitatifs	17	5	12	12
Régions explorées	Laryngée Supralaryngée	Laryngée	Laryngée	Laryngée Supralaryngée Comportement vocal
Critères quantitatifs	6	4	5	5
Corpus	Parole spontanée Lecture	Parole spontanée	Lecture	Parole spontanée Comptage Voyelle tenue

TAB. *1.6* – *Comparaison de quatre échelles d'évaluation perceptive [De Bodt et al., 1996]*

• **L'échelle Vocal Profile Analysis Scheme (VPAS)** de Laver (1980) [Laver et al., 1985; De Bodt et al., 1996]

L'analyse perceptive par VPAS nécessite une connaissance approfondie du modèle descriptif de la qualité vocale proposé par Laver (1980). Sa particularité est de décrire les caractéristiques phonétiques des qualités de la voix. Chaque caractéristique est évaluée par rapport à une position articulatoire neutre définie par la physioacoustique.

Elle se propose d'analyser les caractéristiques vocales laryngées et supralaryngées (conduit vocal) à l'aide de 17 paramètres répartis en 3 catégories :

1. la qualité vocale : les qualités laryngées et supralaryngée des sons ;
2. la qualité prosodique : la hauteur et l'intensité ;
3. l'organisation temporelle : le rythme respiratoire, la continuité, le débit, la cadence.

Les paramètres sont à évaluer sur de la lecture ou de la parole spontanée selon une cotation en 6 niveaux.

Le VPAS nécessite une formation sur cassettes audio avec laquelle un entraînement de 12 heures permettrait d'atteindre un taux de concordance entre 65-75 % des jugements [Wirz & Beck, 1995]. De part sa complexité d'utilisation, le VPAS est une échelle dont la mise en œuvre en milieu clinique est difficile.

• **L'échelle GRBAS** de Hirano (1981) [De Bodt et al., 1997]

Inspirée de Isshiki et al. (1969), l'échelle GRBAS de Hirano (1981) est une échelle compacte et simple d'utilisation, composée de 5 paramètres décrits dans le tableau 1.7.

Initale	Terme	Signification	Définition
G	Grade	Grade global de dysphonie	Impression globale du degré d'anormalité de la voix
R	Rough	Raucité	Impression d'irrégularité des vibrations des cordes vocales qui correspond aux fluctuations irrégulières de la Fo et/ou à l'amplitude du son glottique
B	Breathy	Caractère soufflé	Impression d'une fuite d'air assez importante à travers les cordes vocales, relative à des turbulences
A	Asthenic	Asthénie	Manque de puissance de la voix relatif à une intensité faible du son et/ou un manque des harmoniques élevés
S	Strained	Forçage	Impression d'un état hyperfonctionnel de phonation relatif à une Fo anormalement haute (bruit dans les hautes fréquences et/ou richesse en harmoniques dans les hautes fréquences)

TAB. *1.7* – *Définition des paramètres de l'échelle GRBAS*

Actuellement, c'est la méthode la plus utilisée en pratique clinique quotidienne [Wirz & Beck, 1995]. Elle permet d'évaluer la fonction laryngée d'un patient sur de la lecture ou de la parole spontanée, suivant 5 paramètres qualitatifs à quantifier selon 4 niveaux de sévérité de la dysphonie :

0 = voix normale	1 = altération légère	2 = altération moyenne	3 = altération sévère

Cette méthode d'évaluation ne prend pas en compte la fonction supralaryngée, la hauteur tonale et la sonie[8] du patient. Il faut noter que les paramètres de grade global (G), de raucité (R) et de souffle (B) sont considérés comme étant les plus fiables et moins soumis à la variabilité d'un jury d'écoute que les paramètres d'asthénie (A) et de forçage (S) [Millet & Dejonckere, 1998].

• **L'échelle Hammarberg Scheme** de Hammarberg (1986)

Hammarberg Scheme est une échelle d'évaluation perceptive à 12 paramètres décrits dans le tableau 1.8 qui sont évalués sur 5 niveaux de sévérité (de 0 = normal à 4 = très sévère).

[8] mesure l'impression subjective d'intensité percue chez l'être humain

Paramètres de qualité vocale	Définition
Aphonie / Aphonie intermittente	Désonorisations intermittentes, voix parfois chuchotée
Voix Soufflée	Fuite d'air glottique audible à cause d'une fermeture glottique insuffisante
Hyperfonctionnel / Tendu	Voix serrée, impression de compression des cordes vocales en phonation
Hypofonctionnel / Détendu	tension insuffisante des cordes vocales, voix faible et lâche
Voix grinçante (fry)	Série rapide de coups ; comme à l'ouverture d'une porte rouillée Vibration périodique de basses fréquences
Voix rauque	Voix grave avec vibration irrégulière des cordes vocales
Voix éraillée	Voix aiguë avec vibration irrégulière des cordes vocales
Cassures vocales	Ruptures intermittentes de fréquence
Diplophonie / Bitonalité	Deux fréquences différentes perçues simultanément
Registre	Mode de vibration des cordes vocales : lourd (poitrine) et léger (tête)
Hauteur	Perception auditive de la fréquence fondamentale
Sonie (intensité)	Perception auditive de l'intensité acoustique

TAB. *1.8* – *The Hammarberg Scheme [Hammarberg, 1986]*

Le matériau phonétique est la lecture d'un texte d'une durée de 40 secondes environ. Cette échelle d'évaluation n'est utilisée qu'en Suède malgré les tentatives et efforts de standardisation à d'autres pays.

• **L'échelle Buffalo Voice Profile Systeme (BVP)** de Wilson (1987) [De Bodt et al., 1996]

Le système BVP est une méthode d'évaluation à 12 paramètres (tableau 1.9) qui vise à analyser les caractéristiques vocales laryngées et supralaryngées, ainsi que le comportement vocal :

le timbre laryngé	la hauteur	l'intensité
la résonance orale	la résonance nasale	le volume respiratoire
le tonus musculaire	le forçage	le débit
l'anxiété perçue dans la voix	l'intelligibilité	le grade global de la dysphonie

TAB. *1.9* – *Les 12 paramètres du système Buffalo Voice Profile (BVP) de Wilson (1987)*

Le jury évalue les paramètres selon 5 niveaux de sévérité :

0 = normal	1 = peu sévère	2 = moyen	3 = sévère	4 = très sévère

Le matériau phonétique comprend la lecture, la parole spontanée, la voyelle tenue et une épreuve de comptage. L'évaluation d'une voix dysphonique avec BPV informe sur la sévérité du dysfonctionnement vocal, son impact sur la communication et sur l'intervention à envisager.

Le jury d'écoute

Pour un usage clinique quotidien, une échelle d'évaluation doit être simple, reproductible et pratique. Néanmoins, son utilisation n'est pas sans comporter une part d'imprécision. En effet, chaque spécialiste définit à l'usage (ou inconsciemment) ses propres critères subjectifs. La constitution de jury d'écoute tente de réduire cette imprécision avec une prise de décision par consensus lors de séances d'évaluation vocale. Constitué de plusieurs auditeurs, un jury d'écoute évalue à l'oreille et caractérise le degré de dysphonie d'une voix. Son but est d'obtenir un jugement fiable et reproductible sur la qualité vocale de chaque échantillon présenté.

En ce sens, la constitution du jury d'écoute doit prendre en compte plusieurs éléments tels que l'expérience des auditeurs en terme d'évaluation de la voix pathologique, leurs stratégies d'écoute ou la cohérence du groupe. En effet, un auditeur expérimenté peut ressentir des sensations kinesthétiques (ou physiques) à l'écoute d'une voix dysphonique [Moses, 1954], ce qui n'est pas forcément le cas pour un auditeur naïf. De même, sa capacité d'analyser une voix en discernant les mécanismes de production et le fonctionnement vocal est une élément important afin d'évaluer l'efficacité dans la communication. Enfin, l'origine socio-culturelle, la profession (phoniatre, médecin ORL, orthophoniste, ...), le secteur d'activité (hôpital, cabinet médical, libéral, ...), ..., constituent des informations pertinentes pour obtenir une structure de groupe homogène et harmonieuse

Généralement, un jury est considéré comme fiable si ses évaluations sont reproductibles. Mais alors, suffirait-il de multiplier l'effectif du jury pour augmenter sa fiabilité ? Ou alors, existe-il un nombre optimal de membres garantissant la fiabilité d'un jury ? Plusieurs travaux ont porté sur la problématique du recrutement des auditeurs de jury d'écoute [Kreiman et al., 1992; Gerratt et al., 1993; Wolfe et al., 2000] sans vraiment répondre consensuellement à ces interrogations. Des études contradictoires ont été publiées comme par exemple, Anders et al. (1988) pour qui, l'expérience des auditeurs peut influencer la fiabilité du jury d'écoute, ce qui est infirmé par De Bodt et al. (1997) en situation de « test-retest ».

Ainsi, il semble que la constitution de jury d'écoute n'est pas aussi triviale que cela puisse paraître et que, même si elle améliore la fiabilité de l'analyse perceptive individuelle, cela nécessite la mise en place de certaines règles permettant de s'affranchir d'un certain nombre de biais. Pour cela, d'autres facteurs sont à prendre en considération comme les références auditives du jury, sa formation, son entraînement, le déroulement des séances d'écoute, le nombre de sessions d'écoute, la présentation des stimuli sonores, ... , afin de définir des protocoles expérimentaux robustes et fiables. Le lecteur pourra se référer à [Revis, 2004] pour obtenir plus de détails sur le sujet.

D'une manière générale, il est admis que l'analyse perceptive doit être menée par plusieurs auditeurs experts et durant plusieurs sessions d'écoute pour qu'elle puisse atteindre une fiabilité raisonnable. La mise en place d'un tel protocole peut s'avérer fina-

lement contraignante et difficilement applicable. En effet, les inconvénients majeurs de cette approche restent le coût humain non négligeable (réunions périodiques de plusieurs experts, durée des séances d'écoute, disponibilité simultanée des experts, ...) et le manque de fiabilité dû à différents facteurs liés aux caractéristiques intrinsèques du jugement perceptif :

◦ **Subjectivité** : le caractère «*agréable/désagréable*» est une sensation subjective par laquelle aucun objet n'est représenté. En particulier, le caractère esthétique ou dysharmonieux de la voix pathologique dépend du goût de chacun.

◦ **Variabilité intra-individuelle** : un même individu peut ressentir ou interpréter les objets différemment en fonction de son état psychologique (de son humeur, de son état d'esprit, de son stress, de sa disponibilité, ...) ou à des moments différents. Cette variabilité se retrouve lors de l'analyse perceptive de dysphonie vocale, lorsqu'un auditeur du jury n'attribue pas la même note à un même stimulus d'une séance à l'autre.

◦ **Variabilité inter-individuelle** : elle est due essentiellement à la connaissance et/ou méconnaissance du phénomène dysphonique, à l'expérience des auditeurs. Cette variabilité est d'autant plus importante si le jury d'écoute est de culture et d'écoles cliniques différentes.

Le matériau phonétique

Le matériau phonétique est le type de fragment de parole qui sera utilisé par un jury d'écoute et par les méthodes objectives, comme support à l'évaluation de la qualité de la voix et de la parole. Le choix du matériel sonore est important de par les intérêts et les limites qui les caractérisent.

Actuellement, les matériaux phonétiques les plus utilisés sont la voyelle tenue et la parole (spontanée, chantée, lue, mots répétés, syllabes répétées, nombres énumérés, ...).

• La voyelle tenue

Comme matériau phonétique, la voyelle la plus utilisée est le /a/ sachant qu'une autre voyelle peut être préférée comme /i/ ou /ou/. Son utilisation présente comme intérêt :

- facile à produire dans la pratique clinique ;
- peu sujette à l'émotion du patient ;
- peu affectée par les accents régionaux, les contextes prosodiques ;
- ne contient pas d'événement articulatoire (transitions « CV » ou « Sourde/Sonore ») ce qui permet une meilleure concentration de l'auditeur sur les caractéristiques laryngées [De Krom, 1994] ;
- renseigne sur le fonctionnement du vibrateur dans la mesure où il s'agit d'une tâche de stabilité dans laquelle toute instabilité est synonyme de dysfonctionnement.

Pour De Krom (1995), les portions transitoires « attaque/finale » de la voyelle tenue (lieux de « démarrage/arrêt » du voisement) contiendraient une part importante de l'information sur la raucité, et devraient donc être prises en compte durant l'analyse perceptive. Cette observation intéresse principalement les voix normales, car pour les dysphonies sévères, les altérations vocales se manifestent sur la totalité de la voyelle c-à-d les portions transitoires et stables.

Néanmoins, les voix dysphoniques éprouvent plus de difficulté dans la mise en action des cordes vocales que les voix normales. Les portions transitoires s'allongent et deviennent d'autant plus instables que la mise en vibration est perturbée. Dans ce sens, Giovanni et al. (1995) propose d'inclure la mesure correspondant au temps de stabilisation de la vibration pour l'évaluation des dysphonies. En pratique, la plupart des méthodes objectives sur la voix pathologique ne s'intéresse qu'à la partie stable de la voyelle tenue *i.e.* la portion du signal où les cordes vocales atteignent un régime vibratoire permanent. En effet, les portions transitoires instables rendent impossible l'estimation d'indices de fluctuation.

De plus, ce support phonétique reste controversé dans la littérature car il tend à sous-estimer la dysphonie. Un patient dysphonique peut compenser ses déficits vocaux et produire la voyelle avec une émission satisfaisante. Par ailleurs, [Revis et al., 2002] a montré que certains phénomènes vocaux issus de la parole spontanée comme l'attaque, sont pertinents pour l'évaluation des dysphonies. Finalement, ce matériau phonétique ne représente pas l'expression de la communication au quotidien, à la différence de la parole.

• La parole

La parole est le matériau phonétique le plus proche de la voix conversationnelle sur laquelle le médecin se forge une première impression durant l'examen initial avec le patient, l'interrogatoire. A la différence de la voyelle tenue, la parole rend compte des phénomènes phonémiques (coarticulation, transitions formantiques, attaques/sorties, interruptions vibratoires, ...) et prosodiques (rythme, intonation, mélodie, ...). De plus, la production des phonèmes voisés et la dynamique de la parole impliquent la participation de la musculature laryngée induisant les passages successifs et alternés entre les modes glottiques d'adduction et d'abduction. Il est difficile pour un sujet dysphonique d'être tenté de compenser ses déficits vocaux dans une situation aussi proche de la phonation normale. Pour Hammarberg (2000) : « *les variations survenant dans la parole, comme l'attaque vocale, l'arrêt vibratoire, les cassures de la phonation, ..., sont des éléments cruciaux de la qualité de la voix* ». C'est la raison pour laquelle la parole semble être le matériau privilégié pour l'analyse perceptive. La lecture d'un texte permet de disposer d'un matériel sonore « phonétiquement équilibré »[9] facilitant la constitution de corpus homogène ainsi que la standardisation des protocoles d'évaluation.

[9] censé contenir l'ensemble des phonèmes de la langue dans toutes les combinaisons possibles

1.5.2 L'analyse objective et les méthodes instrumentales

L'analyse objective est la deuxième méthodologie proposée comme une alternative à l'évaluation perceptive pour pallier ses inconvénients et faiblesses précédemment décrits. Depuis plusieurs années, une multitude de travaux se sont consacrés à la détection et classification de la voix pathologique au moyen d'analyses acoustiques et paramétriques, de méthodes statistiques et de techniques de modélisation issues des domaines de la reconnaissance des formes et du traitement automatique de la parole [Kasuya et al., 1986; Gavidia-Ceballos & Hansen, 1996; Ritchings et al., 2002; Maguire et al., 2003; Alonso et al., 2005; Fredouille et al., 2005; Saenz-Lechon et al., 2006; Yu et al., 2007].

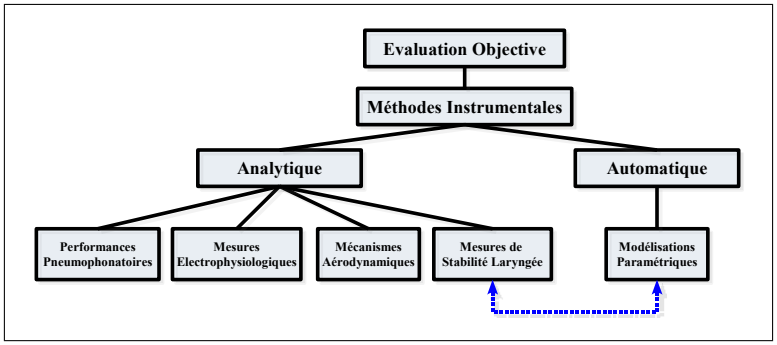


FIG. *1.8* – *Les méthodes instrumentales pour l'évaluation objective de la voix dysphonique.*

Les méthodes objectives instrumentales peuvent être divisées en deux catégories d'approches présentées sous les appellations « analytique » et « automatique » (figure 1.8).

Les méthodes instrumentales dites « analytiques »

De nombreux travaux [Teston & Galindo, 1995; Piccirillo et al., 1998b] ont porté sur l'élaboration d'outils d'évaluation objective standardisés et applicables aux voix dysphoniques. Ces derniers reposent sur l'acquisition de mesures acoustiques, aérodynamiques et/ou électrophysiologiques par le biais de capteurs. Pour la plupart des auteurs, l'élaboration de méthodes objectives multiparamétriques se justifie par le caractère multidimensionnel [10] de la production vocal [Wuyts et al., 2000]. Leur but est de qualifier et surtout de quantifier les troubles vocaux à partir de mesures multiparamétriques. Dans ce contexte, les mesures sont associées au système qui applique sur ces dernières une technique statistique (analyse dyscriminante, étude de corrélation, régression linéaire, ...) pour fournir une décision. Ces méthodes sont appelées « analytiques » dans le sens où les mesures estimées - acoustique, aérodynamique ou physiologique - apportent un ensemble restreint d'indices susceptibles d'être caractéristiques de dysfonctionnements vocaux. La dimension « multiparamétrique » de la méthode permet la constitution d'un ensemble d'indices facilitant ainsi la prise de décision. On peut distinguer différentes classes d'indices.

[10] comme la raucité, le souffle ou le chevrotement

• Les performances pneumophonatoires

Les performances pneumophonatoires donnent une indication générale sur l'étendue vocale (le **Phonétogramme** ou Voice Range Profile) et sur l'endurance phonatoire (le **Temps Maximal de Phonation (TMP)**) d'un sujet en phonation. Ces performances s'évaluent avec des mesures acoustiques, aérodynamiques ou physiologiques selon la technique utilisée.

Le **Phonétogramme** (ou aire dynamique vocale) est une figure représentant l'étendue vocale dans un plan cartésien sous la forme d'une aire en forme de patatoïde avec les fréquences en abscisse et les intensités en ordonnée. Il permet de caractériser les faiblesses de la fonction vocale et les performances phonatoires puisque les notes sont émises aux intensités les plus basses et les plus fortes possibles. Le phonétogramme est un outil d'aide au diagnostic des troubles de la voix [Schutte & Seidner, 1983].

Le **TMP** consiste à mesurer le temps de tenue maximale d'une émission vocalique (généralement le/a/) à une hauteur et intensité donnée. L'évaluation de la **Capacité Vitale**[11] pulmonaire divisée par le TMP représente une mesure de rendement vocal, le **Quotient Phonatoire (QP)** [Hirano et al., 1968]. Le **QP** augmente avec le degré de désordre vocal du sujet dysphonique. Une valeur élevée supposerait une fuite glottique pouvant avoir comme origine un mauvais accolement des cordes vocales.

• Les mesures de stabilité laryngée

Majoritairement de nature acoustique, ces évaluations s'intéressent à calculer des indices spécifiques à partir des signaux de parole supposés contenir de l'information sur les dysfonctionnements vocaux. Parmi le grand nombre d'indices acoustiques décrits dans la littérature (une cinquantaine dans [Hirano, 1989]), nous présenterons ici les principaux indices liés à la stabilité vibratoire de la glotte et au souffle de la voix. Sachant que l'instabilité vibratoire de la glotte est une cause essentielle des dysphonies et que la sévérité du trouble vocal peut dépendre de l'importance de la fuite glottique durant la phonation, ces deux catégories d'indices acoustiques (acoustiques et aérodynamiques) constituent des élements très importants pour l'évaluation de la dysphonie.

En acoustique, la fréquence fondamentale (**F0**) est la vitesse de vibration des cordes vocales. Elle correspond à la hauteur tonale qui se modifie généralement pour les personnes atteintes d'une pathologie vocale. Isolée, cette mesure est peu informative. Cependant, la plupart des indices utilisés pour l'évaluation de la dynamique ou de l'instabilité laryngée, dérive de cette mesure comme l'**écart type de la F0**, le **jitter**, ...

Les mesures de stabilité laryngée peuvent être définies selon la durée des fluctuations ou le souffle vocal.

[11] elle représente le volume d'air maximal mesuré sur une expiration à l'aide d'un spiromètre

les fluctuations à court terme

Elles correspondent aux variations d'une durée d'un cycle glottique et caractérisent surtout les dysphonies avec lésions des cordes vocales.

L'instabilité à court terme de la **F0** se traduit par des variations de durée entre chaque cycle d'oscillation et se calcule par le **jitter** avec ses différentes déclinaisons : le **jitter absolu moyen**, le **jitter factor**, le **jitter ratio** et **Relative Average Perturbation (RAP)**. On précisera que le **jitter factor** est un bon indicateur de la stabilité du vibrateur laryngé, assez bien corrélé avec la raucité.

L'instabilité à court terme de l'amplitude se traduit par des variations d'amplitude entre chaque cycle d'oscillation et se calcule par le **shimmer** avec ses différentes déclinaisons : le **shimmer moyen**, le **shimmer factor** et **Amplitude Perturbation Quotient (APQ)**.

L'instabilité du signal glottique se manifeste par la présence d'une composante bruitée qui lui est superposée. L'apériodicité du signal vocal peut être mesurée par plusieurs méthodes comme la méthode **Harmonic Noise Ratio (HNR)** proposée par Yumoto & Gould (1982) ou la méthode **Normalized Noise Energy (NNE)** proposée par Kasuya et al. (1986).

les fluctuations à moyen terme

Elles correspondent aux variations d'une durée maximale de l'ordre d'un phonème et caractérisent généralement les atteintes neurologiques (tremor).

L'instabilité à moyen terme de la **F0** se traduit par des variations de fréquence au cours du temps et se mesure par l'**écart type de la F0**. La **F0 moyenne** fournit une mesure globale de la hauteur tonale perçue (aiguë, grave, ...). Pour les sujets dysphoniques, la **F0 moyenne** des femmes diminue globalement avec le degré de sévérité du trouble vocal, alors que pour les hommes, cette tendance s'inverse. Le **coefficient de variation de la F0** est un bon indice d'évaluation globale de la dynamique et de l'instabilité laryngée à moyen terme. Sur du matériau phonétique comme la voyelle tenue, sa pertinence est d'autant plus importante si la voix est perçue avec du vibrato ou du chevrotement d'origine neurologique.

Dans le cadre de la dysphonie, la mesure du **jitter** a tendance à considérer certains aspects du signal vocal (la bitonalité par exemple) comme une instabilité sans les différencier des instabilités « vraies » *i.e.* fluctuations aléatoires. Contre cela, Giovanni et al. (1999) proposent d'utiliser une mesure issue de la dynamique non-linéaire, le **coefficient de Lyapounov (CL)**, et plus particulièrement de calculer le **Plus Grand Exposant de Lyapounov (PGEL)** pour évaluer la non linéarité du signal glottique. Une grande valeur du **PGEL** traduit un comportement chaotique du vibrateur laryngé. Les travaux de [Yu et al., 2000] ont montré la pertinence du **CL** par rapport au **jitter** pour la discrimination des grades perceptifs GRBAS.

le souffle de la voix

Le souffle de la voix se définit comme un bruit venant se superposer au signal vocal de la source laryngée. Ce bruit « additif » peut avoir comme origine une constriction du conduit vocal provoquant un écoulement aérodynamique ou un défaut d'accolement des cordes vocales créant alors un débit d'air trop important. Sa prise en compte pour l'évaluation de la dysphonie est importante.

Précédemment, nous avons mentionné que l'instabilité du signal glottique se manifestait comme un bruit « additif ». Ainsi, pour obtenir une bonne évaluation du bruit de souffle avec le **HNR** ou le **NNE**, il est nécessaire de minimiser les effets de l'instabilité vibratoire (**jitter**) [Qi et al., 1995].

La mesure du **Signal Ratio (SR)** [Hiraoka et al., 1984] devient particulièrement pertinente pour la dysphonie si son calcul est effectué, non pas sur l'ensemble du spectre, mais uniquement sur des zones de fréquences supérieures à 1 Khz. En effet, sa valeur devient alors très faible pour des sujets dysphoniques dont la composition spectrale est appauvrie dans les aigus (pauvreté harmonique et présence de bruit).

• **Les mécanismes aérodynamiques**

On dénombre quatre paramètres aérodynamiques : la **pression sous glottique**, la **Pression Intra-Orale (PIO)**, le **débit d'air oral** et le **débit d'air nasal**. Dans le cadre de l'évaluation des dysphonies, seul le **débit d'air oral** et la **PIO** sont utilisés.

Les **débits d'air oral/nasal** sont mesurés à l'aide de capteurs de débit adaptés à la prise de mesures en phonation. La **PIO** est mesurée avec un capteur de pression placé dans la bouche. En ce qui concerne la **pression sous glottique**, Smitheran & Hixon (1981) ont mis au point une méthode non invasive permettant de l'estimer de façon indirecte par déduction à partir de la **PIO** (la **Pression Sous Glottique Estimée (PSGE)**). Pour [Giovanni et al., 2000], la **PSGE** est fortement corrélée au forçage vocal.

Permettant une bonne évaluation du rendement laryngien, le **débit d'air oral** permet d'estimer également la **fuite glottique** durant la phonation par le calcul d'un indice d'estimation de la fuite d'air à travers la glotte.

Plusieurs indices sont déduits de ces mesures permettant d'évaluer le **rendement glottique**, l'**efficacité glottique** et la **résistance glottique**. Il faut préciser que la **résistance glottique** est l'indice qui paraît le plus corrélé avec la force de contact des cordes vocales, induisant indirectement une relation avec le forçage vocal.

• **Les mesures électrophysiologiques**

L'**électroglottographie (EGG)** est une méthode d'exploration du fonctionnement laryngien. De façon non invasive, elle est basée sur la mesure de l'impédance électrique translaryngée. Elle permet d'étudier la périodicité du signal glottique de manière quantitative et qualitative. En effet, l'**EGG** rend compte des modifications de la surface d'accolement des plis vocaux au cours du cycle vibratoire, permettant ainsi l'étude du décours des phases d'adduction et d'abduction glottiques. L'**EGG** est aussi un excellent moyen d'observer la fréquence fondamentale (**F0**), pouvant être utilisée pour le calcul de l'instabilité vibratoire de cycle à cycle (**jitter**) [Dejonckere, 1996].

L'**électromyographie (EMG)** du larynx est également une méthode d'exploration de la fonction laryngée. Elle apporte des informations sur l'activité musculaire, notamment des muscles intrinsèques laryngés. L'**EMG** enregistre le potentiel d'action résultant à la fois des muscles adducteurs et abducteurs du larynx. Elle est nécessaire pour l'établissement du diagnostic différentiel des troubles de mobilité laryngée : paralysie, ankylose, myasthénie, ... Elle est utile aussi pour le pronostic des atteintes neurogènes périphériques du larynx et du pharynx.

Quelques exemples de travaux en évaluation instrumentale analytique

Plusieurs travaux se sont consacrés à confronter les résultats de l'évaluation instrumentale analytique à l'évaluation perceptive. Les études jugées les plus significatives sont :

- Les travaux présentés dans [Wolfe et al., 1995] ont porté sur l'étude de 4 mesures acoustiques (F0 moyenne, jitter, shimmer, HNR) calculées sur un corpus de 20 sujets témoins et 60 patients dysphoniques (lésions par nodules, paralysie laryngée unilatérale, dysphonie dysfonctionnelle).
  A travers une analyse de régression, la combinaison des 4 paramètres a obtenu une corrélation de 0.56 entre l'évaluation perceptive et l'évaluation acoustique. L'étude a montré aussi que, parmi les 4 paramètres retenus, seul le shimmer présentait une corrélation significative de 0.54 avec l'évaluation perceptive, valeur très proche de celle obtenue avec la combinaison des 4 paramètres.

- Dans [Wuyts et al., 2000], l'étude repose sur un corpus de 68 sujets témoins et 319 sujets dysphoniques, évalués perceptivement selon le grade G de Hirano (1981).
  Les auteurs proposent un index de sévérité de la dysphonie (Dysphonia Severity Index, DSI) comme indice objectif et quantitatif de la qualité vocale d'un sujet dysphonique. Il se définit comme la combinaison pondérée[12] de 4 mesures acoustiques sélectionnées par une analyse multivariée parmi un ensemble de 13 paramètres : F0 la plus haute (Hz), intensité la plus basse (dB), TMP (s) et jitter (%).

[12] $DSI = (0.13 * TMP) + (0.0053 * F0_{haute}) - (0.26 * I_{basse}) - (1.18 * jitter) + 12.4$

A travers une analyse discriminante et la combinaison des 4 variables, une concordance de 56 % a été obtenue avec l'évaluation perceptive.

- Les travaux présentés dans [Piccirillo et al., 1998b,a] reposent sur les paramètres suivants : l'étendue vocale, le débit d'air oral, le temps maximal de phonation et la pression sous glottique. Ces 4 paramètres ont été sélectionnés pour leur pertinence parmi un ensemble de 14 mesures acoustiques, aérodynamiques et électrophysiologiques.
  Les mesures extraites sur un corpus de 97 sujets dysphoniques et 35 sujets témoins ont permis d'établir une corrélation significative entre les mesures sélectionnées et les grades de l'échelle GRBAS.

- Dans Yu et al. (2001), une analyse multiparamétrique des dysphonies a été réalisée sur un corpus masculin de 21 sujets témoins et 63 sujets dysphoniques, évalués par un jury d'experts selon le grade G de l'échelle GRBAS.
  Les 10 paramètres retenus pour cette étude sont décrits dans le tableau 1.10. Ils ont été mesurés à l'aide du système EVA[13] [Teston & Galindo, 1995] sur du /a/ tenu, à l'exception de la pression sous glottique qui a été estimée sur une série de /pa/.
  Une analyse discriminante a permis de détecter des corrélations entre les jugements perceptifs et différentes combinaisons de paramètres. En particulier, la combinaison (F0, signal sur bruit, CL, PSGE, étendue vocal, TMP) a atteint 86 % de concordance avec le jury d'experts.
  Une étude similaire a été menée dans Yu et al. (2002), à l'exception du corpus qui était cette fois constitué de 74 femmes réparties en 6 sujets témoins et 68 sujets dysphoniques. Les résultats ont montré une corrélation entre les jugements perceptifs et objectifs de :
  ◦ 64 % avec une échelle ordinale classique du grade G de Hirano (1981) ;
  ◦ 88 % avec une échelle visuelle analogique discrétisée avec une segmentation non linéaire.

1	Fréquence fondamentale (F0)	6	coefficients de Lyapounov (CL)
2	jitter	7	débit d'air oral
3	intensité	8	pression sous glottique estimée (PSGE)
4	rapport signal sur bruit	9	étendue vocale
5	rapport signal sur bruit (f > 1kHz)	10	temps maximal de phonation (TMP)

TAB. *1.10* – *Les 10 paramètres utilisés dans [Yu et al., 2001, 2002]*

[13] pour Evaluation Vocale Assistée

En résumé, l'approche instrumentale analytique offre des résultats très acceptables mais encore insuffisants pour les praticiens pour être considérée comme un outil de diagnostic. A ce jour, elle ne peut être utilisée en pratique clinique qu'à titre indicatif et/ou expérimental. Les principales limites et contraintes de cette méthode sont :

[1] La majeure partie des analyses objectives repose sur l'acquisition des mesures sur des voyelles tenues (généralement le /a/), matériau phonétique qui reste controversé dans la littérature [Revis et al., 1999; Parsa & Jamieson, 2001] par sa tendance à sous-estimer la dysphonie. Il constitue un contexte d'élocution très éloigné de la parole continue, ne permettant pas de prendre en compte par exemple, les phénomènes vocaux de l'attaque reconnus comme pertinents dans l'évaluation des dysphonies.

[2] L'analyse objective des données repose généralement sur des méthodes statistiques (analyse discriminante, analyse de régression, corrélation, ...) pouvant fournir des résultats dépendants fortement des patients observés en termes de quantité et de qualité.

[3] L'acquisition de certaines mesures nécessite d'utiliser des équipements spécifiques pouvant se révéler onéreux. Cet aspect financier peut constituer une entrave au déploiement de ces méthodes instrumentales pour un usage journalier en milieu clinique.

Les méthodes instrumentales dites « automatiques »

Parallèlement, une deuxième approche d'évaluation objective a été proposée dans la littérature, les méthodes instrumentales dites « automatiques ». Elles reposent sur une analyse automatique de la parole pour la tâche d'évaluation de la dysphonie. En ce sens, l'intérêt s'est plus particulièrement porté sur les techniques utilisées en Traitement Automatique de la Parole (TAP) et à leur adaptation à la reconnaissance de la voix pathologique.

Les principales thématiques du domaine du Traitement Automatique de la Parole dont les techniques associées sont utilisées dans un cadre pathologique, sont :

- la Reconnaissance Automatique de la Parole (RAP) qui consiste en l'étude du contenu linguistique d'un énoncé observé [Haton et al., 2006] ;
- la Reconnaissance Automatique du Locuteur (RAL) qui consiste à reconnaître l'identité d'une personne par analyse de sa voix [Bimbot et al., 2004] ;
- l'Identification Automatique de la Langue (IAL) qui consiste à déterminer la langue parlée [Lamel & Gauvain, 1994] ou les accents régionaux [Ferragne & Pellegrino, 2006] à partir d'un échantillon de parole.

Ces différentes technologies ont prouvé leur pertinence dans l'extraction des informations, linguistiques et extra-linguistiques, véhiculées par la parole et la voix. Ainsi, hormis le message linguistique porté par un signal de parole, d'autres informations sur les spécificités d'un individu peuvent en être extraites telles que son identité, son émotivité, son état pathologique (rhume, rhinolalie, ...) ou ses particularités régionales.

La première composante d'un système de reconnaissance automatique est l'extraction des caractéristiques d'un signal de parole. Cette étape de paramétrisation consiste à transformer le signal acoustique en une séquence de vecteurs de paramètres afin d'en obtenir une représentation simplifiée nécessaire avant les phases d'apprentissage et de test. Cette phase permet de réduire la redondance du signal de parole et d'en extraire les informations pertinentes en vue de la reconnaissance. Différentes représentations paramétriques sont proposées. Les plus couramment utilisées sont issues de l'analyse en banc de filtres avec les coefficients :

- spectraux : LFSC et MFSC (Linear/Mel Frequency Spectral Coefficients) ;
- cepstraux : LFCC et MFCC (Linear/Mel Frequency Cepstral Coefficients).

Le lecteur pourra se référer à la section 3.1 pour plus de détails sur le sujet.

Cependant, dans la mesure où la dysphonie apparaît exclusivement sur la vibration glottique, des mesures acoustiques - caractéristiques de la stabilité laryngée comme F0, jitter, shimmer, HNR - peuvent être estimées sur des trames successives de parole (à court ou moyen terme) afin d'adopter une représentation de séquences de vecteurs, nécessaire pour les phases de modélisation et de test [Wester, 1998; Dibazar et al., 2002].

Différentes techniques de modélisation statistique sont proposées dans la littérature. Ici, seules les principales modélisations paramétriques déjà utilisées dans le cadre de la voix pathologique, sont présentées brièvement.

• les modèles de Markov cachés

Dans les systèmes de RAP, la technique de modélisation acoustique la plus utilisée est le modèle de Markov caché (HMM pour Hidden Markov Model [Rabiner, 1989]). Le principe est de modéliser chacune des sous-unités lexicales (généralement des phonèmes) d'un mot par un HMM. Une chaîne de Markov cachée est un automate à N états caractérisés chacun par une fonction de densité de probabilité. Couramment, ces fonctions sont des modèles de mélange de gaussiennes (GMM pour Gaussian Mixture Model [Reynolds, 1992]). Une matrice de transitions regroupe les probabilités de passage d'un état i vers un état j pour l'ensemble des liaisons inter-états du HMM. Un HMM est caractérisé par l'ensemble des paramètres $\lambda = \{ S , M , A , B , \pi \}$:

1. S = l'ensemble des N états du modèle ;
2. M = le nombre fini de symboles d'observation ;
3. A = la matrice des transitions inter-états ;
4. B = l'ensemble des probabilités d'émission de l'observation o dans les états[14] ;
5. π = distribution de la probabilité de l'état initial.

L'apprentissage des paramètres du modèle HMM consiste à estimer les transitions inter-états et les probabilités d'émissions des états. L'estimation des paramètres λ repose classiquement sur la technique visant à maximiser la vraisemblance[15].

[14] ou ensemble des fonctions de densité de probabilité associées à chaque état

[15] selon le critère du maximum de vraisemblance nommé aussi Maximum Likelihood Estimation (MLE)

• les modèles de mélange de gaussiennes

Les modèles de mélange de gaussiennes (GMM) peuvent être considérés comme des modèles de Markov cachés (HMM) à un seul état où la fonction de densité est un mélange de gaussiennes. Ils sont classiquement utilisés en RAL [Reynolds, 1992]. Pour plus de détails, le lecteur pourra se référer à la section 3.2.

• les machines à vecteurs supports

Les machines à vecteurs supports (SVM pour Support Vector Machines [Vapnik, 1998]) sont des techniques discriminantes dans la théorie de l'apprentissage statistique. L'approche SVM a comme but de séparer un ensemble d'individus en plusieurs classes en maximisant la largeur des marges inter-classes *i.e.* en construisant des frontières de décision optimales (hyperplans à marge maximale).

Cependant, comme la plupart des classes ne sont pas linéairement séparables dans la réalité, le problème de classification est rendu beaucoup plus complexe. L'objectif devient alors de transformer l'espace initial en un espace de plus grande dimension dans lequel le problème de classification redevient linéaire *i.e.* les données non-linéairement séparables sont projetées dans un espace de grande dimension de façon à ce qu'elles deviennent linéairement séparables.

• les modèles neuronaux

Utilisées dans le domaine de la classification et de l'identification, les réseaux de neurones artificiels (ANN pour Artificial Neural Network [Nocera, 1992]) est une technique qui s'inspire des systèmes neuronaux biologiques. Parmi les modèles neuronaux les plus utilisés en RAP[16], seuls les perceptrons multicouches (MLP pour Multi-Layer Perceptron) seront présentés ici. Le perceptron multicouches se compose généralement de 3 couches :

- la couche d'entrée comprenant d cellules et transmettant les caractéristiques x_i en sortie avec $i \in \{1, ..., d\}$;
- la couche cachée comprenant n cellules entièrement connectées aux cellules en entrée et transmettant des signaux y_j en sortie avec $j \in \{1, ..., n\}$;
- la couche de sortie avec c cellules (classes) de sorties connectées entièrement aux cellules cachées et générant des signaux z_k en sortie avec $k \in \{1, ..., c\}$.

L'apprentissage consiste à fixer les paramètres du perceptron multicouches *i.e.* les poids de connexion w_{ji} et w_{kj}, à partir des données du corpus d'apprentissage. En ce sens, l'algorithme itératif d'apprentissage par rétro-propagation du gradient de l'erreur a pour objectif de fixer le poids des connexions qui minimise l'erreur quadratique moyenne commise par le réseau sur l'ensemble d'apprentissage [Rumelhart & McClelland, 1986]. La décision consiste à choisir la classe qui maximise z_k.

[16] on y trouve aussi les réseaux récurrents et les cartes auto-organisatrices

Quelques exemples de travaux en évaluation instrumentale automatique

Plusieurs travaux se sont consacrés à confronter les résultats de l'évaluation instrumentale automatique à l'évaluation perceptive. Les études jugées les plus significatives sont présentées ici :

- Les travaux de Wester (1998) comparent deux méthodes de classification, les techniques de régression linéaire *versus* les modèles de Markov cachés (HMM), pour la tâche d'évaluation de la qualité de la voix. L'analyse comparative repose sur le corpus Kay Elemetrics Corporation (1995) de 607 sujets pathologiques et 36 sujets témoins, évalués perceptivement par un jury de 3 experts sur 3 critères qualitatifs : la raucité, le souffle et le degré global de déviance. Deux types de matériau phonétique sont utilisés pour chaque locuteur : de la parole lue (12 secondes de « Rainbow Passage ») et la séquence de phonèmes /ʌnlai/ du mot « sunlight » extraite du texte lu. Encadré de phonèmes non voisés (/s/ et /t/), le choix de ce segment voisé de parole s'explique par la présence de phénomènes transitoires pertinents qui sont particulièrement importants pour la perception de qualité de la voix.
  Inspirée de De Krom (1993), l'analyse acoustique consiste à extraire de la parole lue un ensemble de paramètres, principalement des mesures Harmonic Noise Ratio (HNR), estimés sur des trames séquencées toutes les 10ms :
  - HNR1 sur [60,400]Hz, HNR2 sur [400,2000]Hz, HNR3 sur [2000,5000]Hz et HNR4 sur [5000,8000]Hz ;
  - high slope=HNR4-HNR3, mid slope=HNR3-HNR2 et low slope=HNR2-HNR1 ;
  - lnF0 = la fréquence fondamentale ;
  - levelDB = l'intensité totale du signal.

  Comparée à l'analyse perceptive, l'étude montre que les HMM obtiennent de meilleures performances (65 %) que l'analyse par régression, sur le caractère soufflée de la voix et le degré global de déviance. Par contre, les deux méthodes présentent des performances semblables sur le critère de la raucité vocale.

- Dans [Dibazar et al., 2002], les auteurs se proposent d'évaluer la méthode HMM pour la détection automatique de la voix pathologique. Issus d'une analyse en banc de filtres, les coefficients cepstraux MFCC (12c + Energie) sont associés à des mesures de F0 pour être modélisés par un classifieur HMM. La méthode est évaluée sur 2 types de matériau différents, voyelle tenue /a/ et parole lue (12 secondes de « Rainbow Passage »), extraits de MEEI Database (2002) sur 53 sujets témoins et 657 sujets présentant des pathologies vocales diverses.
  Les meilleurs résultats sont atteints sur le /a/ tenu avec un taux de 99.44 % de classification correcte sur les données d'apprentissage et 98.30 % sur les données de test. Le texte lu « Rainbow Passage » obtient quant à lui un taux de 98.59 % sur le corpus d'apprentissage et 97.75 % sur le corpus de test. Les auteurs concluent que l'utilisation du support vocalique /a/ constitue un matériau fiable pour détecter si une voix est pathologique ou normale, même si le texte lu reste tout de même pertinent avec des performances légèrements plus faibles.

- A partir d'une paramétrisation acoustique de type MFCC, les travaux de [Godino-Llorente et al., 2006] s'intéressent à la détection de la qualité de la voix pathologique en utilisant les modèles de mélange de gaussiennes (GMM). Les échantillons de parole étudiés sont constitués de voyelles tenues /ah/ d'une durée de 1 à 3 secondes. Les locuteurs regroupent 53 sujets témoins et 173 sujets dysphoniques atteints de pathologies diverses (organiques, neurologiques, traumatiques, psychologiques), sélectionnés dans le corpus [Kay Elemetrics Corporation, 1995]. Issus de l'analyse cepstrale, les coefficients MFCC (n coefficients + Energie avec $n \in \{10, ..., 26\}$) sont associés aux dérivées premières (Δ) et secondes ($\Delta\Delta$) formant des vecteurs de dimension $D = 3n + 3$ paramètres. Afin de réduire la dimension de l'espace paramétrique, les auteurs ont utilisé deux méthodes : le F-Ratio et le Fisher's Discriminant Ratio.
  L'étude montre que chaque test a obtenu un taux de fausse acceptation inférieur au taux de faux rejet avec une efficacité d'environ 94 $\pm$3.2 %. Les auteurs concluent que les paramètres cepstraux complétés des paramètres Δ obtiennent les meilleurs résultats pour la tâche visée (16MFCC et 6 gaussiennes par mixture). La combinaison des dérivées $\Delta\Delta$ ne constituent pas une information pertinente sur les résultats.

Les exemples de travaux présentés ci-dessus constituent principalement des classifications binaires « voix normale vs. voix pathologique » de la qualité de la voix qui s'appuient pour la plupart sur la voyelle tenue comme matériau phonétique, dénotant une forte influence de l'approche analytique.

En résumé, comparées aux méthodes instrumentales analytiques, l'avantage et l'originalité de ces approches reposent sur :

[1] La capacité à analyser de la parole continue proche de l'élocution naturelle (bien que très souvent, les études reposent sur de la voyelle tenue).

[2] La capacité à traiter de grands corpus, permettant de mener des études à grande échelle et d'obtenir des informations statistiques significatives.

[3] Une analyse acoustique, simple et automatique, permettant une utilisation clinique facile à caractère non invasif et à faible coût humain.

Les méthodes instrumentales dites « hybrides »

Des méthodes objectives « hybrides » que l'on pourrait aussi qualifier de « transversales », ont été proposées pour l'évaluation de la voix pathologique. Ces dernières s'appuient sur des mesures « analytiques » - estimées généralement sur de la voyelle tenue avec une seule valeur par mesure étudiée - qui serviront comme données d'apprentissage à des techniques statistiques issues de la TAP (comme les GMM, ANN, SVM, ...) pour exploiter les résultats.

Par exemple, les travaux présentés dans [Wang & Jo, 2006] se concentrent sur la classification de la voix pathologique avec des modèles de mélanges de gaussiennes (GMM) et comparent les résultats obtenus avec ceux d'une étude précédente utilisant des réseaux de neurones artificiels (ANN) selon le même protocole expérimental.

Pour cela, 6 mesures acoustiques (Jitter, RAP, Shimmer, APQ, HNR, SPI [17]) sont calculées sur un corpus de 41 sujets témoins et 111 patients dysphoniques (polypes, kystes, oedèmes, nodules, laryngites, ...) sur de la voyelle tenue /a/.

A partir de ces mesures, les classifications à base de ANN et de GMM sont employées pour discriminer les voix « normales » ou « pathologiques ». En ce sens, la méthode GMM atteint un taux de 98.4 % de classification correcte sur les données d'apprentissage et 95.2 % sur les données de test. Alors que la méthode ANN affiche un taux de 98.0 % sur le corpus d'apprentissage et 94.2 % sur le corpus de test. Dans cette étude, les auteurs montrent que l'approche GMM obtient de « *meilleures performances* » que les modèles neuronaux (ANN). Bien que les différences soient minimes, il est montré que les modèles GMM peuvent être efficaces pour la tâche de classification de la voix pathologique et offrent plus de robustesse pour des applications pratiques.

Dans la même optique, [Chen et al., 2007] introduit l'utilisation des SVM pour la tâche de classification binaire - voix normales contre voix pathologiques. Dans ces travaux, 27 mesures acoustiques, issues d'une analyse analytique, sont extraites d'une voyelle tenue /a/ pour chacun des sujets présents dans le corpus (177 patients et 39 sujets contrôle issus de la base de données Kay Elemetrics Corporation (1995)).

Les auteurs comparent les résultats de classification obtenus sur les 27 mesures acoustiques brutes avec ceux obtenus après application d'une analyse en composante principale (ACP) pour réduire l'espace paramétrique. Des taux de classification corrects de 92.2 % et 98.1 % sont respectivement atteints. Le meilleur résultat étant obtenu avec seulement les 2 premiers axes principaux issus de l'ACP, les auteurs mettent, ainsi, en exergue un fort taux de corrélation entre certaines mesures acoustiques. Ils soulignent finalement l'importance d'une sélection plus pertinente des mesures acoustiques à extraire.

[17] pour Soft Phonation Index

1.5.3 La méthode « Phonetic Labeling »

Habituellement, l'évaluation perceptive se pratique sur de la parole continue et consiste à donner une impression perçue sur la globalité du signal acoustique. La méthode « Phonetic Labeling » propose d'évaluer perceptivement le dégré du trouble vocal sur chaque phonème d'une phrase en les considérant séparément et selon différents paramètres qu'ils soient dysphoniques ou phonétiques. Ainsi, cette méthode fournit un étiquetage phonétique d'un morceau de texte prononcé par un locuteur permettant d'obtenir une cartographie qualitative très précise de sa dysphonie.

La méthode « Phonetic Labeling » s'inspire des travaux de Lorch & Whurr (2003) montrant la pertinence de certains contextes phonétiques favorisant l'apparition d'occurrences pathologiques chez des sujets atteints de dysphonie spasmodique. Les travaux de Revis et al. (2006) proposent une étude élargie à des patients dysphoniques, décrivant les caractéristiques pathologiques de chacun des phonèmes constitutifs d'un échantillon de parole. Pour cela, une phrase extraite de « La chèvre de Monsieur Seguin » d'Alphonse Daudet a été retenue comme matériau phonétique de l'étude :

« Il les perdait toutes de la même façon »
« i l e p ɛ r d e t u t d ø l a m ɛ m f a s ɔ̃ »

Le choix de cette portion de texte se justifie par la présence de configurations textuelles diverses comme la voyelle initiale, des combinaisons de phonèmes voisés, d'occlusives sonores et sourdes, ... Cependant, il faut préciser que cette phrase ne correspond pas à un texte pouvant être qualifié de « phonétiquement équilibré ».

Dans un premier temps, une analyse perceptive, réalisée par un jury sur l'ensemble des échantillons de parole de l'étude suivant les paramètres dysphoniques « GRB »[18], a servi de référence pour l'étiquetage des phonèmes évalués selon 5 paramètres :

- la raucité et le souffle (*paramètres dysphoniques*) ;
- l'aspiration et le paramètre « craqué »[19] (*paramètres phonétiques*) ;
- le dévoisement (*paramètre mixte*).

Dans un deuxième temps, une procédure dite « en entonnoir » a été utilisée par une orthophoniste pour étiqueter, un à un, chacun des phonèmes contenus dans la phrase. Décrite dans [Revis, 2004], cette procédure a été élaborée afin de pallier la difficulté d'évaluer « à l'oreille » un matériau aussi court qu'un simple phonème[20]. Dans ces conditions, des mesures quantitatives issues de l'étiquetage ont été calculées pour chaque patient. Ces derniers correspondent aux nombres d'occurrences pour chacun des 5 paramètres retenus pour l'étude, au nombre total d'occurrences de l'ensemble des paramètres et au nombre total de phonèmes atteints.

[18] de l'échelle GRBAS de Hirano (1981)
[19] creak
[20] de l'ordre de quelques dizaines de millisecondes (de 5 à 50)

Malgré le temps non négligeable de la procédure (environ 15 minutes par stimulus) et la concentration requise durant les écoutes (soutenue et intense), les intérêts suscités par la méthode « Phonetic Labeling » sont nombreux. Les principales observations relevées par l'étude sont les suivantes :

• les sujets sains produisent des occurrences pathologiques (moyenne de 10 % avec un écart type de 6). L'analyse des résultats montrent que les hommes sont sujets naturellement à la raucité et les femmes au souffle.

• le nombre d'occurrences pathologiques est fortement corrélé au degré de sévérité de la dysphonie. Cette observation tend à montrer que le jugement perceptif fonctionne comme la somme de phénomènes dysphoniques surgissant dans le discours, et que la méthode « Phonetic Labeling » est très représentative de l'analyse perceptive « classique » de la dysphonie.

• la représentation en couleur de la répartition des occurrences pathologiques dans la phrase permet d'obtenir la cartographie qualitative de la dysphonie d'un locuteur dont la précision peut se révéler très utile dans certaines situations :
  - deux locuteurs de même grade (équivalents sur un plan quantitatif) peuvent afficher une prévalence du souffle pour l'un et de la raucité pour l'autre (différences sur un plan qualitatif) ;
  - pour des situations de désaccord entre les auditeurs d'un jury, elle fournit une description détaillée des différentes voix litigieuses permettant de s'affranchir de l'approximation du jury.

• certains contextes phonétiques semblent influer sur l'apparition des occurrences pathologiques comme le phonème /u/ encadré par les deux occlusives sourdes /t/ dans le mot « toutes ». De même, l'attaque[21] semble favoriser l'émergence des occurrences pathologiques, et plus particulièrement les voyelles présentant une attaque comme la voyelle initiale /i/ et les phonèmes succédant à une consonne non voisée.

• les consonnes sourdes ne sont pas atteintes selon les critères dysphoniques étudiés dans la mesure où la dysphonie est essentiellement liée à la vibration laryngée. Par contre, elles peuvent être affectées suivant les paramètres phonétiques comme l'aspiration.

• en pratique clinique, la méthode « Phonetic Labeling » permettrait de suivre l'évolution de la qualité vocale d'un patient durant un traitement thérapeutique.

[21] dans une succession de phonèmes voisés, l'attaque n'intervient que sur le premier phonème c-à-d lors de la mise en vibration des cordes vocales qui reste active sur les phonèmes suivants ; par exemple, pour la séquence /dølamɛm/, l'attaque n'intervient que sur le /d/

1.6 Conclusion

Cet état de l'art montre que le jugement perceptif de la qualité de la voix reste une composante essentielle pour la majorité des protocoles d'évaluation de la voix pathologique [Hirano, 1989]. Pour ces évaluations perceptives, plusieurs échelles ont été proposées telles que le Vocal Profile Analysis Scheme (VPAS) de Laver (1980), le GRBAS Scale de Hirano (1981), le Hammarberg Scheme de Hammarberg (1986), le Buffalo Voice Profile Systeme (BVP) de Wilson (1987). Cependant, aucun de ces protocoles n'a été largement accepté pour une pratique clinique quotidienne. Les raisons principales étant :

1. la validité des échelles utilisées ;
2. l'imprécision des auditeurs à qualifier et quantifier une voix.

Néanmoins, l'évaluation perceptive est toujours considérée à l'heure actuelle comme le *Gold Standard* pour l'évaluation de la qualité vocale. Le jugement perceptif reste donc la référence face à laquelle les mesures instrumentales sont confrontées, et indispensable pour suivre l'évolution d'un traitement médical, rééducatif ou post-opératoire.

Afin de pallier les faiblesses et inconvénients de l'évaluation perceptive, les méthodes instrumentales dites « analytiques » ont été proposées dans la littérature. Ces dernières reposent sur l'acquisition de mesures acoustiques, aérodynamiques et/ou électrophysiologiques par le biais de capteurs. Les mesures estimées constituent un ensemble d'indices susceptibles de caractériser les dysfonctionnements vocaux directement liés à la dysphonie comme les fluctuations laryngées, la fuite glottique ou l'activité musculaire laryngée. Cependant, malgré des résultats très acceptables obtenus par cette approche « analytique », ceux-ci restent encore insuffisants pour les praticiens pour qu'elle puisse être considérée comme un outil de diagnostic. De plus, à ce jour encore, certaines limites et contraintes (comme l'utilisation de la voyelle tenue) renforcent l'idée qu'elle ne puisse être utilisée en pratique clinique qu'à titre indicatif et/ou expérimental.

Une deuxième catégorie de méthodes objectives s'est alors intéressée aux techniques utilisées en Traitement Automatique de la Parole (TAP), les méthodes instrumentales dites « automatiques ». Leur adaptation au contexte de la voix pathologique apporte des avantages comme l'utilisation de la parole continue ou une analyse acoustique facile et automatique, permettant un usage journalier en milieu clinique. Plusieurs études reposant sur différentes techniques utilisées en TAP, ont montré l'efficacité de cette approche aux travers de résultats très prometteurs et encourageants, laissant entrevoir de nouvelles possibilités d'amélioration des systèmes. Leur objectif est principalement de fournir un outil d'aide au diagnostic. Le travail de cette thèse s'inscrit dans ce contexte. Néanmoins, son originalité est d'utiliser le système de classification automatique comme un outil pour caractériser les phénomènes liés à la dysphonie dans le signal de parole. La première étape a consisté à adapter le système automatique au contexte pathologique afin de pouvoir utiliser des techniques de RAL à la tâche d'évaluation du degré de sévérité de la voix dysphonique.

Deuxième partie

Le Système et ses Performances

Chapitre 2

Le contexte expérimental

Sommaire

Résumé

Ce chapitre présente les deux corpus utilisés pour valider les méthodes proposées : CVD et BREF. Constitué d'une grande quantité de données, le corpus BREF est utilisé uniquement pour l'apprentissage du modèle du Monde afin de pallier le manque de données du corpus CVD, le corpus des voix pathologiques regroupant les sujets dysphoniques et de contrôle. Pour conclure ce chapitre, la présentation des résultats expérimentaux sera détaillée afin de mieux en comprendre la portée et la signification.

Ce chapitre est consacré à la description du contexte expérimental incluant l'utilisation de deux corpus pour les différentes expériences présentées dans cette thèse. Le premier corpus CVD (Corpus des Voix Dysphoniques) regroupe l'ensemble des voix dysphoniques et de contrôle. Il servira à l'apprentissage des modèles pathologiques, dérivés du *modèle du monde* par l'application de techniques d'adaptation (techniques décrites à la section 3.2). Le deuxième corpus nommé BREF sera utilisé pour l'apprentissage d'un *modèle générique* appelé aussi *modèle du Monde*, indispensable pour pallier le manque de données d'entraînement des *modèles de grade*[1].

Avant d'utiliser les deux corpus dans les différentes étapes du système RAL adapté au contexte de la voix pathologique, des pré-traitements comme la détection de parole, devront être appliqués sur les signaux acoustiques qui les constituent. En effet, seules les portions de « parole » doivent être prises en compte pour la tâche de reconnaissance des degrés de sévérité de la dysphonie. Dans ces conditions, les signaux acoustiques des deux corpus devront être « *nettoyés* » des portions de « non parole/silence ». Pour cela, deux techniques différentes seront utilisées selon la nature des corpus :

- CVD : un alignement phonétique contraint par le texte ;
- BREF : une modélisation de l'énergie à l'aide d'un mélange de gaussiennes.

Pour conclure ce chapitre, la section 2.4 apporte des explications nécessaires sur la présentation des résultats expérimentaux afin de mieux en comprendre la portée et la signification.

[1] dans le contexte pathologique, un modèle correspond à un niveau de sévérité de dysphonie *i.e.* à un grade de l'échelle perceptive GRBAS [Hirano, 1981]

2.1 Le Corpus CVD : Corpus des Voix Dysphoniques

Description

Le Corpus[2] des Voix Dysphoniques mis à disposition par le LAPEC[3] [Briffa, 2004] est constitué de 80 échantillons de voix féminines correspondant à 20 sujets témoins et 60 patientes dysphoniques, âgés de 17 à 50 ans (moyenne de 32.2 ans). L'ensemble des patientes dysphoniques a fait l'objet d'un examen laryngoscopique faisant apparaître les pathologies énumérées dans le tableau 2.1 (dysphonies essentiellement d'origine fonctionnelle décrites en 1.3.1).

Pathologies	Nombre	Pathologies	Nombre	Pathologies	Nombre
épaississement muqueux	1	nodules	23	sulcus	1
fuite glottique	1	oedème	16	serrage	1
kyste	4	polype	10	cordes vocales normales	3

TAB. 2.1 – *Répartition des pathologies vocales des 60 patientes dysphoniques du corpus CVD*

En ce qui concerne le support vocal, chaque sujet a été enregistré sur la lecture d'un paragraphe de « La chèvre de Monsieur Seguin » d'Alphonse Daudet :

« *Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon : un beau matin, elles cassaient leur corde, s'en allaient dans la montagne, et là-haut le loup les mangeait. Ni les caresses de leur maître, ni la peur du loup, rien ne les retenait.* »

Les enregistrements ont été évalués par consensus par un jury d'experts selon le grade global G de dysphonie de l'échelle GRBAS décrite en 1.5.1. L'ensemble du corpus étiqueté se présente de la manière suivante : 80 locuteurs équitablement répartis dans les 4 grades (20 voix normales, 20 voix avec dysphonie légère, 20 voix avec dysphonie moyenne, 20 voix avec dysphonie sévère) dont les durées s'étalent de 13.5 à 77.7 secondes avec une moyenne de 18.9 secondes et un écart type de 7.6 secondes.
.
Il est à noter que seul le critère G de l'échelle GRBAS est exploité dans toutes les études menées dans cette thèse. Deux raisons principales expliquent ce choix :

1. Le critère G est considéré comme l'un des critères les plus fiables de l'échelle GRBAS au cours d'une évaluation perceptive ;
2. Le critère G est mieux adapté aux analyses paramétriques proposées dans cette thèse. En effet, il ne porte pas sur un phénomène particulier contrairement aux autres critères de l'échelle GRBAS qui requièrent des analyses spécifiques ; il peut donc être traité plus efficacement par des analyses à plus large spectre.

[2] nommé CVD dans le reste du document

[3] Laboratoire d'Audio-Phonologie Expérimentale et Clinique de la Faculté de Médecine de Marseille dont les membres ont rejoint le Laboratoire Parole et Langage (LPL) en 2007

Travail manuel réalisé

En raison du cadre pathologique, un travail manuel sur le corpus CVD s'est révélé nécessaire pour améliorer la qualité de l'alignement phonétique qui lui sera appliqué par la suite. Cet alignement permettra, d'une part à « nettoyer » les signaux acoustiques des portions de « non parole/silence » et d'autre part à analyser des caractéristiques de la dysphonie selon les différentes classes de phonèmes (étude décrite au chapitre 7). Cette intervention manuelle a été principalement motivée par l'écoute des enregistrements des patientes dysphoniques laissant apparaître des situations peu habituelles, notamment pour les dysphonies les plus sévères de grade 3, en comparaison avec les voix « normales » de grade 0.

Voici quelques exemples de ces phénomènes observés :

- des portions de désonorisation ou des phénomènes de réduction importante comme l'absence de réalisation d'une voyelle se réduisant à un souffle
- des dysfonctionnements de l'attaque vocale : « soufflée » ou « en coup de glotte »
- des hésitations provoquant des répétitions ou des substitutions de mots
- des erreurs de lecture
- des ajouts d'expressions tels que « euh », « excusez-moi », « donc », ...
- des râlements
- un accent régional prononcé comme « segu*ing* », « s*é*guin », ...

Afin d'améliorer la qualité de l'alignement phonétique des signaux de parole et donc, de minimiser ses erreurs et leur incidence sur les performances du système adapté, il a semblé nécessaire d'intervenir manuellement et individuellement sur l'ensemble du corpus CVD.

Pour chaque voix du corpus CVD, il a fallu :

1. retranscrire le texte réellement prononcé par le locuteur afin d'y reporter les particularités non présentes dans le texte original ;
2. prendre en compte les nouvelles variantes phonétiques/phonologiques (accent prononcé, pause dans un mot, ...) ;
3. enrichir le lexique de mots en vue de l'alignement automatique (mots hors vocabulaire, erreurs de prononciation, erreurs de diction, ...).

Alignement phonétique

Dans le cadre des travaux menés dans cette thèse, une segmentation phonétique est nécessaire pour chaque signal de parole du corpus CVD. Une segmentation par phonème a donc été extraite automatiquement par un alignement phonétique contraint sur le texte, grâce au système d'alignement du LIA [Linarès et al., 2007; Bürki et al., 2008]. Ce dernier est basé sur un algorithme de décodage Viterbi (1967), un lexique de mots avec leurs variantes phonologiques et un ensemble de 36 phonèmes du français.

La figure 2.1 schématise les différentes étapes de l'alignement automatique produisant la segmentation phonétique d'un signal de parole préalablement transformé en coefficients PLP (Perceptual Linear Predictive). Les modèles acoustiques correspondent à des modèles de Markov cachés (HMM) non contextuels de phonèmes appris avec 200 heures de parole féminine extraite du corpus BREF.

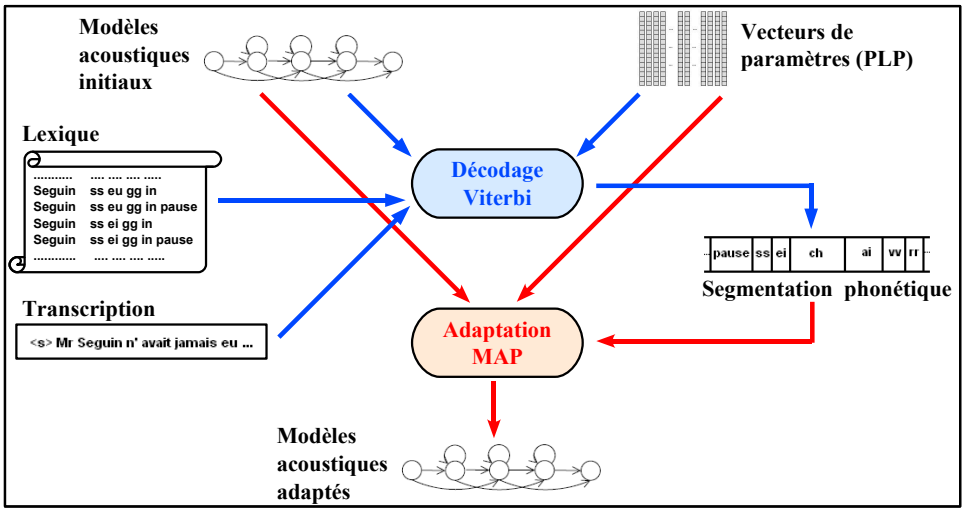


FIG. 2.1 – *Schéma décrivant un alignement phonétique contraint sur le texte.*

Afin d'affiner la qualité de l'alignement phonétique, une phase d'adaptation de type *MAP* des modèles acoustiques initiaux - utilisés durant le décodage - peut être effectuée à partir de la segmentation phonétique extraite. On peut alors ré-itérer le traitement complet en utilisant les modèles acoustiques adaptés pour obtenir une segmentation phonétique plus précise.

Les tableaux (2.2, 2.3, 2.4) présentent l'ensemble des phonèmes utilisés par l'alignement automatique (dénoté AUTO) avec leur correspondance dans Alphabet Phonétique International (API).

Classes de phonèmes	Phonèmes AUTO	Phonèmes API	Exemples
Semi-consonnes	[yy]	[j]	paille
	[ww]	[w]	poire
	[uy]	[ɥ]	huile

TAB. 2.2 – *Correspondance entre les symboles AUTO et API : les semi-consonnes*

Classes de phonèmes	Phonèmes AUTO	Phonèmes API	Exemples
Voyelles orales	[aa]	[ɑ] ou [a]	pâte ou patte
	[ai]	[ɛ]	père
	[au]	[o]	mot
	[ei]	[e]	blé
	[eu]	[ø] ou [ə]	peu ou je
	[ii]	[i]	vie
	[oe]	[œ]	peur
	[oo]	[ɔ]	botte
	[ou]	[u]	roux
	[uu]	[y]	rue
Voyelles nasales	[an]	[ɑ̃]	banc
	[in]	[ɛ̃]	matin
	[on]	[ɔ̃]	bon
	[un]	[œ̃]	lundi

TAB. 2.3 – *Correspondance entre les symboles AUTO et API : les voyelles*

Classes de phonèmes	Phonèmes AUTO	Phonèmes API	Exemples
Consonnes liquides	[ll]	[l]	sol
	[rr]	[r]	rose
Consonnes nasales	[mm]	[m]	main
	[nn]	[n]	nous
	[nnyy]	[ɲ]	vigne
Consonnes fricatives	[jj]	[ʒ]	jolie
	[vv]	[v]	valise
	[zz]	[z]	rose
	[ch]	[ʃ]	chat
	[ff]	[f]	feu
	[ss]	[s]	tasse
Consonnes occlusives	[bb]	[b]	bon
	[dd]	[d]	dans
	[gg]	[g]	gare
	[OkkBkk]	[k]	sac
	[OppBpp]	[p]	patte
	[OttBtt]	[t]	toit

TAB. 2.4 – *Correspondance entre les symboles AUTO et API : les consonnes*

Il faut souligner que certains phonèmes de la classification phonétique du français ont été regroupés en un seul dans le symbolisme informatique comme [ɑ] et [a] en [aa] ou [ø] et [ə] en [eu]. De même, le phonème [ŋ] qui est plutôt rare en français (par exemple parki**ng**), n'a pas été modélisé mais, le cas échéant, il sera représenté comme la succession des phonèmes [nngg] comme cela est le cas pour [ɲ] avec [nnyy] (par exemple monta**gn**e).

Le tableau 2.5 affiche les durées en secondes par classe de phonèmes et par grade, extraites automatiquement suivant la méthode d'alignement décrite ci-dessus et trois adaptations successives des modèles acoustiques.

Classes	**Grades**				**Effectifs Totaux**		
Phonétiques	*G0*	*G1*	*G2*	*G3*	*nb*	μ	σ
Consonne	**135.13**	**139.21**	**149.83**	**167.28**	**6395**	**0.092**	**0.045**
Liquide	34.56	34.01	36.04	43.03	2181	0.068	0.033
Nasale	29.72	30.17	31.85	33.42	1279	0.098	0.039
Fricative	31.77	32.32	35.07	40.70	1144	0.122	0.057
Occlusive	39.08	42.71	46.87	50.13	1791	0.100	0.039
Voyelle	**103.58**	**98.77**	**103.46**	**109.79**	**5586**	**0.074**	**0.046**
Orale	84.37	80.45	85.22	93.66	4862	0.071	0.044
Nasale	19.21	18.32	18.24	16.13	724	0.099	0.046
Semi-voyelle	**2.80**	**2.98**	**3.37**	**3.45**	**159**	**0.079**	**0.040**
Tous phonèmes	**241.51**	**240.96**	**256.66**	**280.52**	**12140**	**0.084**	**0.046**

TAB. 2.5 – *Durée en secondes par classe phonétique (version AP2) et par grade - Informations quantitatives sur les phonèmes d'une classe phonétique : nombre (nb) avec durée moyenne (μ) et écart-type (σ) associé*

2.2 Le Corpus BREF

Description

Afin de fournir un ensemble de données françaises nécessaire pour le développement et l'évaluation de systèmes de dictée vocale, le LIMSI a développé le corpus BREF [Lamel et al., 1991] en 1991. Le financement de ce projet a été pris en charge par plusieurs partenaires : le GDR-PRC Communication Homme/Machine, la CEE (projet ESPRIT Polyglot) et l'Aupelf-Uref. Le corpus BREF représente plus de 100 heures de parole lues dans le journal « Le Monde » par un ensemble de 120 locuteurs (65 femmes et 55 hommes). Son style est donc de lecture à contenu journalistique et les conditions d'enregistrement sont optimales *i.e.* enregistrement effectué en chambre sourde.

Pour l'ensemble des expériences présentées dans cette thèse, ce corpus a été utilisé uniquement pour l'apprentissage du *modèle du monde*. Son utilisation pour l'entraînement des modèles pathologiques s'est révélée indispensable en raison de la faible quantité de données du corpus CVD ne permettant pas, à lui seul, d'apprendre des modèles statistiques « robustes » et « satisfaisants ». L'utilisation conjointe du corpus CVD et des techniques d'adaptation permet de dériver les modèles pathologiques du *modèle du monde* de manière efficace. Le modèle générique a été appris avec un ensemble de 76 enregistrements de 2 mn chacun de voix exclusivement féminines (le corpus CVD étant constitué de voix de femmes uniquement), ce qui représente environ 2 h 30 mn de durée totale (parole et silence).

Détection « parole/non parole »

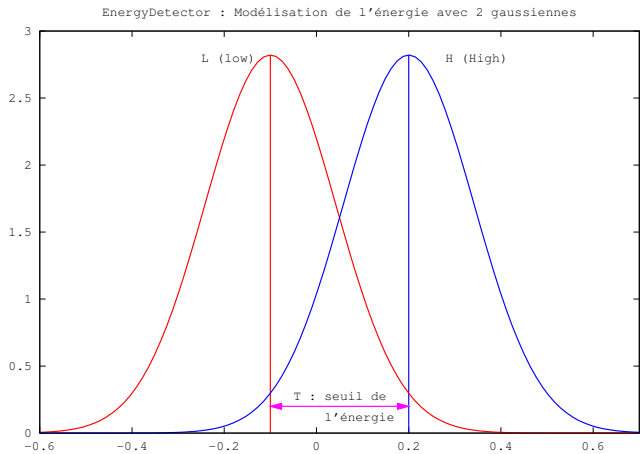


FIG. 2.2 – *Utilitaire « Energy Detector » : détection automatique des portions de parole par modélisation statistique de l'énergie.*

Afin de ne considérer que les portions de parole, les signaux du corpus BREF ont été « nettoyés » à l'aide de l'utilitaire « Energy Detector » de la boîte à outils (LIA_SpkDet

et ALIZE [Bonastre et al., 2005]) développée au laboratoire LIA et disponible en « open source ». Le principe de cet outil repose sur la modélisation non supervisée de l'énergie par un mélange de 2 ou 3 composantes gaussiennes comme le montre la figure 2.2.

Un seuil d'énergie est estimé de manière empirique afin d'écarter les trames ayant une énergie inférieure à celui-ci. Dans son mode le plus simple appelé *MeanStd*, ce seuil est déterminé par l'entraînement d'une bigaussienne sur la composante d'énergie et la recherche de la meilleure distribution *i.e.* composante gaussienne avec le plus fort poids. Cette distribution est utilisée pour calculer le seuil d'énergie : $\tau = \mu_T - \alpha\ \sigma_T$ où τ est le seuil, (μ_T,σ_T) sont les paramètres de la meilleure distribution et α une constante déterminée de manière empirique. L'application du module « Energy Detector » sur l'ensemble des signaux de parole du corpus BREF sélectionnés dans ce travail a permis de détecter automatiquement les portions de parole, totalisant une durée de 1 h 20 mn.

2.3 L'exploitation du corpus CVD

Les systèmes automatiques de traitement de la parole, basés sur une approche statistique, nécessitent une grande quantité de données pour l'apprentissage des modèles. En d'autres termes, plus la taille du corpus d'apprentissage est importante, mieux les paramètres des modèles statistiques sont estimés et par conséquent, plus performant est le système de reconnaissance.

L'approche statistique mise en œuvre ici pour caractériser les voix dysphoniques, nécessiterait donc une grande quantité de données pour le corpus CVD. Néanmoins d'un point de vue statistique, celui-ci est plutôt considéré « de petite taille ». Dans le contexte clinique, le corpus CVD est relativement important si l'on considère la difficulté de constituer un corpus de 80 voix féminines dont 60 sont atteintes de dysphonies dysfonctionnelles[4], enregistrées sur le même matériau phonétique, évaluées perceptivement par le même jury d'experts et réparties équitablement entre les 4 grades de l'échelle GRBAS. Aussi l'exploitation du corpus CVD oblige à prendre des précautions afin que les résultats, obtenus lors des différentes expérimentations, soient qualifiés de « robustes » et interprétables d'un point de vue statistique. Cela implique la mise en place d'une méthodologie basée sur la technique « leave_x_out » visant à respecter deux objectifs :

1. une séparation totale des données d'apprentissage et de test ;
2. une augmentation de la robustesse de la décision.

Le principe de la technique « leave_x_out » consiste à effectuer une évaluation circulaire sur le corpus : x exemples du corpus sont exclus à chaque entraînement d'un modèle ; ils seront utilisés ensuite comme éléments de test sur ce même modèle. Outre les objectifs cités ci-dessus, la mise en place de la technique « leave_x_out » permet de limiter l'influence de voix particulières au sein des modèles et de s'assurer que les voix utilisées pour l'apprentissage des modèles soient exclues des jeux de tests afin de différencier la détection de la pathologie de la reconnaissance du locuteur.

Séparation totale des données d'apprentissage et de test

A partir d'un même corpus, il est possible de constituer deux corpus distincts, d'apprentissage et de test, et de s'assurer qu'une voix testée sur un modèle n'a pas servi à son apprentissage. Le principe général est illustré par la figure 2.3. On s'assurera que tous les modèles sont entraînés avec un même nombre de voix afin d'obtenir une homogénéité dans leur structure.

[4] même s'il ne s'agit pas d'une même pathologie : polype, oedème, kyste, ...

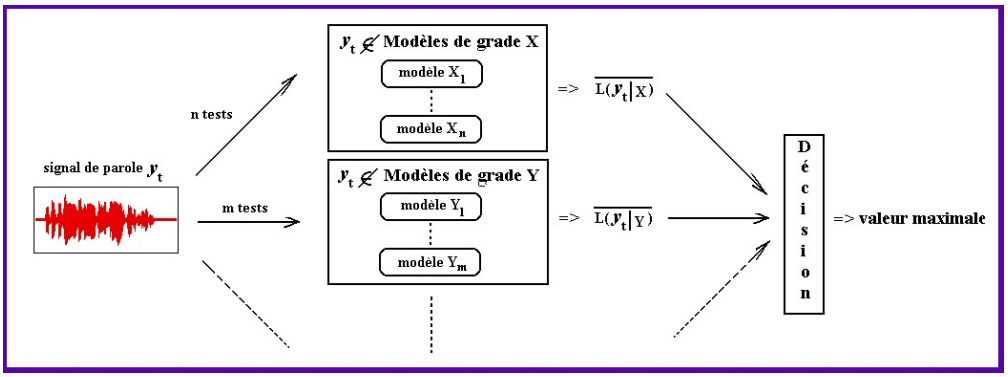


FIG. 2.3 – *Technique « leave_x_out » : le signal y_t de grade Y est testé sur les n modèles de grade X et sur les m modèles de son propre grade Y dont il a été exclu durant leur apprentissage. Une vraisemblance moyenne est calculée sur chacun des grades. La décision correspond au grade du modèle sur lequel la plus grande vraisemblance est obtenue.*

Augmentation de la robustesse de la décision

De plus, pour que les différents protocoles expérimentaux mis en œuvre dans cette thèse soient qualifiés de robustes, il est nécessaire d'augmenter le nombre de tests et ainsi la pertinence au sens statistique des résultats obtenus. En présence d'un corpus d'apprentissage de taille réduite comme c'est le cas ici avec le corpus CVD, la technique « leave_x_out » permet l'apprentissage d'un plus grand nombre de modèles statistiques et par conséquent, d'augmenter le nombre de tests pour chacun des signaux de parole. Il faut préciser que malgré l'augmentation du nombre de tests, chaque locutrice du corpus CVD obtiendra à l'issue de la phase de test, une seule valeur de vraisemblance pour chacun des grades qui correspondra à la moyenne calculée sur N tests effectués dans le grade correspondant.

Exemple pratique

Si l'on met en œuvre la technique « leave_one_out » *i.e.* exclusion d'une seule voix lors de l'apprentissage des modèles de grade, il sera appris 80 modèles de 19 voix chacun (soit 20 modèles par grade). Un signal y_t de grade $g = \{2\}$ sera testé sur les 60 modèles de grade $\overline{g} \in \{0, 1, 3\}$ et sur le seul modèle de grade 2 dont il aura été exclu de l'apprentissage.

A l'issue de la phase de test, le signal y_t comptabilisera 4 valeurs de vraisemblance :
◦ 1 vraisemblance obtenue sur le modèle de grade 2 : $L(y_t|X_g)$ où $g = \{2\}$;
◦ 3 vraisemblances correspondant chacune à la moyenne des vraisemblances obtenues sur les 20 modèles de même grade : $L(y_t|X_{\overline{g}}) = \frac{1}{20} \Sigma_{i=1}^{20} L(y_t|X_{\overline{g}}^i)$ où $\overline{g} \in \{0, 1, 3\}$. La décision correspond au grade du modèle sur lequel la plus grande vraisemblance est obtenue.

2.4 Présentation des résultats

Dans la partie III, les résultats expérimentaux sont exprimés en terme de Taux Correct de Classification noté *TCC* dans le tableau 2.6.

Grade 0	Grade 1	Grade 2	Grade 3	Global
% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/80)

TAB. 2.6 – *Entête d'un tableau de résultats de la classification 4-G en terme de % TCC*

Comme le montre le tableau 2.6, deux valeurs de *TCC* sont à distinguer et se définissent de la manière suivante :

$$\%\ TCC\ Grade = \frac{nb\ locuteurs\ du\ grade\ reconnus\ dans\ leur\ grade\ perceptif}{nb\ locuteurs\ dans\ le\ grade\ soit\ 20} \tag{2.1}$$

$$\%\ TCC\ Global = \frac{nb\ locuteurs\ reconnus\ dans\ leur\ grade\ perceptif}{nb\ total\ de\ locuteurs\ soit\ 80} \tag{2.2}$$

Afin de mesurer la qualité du système et d'analyser les erreurs de classification, une matrice de confusion pourra être présentée sous la forme du tableau 2.7 :

	S_G0	S_G1	S_G2	S_G3
P_G0	**...**	...	...	...
P_G1	...	**...**	...	...
P_G2	...	...	**...**	...
P_G3	...	...	...	**...**

TAB. 2.7 – *Exemple de matrice de confusion d'une classification 4-G*

Elle fournit le nombre d'erreurs et le type de confusion entre la réponse donnée par le système - noté S_Gx - et la référence perceptive - notée P_Gx. La colonne S_Gx correspond au nombre de locuteurs classés par le système dans le grade x et la ligne P_Gx correspond au nombre de locuteurs de grade perceptif x évalués par le système dans les différents grades. La diagonale de la matrice fournit le nombre de réponses correctes *i.e.* le nombre de locuteurs classés par le système dans leur grade perceptif.

Note : tous les résultats présentés dans la partie III sont issus du classifieur GMM et doivent être interprétés d'un point de vue statistique.

Intervalle de confiance

En raison de la taille réduite du corpus CVD, les scores expérimentaux TCC devront être analysés avec précaution. En effet, donner un résultat sans apporter une indication sur sa précision n'a que peu d'intérêt car sa reproductibilité n'est pas confirmée. L'une

des méthodes classiques utilisées est le calcul de l'intervalle de confiance (noté IC) pour juger de l'apport significatif d'une approche en terme de performance. Le calcul de l'intervalle de confiance se définit selon la formule suivante :

$$IC = \pm 1.96 \sqrt{\frac{\%\ \text{TCC}\ (1 - \%\ \text{TCC})}{\text{N - 1}}} \qquad \text{avec N = Nb de tests} \tag{2.3}$$

sur une plage de valeurs variant de 0 % à 100 % avec une base de 80 tests, on obtient la figure 2.4 qui illustre les variations de l'intervalle de confiance à considérer selon les différents scores TCC obtenus par des expériences de type 4-Grades.

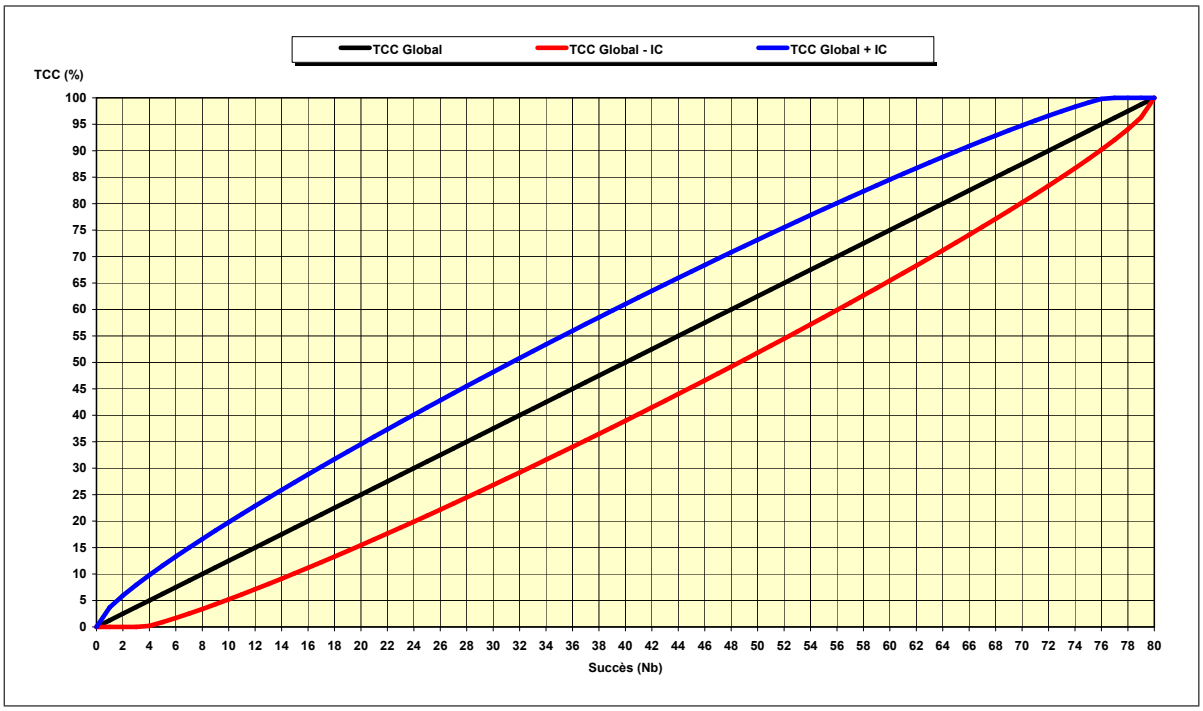


FIG. 2.4 – *Représentation graphique de la courbe TCC Global avec ± IC (Intervalle de Confiance) calculés avec un nombre de 80 tests.*

On notera qu'un TCC Global de 85.0 % (soit 68 locutrices bien classées dans leur grade perceptif) devra être interprété sur un intervalle de confiance de ± 7.9 % soit la plage TCC de $[77.1\ \%, 92.9\ \%]$. La valeur maximale est atteinte pour un TCC de 50.0 % avec un IC de ± 11.0 %.

La difficulté à contrôler le niveau de confiance des résultats et d'afficher des scores TCC significatifs est due essentiellement à la quantité réduite des données du corpus CVD. Néanmoins, l'objectif principal des travaux de cette thèse n'est pas d'obtenir les meilleures performances sur un système optimisé pour une tâche donnée mais plutôt, de mieux appréhender et comprendre les phénomènes acoustiques liés à la dysphonie.

2.5 Conclusion

Dans cette section, les deux corpus - CVD et BREF - utilisés dans le contexte expérimental de cette thèse, ont été présentés ainsi que les différentes techniques mises en œuvre pour extraire l'information utile - « phonétique » pour CVD et « parole » pour BREF - pour les différentes phases du système RAL.

Concernant le corpus des voix dysphoniques CVD, deux versions d'alignement phonétique seront utilisées dans les différentes expériences présentées dans la partie III. Dans un premier temps, l'alignement phonétique (décrit en section 2.1) a été appliqué sur le corpus CVD en utilisant la même transcription du texte de « La chèvre de Monsieur Seguin » pour l'ensemble des locutrices. Or il est apparu nécessaire d'effectuer un travail manuel sur le corpus CVD en raison des observations relevées dans la section 2.1.

On distinguera comme version d'alignement du corpus CVD :

AP1 : Alignement Phonétique 1re version
- « La chèvre de Monsieur Seguin » identique pour l'ensemble des locutrices
- ajout de mots manquants dans le lexique comme « Seguin »
- alignement phonétique décrit en section 2.1

AP2 : Alignement Phonétique 2e version
- « La chèvre de Monsieur Seguin » personnalisée pour chacune des locutrices
- interventions manuelles décrites en section 2.1
- alignement phonétique décrit en section 2.1

Lors de la présentation de résultats issus d'expériences utilisant pour un même protocole expérimental des versions différentes d'alignement, de légères différences dans les scores pourront apparaître selon le niveau de version de l'alignement utilisé ; les performances sont généralement légèrement meilleures pour la version AP2.

Il faut préciser que le tableau 2.5 affiche les durées par classe de phonèmes et par grade issues de l'alignement phonétique version AP2.

Chapitre 3

Le système RAL adapté au contexte pathologique

Sommaire

Résumé

Dans ce chapitre, nous présentons le système de reconnaissance automatique du locuteur adapté à la tâche de classification des voix pathologiques. Pour cela, les techniques utilisées pour chacune des trois phases - la paramétrisation, la modélisation et la décision - qui constituent classiquement un système de RAL, seront décrites ainsi que les spécificités apportées en raison du contexte pathologique.

Avant d'aborder la problématique de l'évaluation « objective » de la dysphonie, un système de RAL a dû être adapté à différents niveaux pour la tâche de reconnaissance du niveau de sévérité du trouble vocal. Dans ce chapitre, les différentes techniques issues de l'approche statistique et mises en œuvre dans les expériences présentées dans cette thèse, sont décrites avec, le cas échéant, les spécificités qui ont dû être apportées en raison du contexte pathologique.

Pour le système de RAL, nous nous intéressons à trois phases :

1. la phase de paramétrisation (décrite en 3.1)
2. la phase de modélisation (décrite en 3.2)
3. la phase de décision (décrite en 3.3)

Le système adapté sera décrit suivant cet ordre chronologique.

Le processus de décision en RAL est dépendant de la tâche visée. En VAL, la décision correspond à une décision binaire, « Acceptation » ou « Rejet », suivant que la mesure de ressemblance est supérieure à un seuil de décision. Le but est d'associer à la décision une fonction de coût qui rende le système performant. Différentes techniques de normalisation des scores de confiance sont proposées dans la littérature [Scheffer, 2006] dont l'objectif est de rendre plus robustes les décisions prises par le système de VAL *i.e.* déterminer un seuil robuste pour la fonction de décision. Parmi les techniques de normalisation les plus connues, nous pouvons citer « Znorm ou Z-normalisation » [Reynolds, 1997] ou encore « Tnorm ou T-normalisation » [Auckenthaler et al., 2000]. Par contre en IAL (milieu fermé), elle correspondra à une décision « 1 parmi N » basée sur le « Maximum de Vraisemblance » désignant le locuteur le plus probable parmi les N connus du système.

Il est important de relever que la plateforme informatique utilisée par le système, repose sur des boîtes à outils disponibles en « open source » et dédiées à des tâches spécifiques telles que :

• **analyse acoustique :** SPRO developpée à l'IRISA [Gravier, 2003] ;

• **analyse statistique :** ALIZE[1] et SpkDet développées au LIA [Bonastre et al., 2005].

[1] qui fait désormais partie du projet Mistral ; pour plus de renseignements voir le site http ://mistral.univ-avignon.fr

3.1 La paramétrisation acoustique

La phase de paramétrisation permet de réduire la redondance du signal de parole et d'en extraire les informations pertinentes en vue de la reconnaissance. Comme le montre la figure 3.1, il s'agit d'une analyse acoustique qui transforme un signal de parole en un ensemble de vecteurs de paramètres [2]. Ces derniers doivent caractériser au mieux et à chaque instant, le signal de parole afin d'être pertinents, robustes au bruit et discriminants pour en faciliter la reconnaissance. Cette transformation fournit ainsi une représentation simplifiée du signal nécessaire avant les phases d'apprentissage et de test.

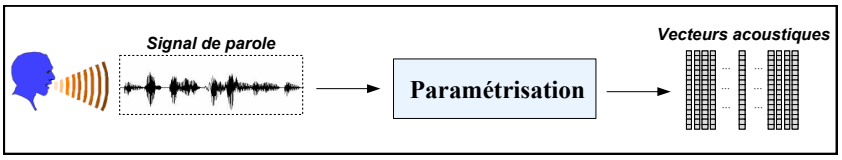


FIG. 3.1 – *Phase de paramétrisation des signaux de parole du système de RAL adapté à la classification des voix pathologiques.*

Il existe un certain nombre de techniques de représentation du signal de parole. Nous présenterons ici les principaux paramètres issus de l'analyse spectrale utilisés dans le cadre de cette thèse :

1. les coefficients de prédiction linéaire ;
2. les coefficients issus de l'analyse en banc de filtres.

Ils constituent les principaux paramètres utilisés en reconnaissance vocale [Furui, 1981; Hermansky, 1990].

3.1.1 Le pré-traitement acoustique

Les signaux de parole sont échantillonnés à 16kHz. Ils subissent un traitement de pré-accentuation couramment utilisé afin de compenser le fait que l'amplitude des pics harmoniques décroît en fonction de la fréquence avec une pente spectrale de l'ordre de -12dB par octave du signal de parole [Picone, 1993] à la source qui est relevée par le phénomène dit de « radiation aux lèvres » de +6dB. Cette étape consiste à filtrer le signal avec un filtre passe-haut du premier ordre : $H(z) = 1 - k\, z^{-1}$ avec $k \in [0,1[$ afin de renforcer la contribution des hautes fréquences [3]. Elle peut aussi être définie algorithmiquement de la manière suivante :

$$\hat{s}(n) = s(n) - k\, s(n-1) \tag{3.1}$$

où le coefficient de pré-accentuation k est généralement compris entre $[0.90, 1[$. Le filtre de pré-accentuation est appliqué sur le signal avant le fenêtrage. Ici, k est fixé empiriquement à 0.95.

[2] appelés aussi indifféremment vecteurs acoustiques ou vecteurs d'observations.
[3] les sons aigus, toujours plus faibles en énergie que les sons graves, sont « avantagés ».

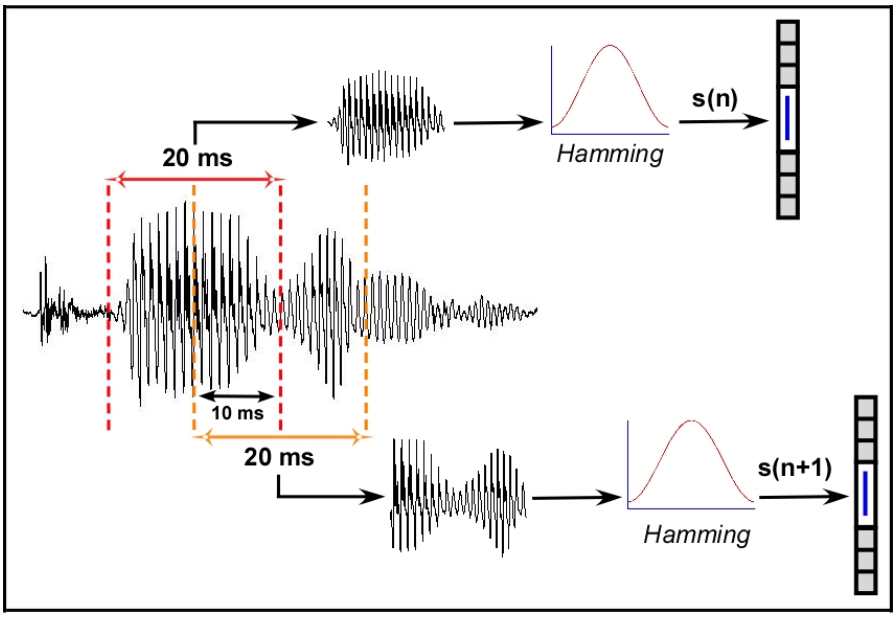


FIG. 3.2 – *Toutes les 10ms, extraction des trames de 20ms sur lesquelles le signal peut en général être considéré comme quasi stationnaire, avec un recouvrement de 50 % entre deux fenêtres successives temporelles. Chaque trame est pondérée par une fenêtre de type Hamming pour améliorer l'analyse et limiter les effets de bord.*

Comme le montre la figure 3.2, chaque signal est ensuite décomposé en trames de 20ms, extraites toutes les 10ms, sur lesquelles une fenêtre de pondération est appliquée. Différents types de pondération existent : Hamming, Hanning, Blackman, Kaiser-Bessel. Ici, une fenêtre de type Hamming est utilisée et décrite par :

$$Hamming(i) = 0.54 - 0.46\ cos(\frac{2\pi i}{N}) \quad avec \quad i \in [0, N-1] \tag{3.2}$$

où N est la taille de la fenêtre en nombre d'échantillons de signal. L'application sur chaque trame d'une fenêtre de pondération a pour but d'une part, de concentrer la répartition de l'énergie sur les basses fréquences et d'autre part, d'amoindrir les fortes variations du signal sur les bords de la fenêtre, variations qui entraînent une mauvaise estimation des coefficients du filtre si elles n'étaient pas atténuées[4]. En effet, le découpage du signal produit des discontinuités[5] aux frontières des trames susceptibles d'introduire des artefacts dans les spectres [Harris, 1978].

3.1.2 L'analyse par prédiction linéaire

La méthode d'analyse par prédiction linéaire se fonde sur les connaissances de production de la parole en partant de l'hypothèse d'un modèle linéaire [Fant, 1960]. Le conduit vocal est modélisé par un filtre auto-régressif (AR) excité soit par un bruit blanc

[4] réduction des effets de bord résultant de la segmentation en trames
[5] qui se manifestent par des lobes secondaires dans le spectre

(pour les frications), soit par un peigne de Dirac (pour les sons voisés).

L'analyse par prédiction linéaire [Markel & Gray, 1976] suppose que les échantillons $s(n)$ du signal de parole sont corrélés et que le signal peut être prédit par une combinaison linéaire des p échantillons précédents :

$$\hat{s}(n) = \sum_{i=1}^{p} a_i \, s(n-i) \tag{3.3}$$

où $\hat{s}(n)$ est la valeur du signal de prédiction, $s(n-i)$ sont les valeurs des observations précédentes, p est l'ordre de prédiction et a_i sont les coefficients de prédiction.

La différence entre le signal $s(n)$ et sa valeur prédite $\hat{s}(n)$ constitue l'erreur de prédiction (ou résidu) du modèle :

$$e(n) = s(n) - \hat{s}(n) \tag{3.4}$$

De l'équation 3.4, on voit que chaque échantillon $s(n)$ peut s'écrire sous la forme d'une combinaison linéaire des p échantillons précédents, à laquelle s'ajoute un bruit blanc gaussien $e(n)$ de variance σ^2 :

$$s(n) = e(n) + \sum_{i=1}^{p} a_i \, s(n-i) \tag{3.5}$$

On cherche donc les coefficients a_i qui minimisent la puissance de l'erreur de prédiction $E[e(n)^2]$. Cette minimisation par rapport aux a_i conduit au système d'équations dit de Yule-Walker réduites de $p+1$ équations et de $p+1$ inconnues, dans lequel le vecteur des coefficients $A = (1, a_1, ..., a_p)^T$ est solution :

$$R \, A = (\sigma^2, 0, ..., 0)^T \tag{3.6}$$

où R est la matrice d'autocorrélation du signal et correspond à une matrice de Tœplitz constituée des $p+1$ premiers coefficients d'autocorrélation :

$$R = [R_{ij}]_{1 \leq i,j \leq p+1} \qquad avec \; R_{ij} = \hat{r}_{|i-j|} \tag{3.7}$$

Dans l'ouvrage de Boite & Kunt (1987) ou de Moreau (1995), on peut trouver l'algorithme développé par Levinson en 1947 puis modifié en 1960 par Durbin, qui permet de calculer rapidement les coefficients de prédiction $\{a_i\}_{i=1,...,p}$ en résolvant le système 3.6. Le toolkit SPRO [Gravier, 2003] utilisé dans ces travaux repose sur cette méthode.

Plusieurs approches sont proposées dans la littérature pour estimer l'enveloppe spectrale par une analyse LPC. Deux méthodes sont classiquement utilisées :

1. la méthode de l'autocorrélation (présentée ci-dessus) ;
2. la méthode de la covariance.

Pour plus d'informations, le lecteur pourra consulter [Tubach, 1989; Makhoul, 1975].

3.1.3 L'analyse fréquentielle

Le signal de parole évoluant temporellement, l'analyse fréquentielle du signal de parole permet d'obtenir une représentation « temps-fréquence » du signal afin d'en extraire des informations - autres que temporelles - jugées pertinentes pour la tâche de reconnaissance visée. Pour cela, des spectres à court terme sont calculés sur des portions de signal de faibles durées (également appelées trames) partant de l'hypothèse que le signal y est « quasi stationnaire » *i.e.* de l'ordre de la dizaine de millisecondes. Cette section détaille le principe de l'analyse.

Chaque trame (de taille $\Delta t = 20\ ms$ soit 320 points) est suivie d'une analyse fréquentielle par application d'une Transformée de Fourier Discrète (TFD sur N = 512 points) sur les trames temporelles suivant l'équation :

$$S(f) = \sum_{n=0}^{N-1} s(n)\ e^{-\,i\,2\,\pi\,f\,n} \tag{3.8}$$

Le nombre de points du signal issu de la fenêtre de pondération détermine la qualité de l'analyse. Les valeurs de la trame sont complétées par des zéros (méthode « zero-padding ») à hauteur des 512 points. A l'issue de cette transformation, 256 amplitudes $a(n)$ et 256 phases $p(n)$ sont calculées à partir des valeurs complexes fournies par la FFT[6] suivant l'équation :

$$a(n) = \sqrt{Re^2(n) + Im^2(n)} \quad et \quad p(n) = \arctan(\frac{Im(n)}{Re(n)}) \quad avec\ n\ =\ 0\ ..\ N\ -\ 1 \tag{3.9}$$

On obtient alors un spectogramme à « bande étroite » avec une résolution fréquentielle de $\Delta f = 16000/512 = 31,25\ Hz$ sur une bande de fréquences [7] de [0-8000]Hz et une résolution temporelle de $\Delta t = 20\ ms$.

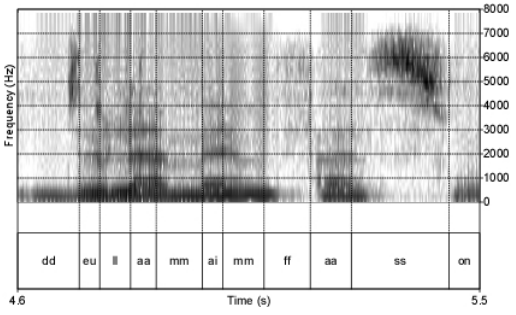


FIG. 3.3 – *Spectrogramme à « large bande » ($\Delta t = 4\ ms$) de la phrase « de la même façon ».*

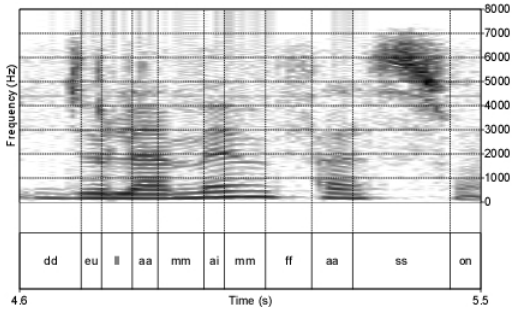


FIG. 3.4 – *Spectrogramme à « bande étroite » ($\Delta t = 20\ ms$) de la phrase « de la même façon ».*

La caractéristique du spectrogramme à « bande étroite » est de bien faire apparaître les composantes harmoniques du signal, à la différence d'un spectrogramme à « large

[6] Fast Fourier Transform : terme générique qui regroupe les différents algorithmes de calcul d'une TFD
[7] Limite de Shanon ou critère de Nyquist : la fréquence la plus élevée = $Fe/2$ où $Fe = 16Khz$

bande » (par exemple, avec $\Delta f = 250\ Hz$ et $\Delta t = 4\ ms$) qui rend compte des événements temporels brefs (comme par exemple, dans les explosions de consonnes occlusives ou dans les transitions entre phonèmes), au détriment de la résolution fréquentielle.

Comme cela est illustré sur les figures 3.3 et 3.4 qui représentent les spectrogrammes en « large bande » et « bande étroite » de la phrase « de la même façon », le compromis entre la résolution temporelle et la résolution fréquentielle y apparaît déterminant selon celle que l'on souhaite privilégier.

En résumé, pour chaque trame extraite du signal de parole, 256 valeurs d'amplitudes sont calculées sur une plage de fréquences de [0-8000]Hz avec un pas de discrétisation de $31,25\ Hz$.

3.1.4 L'analyse en banc de filtres

L'analyse en banc de filtres est une méthode d'estimation de l'enveloppe spectrale du signal simple et peu coûteuse. Le principe est d'obtenir une approximation du spectre de puissance de Fourier par l'évaluation de l'énergie du signal dans différentes sous-bandes contiguës découpées dans la bande passante utile.

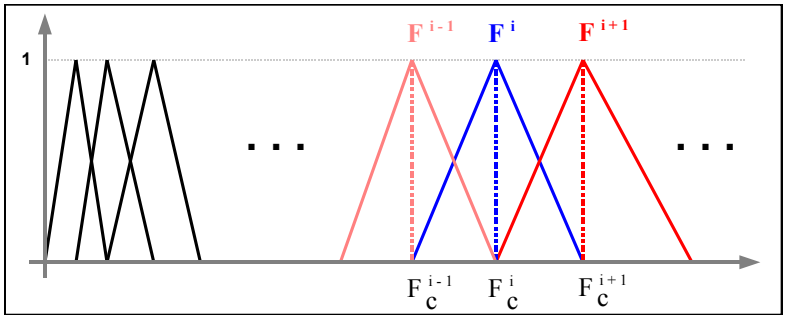


FIG. 3.5 – *Structure du banc de filtres distribués selon sur une échelle de Mel.*

Le banc de filtres (figure 3.5) est généralement constitué de 24 filtres triangulaires [8] dont les fréquences centrales sont réparties sur une échelle linéaire ou une échelle logarithmique issue de la psycho-acoutique comme Mel ou Bark. Chaque filtre F^i couvre un intervalle fréquentiel $[F^i_{min}, F^i_{max}] = [F^{i-1}_c, F^{i+1}_c]$, avec une valeur maximale de 1 pour sa propre fréquence centrale. Pour une valeur de fréquence donnée, il y a au plus 2 filtres avec des valeurs non nulles et la somme de ces valeurs vaut toujours 1.

Ainsi pour chaque trame issue de l'analyse fréquentielle, nous obtenons un vecteur de 24 coefficients. Chacun de ces coefficients correspond finalement au logarithme de l'énergie du signal contenue dans les 24 canaux de fréquences. Un sous-ensemble de coefficients extrait du vecteur de 24 paramètres de l'analyse en banc de filtres est directement interprétable en terme de sous-bande fréquentielle.

[8] le i^{ieme} filtre est noté F^i et sa fréquence centrale F^i_c avec $i = 1$ .. 24

L'échelle de Mel qui tient compte des particularités de l'oreille humaine [Davis & Mermelstein, 1980], permet donc de rapprocher l'analyse en banc de filtres de la perception de l'audition humaine.

L'échelle de fréquence Mel est définie par : $F_{Mel} = 2595\ log_{10}(1 + \frac{F_{Hz}}{700})$

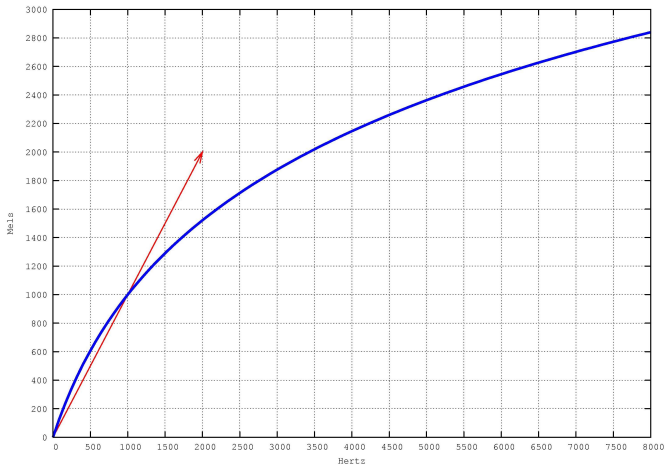


FIG. 3.6 – *Echelle de Mel : correspondance entre les fréquences en Hz et en Mel, la courbe peut être considérée comme quasi-linéaire dans les basses fréquences ($\leq$ 1000Hz) et logarithmique dans les hautes fréquences ($\geq$ 1000Hz).*

Comme cela apparaît sur la figure 3.6, l'échelle de Mel a comme particularité d'être quasi-linéaire en basse fréquence ($\leq$ 1000Hz) et logarithmique en haute fréquence ($\geq$ 1000Hz).

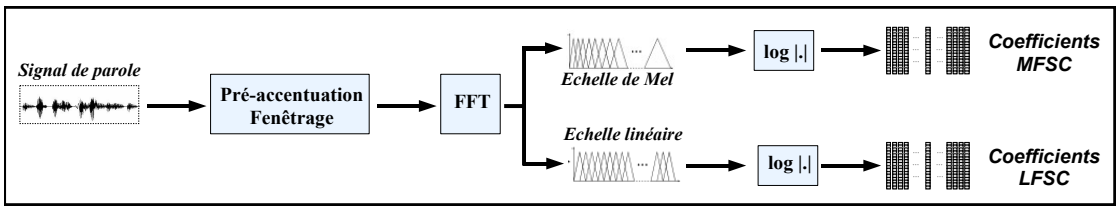


FIG. 3.7 – *Analyse en banc de filtres : paramètres MFSC et LFSC.*

Comme l'illustre la figure 3.7, les coefficients spectraux sont obtenus par le calcul d'une transformée de Fourier à court terme sur chaque trame, suivie d'une analyse en banc de 24 filtres triangulaires dont les fréquences centrales sont réparties sur une échelle :

1. Mel résultant en un vecteur de 24 coefficients MFSC
2. linéaire résultant en un vecteur de 24 coefficients LFSC

3.1.5 L'analyse cepstrale

Dans le cadre théorique du modèle « source-filtre », l'intérêt majeur de la représentation cepstrale est de dissocier l'excitation glottique (la source) des résonances du conduit vocal (le filtre). Ainsi dans un modèle acoustique de production « source-filtre », un signal de parole s_n peut être considéré comme issu de la convolution :

- d'une source sonore x_n (le fondamental F_0)
- du conduit vocal h_n (les fréquences de résonance formantique).

Le résultat de cette convolution s'écrit :

$$s_n = x_n \otimes h_n \tag{3.10}$$

Afin de séparer la contribution de la source et du conduit (déconvolution), il est plus intéressant de transformer l'équation 3.10 en une opération de somme par homomorphisme *i.e.* passage dans le domaine « log-spectral » qui est le principe même de l'analyse cepstrale [Rabiner & Schafer, 1978] :

$$\hat{s}_n = \hat{x}_n \oplus \hat{h}_n \tag{3.11}$$

Cette propriété de transformer le produit $\otimes$ en une somme $\oplus$, rend le cepstre très intéressant dans la mesure où il transforme un problème non linéaire en un problème linéaire.

A partir d'une analyse LPC

Il est possible de transformer directement les coefficients de prédiction linéaire en coefficients cepstraux. Le toolkit SPRO [Gravier, 2003] implémente la méthode définie dans [Miet, 2001] qui calcule directement les coefficients cepstraux LPCC à partir des coefficients LPC selon la formule suivante :

$$LPCC_i = -\ LPC_i + \frac{1}{i} \sum_{j=1}^{i-1} (i-j)\ LPC_j\ LPCC_{i-j} \quad avec \quad i \in [1, N] \tag{3.12}$$

où N est le nombre de coefficients LPC.

A partir d'une analyse en banc de filtres

En pratique, les coefficients cepstraux c_i peuvent être obtenus à partir des énergies e_j issues du banc de filtres par la Transformée en Cosinus Discrète Inverse[9] donnée par :

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} e_j\ cos(\frac{\pi\ i\ (j\ -\ 0.5)}{N}) \quad avec \quad i \in [1, M]\ \ et\ \ M \leqslant N \tag{3.13}$$

où M est le nombre de coefficients cepstraux, N est le nombre de canaux du banc de filtres.

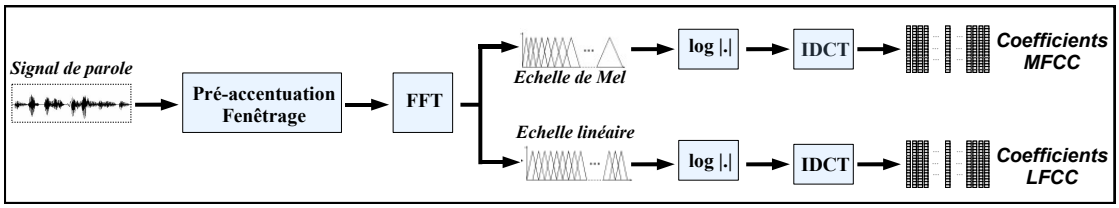


FIG. 3.8 – *Analyse en banc de filtres : paramètres MFCC et LFCC.*

Comme le montre la figure 3.8, les coefficients cepstraux sont obtenus par le calcul d'une transformée de Fourier à court terme sur chaque trame, suivie d'une analyse en banc de 24 filtres triangulaires dont les fréquences centrales sont réparties sur une échelle linéraire ou de Mel, et finalement par application d'une Transformée en Cosinus Discrète Inverse :

1. échelle de Mel résultant en un vecteur de N coefficients MFCC
2. échelle linéaire résultant en un vecteur de N coefficients LFCC

où ici $N = 16$.

3.1.6 Les paramètres dynamiques

En RAL, par opposition aux informations de nature statique décrites précédemment, les informations dynamiques véhiculées par le signal de parole sont considérées comme une source potentielle d'informations caractéristiques du locuteur. Les informations de nature dynamique reflètent les phénomènes de co-articulation, les trajectoires formantiques ainsi que des informations temporelles (vitesse d'élocution, distribution des pauses, ...). Une méthode d'extraction des informations dynamiques repose sur le calcul des coefficients delta [Fredouille, 2000] qui peuvent se calculer de la manière suivante :

$$\Delta C_{k,t} = \frac{\sum_{\theta=1}^{\Theta} \theta \, (C_{k,t+\theta} - C_{k,t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2} \tag{3.14}$$

où $\Delta C_{k,t}$ est le coefficient delta (dérivée de 1er ordre) du coefficient d'indice k de la trame t, calculé en utilisant la fenêtre centrée de largeur [-Θ , +Θ] trames. Généralement, Θ est une valeur comprise entre 2 et 4 afin d'obtenir un contexte de 5, 7 ou 9 trames. La même formule peut être appliquée aux coefficients delta pour obtenir les coefficients delta-delta (dérivée de 2^{e} ordre)[10]. Ainsi à chaque trame, il sera associé aux coefficients statiques une information sur la dynamique temporelle du signal à court terme *i.e.* leurs valeurs dynamiques locales.

Furui (1986) a montré que les performances en RAP peuvent être améliorées en ajoutant les dérivées temporelles des coefficients statiques aux vecteurs de paramètres initiaux.

[9] Inverse Discrete Cosine Transform en anglais (IDCT)
[10] une autre méthode a proposé leur calcul par approximation [Deller et al., 1999]

Les travaux de Misra et al. (2003) ont montré que les gains ainsi observés sont généralement d'autant plus importants que les données sont bruitées ; de même, avec Yang et al. (2005) pour qui « l'information dynamique est plus robuste au bruit additif que son équivalent statique ». Des observations similaires ont été reportées en RAL [Fredouille & Bonastre, 1998; Besacier, 1998].

Dans cette approche d'analyse de l'information de nature dynamique, nous proposons d'utiliser les dérivées de 3^{e} ordre (les coefficients delta-delta-delta) qui correspondent en physique à la dérivée de l'accélération par rapport au temps. Egalement utilisée en RAL [Kajarekar, 2005], cette grandeur physique est particulièrement utilisée dans des secteurs tels que le ferroviaire faisant intervenir le confort dans les trains, la limitation de la brutalité des accélérations et des freinages des métros, ... En effet, sachant qu'une accélération génère des efforts, la variation temporelle de l'accélération (Jerk ou secousse) génére des variations d'efforts pouvant provoquer de l'inconfort pour les occupants d'un véhicule par exemple. Ainsi lorsque les phénomènes de résonances vibratoires sont un problème à prendre en compte, le Jerk peut également être consideré comme un critère important.

3.1.7 Le post-traitement acoustique

Durant l'enregistrement de signaux acoustiques, des effets environnementaux peuvent dégrader fortement les résultats en RAP et en RAL. Pour réduire les effets négatifs dus à certaines sources de variabilité du signal acoustique (les conditions d'enregistrement, le canal de transmission, ...), et rendre les paramètres plus robustes au bruit, un traitement de normalisation est appliqué sur les vecteurs acoustiques. Différentes techniques de normalisation sont proposées dans la littérature : CMS [Furui, 1981], normalisation de la variance [Openshaw & Mason, 1994], « feature warping » ou « Gaussianisation » [Pelecanos & Sridharan, 2001], « feature mapping » [Reynolds, 2003]. Ici, une normalisation par retrait de la moyenne et de la variance est utilisée.

Si l'on considère le signal s_n et $s_n[t]$ le vecteur de paramètres d'indice t avec $t \in [1, T]$, la normalisation par soustraction de la moyenne μ_n est définie par :

$$s_n^{'}[t] = s_n[t] - \mu_n \quad avec \quad \mu_n = \frac{1}{T} \sum_{\tau=1}^{T} s_n[\tau] \tag{3.15}$$

Pour la normalisation par la variance σ_n, la règle devra respecter :

$$s_n^{''}[t] = \frac{s_n^{'}[t]}{\sigma_n} \quad avec \quad \sigma_n = \sqrt{\frac{1}{T} \sum_{\tau=1}^{T} (s_n^{'}[\tau])^2} \tag{3.16}$$

Ainsi les vecteurs de paramètres sont normalisés pour obtenir une distribution de moyenne 0 et de variance 1. La moyenne et la variance sont estimées uniquement sur les portions de parole contenues dans le signal nettoyé préalablement des portions de « non-parole » (voir les sections 2.1 et 2.2 consacrées à l'alignement des corpus CVD et BREF).

3.2 La modélisation statistique

Contexte général en RAL

Le système adapté à la classification des voix dysphoniques et dérivé d'un système classique de RAL, est basé sur une modélisation statistique reposant sur un mélange de gaussiennes (GMM). Depuis son introduction par Reynolds (1992), la modélisation par GMM reste «l'état de l'art» de la RAL en terme de modélisation du locuteur en mode indépendant du texte, malgré l'émergence d'autres techniques telles que les SVM ou des systèmes hybrides SVM/GMM [Scheffer, 2006].

Dans cette approche, un locuteur est modélisé par un GMM qui est une somme pondérée de M distributions gaussiennes multi-dimensionnelles, chacune caractérisée par un vecteur moyen $\overline{x}$ (dimension d), une matrice de covariance Σ généralement diagonale (dimension d x d) et le poids p de la composante gaussienne dans le mélange [Reynolds, 1995]. Un modèle GMM est appris avec l'ensemble des données d'apprentissage en estimant les paramètres ($\overline{x}$, Σ, p) du GMM maximisant la vraisemblance des données d'apprentissage grâce à l'algorithme EM/ML [Dempster et al., 1977].

Suivant la loi de mélange de M gaussiennes, la densité de probabilité d'un vecteur y_t de dimension d s'exprime sous la forme :

$$p(y_t|X) = \sum_{i=1}^{M} p_i \, \mathcal{N}(y_t; \overline{x_i}, \Sigma_i) \quad sous\ la\ contrainte \quad \sum_{i=1}^{M} p_i = 1 \quad et \quad \forall_i : p_i \geq 0 \qquad (3.17)$$

où l'on trouve :

- p_i le poids de la i^{ieme} gaussienne,
- $\mathcal{N}(y_t; \overline{x_i}, \Sigma_i)$ la loi normale de la i^{ieme} gaussienne de moyenne $\overline{x_i}$ et de variance Σ_i,
- $X = \{\overline{x}, \Sigma, p\}$ l'ensemble des paramètres du GMM.

Ainsi la vraisemblance pour que le signal Y (constitué de T vecteurs) soit produit par le modèle X est donnée par :

$$p(Y|X) = \prod_{t=1}^{T} p(y_t|X) \qquad (3.18)$$

Remarque : les modèles de mélange de gaussiennes peuvent être considérés comme des modèles HMM à un seul état où la fonction de densité est un mélange de gaussiennes.

Classiquement, deux phases d'apprentissage sont nécessaires en RAL pour pallier le manque fréquent de données d'apprentissage disponible pour un locuteur [Bimbot et al., 2004] :

1. apprentissage d'un modèle de parole générique[11] estimé par l'algorithme EM/ML sur une grande population de locuteurs ;
2. apprentissage d'un modèle de locuteur, dérivé du modèle générique en appliquant des techniques d'adaptation comme par exemple la technique du *Maximum A Posteriori* (MAP).

Selon le critère du Maximum de Vraisemblance (ML), l'estimation des paramètres X d'un modèle à partir des données Y que ce modèle est censé représenter, repose sur l'estimateur ML qui maximise la probabilité des données dans le modèle selon :

$$X_{ML} = \underset{X}{argmax}\ p(Y|X) \tag{3.19}$$

Introduite par Gauvain & Lee (1994), l'adaptation MAP vise à obtenir une estimation bayésienne pour les paramètres du modèle grâce à l'apport de nouvelles données. Cette technique est devenue la référence en VAL [Reynolds et al., 2000]. Elle consiste à estimer les paramètres « les plus probables » du modèle X compte tenu des observations Y en maximisant la vraisemblance *a posteriori* $p(X|Y)$ du modèle X par :

$$X_{MAP} = \underset{X}{argmax}\ p(X|Y) = \underset{X}{argmax}\ p(Y|X)\ p(X) \tag{3.20}$$

L'estimation des nouveaux paramètres X_{MAP} selon le critère MAP peut encore s'écrire :

$$X_{MAP} = \eta X_{app}\ +\ (1-\eta)X_{UBM} \tag{3.21}$$

avec X_{app} est le modèle obtenu par une itération sur les données d'apprentissage, X_{UBM} est le modèle générique UBM (Universal Background Model) et η est une constante définie empiriquement.

Contexte pathologique

Dans le contexte pathologique comme le montre la figure 3.9, un modèle ne correspond plus à un locuteur mais à un niveau de sévérité de dysphonie. Il sera appelé **modèle de grade** G_g avec $g \in \{0,1,2,3\}$. Le modèle de grade G_g est appris sur l'ensemble des voix évaluées perceptivement dans le grade g.

Le modèle générique sera appris avec le corpus d'apprentissage BREF (décrit en 2.2). Les modèles de grade seront adaptés à partir du modèle générique et du corpus CVD utilisé comme corpus d'adaptation (décrit en 2.1).

[11] également connu sous la dénomination UBM : Universal Background Model

De plus, on s'assurera que les voix utilisées pour l'apprentissage des modèles de grade sont exclues des jeux de tests afin de différencier la détection de la pathologie de la reconnaissance du locuteur *i.e.* mise en œuvre de la technique « leave_x_out » décrite en 2.3.

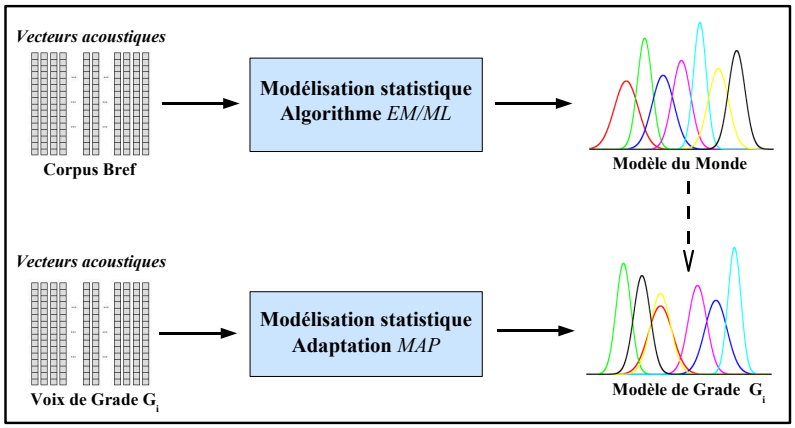


FIG. 3.9 – *Modélisation statistique du système de RAL adapté à la classification des voix pathologiques.*

Tous les modèles GMM sont composés ici de 128 composantes gaussiennes avec des matrices de covariance diagonales. De plus, l'adaptation des modèles de grade est effectuée en appliquant la transformation linéaire (équation 3.21) uniquement sur les moyennes des gaussiennes (les variances et les poids ne sont pas adaptés) avec une valeur de 14 pour le paramètre « regulator factor ». Pour plus de détails, le lecteur pourra se référer à [Reynolds et al., 2000].

3.3 La décision

Contexte général en RAL

Lors de la phase de test, une mesure de similarité entre des vecteurs acoustiques y_t issus d'un signal et un modèle X est calculée suivant :

$$L(y_t|X) = \sum_{i=1}^{M} p_i \, L_i(y_t) \tag{3.22}$$

où $L_i(y_t)$ est la vraisemblance du signal y_t par rapport à la gaussienne i, M le nombre de gaussiennes et p_i le poids de la gaussienne i.

La vraisemblance $L_i(y_t)$ des vecteurs y_t de dimension d s'exprime par :

$$L_i(y_t) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \, e^{-\frac{1}{2}(y_t-\overline{x_i})^T(\Sigma_i)^{-1}(y_t-\overline{x_i})} \tag{3.23}$$

Selon l'équation 3.18, la Log-Vraisemblance moyenne obtenue par un signal Y composé de T vecteurs sur le modèle X correspond à :

$$\overline{LL(Y|X)} = \frac{1}{T}\left(\sum_{t=1}^{T} log(\sum_{i=1}^{M} p_i \, L_i(y_t)) \right) \tag{3.24}$$

Comme précisé en introduction, le processus de décision est dépendant de la tâche visée. En VAL, la décision correspondra à une décision binaire « *Acceptation* » ou « *Rejet* » suivant que la mesure de ressemblance est supérieure à un seuil de décision. Par contre en IAL (milieu fermé), elle correspondra à une décision « 1 parmi N » basée sur le « *Maximum de Vraisemblance* » désignant le locuteur le plus probable parmi les N connus du système.

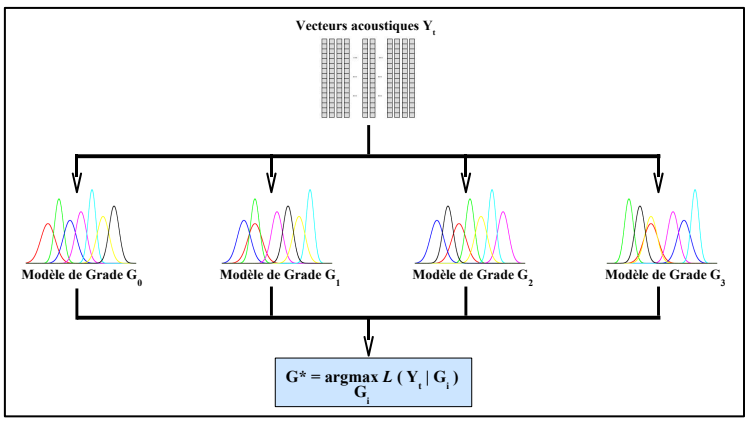


FIG. 3.10 – *Phase de test du système de RAL adapté à la classification des voix pathologiques.*

Contexte pathologique

Dans le contexte pathologique (figure 3.10), la décision correspondra au grade g du modèle G_g pour lequel la plus grande vraisemblance est obtenue. Cette définition de la décision est proche de celle d'IAL. On dira que le système a classé la voix du locuteur Y dans le grade g.

3.4 Conclusion

Dans cette section, le système de RAL adapté à la classification des voix dysphoniques a été présenté. Les trois phases classiques - paramétrisation, modélisation et décision - d'un système dédié à la reconnaissance du locuteur, ont été décrites en spécifiant les adaptations apportées à la problématique des pathologies vocales abordée dans cette thèse.

Il est important de préciser que l'adaptation ne concerne pas la détermination des valeurs des paramètres du système permettant d'optimiser ses performances en terme d'amélioration des taux de reconnaissance. Les paramètres pouvant être facteurs d'amélioration de performances sont : le nombre de gaussiennes par GMM, les seuils de Flooring et/ou de Ceiling, l'adaptation MAP de la moyenne et/ou de la variance et/ou du poids, une distribution GMM diagonale ou pleine, ... L'étape préliminaire a consisté à préparer un système de base avec une configuration minimale permettant d'obtenir des résultats suffisants pour observer les phénomènes pertinents de la dysphonie.

Chapitre 4

L'évaluation du système

Sommaire

Résumé

Evaluer une voix dysphonique selon une échelle ordinale de 4 grades provient de besoins cliniques d'évaluation globale, rapide et simple. Cette catégorisation simpliste n'est pas sans problème si l'on souhaite affiner le niveau de sévérité de la dysphonie ou si l'on considère la subjectivité du jugement perceptif. Dans ce chapitre, nous présentons 3 tâches de classification - 2-Grades, 4-Grades et 7-Grades - élaborées pour évaluer le système automatique selon différentes configurations du grade de la dysphonie.

Dans ce chapitre, nous nous intéressons à trois différentes tâches de classification mises en œuvre afin d'évaluer les performances du système adapté au contexte pathologique. Tout comme précisé en 3.4, ce n'est pas la « performance » dans le sens de « meilleure classification » qui est recherchée ici, mais plutôt de comparer les résultats du système automatique selon différentes configurations de l'échelle ordinale du grade de dysphonie.

Les 3 tâches sont présentées sous les appellations suivantes :

- la classification 2-Grades ou Control/Patho ;
- la classification 4-Grades ou par grade GRBAS ;
- la classification 7-Grades ou par grade intermédiaire.

Pour chacune, le protocole expérimental sera décrit afin de vérifier leur robustesse d'un point de vue statistique. Les règles relatives à l'apprentissage des modèles de grade qui devront être respectées au cours d'une expérience, sont les suivantes :

R1. tous les modèles doivent être entraînés avec un même nombre de locutrices ;
R2. utilisation de la technique « leave_x_out » (décrite en 2.3) ;
R3. utilisation de l'ensemble des locutrices du corpus CVD ;
R4. un modèle « multi-grade » doit être appris avec le même nombre de locutrices de chaque grade qui le constitue.

Finalement, une évaluation expérimentale sera présentée pour chacune des classifications. Pour cela, une même paramétrisation acoustique sera utilisée afin de pouvoir comparer la pertinence des différentes tâches. Celle-ci reposera sur une analyse spectrale de type 72MFSC avec 24 coefficients statiques auxquels seront associés les coefficients dérivés de 1^{er} ordre (24Δ) et de 2^{e} ordre ($24\Delta\Delta$). La segmentation des signaux acoustiques correspondra à la version AP2 (pour plus de détails, voir la section 2.5).

4.1 La classification 2-Grades ou Control/Patho

Cette classification « bi-classe » nous a permis de valider la pertinence du système automatique adapté à la reconnaissance des voix pathologiques.

Protocole expérimental

Cette première classification consiste à observer si un système de RAL, adapté à la reconnaissance des voix pathologiques, réagit favorablement à la classification binaire *i.e.* détecter si une voix donnée est reconnue en tant que « *voix normale* » ou « *voix dysphonique* ». Par conséquent, deux modèles de grade doivent être estimés, $G_{Control}$ et G_{Patho} correspondant respectivement au modèle des voix de contrôle (Grade 0) de l'échelle GRBAS décrite en 1.5.1 et au modèle des voix dysphoniques (Grades 1, 2 et 3).

Chacun des modèles sera appris avec 18 voix. Ce nombre se justifie principalement par la règle R4 qui conduit à apprendre les modèles G_{Patho} avec 6 locutrices de chaque grade pathologique (Grades 1, 2 et 3). Concernant les modèles $G_{Control}$, toutes les voix seront utilisées de manière cyclique conduisant à l'apprentissage de $C_{20}^{18} = 190$ modèles. Par contre, en raison du nombre combinatoire de modèles pathologiques G_{Patho} pouvant être estimé sur la base de 3 x 6 voix pathologiques, une limite a été fixée ici à 9 modèles. On s'assurera que chaque locutrice dysphonique ne participe pas à l'apprentissage d'au moins 6 modèles G_{Patho} parmi les 9 afin que chacune des voix du corpus CVD puisse être testée sur 6 modèles $G_{Control}$ et 6 modèles G_{Patho}.

Lors de la phase de test, chaque voix y_t sera donc comparée avec :

- 6 modèles $G_{Control}$ appris chacun à partir de 18 voix normales et $y_t \notin G_{Control}$
- 6 modèles G_{Patho} appris chacun à partir de 18 voix dysphoniques équitablement réparties entre les 3 grades pathologiques (6 voix par grade) et $y_t \notin G_{Patho}$

Le tableau 4.1 donne des informations quantitatives sur les tests effectués par l'ensemble des locutrices du corpus CVD lors d'une classification Control/Patho. Chaque ligne LG_g correspond au nombre de tests effectués par les locutrices de grade g sur les modèles *Control* et *Patho*.

	$MG_{Control}$	MG_{Patho}	**Total**
LG0	**120**	120	240
LG1	120	**120**	240
LG2	120	**120**	240
LG3	120	**120**	240
Total	480	480	**960**

TAB. 4.1 – *Nombre de tests effectués lors d'une classification 2-Grades (Control/Patho)*

A l'issue de ces comparaisons, les moyennes des vraisemblances des tests *Control* (6 modèles $G_{Control}$) et *Patho* (6 modèles G_{Patho}) sont calculées et comparées pour fournir une unique décision pour la voix y_t :

$$\widetilde{g} = \underset{g}{argmax}\ L(y_t|G_g) \qquad avec\ g\ \in\ \{\ Control\ ,\ Patho\ \} \tag{4.1}$$

Evaluation expérimentale

Le tableau 4.2 présente les résultats de la classification Control/Patho. La colonne « IC » désigne l'intervalle de confiance accordé aux résultats et se calcule suivant l'équation définit en 2.3. Cette information figurera dans l'ensemble des tableaux du chapitre.

Control	**Patho**	**Global**	
% TCC (nb/20)	% TCC (nb/60)	% TCC (nb/80)	± IC
95.0 (19)	91.7 (55)	92.50 (74)	5.8

TAB. 4.2 – *Résultats de la classification Control/Patho en terme de % TCC - Paramétrisation 72MFSC (24c + 24Δ + 24ΔΔ) - Bande totale [0-8000]Hz*

Nous obtenons un score Global de 92.50 % qui constitue un résultat très encourageant et prometteur, nous permettant de constater qu'un système de RAL, adapté au contexte pathologique, réagit favorablement à la classification binaire.

72MFSC (24c + 24Δ + 24ΔΔ) en [0-8000]Hz

	S_Control	**S_Patho**
P_G0	**95.0 % (19)**	5.0 % (1)
P_G1	20.0 % (4)	**80.0 % (16)**
P_G2	5.0 % (1)	**95.0 % (19)**
P_G3	0.0 % (0)	**100.0 % (20)**

TAB. 4.3 – *Matrice de confusion de la classification Control/Patho*

La matrice de confusion 4.3 permet d'identifier les erreurs de classification suivant le grade de chaque voix :

1. la totalité des voix de grade 3 est bien classée (100.0 %)
2. les voix de grade 0 et de grade 2 sont majoritairement bien classées (95.0 %)
3. le grade 1 présente le plus fort taux de confusion malgré un score acceptable (80.0 %)

Des rapprochements avec l'analyse perceptive peuvent être constatés :

• **Capacité à distinguer les voix normales / dysphoniques :** 95.0 % et 91.7 % resp.

Tout auditeur, expérimenté ou non à l'évaluation de la voix dysphonique, est plus habitué à entendre des voix normales que pathologiques dans la vie quotidienne ; il se forge ainsi un prototype « voix normale » stable, lui permettant de distinguer sans difficulté une « voix normale / voix anormale », indépendamment du niveau d'expérience [Kreiman et al., 1992, 1993; De Bodt et al., 1997].

• **Discrimination croissante des grades :** 80.0 % (G1) $\Rightarrow$ 95.0 % (G2) $\Rightarrow$ 100.0 % (G3)

La fiabilité des auditeurs augmente avec la sévérité de la pathologie ; la distance entre le prototype « voix normale » et la voix dysphonique est d'autant plus grande que la sévérité de la pathologie est importante [Rabinov et al., 1995].

4.2 La classification 4-Grades ou par grade GRBAS

Cette classification basée sur le grade global de l'échelle GRBAS de Hirano (1981) constitue l'objectif principal de cette thèse en terme d'évaluation objective des voix dysphoniques selon la sévérité du trouble vocal.

Protocole expérimental

Son principe consiste en une classification par grade selon l'échelle GRBAS [Hirano, 1981] afin d'analyser le comportement du système ; l'ensemble des voix est testé à travers le système et les résultats, comparés à ceux de l'analyse perceptive, permettent de mesurer les performances de la méthode objective. En d'autres termes, il s'agit de classer une voix suivant les 4 niveaux du grade global de l'échelle GRBAS. Par conséquence, quatre modèles de grade G_g sont à estimer avec $g \in \{0, 1, 2, 3\}$.

Chacun des modèles de grade sera appris avec 19 voix de même grade. Ce nombre se justifie par la mise en œuvre de la technique « leave_one_out » dont le principe est de retirer une locutrice (de test) de l'ensemble d'apprentissage, à chaque entraînement d'un modèle. Dans ces conditions, 20 modèles de 19 voix seront appris pour chaque grade soit un total de 80 modèles. Chaque locutrice sera testée sur le modèle de son grade perceptif dont elle aura été exclue lors de l'apprentissage et sur les 60 modèles des 3 autres grades.

Lors de la phase de test, chaque voix y_t de grade g sera comparée avec :

- 1 modèle G_g appris à partir de 19 voix de grade g et $y_t \notin G_g$
- 3 x 20 modèles $G_{\overline{g}}$ appris chacun à partir de 19 voix de grade $\overline{g}$ et $y_t \notin G_{\overline{g}}$ avec $\overline{g} \in \{0, 1, 2, 3\} - \{g\}$

Le tableau 4.4 donne des informations quantitatives sur les tests effectués par l'ensemble des locutrices du corpus CVD lors d'une classification 4-Grades. Chaque ligne LG_g correspond au nombre de tests effectués par les locutrices de grade g sur les modèles G_g avec $g \in \{0, 1, 2, 3\}$.

	MG0	**MG1**	**MG2**	**MG3**	**Total**
LG0	**20**	400	400	400	1220
LG1	400	**20**	400	400	1220
LG2	400	400	**20**	400	1220
LG3	400	400	400	**20**	1220
Total	1220	1220	1220	1220	**4880**

TAB. 4.4 – *Nombre de tests effectués lors d'une classification 4-Grades (4-G)*

A l'issue de ces comparaisons, les moyennes des vraisemblances des tests sur chaque grade (1 modèle G_g et 3 x 20 modèles $G_{\overline{g}}$) sont calculées et comparées pour fournir une

unique décision pour la voix y_t de grade g :

$$\widetilde{g} = \underset{g}{argmax}\ L(y_t|G_g) \qquad avec\ g\ \in\ \{\ 0\ ,\ 1\ ,\ 2\ ,\ 3\ \} \tag{4.2}$$

Evaluation expérimentale

Le tableau 4.5 affiche les résultats obtenus par la classification 4-Grades.

Grade 0	**Grade 1**	**Grade 2**	**Grade 3**	**Global**	
% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/80)	± IC
95.0 (19)	65.0 (13)	70.0 (14)	85.0 (17)	78.75 (63)	9.0

TAB. 4.5 – *Résultats de la classification 4-G en terme de % TCC - Paramétrisation 72MFSC (*$24c + 24\Delta + 24\Delta\Delta$*) - Bande totale [0-8000]Hz*

Nous obtenons un score Global de 78.75 % qui constitue un résultat acceptable malgré un intervalle de confiance (colonne « IC ») élevé de 9.0 %. Il apparaît donc que, même si les résultats sont très encourageants, ils restent encore insuffisants pour les praticiens pour son adoption dans une pratique clinique quotidienne.

Au regard des scores TCC obtenus par les différents grades, on constate que :

1. le grade 0 est le mieux reconnu (95.0 %)
2. les grades 1 et 2 présentent le plus fort taux de confusion (65.0 % et 70.0 % resp.)
3. le grade 3 obtient un résultat très acceptable (85.0 %)

Concernant les erreurs de classification reportées dans la matrice de confusion 4.6, un comportement similaire entre les grades 1 et 2 peut être observé avec 6 voix G1 classées en G2 et 5 voix G2 classées en G1.

72MFSC ($24c + 24\Delta + 24\Delta\Delta$) en [0-8000]Hz

	S_G0	**S_G1**	**S_G2**	**S_G3**
P_G0	**19**	0	1	0
P_G1	1	**13**	6	0
P_G2	0	5	**14**	1
P_G3	0	0	3	**17**

TAB. 4.6 – *Matrice de confusion de la classification 4-G*

Ce report mutuel de la majorité des erreurs de classification entre ces deux grades montre la difficulté du système à discriminer les voix « légèrement » et « modérément » dysphoniques.

4.3 La classification 7-Grades ou par grade intermédiaire

A la vue des résultats précédents, nous proposons une redéfinition affinée des frontières des grades dysphoniques en raison du caractère subjectif de la référence qu'ils constituent. Il s'agira alors de répondre à la question : « *une configuration plus fine des niveaux de sévérité de la dysphonie améliore-t-elle la discrimination objective du trouble vocal ?* ».

Juger une voix dysphonique en utilisant la catégorisation simplifiée de l'échelle GRBAS en 4 grades renforce le caractère subjectif de l'évaluation perceptive. La difficulté est d'autant plus importante si la voix est perçue comme « ambiguë ». Par exemple, considérons une voix classée dans le grade G2 avec un niveau 2.1 sur une échelle analogique. Cette voix n'est-elle pas plus proche d'une voix classée en G1 avec un niveau analogique 1.9 que d'une voix classée dans le même grade G2 avec un niveau analogique 2.9 ? Nous proposons ici de répondre au problème de frontière catégorielle en utilisant des grades intermédiaires offrant ainsi une meilleure granularité.

La classification en 7 grades est par conséquent une classification plus fine que celle en 4-Grades dans le sens où, aux 4 grades définis dans l'échelle GRBAS, il est rajouté 3 nouveaux grades ($G01$,$G12$,$G23$) nommés « intermédiaires » en raison de leur composition regroupant des locutrices de grades adjacents. Cette approche nous permettra d'observer si des erreurs de la classification 4-Grades peuvent être « *récupérées* » par une telle configuration de l'échelle GRBAS *i.e.* si des voix mal classées en 4-Grades, peuvent être classées dans un grade intermédiaire et adjacent à leur grade perceptif. Par conséquent, sept modèles de grade G_g sont à estimer avec $g \in \{0, 01, 1, 12, 2, 23, 3\}$.

Protocole expérimental

Chaque modèle de grade sera appris avec 18 voix. Ce nombre se justifie principalement par la règle R4 qui conduit à apprendre les modèles intermédiaires ($G01$, $G12$, $G23$) avec 9 locutrices de chaque grade adjacent. Ainsi il sera appris $C_{20}^{18} = 190$ modèles ($G0$, $G1$, $G2$, $G3$) et 18 modèles ($G01$, $G12$, $G23$). On s'assurera que chaque locutrice d'un grade intermédiaire ne participe pas à l'apprentissage d'au moins 9 modèles de même grade intermédiaire parmi les 18 afin que chaque voix soit testée sur 9 modèles ($G0$, $G1$, $G2$, $G3$) et 9 modèles ($G01$, $G12$, $G23$).

Durant la phase de test, chaque voix y_t sera donc comparée avec :

- 4 x 9 modèles G_g appris chacun à partir de 18 voix et $y_t \notin G_g$ avec $g \in \{0, 1, 2, 3\}$
- 3 x 9 modèles G_g appris chacun à partir de 18 voix équitablement réparties entre 2 grades adjacents (9 voix par grade) et $y_t \notin G_g$ avec $g \in \{01, 12, 23\}$

Le tableau 4.7 donne des informations quantitatives sur les tests effectués par l'ensemble des locutrices du corpus CVD lors d'une classification 7-Grades. Chaque ligne LG_g correspond au nombre de tests effectués par les locutrices de grade perceptif $g \in \{0, 1, 2, 3\}$ sur les modèles de grade intermédiaire G_g où $g \in \{0, 01, 1, 12, 2, 23, 3\}$.

	MG0	MG01	MG1	MG12	MG2	MG23	MG3	Total
LG0	**180**	**180**	180	180	180	180	180	1260
LG1	180	**180**	**180**	**180**	180	180	180	1260
LG2	180	180	180	**180**	**180**	**180**	180	1260
LG3	180	180	180	180	180	**180**	**180**	1260
Total	720	720	720	720	720	720	720	**5040**

TAB. 4.7 – *Nombre de tests effectués lors d'une classification 7-Grades (7-G)*

A l'issue de ces comparaisons, les moyennes des vraisemblances des tests obtenues sur chacun des grades (7 x 9 modèles G_g où $g \in \{0, 01, 1, 12, 2, 23, 3\}$) sont calculées et comparées pour fournir une unique décision pour la voix y_t :

$$\widetilde{g} = \underset{g}{argmax}\ L(y_t|G_g) \qquad avec\ g\ \in\ \{\ 0\ ,\ 01\ ,\ 1\ ,\ 12\ ,\ 2\ ,\ 23\ ,\ 3\ \} \tag{4.3}$$

Evaluation expérimentale

Ici, on comptabilisera les locutrices classées dans un grade intermédiaire et adjacent à leur grade perceptif comme des « succès ».

Le tableau 4.8 présente les résultats de la classification 7-Grades. Un score Global de 86.25 % est atteint ce qui constitue un résultat intermédiaire à ceux obtenus par les classifications Control/Patho et 4-Grades (92.50 % et 78.75 % respectivement).

Grade 0	Grade 1	Grade 2	Grade 3	Global	
% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/80)	± IC
95.0 (19)	70.0 (14)	80.0 (16)	100.0 (20)	86.25 (69)	7.6

TAB. 4.8 – *Résultats de la classification 7-G en terme de % TCC - Paramétrisation 24MFSC (*$24c + 24\Delta + 24\Delta\Delta$*) - Bande totale [0-8000]Hz*

Concernant les résultats obtenus dans les différents grades, on notera que :

1. le grade 0 est très bien reconnu (95.0 %)
2. les grades 1 et 2 présentent le plus fort taux de confusion (70.0 % et 80.0 % respectivement)
3. le grade 3 est entièrement reconnu (100.0 %)

La matrice de confusion 4.9 permet d'obtenir une représentation plus fine des erreurs de classification. On constate que les 3 grades intermédiaires ($G01$, $G12$, $G23$) récupèrent 16 voix soit 20 % du total des voix. Parmi celles-ci, 6 voix ont permis d'améliorer le score Global de l'expérience : 1 pour G1, 2 pour G2 et 3 pour G3.

72MFSC ($24c + 24\Delta + 24\Delta\Delta$) en [0-8000]Hz

	S_G0	**S_G01**	**S_G1**	**S_G12**	**S_G2**	**S_G23**	**S_G3**
P_G0	**19**	**0**	0	0	1	0	0
P_G1	1	**1**	**9**	**4**	3	2	0
P_G2	0	1	3	**1**	**13**	**2**	0
P_G3	0	0	0	0	0	**5**	**15**

TAB. *4.9* – *Matrice de confusion de la classification 7-G*

Il apparaît qu'une voix mal classée dans l'expérience 4-Grades, en raison d'une vraisemblance très proche entre plusieurs grades, puisse être classée dans un grade intermédiaire et adjacent à son grade perceptif. Cette expérience soulève la difficulté d'un système automatique à trancher entre 2 grades potentiels (pour lesquels la vraisemblance est très proche) à la différence d'un jury d'écoute.

4.4 Discussion

Choix de la tâche

Dans ce chapitre, les différentes expériences d'évaluation du système montrent que la redéfinition des frontières de grades de l'échelle perceptive GRBAS, qu'elle soit « binaire » avec une classification Control/Patho ou « plus fine » avec une classification par grades intermédiaires, permet une amélioration des performances de la classification 4-Grades avec des gains relatifs de +17.46 % et +9.52 % respectivement. Cette constatation corrobore la difficulté d'évaluer, perceptivement ou objectivement, une voix dysphonique avec une échelle ordinale à 4 niveaux qui, à l'origine, constituait une catégorisation simplifiée répondant partiellement aux besoins cliniques d'évaluation globale et rapide.

La bonne performance de la classification Control/Patho (TCC global de 92.5 %) sur un système non optimisé[1], apparaît comme une méthode intéressante pour la détection de la dysphonie à titre préventif, dans le cadre de la médecine du travail par exemple. Il en est de même pour la classification par grade intermédiaire (TCC global de 86.25 %) qui offre une alternative intéressante à la classification 4-Grades en redéfinissant plus finement l'échelle conventionnelle entachée d'une variabilité importante inter/intra auditeur. Les grades intermédiaires permettent d'obtenir une meilleure granularité afin de répondre au problème de frontières catégorielles, et ainsi d'appréhender l'évaluation de la dysphonie comme une manifestation graduelle de type analogique, plutôt que comme un phénomène catégoriel comme le laisserait entrevoir la notion de grade.

La question sous-jacente que l'on peut se poser est : « Comment utiliser les résultats des classifications Control/Patho et 7-Grades pour redéfinir plus efficacement la décision de la classification 4-Grades ? »

Choix de la décision

Comme pour une tâche d'IAL en « milieu fermé », le processus de décision dans le contexte pathologique correspond à une décision « 1 parmi N » basée sur le critère du « *Maximum de Vraisemblance* » désignant le modèle de grade le plus probable parmi ceux connus du système *i.e.* le grade g du modèle G_g sur lequel la plus grande vraisemblance est obtenue par le signal y_t.

$$\widetilde{g} = \underset{g}{argmax}\ L(y_t|G_g)$$

Cette décision est qualifiée de « globale » lorsqu'elle est prise sur la totalité du signal de parole comme cela est le cas dans les 3 protocoles expérimentaux présentés dans cette section.

[1] l'amélioration des performances du système n'est pas l'objectif recherché. Il s'agit d'appréhender les phénomènes acoustiques liés à la dysphonie afin de mieux en comprendre les caractéristiques.

Cependant, on peut se demander si la dysphonie est un phénomène qui se manifeste de manière constante sur l'ensemble du discours. L'autre alternative serait qu'il existe des contextes phonétiques qui favorisent l'émergence de la dysphonie comme des structures phonologiques « $\text{Occlusive}_{\text{sourde}}\, V_{\text{fermée}} \text{Occlusive}_{\text{sourde}}$ »[2] où la voyelle fermée serait incitée au dévoisement. Les travaux de Revis (2004) sur la méthode « Phonetic Labeling » ont montré que l'attaque favorise l'émergence de la dysphonie. Voici un exemple de phrase du corpus CVD où les phonèmes soulignés présentent une attaque :

« Il les perdait toutes de la même façon »
« i l e p ɛ r d e t u t d ø l a m ɛ m f a s ɔ̃ »

Or, sachant que la dysphonie est un trouble directement lié à la perturbation de la vibration des cordes vocales, seuls les phonèmes voisés peuvent être affectés *a priori* par des occurrences pathologiques. Doit-on alors considérer les consonnes sourdes (non-voisées) dans le processus de décision ?

De plus, parmi les résultats obtenus par la méthode « Phonetic Labeling », Revis et al. (2006) a montré que le nombre d'occurrences de phonèmes « pathologiques » est fortement corrélé avec le niveau de sévérité de la dysphonie attribué au patient par l'analyse perceptive « classique ». Est-ce que la perception de la dysphonie fonctionne comme une somme d'événements pathologiques apparaissant au cours de la parole continue ? On se rend compte qu'une décision « globale » n'est pas forcément la mieux appropriée à notre problématique. Une décision « locale » sur des portions de signal de parole (comme sur les phonèmes, syllabes, ...) serait mieux adaptée et permettrait de définir un paradigme de décision basé sur les informations phonétiques, permettant d'améliorer les performances du système automatique. La définition d'un arbre de décision phonétique constitue une voie intéressante pour améliorer la fiabilité de la classification. Pour un système fournissant des décisions « locales » (par phonème ou classe phonétique), d'autres redéfinitions de la décision « globale » peuvent être envisagées comme le vote majoritaire, des opérations arithmétiques, ...

Une autre manière de définir une stratégie de décision consiste à construite un système de décision à partir de plusieurs experts ou sources d'informations. Dans notre cas, cela consisterait à considérer les 3 classifications (2-Grades, 4-Grades, 7-Grades) comme des sous-systèmes experts (ou sources d'informations) et composantes du système de décision. Différentes approches peuvent être utilisées pour fusionner les décisions issues de différents experts comme des modèles statistiques (GMM, SVM, MLP, ...) ou encore, des opérations arithmétiques appliquées de façon empirique (addition, multiplication, moyenne, vote majoritaire, ...). Classiquement, les systèmes multi-experts s'appuient sur une stratégie de pondération par des indices représentatifs du degré de confiance des différents experts [Rahman & FairHurst, 1998]. Les indices constituent une connaissance *a priori* du pouvoir discriminant de chaque expert pour la tâche de reconnaissance.

[2] comme pour le mot « toutes » avec la voyelle fermée antérieure arrondie [ou]

Dans le cadre de la thèse, nous utiliserons la classification 4-Grades basée sur le grade global de l'échelle GRBAS de [Hirano, 1981] car elle reste la seule pour laquelle nous disposons d'une référence actuellement. La décision « globale » ou « locale » correspondra au grade du modèle sur lequel la plus grande vraisemblance aura été estimée. La redéfinition d'une décision mieux adaptée au contexte pathologique peut être envisagée avec l'analyse phonétique décrite en 7.1. En effet, l'analyse des décisions « locales »[3] fournies par la méthode devraient permettre de décrire une stratégie de décision mieux adaptée et plus efficace pour les voix pathologiques. Cependant, sortant du cadre de cette thèse, cette étude ne sera pas réalisée ici.

[3] catégorisées par classe de phonèmes

Troisième partie

La Recherche des Informations Pertinentes

Chapitre 5

L'étude paramétrique

Sommaire

Résumé

Ce chapitre est consacré à l'étude de différentes représentations paramétriques du signal de parole utilisées classiquement en RAL et appliquées ici dans un contexte pathologique. Une première étude comparative portant uniquement sur les coefficients statiques, permettra de relever la technique de paramétrisation la plus pertinente. Dans un deuxième temps, les coefficients dynamiques seront ajoutés aux coefficients statiques afin d'évaluer l'apport de l'information dynamique pour la tâche de classification par grade de la dysphonie.

Le premier travail entrepris dans cette thèse consacrée à l'évaluation objective de la voix pathologique concerne une analyse comparative entre différentes représentations du signal de parole classiquement utilisées en RAL. L'objectif est de relever la pertinence des techniques de paramétrisation pour la tâche de classification de la voix dysphonique.

Tout d'abord, une étude portant uniquement sur les coefficients statiques pour la tâche de classification 4-Grades est proposée. Contrairement aux coefficients dynamiques, les coefficients statiques représentent uniquement l'information contenue dans la trame courante et ne portent aucune information de nature temporelle ou dynamique. Les coefficients dynamiques capturent la trajectoire (ou l'évolution) temporelle à court terme des valeurs des coefficients statiques. La prise en compte de l'information dynamique est effectuée dans un second temps avec l'ajout des dérivées temporelles des coefficients statiques aux vecteurs de paramètres afin d'observer si le gain, généralement constaté en RAP et en RAL, se vérifie aussi dans le cadre de la voix pathologique.

Pour l'analyse des coefficients statiques et dynamiques, nous comparons deux techniques de paramétrisation des signaux de parole (le lecteur pourra se référer à la section 3.1 pour plus de détails sur la paramétrisation acoustique) :

1. analyse en banc de filtres
   - 24LFSC et 16LFCC (échelle linéaire)
   - 24MFSC et 16MFCC (échelle Mel)
2. analyse par prédiction linéaire
   - 12LPC
   - 12LPCC

Face à la problématique pathologique, le choix de la représentation acoustique des signaux de parole apparaît comme très importante sachant que la dysphonie est un trouble directement lié à la perturbation du son laryngé. A travers l'étude menée dans ce chapitre, les performances obtenues par les différentes paramétrisations seront analysées en regard de leurs propriétés acoustiques spécifiques afin d'estimer si la manifestation des phénomènes dysphoniques ne se localise pas uniquement dans la région de la source glottique mais s'étend aussi à un niveau supra-laryngé. En d'autres termes, est-ce que l'information contenue dans le conduit vocal est complémentaire, voire prépondérante, à celle issue de la source laryngée pour discriminer le trouble vocal ?

Nous précisons que la segmentation des signaux acoustiques utilisée pour l'ensemble des expériences présentées dans ce chapitre correspond à la version AP2 (détails en 2.5).

5.1 L'analyse des coefficients statiques

Portant uniquement sur les coefficients statiques, cette première étude va permettre à travers une analyse comparative de relever la technique de paramétrisation qui semble la plus pertinente pour la tâche de classification 4-Grades de la voix dysphonique.

Résultats

Le tableau 5.1 et la figure 5.1 donnent les résultats obtenus par les différentes paramétrisations utilisant uniquement les coefficients statiques. La paramétrisation spectrale de type MFSC obtient le meilleur TCC Global avec 73.75 % qui correspond à 59 locuteurs sur 80 classés par le système dans leur grade perceptif.

	Grade 0	**Grade 1**	**Grade 2**	**Grade 3**	**Global**	
Paramètres statiques	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/80)	± IC
12LPC	75.0 (15)	50.0 (10)	50.0 (10)	50.0 (10)	56.25 (45)	10.9
12LPCC	65.0 (13)	50.0 (10)	65.0 (13)	65.0 (13)	61.25 (49)	10.7
16LFCC	80.0 (16)	**55.0 (11)**	60.0 (12)	60.0 (12)	63.75 (51)	10.6
16MFCC	**95.0 (19)**	**55.0 (11)**	55.0 (11)	**75.0 (15)**	70.00 (56)	10.1
24LFSC	85.0 (17)	**55.0 (11)**	50.0 (10)	70.0 (14)	65.00 (52)	10.5
24MFSC	90.0 (18)	**55.0 (11)**	**75.0 (15)**	**75.0 (15)**	**73.75 (59)**	9.7

TAB. *5.1* – *Résultats de la classification 4-G selon différentes paramétrisations en terme de % TCC.*

Les coefficients obtenus par prédiction linéaire, LPC et LPCC, affichent des performances globales nettement inférieures à celles de l'analyse en banc de filtres. Les paramètres LPC atteignent leur meilleur score (75.0 %) sur les voix normales et un TCC moyen (50.0 %) sur les voix dysphoniques (G1/G2/G3), alors que les LPCC obtiennent leur meilleur score (65.0 %) sur l'ensemble des grades sauf pour le grade 1 (50.0 %).

Concernant l'analyse en banc de filtres, qu'elle soit cepstrale ou spectrale, l'utilisation de l'échelle Mel a tendance à améliorer les scores TCC globaux. La particularité de l'échelle Mel qui est de mieux représenter la sélectivité de l'oreille humaine en offrant une meilleure résolution dans les basses fréquences, apparaît comme étant pertinente pour la classification des voix dysphoniques. L'association de l'échelle Mel à une analyse en banc de filtres permet donc une amélioration des performances vérifiée ici par le comportement des scores globaux :

- **analyse cepstrale :** de 63.75 % (LFCC) à 70.00 % (MFCC) ⇒ gain de 5 locuteurs ;
- **analyse spectrale :** de 65.00 % (LFSC) à 73.75 % (MFSC) ⇒ gain de 7 locuteurs.

Cette constatation nous amènera à étudier plus précisément les caractéristiques acoustiques de la dysphonie dans le domaine fréquentiel à travers une approche par sous-bandes de fréquences décrite au chapitre 6.

En ce qui concerne les scores TCC par grade, les voix de grade G1/G2 semblent ne pas profiter du bénéfice apporté par l'échelle logarithmique comme cela apparaît pour les grades G0/G3 (spécialement pour les coefficients cepstraux) :

- **les voix G1 :** TCC de 55.0 % sur les 4 paramétrisations (LFCC/MFCC/LFSC/MFSC) ;
- **les voix G2 :** TCC de 60.0 % (LFCC) qui régresse à 55.0 % (MFCC) ;

à l'exception des coefficients spectraux pour le grade G2 qui affichent un gain de 4 locuteurs avec un TCC de 50.0 % (LFSC) qui progresse à 75.0 % (MFSC).

De plus, il faut souligner que la paramétrisation MFSC se distingue des coefficients MFCC uniquement par leur bonne performance obtenue sur les voix de grade 2 (TCC de 75.0 % et de 55.0 % respectivement).

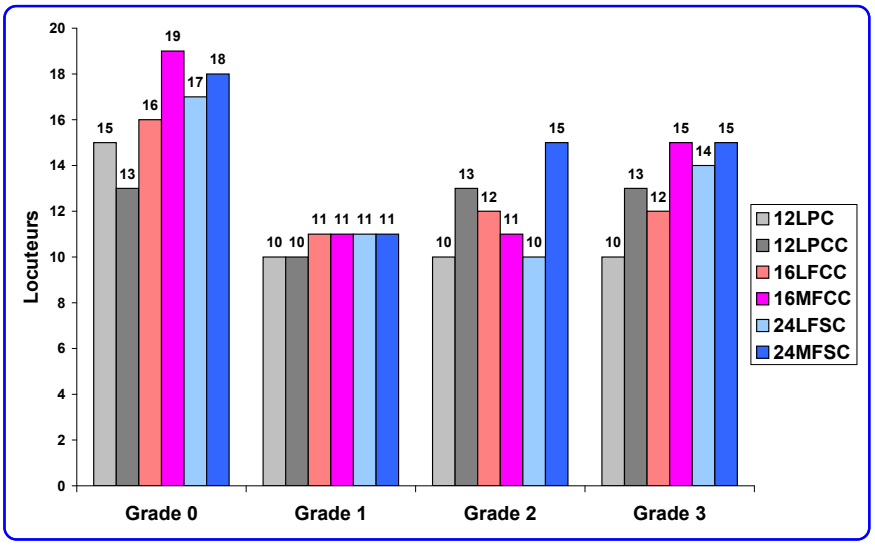


FIG. 5.1 – *Résultats de la classification 4-G selon différentes paramétrisations en terme de nombre de locuteurs (sur 20) correctement classés dans leur grade perceptif.*

Discussion

D'une manière générale, on observe que :

1. les voix normales G0 sont les mieux reconnues avec les paramétrisations MFCC et MFSC (95.0 % et 90.0% resp.)
   ⇒ *cela confirme que la paramétrisation MFCC est « l'état de l'art » en RAL*
2. la discrimination est la plus faible sur les voix légèrement dysphoniques G1
   ⇒ *aucune paramétrisation ne permet d'extraire de l'information statique suffisamment pertinente pour caractériser les dysphonies légères*

3. pour les voix G2, seuls les paramètres MFSC se distinguent des autres paramétrisations, affichant un écart très important avec les MFCC (75.0 % et 55.0 % resp.)
   ⇒ *l'apport des informations issues de la source laryngée semble importante pour distinguer les dysphonies modérées*
4. les voix G3 les plus dysphoniques obtiennent un résultat satisfaisant avec un TCC de 75.0 % pour les paramètres MFCC et MFSC
   ⇒ *les informations issues du conduit vocal suffisent-elles à discriminer les voix les plus sévèrement dysphoniques ?*

Les points 3 et 4 peuvent être justifiés par les propriétés des paramètres MFCC pour lesquels une déconvolution du signal est appliquée pour réduire l'influence de la source glottale et ne garder que l'information résonantielle du conduit vocal (pour plus de détails, voir la section 3.1.5).

La première observation concerne les paramètres LPC et LPCC obtenus par prédiction linéaire. Comparés aux autres paramétrisations, ils obtiennent les plus faibles résultats excepté pour le grade 2 où les LPCC présentent un score TCC de 65.0 % devancés uniquement par les MFSC avec un TCC de 75.0 %. Même pour les voix normales, les résultats sont plutôt faibles voire décevants. Provenant de la téléphonie et de la synthèse de la parole, cette technique permet une réduction des données pour la transmission d'un signal sur les lignes téléphoniques. Elle est aussi utilisée pour le calcul des formants. Reposant sur l'estimation de chaque échantillon du signal en fonction des p échantillons précédents, la question que l'on peut se poser sur cette technique prédictive est la suivante : « est-elle vraiment adaptée au contexte dysphonique où le trouble/désordre vocal peut apparaître sous des *formes chaotiques* d'un point de vue acoustique et donc difficilement prédictible ? » En effet, il est reconnu que le principal inconvénient de la technique LPC est d'estimer de manière uniforme le spectre sur l'ensemble des fréquences de la bande passante audible [Hermansky, 1990; Hajaiej et al., 2006]. Dans un contexte pathologique, il peut arriver que des particularités spectrales liées directement à la dysphonie soient écartées ou excessivement prédominantes dans le processus de paramétrisation. Les faibles résultats expérimentaux observés par la prédiction linéaire reflètent sans doute ce défaut dans la modélisation du spectre auditif, surtout si l'on considère que les caractéristiques acoustiques de la dysphonie ne sont pas distribuées uniformément sur l'ensemble du spectre.

La deuxième observation concerne le comportement plutôt inattendu des coefficients cepstraux par rapport aux coefficients spectraux. En effet, l'analyse cepstrale est une méthode qui vise à séparer la contribution de la source et du conduit vocal par déconvolution, en prenant comme hypothèse que « *le signal vocal est produit par un signal excitateur (source glottique) traversant le conduit vocal* ». Le spectre ainsi débarrassé des informations relatives à la source glottale, ne contient « *théoriquement* » que des informations sur le conduit vocal. Sachant que la dysphonie concerne essentiellement la source glottale, les paramètres directement liés au vibrateur laryngé sont donc d'une importance primordiale pour l'évaluation du trouble vocal. Or, à travers les résultats présentés, il apparaît que cette nature d'information d'origine laryngée semble ne faire défaut aux coefficients cepstraux que pour les voix de grade 2.

Par contre, les meilleures performances obtenues par les coefficients spectraux MFSC semblent se justifier par la nature de l'information véhiculée, information sur la source laryngée dont les caractéristiques sont :

- la fréquence fondamentale ;
- la vibration des cordes vocales pour les sons voisés ;
- les bruits de friction pour les fricatives et d'explosion (burst) pour les occlusives.

Cependant, le comportement des coefficients cepstraux peut être attribué à leur particularité d'être fortement décorrélés et de représenter une information proche des formants[1]. Or, sachant que la dysphonie est un trouble du timbre de la voix et que la configuration formantique est en relation avec le timbre vocalique, ces particularités portées par les paramètres MFCC semblent être importantes dans les résultats observés que l'on peut interpréter de deux façons :

- soit les coefficients MFCC contiennent de l'information sur la source ;
- soit la dysphonie n'est pas uniquement un problème de source laryngée et alors, elle se manifeste aussi au niveau supra-laryngé.

[1] renforcements spectraux qui correspondent aux fréquences de résonance du conduit vocal et qui caractérisent l'évolution de sa forme

5.2 L'analyse des coefficients dynamiques

Cette section est consacrée aux informations dynamiques caractéristiques du locuteur et considérées comme une source pertinente pour la RAL. Classiquement, l'extraction des informations dynamiques s'effectue au moyen d'une fenêtre temporelle glissant le long du signal de parole (pour plus de détails, voir la section 3.1.6).

Dans le cadre de la voix pathologique, nous proposons d'utiliser l'information dynamique à court terme afin d'apprécier sa capacité de discrimination. Pour cela, les dérivées de 1er ordre (Δ), de 2^{e} ordre ($\Delta\Delta$) et de 3^{e} ordre ($\Delta\Delta\Delta$) sont calculées et ajoutées aux vecteurs acoustiques. De plus, nous proposons la prise en compte d'une fenêtre temporelle suffisante pour exploiter correctement ces informations dynamiques en faisant varier la variable Θ de l'équation 3.14 de 2 à 4 afin d'obtenir un contexte de 5, 7 ou 9 trames pour le calcul des coefficients dynamiques.

5.2.1 L'analyse paramétrique comparative

Dans un premier temps, nous allons analyser les performances du système sur différentes paramétrisations auxquelles seront associés successivement les coefficients dynamiques Δ, $\Delta\Delta$ et $\Delta\Delta\Delta$ calculés avec une fenêtre de 5 trames. Pour cette analyse comparative, les paramètres utilisés seront du type LFCC/MFCC et LFSC/MFSC.

Résultats

Le tableau 5.2 et la figure 5.2 affichent les résultats globaux obtenus par l'ajout des coefficients dynamiques calculés sur une fenêtre centrée de 5 trames[2]. La première ligne du tableau intitulée «*Coefficients statiques*» rappelle les scores obtenus par les coefficients statiques sur les différentes paramétrisations (tableau 5.1). Elle figurera dans l'ensemble des tableaux relatifs à l'information dynamique et présentés dans cette section.

Paramètres dynamiques	**16LFCC**	**16MFCC**	**24LFSC**	**24MFSC**
Coefficients statiques	*63.75 (51)*	*70.00 (56)*	*65.00 (52)*	*73.75 (59)*
Δ	66.25 (53)	72.50 (58)	70.00 (56)	**75.00 (60)**
Δ + $\Delta\Delta$	65.00 (52)	73.75 (59)	70.00 (56)	**78.75 (63)**
Δ + $\Delta\Delta$ + $\Delta\Delta\Delta$	67.50 (54)	75.00 (60)	70.00 (56)	**77.50 (62)**

TAB. 5.2 – *Résultats globaux de la classification 4-G en terme de % TCC (nb/80).*
Ajout des paramètres dynamiques en contexte de 5 trames selon différentes paramétrisations.

Confirmant ce qui a été mentionné dans l'introduction de cette section, l'ajout des coefficients dynamiques améliorent les performances obtenues uniquement avec les in-

[2] taille de fenêtre classiquement utilisée en RAL

formations statiques. Quelle que soit la nature des coefficients dynamiques ajoutés, les gains absolus obtenus par les différentes paramétrisations sont de :

- 3.75 % (soit 3 locuteurs) pour les paramètres LFCC
- 5.00 % (soit 4 locuteurs) pour les paramètres MFCC/LFSC/MFSC

Au regard des valeurs présentées dans le tableau 5.2, on observe que :

1. la paramétrisation de type LFCC affiche les plus faibles résultats avec un meilleur TCC de 67.50 % à l'ajout des coefficients ($\Delta + \Delta\Delta + \Delta\Delta\Delta$) ;
2. les paramètres MFCC améliorent leur TCC à chaque ajout d'un type de coefficients dynamiques progressant de 72.50 % (Δ) $\rightsquigarrow$ 75.00 % ($\Delta + \Delta\Delta + \Delta\Delta\Delta$) ;
3. la paramétrisation LFSC atteint un TCC de 70.00 % dès l'ajout des coefficients Δ et se stabilise à cette valeur malgré l'ajout des coefficients $\Delta\Delta$ puis $\Delta\Delta\Delta$;
4. comme observé en section 5.1, les paramètres MFSC affichent encore les meilleurs résultats avec notamment un TCC de 78.75 % à l'ajout des coefficients ($\Delta + \Delta\Delta$) ;
5. la pertinence de l'échelle Mel est encore observée ici avec de meilleures performances pour les paramétrisations MFCC et MFSC qui atteignent respectivement les scores TCC globaux de 75.0 % et 78.75 %.

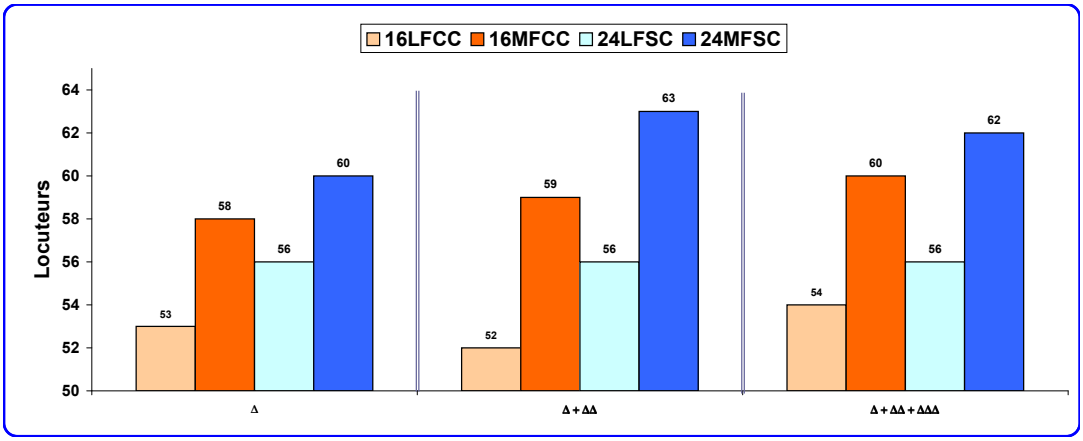


FIG. 5.2 – *Résultats de la classification 4-G selon différentes paramétrisations en terme de nombre de locuteurs correctement classés sur 80 - Ajout des paramètres dynamiques en contexte de 5 trames.*

Sur la figure 5.2, on observe que :

1. les paramétrisations à échelle Mel obtiennent les meilleurs résultats quelle que soit la nature des coefficients dynamiques ;
2. les coefficients cepstraux obtiennent leur meilleur résultat dans la configuration ($\Delta + \Delta\Delta + \Delta\Delta\Delta$) : 67.50 % pour LFCC et 75.00 % pour MFCC ;
3. les coefficients spectraux obtiennent leur meilleur résultat dans la configuration ($\Delta + \Delta\Delta$) : 70.0 % pour LFSC et 78.75 % pour MFSC.

Discussion

D'une manière générale, l'ajout de l'information dynamique aux paramètres statiques tend à améliorer les performances des différentes paramétrisations. Comme le montre la série en jaune sur la figure 5.3, le gain absolu variant de 3.75 % à 5.00 % (soit de 3 à 4 locuteurs) selon la paramétrisation, montre que ces dernières ont un comportement presque identique si l'on ne considère par la nature de l'information dynamique ajoutée.

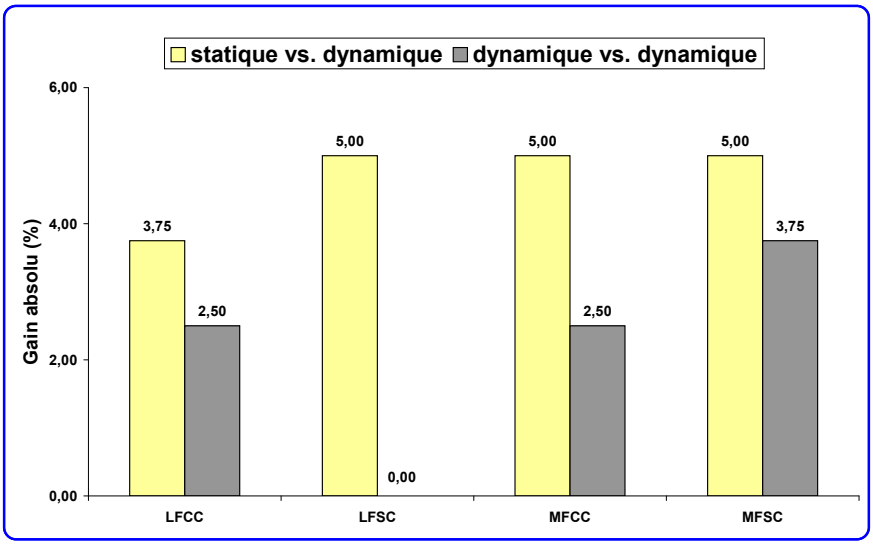


FIG. 5.3 – *Gains absolus par type de paramétrisation : statique vs. dynamique (en jaune) et dynamique vs. dynamique (en gris).*

En d'autres termes, le classement[3] des paramétrisations observé durant l'étude sur les coefficients statiques, est conservé avec une amélioration de 4 locuteurs pour les 3 meilleures (soit LFSC/MFCC/MFSC par ordre croissant). Aucune paramétrisation ne se distingue des autres de par ses gains obtenus à l'ajout de l'information dynamique.

Si l'on s'intéresse maintenant à l'accroissement entre les différentes configurations de paramètres dynamiques au sein d'une même paramétrisation (série en gris sur la figure 5.3), on constate :

1. aucune amélioration pour les coefficients spectraux LFSC ;
2. un accroissement de 2 locuteurs pour les coefficients cepstraux LFCC/MFCC
   $\Rightarrow$ soit un gain absolu de 2.50 % ;
3. un accroissement de 3 locuteurs pour les paramètres MFSC
   $\Rightarrow$ soit un gain absolu de 3.75 %.

[3] rappelé à la ligne « *Coefficients statiques* » du tableau 5.2

L'ajout successif des dérivées temporelles de 1er, 2^{e} et 3^{e} ordre aux coefficients statisques

$$\Delta \quad \rightarrow \quad \Delta + \Delta\Delta \quad \rightarrow \quad \Delta + \Delta\Delta + \Delta\Delta\Delta$$

ne montre pas des gains aussi « significatifs » que ceux observés avec les paramètres statiques augmentés de l'information dynamique. A l'exception peut-être des paramètres spectraux de type MFSC qui, avec un gain absolu de 3.75 % à l'ajout des coefficients $\Delta\Delta$ aux Δ, conservent leur statut de « meilleure paramétrisation » constaté durant l'étude sur les coefficients statiques (section 5.1).

5.2.2 Le contexte temporel variable

Dans cette étude, nous allons analyser l'impact que peut produire l'utilisation d'une fenêtre temporelle de longueur variable pour le calcul des coefficients dynamiques sur la discrimination des voix dysphoniques. Pour cela, les performances du système seront observées sur l'ajout successif des paramètres dynamiques (Δ, $\Delta\Delta$, $\Delta\Delta\Delta$) aux vecteurs acoustiques de type 24MFSC dans un contexte de 5, 7 et 9 trames (d'autres résultats sont présentés en annexe V avec notamment l'ajout des coefficients dynamiques estimés sur des fenêtres de 7 et 9 trames aux différentes paramétrisations étudiées).

Analyse globale

La figure 5.4 présente les résultats globaux, obtenus par les coefficients spectraux MFSC sur l'ajout des coefficients dynamiques selon les différents contextes de 5, 7 et 9 trames.

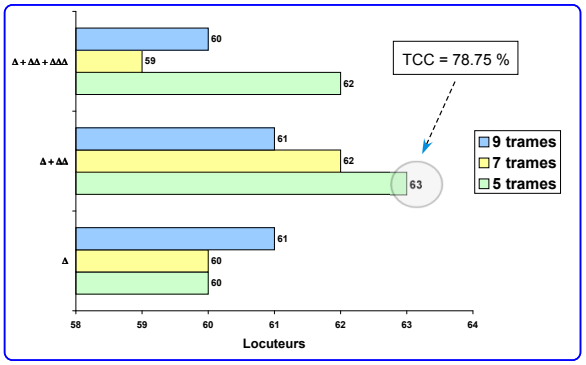


FIG. 5.4 – *Ajout des paramètres dynamiques à la paramétrisation 24MFSC - Résultats Globaux.*

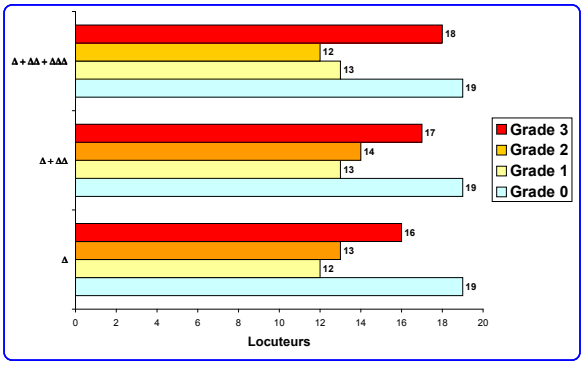


FIG. 5.5 – *Ajout des paramètres dynamiques à la paramétrisation 24MFSC - Contexte de 5 trames.*

En considérant individuellement chaque configuration de coefficients dynamiques, les meilleurs résultats sont obtenus dans un contexte de 5 trames. Le score TCC global de 78.75 % atteint par la configuration dynamique (Δ + $\Delta\Delta$) calculée dans un contexte de 5 trames, reste le meilleur résultat obtenu par les paramètres MFSC. La prise en compte d'une fenêtre temporelle plus grande pour le calcul des coefficients dynamiques, n'apporte aucune information supplémentaire, susceptible d'augmenter la discrimination des grades dysphoniques.

Analyse par grade

Nous allons maintenant analyser la sensibilité des grades selon les différentes configurations dynamiques structurelles (type de coefficients dynamiques) et temporelles (contexte temporel pour le calcul des coefficients dynamiques).

Pour cela, les figures de 5.5 à 5.7 présentent les résultats par grade, obtenus par les coefficients spectraux MFSC sur l'ajout des coefficients dynamiques selon les différents contextes de 5, 7 et 9 trames.

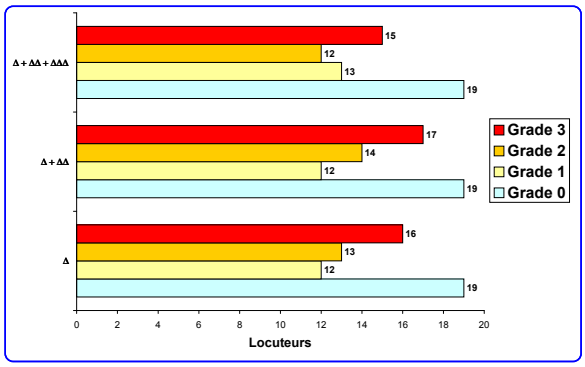


FIG. 5.6 – *Ajout des paramètres dynamiques à la paramétrisation 24MFSC - Contexte de 7 trames.*

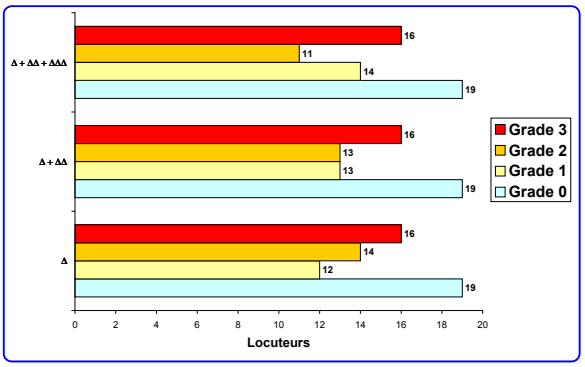


FIG. 5.7 – *Ajout des paramètres dynamiques à la paramétrisation 24MFSC - Contexte de 9 trames.*

L'observation des performances par niveau de sévérité de la dysphonie, montre que :

- **Les voix de grade 0** affichent le même TCC de 95.0 % (19) quel que soit le contexte ou le type d'information dynamique ajoutée.
- **Les voix de grade 1** obtiennent des valeurs TCC variant principalement entre 60.0 % (12) et 65.0 % (13) suivant le contexte et la structure des coefficients dynamiques. On remarquera que le contexte de 9 trames affiche une amélioration croissante d'un locuteur à chaque ajout d'un type de coefficient dynamique, partant de la valeur TCC de 60.0 % (12) avec les coefficients Δ pour atteindre 70.0 % (14) à l'ajout des coefficients ΔΔΔ.
- **Les voix de grade 2** atteignent, à l'ajout des coefficients Δ et ΔΔ, des valeurs de TCC variant de 65.0 % (13) à 70.0 % (14) selon le contexte temporel. Par contre, à l'ajout des coefficients ΔΔΔ, les performances se dégradent pour atteindre un TCC de 55.0 % (11) dans un contexte de 9 trames.
- **Les voix de grade 3** obtiennent leur meilleure performance en contexte de 5 trames à l'ajout des coefficients ΔΔΔ avec un TCC de 90.0 % (18). Considérant uniquement les voix dysphoniques, ce sont les voix de grade 3 (les plus dégradées) qui sont les mieux reconnues par le système automatique.

Concernant les observations faites pour les grades 1 et 2, il faut préciser qu'en raison du peu de données disponibles (20 locuteurs par grade), des écarts constatés d'un seul locuteur (voire de deux) entre deux expériences, ne sont pas considérés comme « significatifs » i.e. comme pertinents.

5.2.3 Discussion - Synthèse

Parmi les différentes paramétrisations éprouvées sur le système de classification automatique, la paramétrisation de type MFSC obtient les meilleures performances avec un TCC global de 78.75 % (soit 63 locuteurs sur 80) à l'ajout des coefficients (Δ + $\Delta\Delta$) calculés avec une fenêtre de 5 trames (figure 5.4) qui constitue une configuration plutôt classique en RAL.

Grades	Meilleures configurations minimales dynamiques			% TCC (nb/20)	
	coefficients ajoutés	trames	coefficients	dynamique	statique
G0	Δ	5	48	95.0 (19)	90.0 (18)
G1	Δ + $\Delta\Delta$ + $\Delta\Delta\Delta$	9	96	70.0 (14)	55.0 (11)
G2	Δ	9	48	70.0 (14)	75.0 (15)
G3	Δ + $\Delta\Delta$ + $\Delta\Delta\Delta$	5	96	90.0 (18)	75.0 (15)

TAB. 5.3 – *Résultats de la classification 4-G en terme de % TCC (nb/20) pour la paramétrisation 24MFSC. Bilan par grade de l'apport des informations dynamiques aux coefficients statiques.*

Pour chaque grade, le tableau 5.3 présente les meilleurs résultats obtenus par les paramètres MFSC à l'ajout des coefficients dynamiques. Précisons que si plusieurs configurations de paramètres dynamiques donnent le meilleur résultat pour un grade donné, la configuration dynamique retenue sera celle avec un nombre minimal de coefficients et de trames, privilégiant une taille réduite de vecteur à celle de la fenêtre dynamique. Ainsi pour le grade 0, seul l'ajout des coefficients Δ dans un contexte de 5 trames est retenu alors que le score de 95.0 % est obtenu sur l'ensemble des expériences de cette section. De même, le grade 2 obtient le même score de 70.0 % sur l'ajout des coefficients $\Delta\Delta$ en contexte de 5 et 7 trames.

Concernant les voix dysphoniques, on peut observer que :

1. **les voix de grade 1** obtiennent un gain important de trois locuteurs par rapport au score atteint avec les coefficients statiques, avec un nombre maximal de 96 coefficients ainsi qu'une large fenêtre temporelle de 9 trames. Le grade 1 est le plus problématique probablement en raison de sa proximité avec les grades G0 et G2. Les coefficients $\Delta\Delta\Delta$ que l'on peut associer aux variations temporelles de l'accélération générant des phénomènes d'instabilité (Jerk ou secousse), semblent particulièrement intéressants pour discriminer plus efficacement ces voix.
2. **les voix de grade 2** n'améliorent pas le score obtenu uniquement avec les coefficients statiques. A priori, la paramétrisation 24MFSC sans information dynamique semble suffisante pour pouvoir les différencier des grades G1 et G3.
3. **les voix de grade 3** tirent bénéfice, comme les voix G1, de l'apport des coefficients dynamiques $\Delta\Delta\Delta$ dans un contexte plus réduit de 5 trames, pour atteindre un score très prometteur de 90.0 %. Tout comme les voix de grade 1, l'ajout des coefficients $\Delta\Delta\Delta$ permettent de mieux discriminer ces voix qui ont la particularité de présenter des phénomènes caractérisant les dysphonies les plus sévères tels que les tremblements, les attaques vocales moins maîtrisées, ...

5.3 Conclusion

A partir des résultats précédents obtenus dans un contexte de voix pathologiques, nous pouvons conclure que :

► ***Analyse paramétrique statique***

1. les analyses par prédiction linéaire semblent nettement moins performantes que les analyses en banc de filtres, laissant supposer que le trouble vocal se manifeste acoustiquement de manière intermittente donc difficilement prédictible
2. l'échelle Mel, proche de la perception fréquentielle de l'oreille humaine, est un facteur d'amélioration des performances et reste préférable à l'échelle linéaire
3. les coefficients spectraux MFSC donnent les meilleurs résultats avec un TCC global de 73.75 % (soit 59 locuteurs sur 80)
4. les coefficients cepstraux MFCC obtiennent un TCC global de 70.00 % (soit 56 locuteurs sur 80), très proche des MFSC, laissant supposer que :
   - soit ils contiennent plus d'information sur la source laryngée qu'ils ne sont censés contenir
   - soit la dysphonie se manifeste également au niveau supra-laryngé
5. la performance de la paramétrisation MFSC par rapport aux coefficients MFCC, est due essentiellement à leur meilleur résultat atteint sur les voix de grade 2

► ***Analyse paramétrique dynamique***

1. l'apport de l'information dynamique est encore vérifié ici
2. l'ordre de classement observé entre les paramétrisations durant l'étude sur les coefficients statiques, reste le même avec des performances améliorées
3. les coefficients spectraux de type MFSC donnent les meilleurs résultats avec un TCC global de 78.75 % (soit 63 locuteurs sur 80) à l'ajout des coefficients ($\Delta + \Delta\Delta$) calculés avec une fenêtre de 5 trames
4. d'après le tableau 5.3 obtenu à partir des paramètres MFSC, il apparaît que pour atteindre un score TCC optimal :
   - le grade 0 tire le meilleur bénéfice de l'apport des coefficients Δ
   - le grade 2 ne tire pas bénéfice de l'apport de l'information dynamique, telle qu'elle est représentée par les coefficients delta
   - les grades 1 et 3 profitent au mieux de l'apport des coefficients ($\Delta + \Delta\Delta + \Delta\Delta\Delta$)
5. aucune nature d'information dynamique ne se distingue des autres en terme d'amélioration de performance
6. la prise en compte d'une fenêtre temporelle plus grande pour le calcul des coefficients dynamiques n'apporte aucune information supplémentaire permettant d'améliorer la discrimination des grades dysphoniques

Chapitre 6

L'étude fréquentielle

Sommaire

Résumé

Ce chapitre est dédié à la caractérisation de la dysphonie dans le domaine fréquentiel. Dans ce contexte, nous présentons le système de classification automatique associé à une architecture en sous-bandes de fréquences afin d'évaluer si des plages fréquentielles sont plus pertinentes que d'autres pour la reconnaissance de la dysphonie.
A travers plusieurs expériences, il apparaît que les basses fréquences [0-3000]Hz ont tendance à être plus intéressantes que les plus hautes fréquences pour discriminer la dysphonie.

Quelques études ont été consacrées à l'analyse acoustique des effets de la dysphonie sur le signal de parole [Wester, 1998; Maguire et al., 2003; Kacha et al., 2005]. En effet, si un expert est capable d'évaluer une voix dysphonique avec une échelle de qualité vocale comme l'échelle GRBAS de Hirano (1981), il lui est plus difficile d'apporter la justification acoustique de son choix.

Comme la dysphonie concerne essentiellement la source laryngée, la plupart des études se sont concentrées sur des paramètres directement liés à ce vibrateur : stabilité FO, jitter, shimmer, HNR, ... [Schoentgen & Bucella, 1997; Wuyts et al., 2000; Yu et al., 2001]. D'autres études ont porté sur le timbre global de la voix, en supposant que les caractéristiques acoustiques de la dysphonie sont distribuées uniformément sur l'ensemble du spectre. Finalement, l'information issue d'une analyse spectrale à long terme a aussi été étudiée, aboutissant à différentes classifications de voix pathologiques [Yanagihara, 1967; Dejonckere & Villarosa, 1986].

Dans ce chapitre, nous proposons d'étudier les caractéristiques de la dysphonie dans le domaine fréquentiel par une étude de ces phénomènes à travers une analyse par sous-bande de fréquences. Au système automatique sera associée une architecture en sous-bandes afin d'analyser la pertinence de certaines plages de fréquences pour la caractérisation des voix dysphoniques.

6.1 L'approche par sous-bande de fréquences

Utilisée en RAP [Bourlard & Dupont, 1996; Hermansky et al., 1996] et en RAL [Auckenthaler & Mason, 1997; Besacier & Bonastre, 1998; Fredouille, 2000], cette approche consiste à diviser l'espace fréquentiel en plusieurs sous-bandes qui seront traitées séparément les unes des autres par le système de reconnaissance. Sa mise en œuvre dans le contexte pathologique se justifie par l'hypothèse que la qualité des informations contenues dans une bande de fréquences particulière peut être caractéristique du niveau de sévérité de la dysphonie. En effet, les auteurs dans [Besacier et al., 2000] ont montré que certaines sous-bandes fréquentielles apparaissent comme plus pertinentes que d'autres pour la tâche de RAL. De plus, Besacier (1998) a montré dans son étude portant sur les systèmes multi-bandes que l'information utile à la caractérisation du locuteur est surtout présente dans les basses ($f \leqslant 500$Hz) et hautes fréquences ($f \geqslant 2500$Hz). Reposant sur la notion d'énergie variable selon les bandes fréquentielles, les travaux de Kitzing (1986) sur le spectre moyen à long terme (LTAS[1]) ont montré la pertinence du critère LTAS pour quantifier la qualité de la voix. Dans le même sens, l'analyse LTAS montre un accroissement d'énergie dans les hautes fréquences en présence de fatigue vocale [Dejonckere, 1986; Crevier-Buchman et al., 1993a]. Dans le contexte de la voix dysphonique, les auteurs dans [Alpan et al., 2008] analysent les dyspériodicités du signal de parole à travers une approche par sous-bandes de fréquences. Les travaux de McCowan & Sridharan (2001) montrent que certains bruits peuvent affecter différemment une ou plusieurs sous-bandes de fréquences. Il apparaît donc intéressant d'exploiter les caractéristiques de chacune des sous-bandes de fréquences en les traitant séparement les unes des autres que de considérer l'espace fréquentiel dans sa globalité.

Dans ce chapitre, l'analyse par sous-bande est utilisée afin d'étudier la manière dont les caractéristiques acoustiques de la dysphonie sont dispersées sur l'ensemble des différentes bandes de fréquences selon le niveau de sévérité du trouble vocal : « est-ce qu'une sous-bande de fréquences est plus discriminante qu'une autre pour la classification des voix dysphoniques ? »

6.1.1 L'architecture en sous-bandes de fréquences

Comme le montre la figure 6.1, les signaux de parole sont filtrés sur des plages de fréquences disjointes de 1 kHz durant la phase de paramétrisation. La bande de fréquences utile s'étendant sur la plage de [0-8000]Hz [2], chaque signal de parole sera représenté par 8 séquences de vecteurs de paramètres représentant chacune une sous-bande fréquentielle de 1 kHz. Pour la tâche de classification, chaque sous-bande de fréquences sera associée à un reconnaisseur qui fournira un score de décision. Les performances obtenues sur chaque sous-bande sont ensuite comparées entre elles afin de déterminer les bandes de fréquences les plus pertinentes selon les grades.

[1] Long Time Average Spectrum
[2] la fréquence d'échantillonnage des signaux de parole est de 16 kHz

Dans ce contexte, les phases d'apprentissage et de test sont réalisées sous-bande par sous-bande dans un espace fréquentiel présentant une résolution plus fine comparée à celle de la bande totale [0-8000]Hz représentant pour sa part le domaine fréquentiel dans sa globalité.

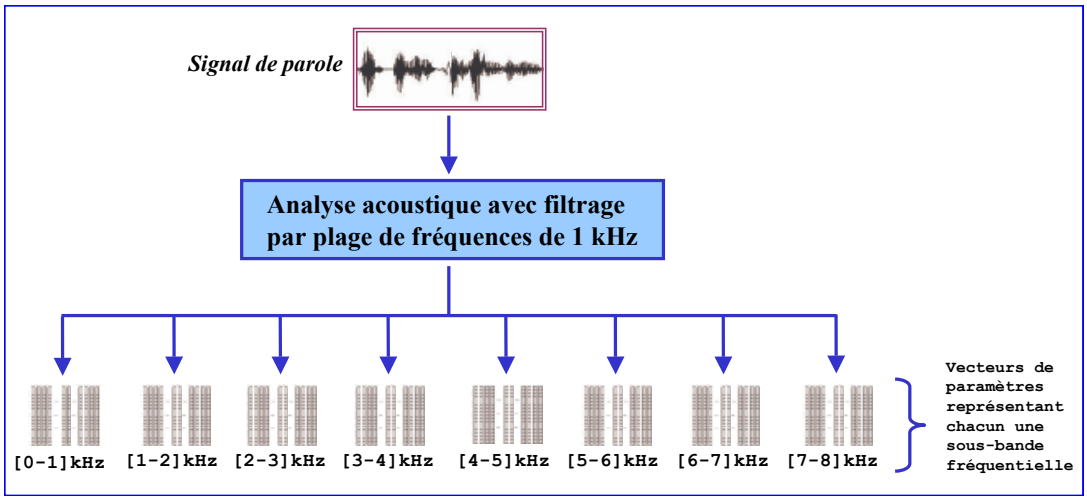


FIG. 6.1 – *Architecture en sous-bandes de fréquences. Durant l'analyse acoustique, le signal de parole est filtré sur des plages de fréquences de 1 kHz pour produire des vecteurs de paramètres qui seront traités individuellement par le système de RAL.*

6.1.2 Le choix de la paramétrisation

Dans le but d'estimer l'importance relative de chaque sous-bande de fréquences pour la tâche de classification des voix pathologiques, le choix de la représentation paramétrique des signaux acoustiques s'est porté sur l'analyse en banc de filtres de type LFSC (24 coefficients spectraux statiques).

Même s'il a été montré dans la section 5.1 que les paramètres LFSC donnent des scores TCC nettement inférieurs à ceux obtenus avec les coefficients MFSC, le choix d'utiliser une échelle linéaire a pour principales motivations :

1. d'obtenir une estimation de l'enveloppe spectrale du signal de parole qui fournit les informations du spectre de la source et de la réponse fréquentielle du conduit vocal *i.e.* tout le contenu spectral du signal source, modifié par le conduit vocal ;
2. d'utiliser une échelle linéaire afin de ne pas déformer les observations fréquentielles par application d'une échelle non-linéaire comme Mel ;
3. de rester cohérent par rapport à la définition de l'échelle Mel étudiée sur l'ensemble de la bande passante audible.

Nous précisons aussi que l'étude, présentée dans cette section, utilise la segmentation de version AP1 (détails à la section 2.5) pour les signaux acoustiques.

6.1.3 L'analyse par sous-bande individuelle

Dans cette première expérience, huit sous-bandes de 1000 Hz sont traitées invidiuellement par le système de classification.

	Grade 0	Grade 1	Grade 2	Grade 3	Global	
24LFSC	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/80)	**± IC**
Bande Totale	85.0 (17)	55.0 (11)	**50.0 (10)**	70.0 (14)	65.00 (52)	10.5
[0-1000]Hz	**85.0 (17)**	**60.0 (12)**	35.0 (7)	**70.0 (14)**	62.50 (50)	10.7
[1000-2000]Hz	**95.0 (19)**	**75.0 (15)**	**50.0 (10)**	60.0 (12)	**70.00 (56)**	10.1
[2000-3000]Hz	80.0 (16)	50.0 (10)	25.0 (5)	65.0 (13)	55.00 (44)	11.0
[3000-4000]Hz	65.0 (13)	35.0 (7)	35.0 (7)	25.0 (5)	40.00 (32)	10.8
[4000-5000]Hz	65.0 (13)	25.0 (5)	20.0 (4)	20.0 (4)	32.50 (26)	10.3
[5000-6000]Hz	40.0 (8)	**65.0 (13)**	20.0 (4)	**70.0 (14)**	48.75 (39)	11.0
[6000-7000]Hz	40.0 (8)	40.0 (8)	35.0 (7)	**70.0 (14)**	46.25 (37)	11.0
[7000-8000]Hz	65.0 (13)	20.0 (4)	30.0 (6)	**80.0 (16)**	48.75 (39)	11.0

TAB. *6.1* – *24LFSC - Résultats de la classification 4-G selon différentes bandes de fréquences de 1000Hz en terme de % TCC*

Le tableau 6.1 compare les performances individuelles des huit sous-bandes et de la bande totale - la figure 6.2 décrit ces mêmes résultats sous une forme graphique. Les matrices de confusion par sous-bande sont fournies dans le tableau 6.2 et celle de la bande totale dans le tableau 6.4.

De ces différents résultats, trois principales tendances peuvent être observées :

► ***Les bandes de fréquences entre 0 et 3000 Hz***

Ces plages de fréquences obtiennent les meilleures performances avec un TCC global qui varie de 55 % à 70 %.

On peut observer également que :

- **la sous-bande [0-1000]Hz**
  affiche un TCC identique à celui de la bande totale pour les voix de grade 0 et de grade 3 (85.0 % et 70.0 % respectivement). De plus, les voix de grade 1 avec un TCC de 60.0 % dépassent légèrement celui obtenu sur la bande totale de 55.0 % ;
- **la sous-bande [1000-2000]Hz**
  obtient un TCC global de 70.0 % qui dépasse légèrement les 65.0 % de la bande totale. De même, les voix de grade 0 et de grade 1 atteignent des valeurs de TCC qui dépassent ceux de la bande totale (95.0 % vs. 85.0 % et 75.0 % vs. 55.0 % respectivement). Cette plage fournit aussi la meilleure performance pour les voix de grade 2 avec un TCC de 50 % ;

[0-1000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**17**	3	0	0
P_G1	2	**12**	5	1
P_G2	0	8	**7**	5
P_G3	2	2	2	**14**

[1000-2000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**19**	1	0	0
P_G1	2	**15**	1	2
P_G2	1	3	**10**	6
P_G3	0	1	7	**12**

[2000-3000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**16**	3	1	0
P_G1	5	**10**	3	2
P_G2	5	3	**5**	7
P_G3	0	0	7	**13**

[3000-4000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**13**	5	2	0
P_G1	7	**7**	4	2
P_G2	4	8	**7**	1
P_G3	0	9	6	**5**

[4000-5000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**13**	4	3	0
P_G1	6	**5**	6	3
P_G2	6	7	**4**	3
P_G3	3	4	9	**4**

[5000-6000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**8**	7	4	1
P_G1	3	**13**	3	1
P_G2	4	7	**4**	5
P_G3	1	1	4	**14**

[6000-7000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**8**	4	6	2
P_G1	5	**8**	3	4
P_G2	4	4	**7**	5
P_G3	0	2	4	**14**

[7000-8000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**13**	2	3	2
P_G1	9	**4**	6	1
P_G2	5	3	**6**	6
P_G3	1	1	2	**16**

TAB. 6.2 – *Matrices de confusion en classification 4-G selon différentes sous-bandes de fréquences de 1000Hz (24LFSC)*

- **la sous-bande [2000-3000]Hz**
  présente des scores TCC inférieurs aux deux sous-bandes précédentes, même si les résultats obtenus pour les voix de grade 0 et de grade 3 restent satisfaisants (80.0 % et 65.0 % respectivement).

En outre, il peut être noté que les erreurs de classification sont distribuées dans les grades adjacents dans la plupart des cas (par exemple sur la sous-bande [0-1000]Hz, les erreurs de classification pour le grade 2 sont réparties respectivement sur les grades 1 et 3 avec 8 et 5 erreurs).

► ***Les bandes de fréquences entre 3000 et 5000 Hz***

Les fréquences entre 3000 et 5000 Hz exhibent les performances globales les plus faibles. Seules les voix normales (grade 0) obtiennent un TCC satisfaisant de 65 %, en dépit d'une perte de performance significative comparée à la bande totale (TCC de 85 %). D'autre part, une forte confusion peut être observée pour les voix dysphoniques, se traduisant par des scores TCC très bas (TCC variant de 20 % à 35 %).

▶ ***Les bandes de fréquences entre 5000 et 8000 Hz***

Les fréquences supérieures à 5000 Hz fournissent de meilleures performances globales comparées à celles de la plage [3000-5000]Hz. Néanmoins, la plupart des erreurs de classification sont très éparpillées entre les grades, démontrant encore une grande confusion.

Il peut être observé que les voix à dysphonies sévères (grade 3) sont très bien classées pour les deux sous-bandes entre 5000 et 7000 Hz (TCC de 70 %) et [7000-8000]Hz (TCC de 80 %, meilleur score).

▷ ***Bilan de l'analyse par sous-bande individuelle***

Finalement, la figure 6.2 qui affiche le nombre de locuteurs bien classés par grade et par sous-bande individuelle, laisse apparaître :

1. des difficultés à reconnaître les voix de grade 2 quelle que soit la sous-bande individuelle considérée ;
2. la capacité des basses fréquences à discriminer la plupart des voix, excepté celles de grade 2 ;
3. la performance « surprenante » des voix de grade 3 dans les hautes fréquences en dépit de la « quasi-absence » de parole dans cette zone.

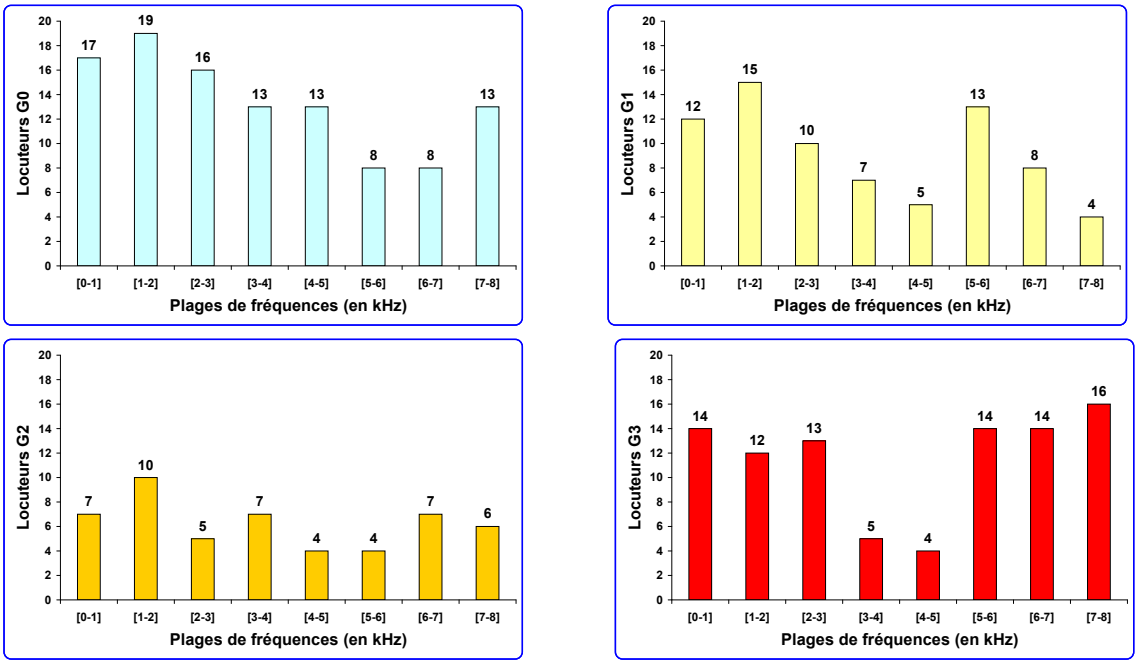


FIG. 6.2 – *Voix bien classées (20 locuteurs par grade) en classification 4-G selon différentes plages de fréquences (LFSC)*

6.1.4 Le regroupement des sous-bandes individuelles

Cette section se concentre sur les trois zones de fréquences mises en valeur dans la section précédente. Ici, la tâche de la classification s'effectue sur les sous-bandes fréquentielles suivantes : [0-3000]Hz, [3000-5400]Hz et [5400-8000]Hz. Cette expérience suppose une complémentarité des sous-bandes de 1 kHz. Les tableaux 6.3 et 6.4 donnent respectivement, les valeurs des TCC et les matrices de confusion obtenues dans les différentes sous-bandes de fréquences regroupées.

	Grade 0	**Grade 1**	**Grade 2**	**Grade 3**	**Global**	
24LFSC	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/80)	**± IC**
Bande Totale	85.0 (17)	55.0 (11)	50.0 (10)	**70.0 (14)**	65.00 (52)	10.5
[0-3000]Hz	**90.0 (18)**	**65.0 (13)**	**65.0 (13)**	65.0 (13)	**71.25 (57)**	10.0
[3000-5400]Hz	65.0 (13)	40.0 (8)	25.0 (5)	65.0 (13)	48.75 (39)	11.0
[5400-8000]Hz	65.0 (13)	35.0 (7)	45.0 (9)	**70.0 (14)**	53.75 (43)	11.0

TAB. 6.3 – *24LFSC - Résultats de la classification 4-G selon différentes bandes de fréquences en terme de % TCC*

Pour chacune des 3 bandes de fréquences considérées, on peut dire que :

► ***La bande de fréquences [0-3000]Hz***

Couvrant «*la région des premiers formants*», cette plage de fréquences est la plus intéressante. On observe principalement que :

1. un meilleur TCC global de 71.25 % est atteint, à comparer aux 65.0 % sur [0-8000]Hz et 70.0 % sur [1000-2000]Hz ;
2. le grade 2 obtient son meilleur résultat avec un TCC de 65.0 % contre 50.0 % pour la meilleure sous-bande individuelle [1000-2000]Hz et la bande totale ;
3. le regroupement de sous-bandes individuelles donne lieu à une performance de classification plus homogène et plus satisfaisante pour les différents grades (TCC de 65.0 % pour les grades dysphoniques), notamment pour les voix de grade 2 qui présentaient les plus faibles performances en sous-bandes individuelles.

► ***La bande de fréquences [3000-5400]Hz***

Apparentée à «*la région des constrictives et occlusives*», cette bande de fréquences obtient le TCC global le plus faible (48.75 %) comparé aux 2 autres bandes.

La confusion observée dans les sous-bandes individuelles est encore présente, à l'exception des voix de grade 3 qui ont tendance à mieux bénéficier des avantages de la complémentarité des sous-bandes individuelles avec un TCC de 65.0 % contre 25.0 % en [3000-4000]Hz et 20.0 % en [4000-5000]Hz.

► ***La bande de fréquences [5400-8000]Hz***

Associée à «*la région résiduelle des constrictives et occlusives*», cette plage de fréquences fournit des performances satisfaisantes pour les voix normales avec un TCC de 65 % et pour les voix à dysphonie sévère avec un TCC de 70 %.

Concernant la nature de l'information de parole portée par cette bande, le TCC des voix de grade 3 peut trouver une explication par la présence de bruit résiduel «voilé» (ou «essoufflé») caractéristique des voix à dysphonie sévère. En effet, la présence de souffle dans la voix dysphonique peut être due à un écoulement turbulent causé par un mauvais accolement des cordes vocales. Ce bruit de turbulence[3] *excessif* peut s'intensifier jusqu'à devenir audible dans le cas de certaines voix pathologiques. Quand la turbulence est pleinement developpée, elle est alors perçue comme un bruit à large bande spectrale, clairement visible sur un spectrogramme. Elle correspond au critère B («Breathy») de l'échelle GRBAS servant à évaluer *l'impression de souffle dans la voix en rapport avec une incompétence glottique avec bruit de turbulence*. Par contre, il est plus difficile d'expliquer le comportement des voix normales dans cette bande de hautes fréquences, excepté par l'absence d'information dans ce grade comparé aux grades dysphoniques où le bruit croît au fur et à mesure que la sévérité du trouble augmente.

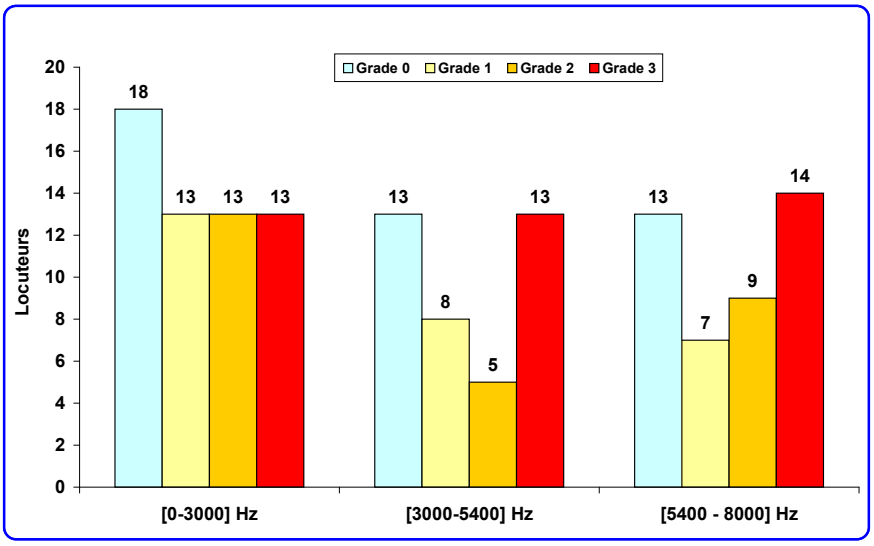


FIG. 6.3 – *Voix bien classées (20 locuteurs par grade) en classification 4-G selon le regroupement des sous-bandes individuelles (24LFSC).*

[3] appelé «bruit additif»

[0-3000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**18**	1	1	0
P_G1	1	**13**	6	0
P_G2	0	6	**13**	1
P_G3	0	2	5	**13**

[3000-5400]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**13**	6	1	0
P_G1	8	**8**	2	2
P_G2	7	6	**5**	2
P_G3	2	1	4	**13**

[5400-8000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**13**	4	3	0
P_G1	8	**7**	4	1
P_G2	5	3	**9**	3
P_G3	0	1	5	**14**

[0-8000]Hz (Bande Totale)

	S_G0	S_G1	S_G2	S_G3
P_G0	**17**	2	1	0
P_G1	2	**11**	5	2
P_G2	2	6	**10**	2
P_G3	0	1	5	**14**

TAB. *6.4* – *Matrices de confusion en classification 4-G selon différentes plages de fréquences (24LFSC)*

6.1.5 L'intégration de l'information utile au système de classification

D'après l'analyse fréquentielle effectuée ci-dessus, la bande de fréquences [0-3000]Hz semble la plus pertinente pour la tâche de reconnaissance des niveaux de sévérité de la dysphonie. L'analyse des coefficients delta décrite en section 5.2 a montré que la configuration dynamique (Δ + $\Delta\Delta$) dans un contexte de 5 trames permettait aux coefficients spectraux du type 24MFSC d'obtenir les meilleures performances sur la bande [0-8000]Hz. Nous allons à présent appliquer la plage de fréquences [0-3000]Hz sur le système de classification en utilisant les analyses spectrales de type LFSC et MFSC avec 24 coefficients statiques auxquels sont associés les coefficients dynamiques (Δ + $\Delta\Delta$). Les résultats seront ensuite comparés avec les performances obtenues sur la bande totale [0-8000]Hz. Cette comparaison permettra de valider l'apport de l'information fréquentielle mis en évidence précédemment dans différentes configurations expérimentales.

	Grade 0	**Grade 1**	**Grade 2**	**Grade 3**	**Global**	
Paramètres [0-8000]Hz	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/80)	**± IC**
24LFSC + 24Δ + 24$\Delta\Delta$	**95.0 (19)**	50.0 (10)	55.0 (11)	**75.0 (15)**	68.75 (55)	10.2
24MFSC + 24Δ + 24$\Delta\Delta$	**95.0 (19)**	**60.0 (12)**	**75.0 (15)**	**75.0 (15)**	**76.25 (61)**	9.4
Paramètres [0-3000]Hz	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/80)	**± IC**
24LFSC + 24Δ + 24$\Delta\Delta$	**95.0 (19)**	**75.0 (15)**	50.0 (10)	**85.0 (17)**	76.25 (61)	9.4
24MFSC + 24Δ + 24$\Delta\Delta$	**95.0 (19)**	70.0 (14)	**70.0 (14)**	**85.0 (17)**	**80.00 (64)**	8.8

TAB. *6.5* – *Comparaison entre LFSC et MFSC - Résultats de la classification 4-G selon les bandes de fréquences [0-8000]Hz et [0-3000]Hz en termes de % TCC*

Le tableau 6.5 donne les performances en terme de TCC pour les différentes configurations paramétriques choisies. Les expériences sont basées sur la version AP1 pour la segmentation des signaux acoustiques (détails à la section 2.5).

Par comparaison avec la bande totale [0-8000]Hz, nous pouvons observer que :

1. la restriction à la bande de fréquences [0-3000]Hz permet d'améliorer les performances de classification pour l'ensemble des grades, quel que soit le type d'échelle utilisé, linéaire ou Mel ;
2. la meilleure performance est atteinte par les coefficients MFSC associés aux coefficients dérivés ($\Delta + \Delta\Delta$) qui obtiennent un TCC global de 80.0 % comparé à 76.25 % atteints sur la bande totale [0-8000]Hz.

De plus, quelle que soit la bande de fréquences considérée, le passage d'une échelle linéaire à Mel ne profite essentiellement qu'aux voix à dysphonie modérée (grade 2) qui affichent un gain de 4 locuteurs. Cette observation soulève la question de la pertinence et de l'efficacité de l'échelle Mel sur la plage de fréquences [0-3000]Hz par rapport à la bande totale [0-8000]Hz.

Filtres	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**
[0-8000]Hz	74.24	156.35	247.17	347.62	458.73	581.62	717.54	867.88	1034.16	1218.08	1421.50	1646.50
[0-3000]Hz	48.21	99.74	154.81	213.68	276.60	343.86	415.75	492.59	574.72	662.51	756.34	856.64
Filtres	**C13**	**C14**	**C15**	**C16**	**C17**	**C18**	**C19**	**C20**	**C21**	**C22**	**C23**	**C24**
[0-8000]Hz	1895.36	2170.61	2475.05	2811.78	3184.23	3596.17	4051.80	4555.75	5113.15	5729.67	6411.57	7165.79
[0-3000]Hz	963.84	1078.43	1200.90	1331.82	1471.74	1621.31	1781.17	1952.05	2134.69	2329.91	2538.57	2761.61

TAB. 6.6 – *Fréquences centrales (en Hertz) de bancs de 24 filtres répartis selon une échelle Mel sur les bandes de fréquences [0-8000]Hz et [0-3000]Hz.*

Afin de répondre à cette question, le tableau 6.6 affiche les fréquences centrales de bancs de 24 filtres répartis selon une échelle Mel sur les 2 bandes de fréquences considérées. Rappelons que l'échelle de Mel a comme particularité d'être linéaire en basse fréquence ($\leq 1000Hz$) et logarithmique en haute fréquence ($\geq$ 1000 Hz). De plus, elle a pour vocation de fournir une meilleure résolution dans les basses fréquences afin de mieux représenter la sélectivité de l'oreille humaine. Cela peut s'observer sur le tableau 6.6 où en [0-8000]Hz, les 9 premiers coefficients représentent les fréquences $\leq$ 1000 Hz (9^e filtre de fréquence centrale 1034.16 Hz) alors qu'en [0-3000]Hz, ce sont les 13 premiers (13^e filtre de fréquence centrale 963.84 Hz). La bande restreinte [0-3000]Hz offre une résolution fréquentielle plus fine sur les basses fréquences ($\leq$ 1000 Hz) que la bande totale [0-8000]Hz. En ce qui concerne les fréquences $\geq$ 1000 Hz, la bande [0-3000]Hz regroupe les 11 derniers coefficients pour représenter la plage [1000-3000]Hz, alors que la bande [0-8000]Hz totalise les 15 derniers coefficients pour représenter la plage de fréquences [1000-8000]Hz. Sur les fréquences $\geq$ 1000 Hz, on observe encore une résolution fréquentielle plus fine pour la bande restreinte que pour la bande totale.

La figure 6.4 dessine les bancs de 24 filtres répartis selon une échelle Mel sur les bandes de fréquences [0-3000]Hz et [0-8000]Hz. On voit clairement que l'échelle Mel offre une résolution plus fine sur la bande [0-3000]Hz qu'en [0-8000]Hz.

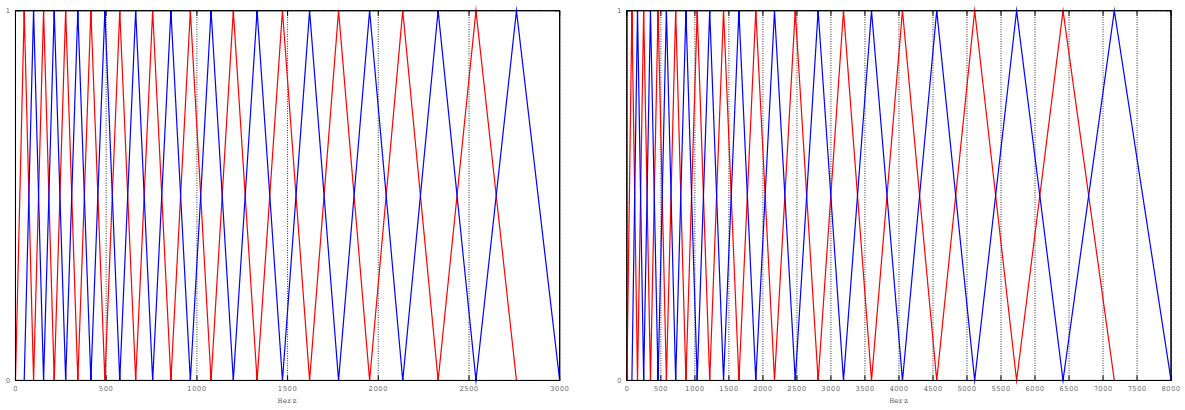


FIG. 6.4 – *Bancs de 24 filtres répartis selon une échelle Mel sur les bandes de fréquences [0-3000]Hz et [0-8000]Hz.*

En conclusion, quelle que soit la bande de fréquences considérée ([0-3000]Hz ou [0-8000]Hz), les coefficients spectraux en échelle Mel associés aux paramètres dynamiques ($\Delta + \Delta\Delta$) permettent d'améliorer les performances de reconnaissance des grades dysphoniques, notamment pour les voix de grade 2 à dysphonie modérée. Par contre, l'utilisation de l'échelle Mel semble être plus profitable sur la bande totale [0-8000]Hz que sur la bande restreinte [0-3000]Hz, si l'on se réfère au gain de 6 locuteurs contre 4 obtenus respectivement.

6.2 L'évaluation perceptive en [0-3000]Hz

Dans la section précédente 6.1, il a été montré que le système de classification a tendance à être plus sensible à la bande de fréquences [0-3000]Hz qu'à la bande totale [0-8000]Hz. La question qui peut être soulevée est la suivante : « Quels effets peut produire cette réduction fréquentielle sur le jugement perceptif d'un jury d'experts ? »

Protocole

Pour répondre à cette question, le corpus CVD filtré[4] sur [0-3000]Hz a été analysé perceptivement selon le même protocole que celui mis en œuvre lors de son évaluation initiale sur la bande totale [0-8000]Hz. Dans ces conditions, les jugements perceptifs ont été votés par consensus par le même jury d'experts au cours d'une même session selon le grade G de dysphonie de l'échelle GRBAS. Le temps séparant l'évaluation initiale du corpus CVD (bande totale [0-8000]Hz en 2003) et cette évaluation du corpus filtré (bande restreinte [0-3000]Hz en 2008) correspond à quatre années. Même si quelques écoutes intermédiaires ont pu être pratiquées durant cette période sur certaines locutrices, on ne peut parler « d'effets mémoire » de la part des membres du jury à l'attribution du grade dysphonique.

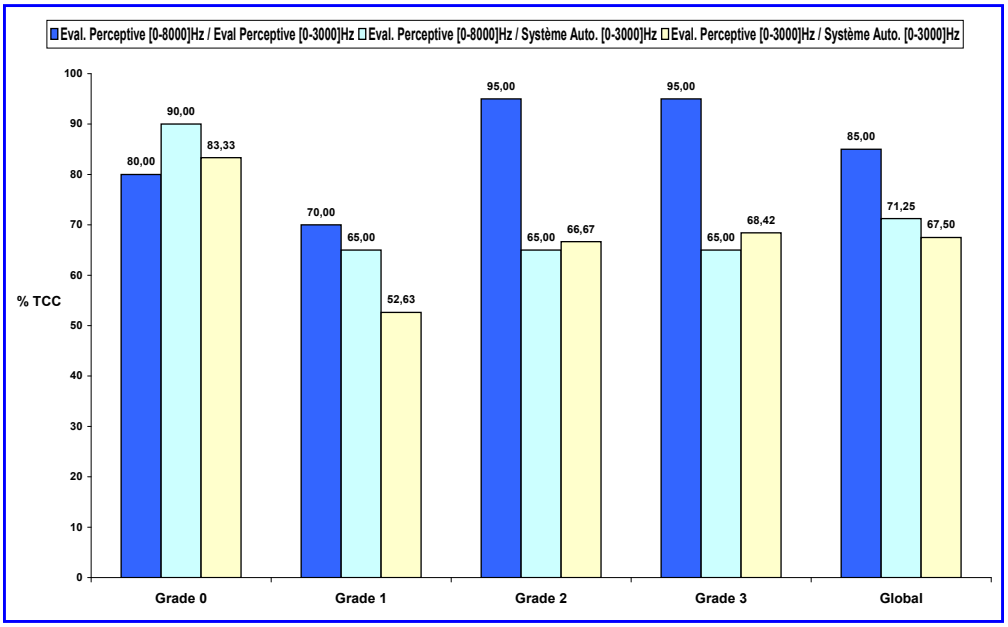


FIG. 6.5 – *Analyse de concordance en terme de % TCC entre :*
(1) l'évaluation perceptive en [0-8000]Hz et l'évaluation perceptive en [0-3000]Hz
(2) l'évaluation perceptive en [0-8000]Hz et le système automatique en [0-3000]Hz
(3) l'évaluation perceptive en [0-3000]Hz et le système automatique en [0-3000]Hz.

[4] par application d'un filtre passe bande avec l'utilitaire sox

En effet, les auditeurs du jury possèdent une solide expérience clinique de l'écoute de la dysphonie et de son évaluation, ainsi qu'un niveau de connaissance très poussé dans le domaine de la pathologie vocale qui permettent de supposer que la variabilité intra-individuelle puisse être la plus réduite possible. L'analyse de l'évaluation faite sur le corpus CVD filtré en [0-3000]Hz permet d'établir différentes comparaisons impliquant d'une part, l'évaluation perceptive initiale du corpus CVD en [0-8000]Hz et d'autre part, l'évaluation du système automatique sur la bande de fréquences restreinte de [0-3000]Hz (décrite en 6.1.4).

Résultats

La figure 6.5 présente les résultats de l'analyse de concordance entre :

1. l'évaluation perceptive en [0-8000]Hz et l'évaluation perceptive en [0-3000]Hz
2. l'évaluation perceptive en [0-8000]Hz et le système automatique en [0-3000]Hz
3. l'évaluation perceptive en [0-3000]Hz et le système automatique en [0-3000]Hz

On peut observer que :

1. le meilleur taux de concordance de 85.0 % est atteint entre les deux évaluations perceptives [0-8000]Hz et [0-3000]Hz, montrant pour chacun des grades :
   - un score très élevé de 95.0 % pour les grades 2 et 3 (les plus dysphoniques) ;
   - un score modéré de 70.0 % pour les voix de grade 1 ;
   - un score élévé de 80.0 % pour les voix normales de grade 0.
2. la classification automatique en [0-3000]Hz obtient un meilleur TCC global lorsque la référence perceptive est évaluée sur [0-8000]Hz, avec 71.25 % contre 67.50 %. Cette tendance est due principalement aux performances obtenues par les voix de grade 0 et 1 qui obtiennent des valeurs TCC de 90.0 % et 65.0 % respectivement. Seuls les grades 2 et 3 obtiennent des scores TCC légèrement plus favorables avec l'évaluation perceptive en [0-3000]Hz avec 66.67 % contre 65.00 % pour le grade 2 et 68.42 % contre 65.00 % pour le grade 3.
   Par contre, le score TCC des voix de grade 1 est relativement faible avec le jugement perceptif en [0-3000]Hz affichant 52.63 % contre 65.00% en évaluation perceptive [0-8000]Hz.

Discussion

Concernant les évaluations perceptives, le taux de concordance atteint de 85.0 % peut être dû à la variabilité intra/inter auditeurs durant les sessions d'écoutes sachant que l'évaluation de la voix est soumise à la variabilité et à la subjectivité intrinsèque du jugement perceptif. En effet, Revis (2004) montre que le niveau de reproductibilité des évaluations perceptives est habituellement d'environ 57 % en inter-individuel et de 62 % en intra-individuel. Toutefois, il est tout de même intéressant de constater que les

désaccords ont lieu principalement sur les voix de grades 0 et 1 comme illustré dans la matrice de confusion 6.7.

	S_G0	S_G1	S_G2	S_G3
P_G0	**16**	4	0	0
P_G1	2	**14**	4	0
P_G2	0	1	**19**	0
P_G3	0	0	1	**19**

TAB. *6.7* – *Matrice de confusion de l'évaluation perceptive du corpus CVD filtré en [0 - 3000]Hz (colonnes S_G) par rapport à l'évaluation perceptive initiale (corpus CVD en [0-8000]Hz)*

L'autre hypothèse serait que la réduction de la bande de fréquences sur des voix normales ou légèrement dysphoniques, donnant des voix de tonalité plus grave (pitch sans harmonique aiguë), puisse affecter les jugements d'un expert. Les dégradations en terme de TCC peuvent l'expliquer par une surestimation de la dysphonie pour les voix normales G0 (4 en grade 1) et les dysphonies légères G1 (4 en grade 2). Par contre, le filtrage des fréquences supérieures à 3000 Hz n'a pour ainsi dire pas « *influencé* » le jury d'écoute pour l'évaluation des voix les plus dysphoniques (grades 2 et 3). Il semblerait que les dégradations en terme de qualité de voix ne soient pas accentuées davantage par le filtrage.

Evaluation perceptive [0-8000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**18**	1	1	0
P_G1	1	**13**	6	0
P_G2	0	6	**13**	1
P_G3	0	2	5	**13**

Evaluation perceptive [0-3000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**15**	3	0	0
P_G1	4	**10**	5	0
P_G2	0	7	**16**	1
P_G3	0	2	4	**13**

TAB. *6.8* – *Matrices de confusion de la classification automatique en [0 - 3000]Hz par rapport à l'évaluation perceptive [0-8000]Hz (à gauche) et l'évaluation perceptive [0-3000]Hz (à droite)*

En ce qui concerne la classification automatique, les résultats sont légèrement meilleurs avec l'évaluation perceptive de même bande de fréquences ([0-3000]Hz) pour les voix de grades 2 et 3. Par contre avec le jugement perceptif en bande totale, cette tendance s'inverse pour les voix normales et plus particulièrement pour les voix légèrement dysphoniques (grade 1). Ce comportement est fortement corrélé avec les observations faites dans la comparaison des 2 évaluations perceptives, spécialement au regard du faible taux de concordance obtenu pour le grade 1. Les matrices de confusion en 6.8 montrent que la confusion provient essentiellement des voix filtrées en [0-3000]Hz et évaluées par le jury en grade 1 qui ont été classées différemment par le système automatique (4 voix en G0 et 5 voix en G2). Notons aussi qu'il semble normal que les valeurs TCC soient globalement plus élevées entre la classification automatique et le jugement perceptif en bande totale car ce dernier a été utilisé pour entraîner les modèles correspondant à chacun des grades. Finalement, une analyse plus approfondie n'a pas permis d'établir une corrélation sur les voix mal classées entre les 2 évaluations concernant la bande de fréquences [0-3000]Hz, le système automatique et le jugement perceptif.

6.3 La bande téléphonique

L'un des problèmes de la parole transmise à travers le canal du téléphone est la restriction à la bande passante normalisée [300-3400]Hz. Il est bien connu que cette limitation perturbe fortement les systèmes automatiques liés à la parole comme par exemple, la reconnaissance de la parole ou du locuteur. L'absence d'information utile à la caractérisation du locuteur dans cette bande de fréquences explique en partie la dégradation des performances des systèmes automatiques [Reynolds, 1994].

Considérant la bande de fréquences [0-3000]Hz, nous allons examiner dans cette section l'impact de la restriction relative à la bande téléphonique sur la classification automatique des voix dysphoniques. Par conséquent, les signaux de parole du corpus CVD ont été filtrés en [300-3000]Hz. La paramétrisation utilisée correspond aux coefficients spectraux de type LFSC (24 coefficients statiques). Il est à noter que l'utilisation de la version AP2 (détails à la section 2.5) pour la segmentation des signaux acoustiques, entraîne une légère différence dans les résultats de l'expérience [0-3000]Hz présentée plus bas et celle affichée en 6.1.4.

Résultats

Toutes les classifications présentées dans cette section ont été comparées avec le jugement perceptif du corpus CVD en [0-8000]Hz. Le tableau 6.9 affiche les résultats obtenus par le système automatique sur la bande de fréquences [300-3000]Hz.

	Grade 0	**Grade 1**	**Grade 2**	**Grade 3**	**Global**	
Paramètres 24LFSC	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/80)	**± IC**
[0-8000]Hz	85.0 (17)	55.0 (11)	50.0 (10)	**70.0 (14)**	65.00 (52)	10.5
[0-3000]Hz	**90.0 (18)**	**65.0 (13)**	**65.0 (13)**	**70.0 (14)**	**72.50 (58)**	9.9
[300-3000]Hz	85.0 (17)	55.0 (11)	55.0 (11)	55.0 (11)	62.50 (50)	10.7

TAB. 6.9 – *Résultats de la classification 4-G selon la bande de fréquences [300-3000]Hz en termes de % TCC (24LFSC)*

Le tableau 6.10 présente les matrices de confusion de la classification automatique des voix dysphoniques sur les bandes fréquentielles, [0-3000]Hz et [300-3000]Hz. Comme attendu, il peut être observé que tous les grades sont affectés par le retrait des basses fréquences [0-300]Hz, mis en évidence par une perte du TCC global absolu de 10.0 % (de 72.5 % à 62.5 %). En effet, la confusion avec les grades adjacents a augmenté radicalement, et plus particulièrement pour les voix de grade 3 dont la valeur du TCC régresse de 70.0 % à 55.0 %.

[0-3000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**18**	1	1	0
P_G1	1	**13**	6	0
P_G2	0	6	**13**	1
P_G3	0	2	4	**14**

[300-3000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**17**	2	1	0
P_G1	3	**11**	5	1
P_G2	0	7	**11**	2
P_G3	0	2	7	**11**

TAB. *6.10* – *Matrices de confusion de la classification automatique des voix dysphoniques filtrés en [0 - 3000]Hz et en [300 - 3000]Hz (24LFSC)*

Discussion

Sachant que l'analyse des caractéristiques spectrales du son laryngé[5] constitue une méthode objective pour l'évaluation de la dysphonie, le retrait des basses fréquences (inférieures à 300 Hz) ne peut que retirer de l'information utile à la caractérisation du trouble vocal et provoquer une baisse des performances du système automatique.

En extrapolant les observations faites sur le jugement perceptif entre les bandes de fréquences [0-8000]Hz et [0-3000]Hz, on peut plus facilement comprendre que la restriction à la plage de fréquences du téléphone peut affecter le jugement perceptif de façon similaire. Dans des conditions réelles, une baisse plus importante des performances peut être attendue si l'on prend en compte les autres problèmes liés au canal téléphonique tels que la distorsion d'amplitude et de phase du signal, la distorsion du temps de propagation, le décalage de fréquences, les sauts de phase, l'écho et le bruit.

Sur la bande [300-3000]Hz, la baisse observée des performances du système, principalement pour les voix dysphoniques, tend à montrer que les fréquences inférieures à 300 Hz contiennent de l'information utile à la caractérisation du trouble vocal. Avec notamment, la présence de la fréquence fondamentale (F0) dont la moyenne pour les femmes dysphoniques diminue globalement avec le degré de sévérité du trouble vocal (tendance inverse chez les hommes). Le retrait de cette plage fréquentielle, pertinente pour la discrimination de la voix dysphonique, corrobore l'impression auditive perçue d'une personne dysphonique au téléphone. En effet, une personne dysphonique paraît avoir une meilleure qualité de voix au téléphone qu'en conversation directe c-à-d le caractère pathologique de sa voix est atténué par la bande passante téléphonique.

En ce sens, une expérience semblable a été menée sur la bande de fréquences [300-3400]Hz, montrant une plus grande confusion entre les grades avec un TCC global de 58.75 % comme le montre les tableaux 6.11 et 6.12.

[5] notamment de la fréquence fondamentale et de ses variations

	Grade 0	Grade 1	Grade 2	Grade 3	Global	
Paramètres 24LFSC	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/20)	**% TCC** (nb/80)	**± IC**
[300-3000]Hz	**85.0 (17)**	**55.0 (11)**	**55.0 (11)**	55.0 (11)	**62.50 (50)**	10.7
[300-3400]Hz	80.0 (16)	45.0 (9)	45.0 (9)	**65.0 (13)**	58.75 (47)	10.9

TAB. *6.11* – *Résultats de la classification 4-G selon les bandes de fréquences [300-3000]Hz et [300-3400]Hz en termes de % TCC (24LFSC)*

[300-3000]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**17**	2	1	0
P_G1	3	**11**	5	1
P_G2	0	7	**11**	2
P_G3	0	2	7	**11**

[300-3400]Hz

	S_G0	S_G1	S_G2	S_G3
P_G0	**16**	3	1	0
P_G1	4	**9**	6	1
P_G2	0	9	**9**	2
P_G3	0	2	5	**13**

TAB. *6.12* – *Matrices de confusion de la classification automatique des voix dysphoniques filtrés en [300 - 3000]Hz et en [300 - 3400]Hz (24LFSC)*

Seules les voix les plus dysphoniques (grade 3) tirent bénéfice de l'aggrandissement de la plage [300-3000]Hz à [300-3400]Hz avec un TCC de 65.0 % contre 55.0 %. La tendance à la baisse observée sur les voix de grade 1 et 2 (TCC de 45.0 % contre 55.0 %) semble trouver une explication dans l'étude menée sur les sous-bandes individuelles (tableau 6.1). En effet, dans la sous-bande [3000-4000]Hz, les voix dysphoniques présentaient une forte confusion avec de très faibles TCC variant de 35.0 % (grade 1 et 2) à 25.0 % (grade 3). Si une explication peut être apportée pour les grades 1 et 2, il n'en est pas de même pour le gain affiché par les voix de grade 3 au regard de la dernière observation contradictoire, si ce n'est que seule la plage [3000-3400]Hz leur est particulièrement profitable.

D'une manière plus générale, face à l'émergence de nouvelles technologies de communication dans notre société, l'affaiblissement observé des performances du système sur la bande passante téléphonique soulève une importante question au sujet des personnes souffrant de dysphonie : « est-ce que les patients dysphoniques faisant usage du téléphone pour converser ou accéder à des services vocaux, sont pénalisés par leur qualité de voix produisant une voix dysharmonieuse et/ou incompréhensible ? »

6.4 Conclusion

Nous avons proposé d'étudier la manière dont les caractéristiques acoustiques de la dysphonie sont réparties sur l'ensemble de l'espace fréquentiel en analysant les performances d'une classification automatique de la voix dysphonique selon différentes plages de fréquences.

A travers cette approche, trois principales tendances peuvent être observées :

1. les fréquences inférieures à 3000Hz affichent les meilleurs résultats. Leur regroupement en [0-3000]Hz constitue la plage de fréquences la plus intéréssante. Les performances améliorées conduisent à une discrimination plus homogène sur l'ensemble des grades ;
2. la plage intermédiaire de [3000-5000]Hz exhibe les plus faibles résultats traduisant une forte confusion entre les voix dysphoniques ;
3. pour les fréquences supérieures à 5000Hz, seules les voix de grade 3 obtiennent des scores TCC satisfaisants (entre 70 % et 80 %) en dépit de la « *quasi-absence* » de parole dans cette zone. Leur regroupement en [5400-8000]Hz semble ne profiter que pour :
   ○ ***les voix normales (G0)*** *pour lesquelles, il est difficile d'apporter une explication ;*
   ○ ***les voix les plus dysphoniques (G3)*** *pour lesquelles la présence de bruit résiduel « voilé » (ou « soufflé »), de bruit de turbulence causé par un mauvais accolement des cordes vocales ou un comportement du forçage vocal intense, peuvent éventuellement en être la raison.*

La pertinence de la bande restreinte [0-3000]Hz relevée dans cette étude, n'apparaît pas comme « inattendue ». Elle représente la région fréquentielle contenant l'information liée à l'instabilité vibratoire de la glotte qui constitue une cause essentielle des dysphonies.

Concernant l'évaluation perceptive du corpus CVD filtré en [0-3000]Hz, trois observations peuvent être faites :

1. une concordance est obtenue de 85 % entre les évaluations perceptives, [0-3000]Hz et [0-8000]Hz ;
2. les désaccords perceptifs portent principalement sur les voix normales ou de grade 1 ;
3. aucune corrélation n'a pu être établie sur les voix mal classées sur la bande de fréquences [0-3000]Hz entre le système automatique et le jugement perceptif.

Finalement, l'analyse de la bande [300-3000]Hz a montré que la plage [0-300]Hz constitue une région fréquentielle contenant de l'information utile à la caractérisation du trouble vocal. Cette observation renforce l'impression auditive perçue par certains spécialistes de la voix pour lesquels, « *au téléphone, une personne dysphonique paraît avoir une meilleure qualité de voix qu'en situation conversationnelle directe.* »

Chapitre 7

L'étude phonétique

Sommaire

Résumé

A propos de l'évaluation de la voix pathologique, ce chapitre poursuit la caractérisation de la dysphonie dans le domaine fréquentiel afin de mieux en comprendre les phénomènes spécifiques. Partant de l'étude décrite dans le chapitre 6 montrant la pertinence des basses fréquences [0-3000] Hz pour la discrimination de la dysphonie par rapport aux plus hautes fréquences [0-8000]Hz, nous proposons ici d'analyser phonétiquement l'impact de la plage de fréquences [0-3000]Hz sur la reconnaissance du degré de sévérité de la voix pathologique.

Ce chapitre poursuit le travail developpé dans le chapitre 6 et publié dans [Pouchoulin et al., 2007], dans lequel une étude a porté sur les caractéristiques de la dysphonie dans le domaine fréquentiel à travers une analyse par sous-bande de fréquences. Dans ce contexte, il a été montré que la plage de fréquences [0-3000]Hz semble plus pertinente en terme de discrimination des voix dysphoniques que les bandes de plus hautes fréquences, voire que la bande totale [0-8000]Hz.

Nous proposons d'étudier ici les manifestations de la dysphonie selon la nature des segments phonémiques, au moyen d'une analyse phonétique sur les plages de fréquences [0-8000]Hz et [0-3000]Hz. Dans ce sens, les décisions fournies par la classification automatique des voix dysphoniques seront analysées selon différentes classes de phonèmes.

Partant d'une observation relevée durant l'analyse phonétique automatique, nous présenterons une analyse statistique manuelle consistant à montrer la corrélation entre l'allongement de la durée du VOT dans un contexte « $C_{sourde}V$ » et le degré de sévérité de la dysphonie.

Présentée en 1.5.3, l'approche « Phonetic Labeling » se définit comme une étude descriptive et perceptive des caractéristiques pathologiques de différents phonèmes. Très proche de l'analyse phonétique décrite dans ce chapitre, nous proposons de comparer les résultats obtenus par le système automatique (appelée « Automatic Phonetic Labeling ») et les observations relevées par Revis et al. (2006) dans l'approche « Phonetic Labeling ».

7.1 L'analyse phonétique : [0-3000]Hz vs [0-8000]Hz

L'analyse phonétique automatique consiste à observer le comportement du système de RAL dans l'attribution du grade de dysphonie selon différentes classes de phonèmes. Ces comportements seront analysés sur les 2 bandes de fréquences, [0-3000]Hz et [0-8000]Hz, par phonème ou classe de phonèmes, afin d'évaluer l'impact des effets de la dysphonie selon les grades. La paramétrisation utilisée est du type LFSC (24 coefficients statiques) dont les vecteurs acoustiques sont normalisés pour obtenir une distribution de moyenne-0 et de variance-1. La version de la segmentation phonétique mise en œuvre pour les signaux acoustiques, sera la version AP2 (détails à la section 2.5).

Il est à noter que cette catégorisation phonétique n'est utilisée que durant la phase de décision. En effet, les phases de paramétrisation et de modélisation utiliseront l'ensemble du matériau phonémique disponible dans chaque signal de parole du corpus. Par contre, lors d'un test de classification, la décision sera prise sur l'ensemble des segments associés à une classe phonétique. Le tableau 7.1 fournit les différentes classes de phonèmes disponibles dans le corpus de voix dysphoniques avec leur durée (en secondes) pour chaque grade, ainsi que des informations quantitatives sur les phonèmes d'une classe phonétique.

Classes	**Grades**				**Effectifs Totaux**		
Phonétiques	*G0*	*G1*	*G2*	*G3*	*nb*	μ	σ
Consonne	**135.13**	**139.21**	**149.83**	**167.28**	**6395**	**0.092**	**0.045**
. Sonore	88.80	90.56	95.36	106.57	4719	0.081	0.039
. Sourde	46.33	48.65	54.47	60.71	1676	0.125	0.046
Liquide	34.56	34.01	36.04	43.03	2181	0.068	0.033
Nasale	29.72	30.17	31.85	33.42	1279	0.098	0.039
Fricative	31.77	32.32	35.07	40.70	1144	0.122	0.057
. Sonore	*10.14*	*10.32*	*10.45*	*11.76*	*436*	*0.098*	*0.056*
. Sourde	21.63	22.00	24.62	28.94	708	0.137	0.052
Occlusive	39.08	42.71	46.87	50.13	1791	0.100	0.039
. Sonore	*14.38*	*16.06*	*17.02*	*18.36*	*823*	*0.080*	*0.030*
. Sourde	24.70	26.65	29.85	31.77	968	0.117	0.038
Voyelle	**103.58**	**98.77**	**103.46**	**109.79**	**5586**	**0.074**	**0.046**
Orale	84.37	80.45	85.22	93.66	4862	0.071	0.044
Nasale	*19.21*	*18.32*	*18.24*	*16.13*	*724*	*0.099*	*0.046*
Semi-voyelle	***2.80***	***2.98***	***3.37***	***3.45***	***159***	***0.079***	***0.040***
Tous phonèmes	**241.51**	**240.96**	**256.66**	**280.52**	**12140**	**0.084**	**0.046**
. Sonore	195.18	192.31	202.19	219.81	10464	0.077	0.043
. Sourde	46.33	48.65	54.47	60.71	1676	0.125	0.046

TAB. 7.1 – *Durée en secondes par classe phonétique (version AP2) et par grade - Informations quantitatives sur les phonèmes d'une classe phonétique : nombre (nb) avec durée moyenne (μ) et écart-type (σ) associé*

Remarque : les classes de phonèmes avec une durée inférieure à 20s (notées en italique) *i.e.* moins de 1 seconde par locuteur, ne sont données que pour information et ne seront pas prises en compte pour l'analyse phonétique, les résultats associés étant jugés ici peu fiables.

7.1.1 Les résultats « bruts »

Les deux tableaux 7.2 fournissent les performances du système de classification automatique des voix dysphoniques selon les différentes classes de phonèmes et les deux bandes de fréquences : totale [0-8000]Hz et restreinte [0-3000]Hz. Les résultats fournis sont exprimés en terme de Taux Correct de Classification (nommé *TCC*) suivi du nombre de voix correctement classées mentionné entre parenthèse. Comme spécifié précédemment, les classes de phonèmes notées en italique dans le tableau 7.1 de durée inférieure à 20s par grade, ne seront pas analysées dans cette section. Ceux-ci apparaîtront tout de même « barrés » dans les tableaux en 7.2 à titre informatif.

[0-8000]Hz	Grade 0	Grade 1	Grade 2	Grade 3	Global	
Classes phonétiques	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/80)	± IC
Consonne	**80.0 (16)**	**50.0 (10)**	**50.0 (10)**	**85.0 (17)**	**66.25 (53)**	**10.4**
. sonore	70.0 (14)	45.0 (9)	60.0 (12)	75.0 (15)	62.50 (50)	10.7
. sourde	80.0 (16)	60.0 (12)	50.0 (10)	75.0 (15)	66.25 (53)	10.4
Liquide	50.0 (10)	35.0 (7)	40.0 (8)	70.0 (14)	48.75 (39)	11.0
Nasale	50.0 (10)	35.0 (7)	50.0 (10)	45.0 (9)	45.00 (36)	11.0
Fricative	80.0 (16)	40.0 (8)	50.0 (10)	85.0 (17)	63.75 (51)	10.6
~~. sonore~~	~~70.0 (14)~~	~~10.0 (2)~~	~~25.0 (5)~~	~~85.0 (17)~~	~~47.50 (38)~~	~~11.0~~
. sourde	80.0 (16)	45.0 (9)	35.0 (7)	90.0 (18)	62.50 (50)	10.7
Occlusive	65.0 (13)	65.0 (13)	50.0 (10)	70.0 (14)	62.50 (50)	10.7
~~. sonore~~	~~65.0 (13)~~	~~55.0 (11)~~	~~40.0 (8)~~	~~75.0 (15)~~	~~58.75 (47)~~	~~10.9~~
. sourde	75.0 (15)	70.0 (14)	50.0 (10)	70.0 (14)	66.25 (53)	10.4
Voyelle	**70.0 (14)**	**55.0 (11)**	**40.0 (8)**	**60.0 (12)**	**56.25 (45)**	**10.9**
Orale	55.0 (11)	60.0 (12)	40.0 (8)	60.0 (12)	53.75 (43)	11.0
~~Nasale~~	~~70.0 (14)~~	~~20.0 (4)~~	~~20.0 (4)~~	~~65.0 (13)~~	~~43.75 (35)~~	~~10.9~~
~~**Semi-voyelle**~~	~~**25.0 (5)**~~	~~**25.0 (5)**~~	~~**10.0 (2)**~~	~~**65.0 (13)**~~	~~**31.25 (25)**~~	~~**10.2**~~
Tous phonèmes	**85.0 (17)**	**55.0 (11)**	**50.0 (10)**	**70.0 (14)**	**65.00 (52)**	**10.5**
. sonore	75.0 (15)	55.0 (11)	45.0 (9)	65.0 (13)	60.00 (48)	10.8
. sourde	80.0 (16)	60.0 (12)	50.0 (10)	75.0 (15)	66.25 (53)	10.4

[0-3000]Hz	Grade 0	Grade 1	Grade 2	Grade 3	Global	
Classes phonétiques	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/80)	± IC
Consonne	**90.0 (18)**	**55.0 (11)**	**70.0 (14)**	**75.0 (15)**	**72.50 (58)**	**9.9**
. sonore	75.0 (15)	70.0 (14)	30.0 (6)	65.0 (13)	60.00 (48)	10.8
. sourde	90.0 (18)	35.0 (7)	75.0 (15)	60.0 (12)	65.00 (52)	10.5
Liquide	80.0 (16)	50.0 (10)	20.0 (4)	45.0 (9)	48.75 (39)	11.0
Nasale	30.0 (6)	35.0 (7)	55.0 (11)	35.0 (7)	38.75 (31)	10.7
Fricative	85.0 (17)	40.0 (8)	35.0 (7)	85.0 (17)	61.25 (49)	10.7
~~. sonore~~	~~80.0 (16)~~	~~20.0 (4)~~	~~25.0 (5)~~	~~60.0 (12)~~	~~46.25 (37)~~	~~11.0~~
. sourde	80.0 (16)	35.0 (7)	30.0 (6)	80.0 (16)	56.25 (45)	11.0
Occlusive	90.0 (18)	60.0 (12)	50.0 (10)	70.0 (14)	67.50 (54)	10.3
~~. sonore~~	~~30.0 (6)~~	~~45.0 (9)~~	~~15.0 (3)~~	~~75.0 (15)~~	~~41.25 (33)~~	~~10.9~~
. sourde	95.0 (19)	50.0 (10)	65.0 (13)	50.0 (10)	65.00 (52)	10.5
Voyelle	**85.0 (17)**	**30.0 (6)**	**70.0 (14)**	**55.0 (11)**	**60.00 (48)**	**10.8**
Orale	75.0 (15)	35.0 (7)	50.0 (10)	50.0 (10)	52.50 (42)	11.0
~~Nasale~~	~~90.0 (18)~~	~~20.0 (4)~~	~~45.0 (9)~~	~~45.0 (9)~~	~~50.00 (40)~~	~~11.0~~
~~**Semi-voyelle**~~	~~**20.0 (4)**~~	~~**20.0 (4)**~~	~~**50.0 (10)**~~	~~**55.0 (11)**~~	~~**36.25 (29)**~~	~~**10.6**~~
Tous phonèmes	**90.0 (18)**	**65.0 (13)**	**65.0 (13)**	**70.0 (14)**	**72.50 (58)**	**9.9**
. sonore	75.0 (15)	70.0 (14)	50.0 (10)	60.0 (12)	63.75 (51)	10.6
. sourde	90.0 (18)	35.0 (7)	75.0 (15)	60.0 (12)	65.00 (52)	10.5

TAB. 7.2 – *Résultats de classification 4-G par classe phonétique en terme de % TCC selon les 2 bandes de fréquences : totale [0-8000]Hz et restreinte [0-3000]Hz (24LFSC)*

7.1.2 Les performances globales

Le tableau 7.3 reprend les résultats obtenus sur l'ensemble des phonèmes qui apparaissent sur la ligne nommée « **Tous phonèmes** » dans les tableaux 7.2. Un TCC global de 65% est obtenu pour la bande totale [0-8000]Hz contre 72.5% pour la sous-bande [0-3000]Hz comme cela est observé dans la section 6.1.4.

	Grade 0	**Grade 1**	**Grade 2**	**Grade 3**	**Global**	
24LFSC	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/20)	% TCC (nb/80)	± **IC**
[0-8000]Hz	85.0 (17)	55.0 (11)	50.0 (10)	**70.0 (14)**	65.00 (52)	10.5
[0-3000]Hz	**90.0 (18)**	**65.0 (13)**	**65.0 (13)**	**70.0 (14)**	**72.50 (58)**	9.9

TAB. 7.3 – *Résultats de classification 4-G selon les bandes de fréquences [0-8000]Hz et [0-3000]Hz en terme de % TCC (24LFSC)*

[0-8000]Hz (Bande Totale)

	RG0	**RG1**	**RG2**	**RG3**
TG0	**17**	2	1	0
TG1	2	**11**	5	2
TG2	2	6	**10**	2
TG3	0	1	5	**14**

[0-3000]Hz (Bande Restreinte)

	RG0	**RG1**	**RG2**	**RG3**
TG0	**18**	1	1	0
TG1	1	**13**	6	0
TG2	0	6	**13**	1
TG3	0	2	4	**14**

TAB. 7.4 – *Matrices de confusion en classification 4-G selon les bandes de fréquences [0-8000]Hz et [0-3000]Hz (24LFSC)*

Il est intéressant de souligner que l'amélioration concerne principalement le grade 1 (de 55% à 65%) et le grade 2 (de 50% à 65%), grades sur lesquels la plus grande confusion de classification est généralement observée. Comme le montre les matrices de confusion présentées en 7.4, la réduction de la plage fréquentielle apporte une meilleure répartition des erreurs *i.e.* localisation des erreurs d'un grade à proximité de celui-ci. En [0-3000]Hz, la majorité des erreurs de classification des grades 1 et 2 (soit 6 locuteurs chacun) sont évaluées en G2 et en G1 respectivement. De plus, aucune voix de grade 0 et 1 n'est évaluée comme une dysphonie sévère (grade 3), ni aucune voix de grade 2 et 3 n'est évaluée comme une voix normale (grade 0).

7.1.3 L'analyse phonétique en [0-8000]Hz

Sur la bande de fréquences [0-8000]Hz, on peut observer que :

• le grade 0 obtient 80% de TCC sur la classe des consonnes et plus particulièrement sur les fricatives sourdes. Un TCC de 70% est obtenu sur la classe des voyelles et cela, malgré le score de 55% atteint par les voyelles orales. Comparé avec les autres grades, le grade 0 fournit le meilleur TCC (85%) sur l'ensemble des phonèmes ;

• le grade 1 affiche des TCC du même ordre pour les classes des voyelles et des consonnes (55% et 50% resp.), avec 55% sur l'ensemble des phonèmes. De meilleurs TCC sont obtenus par les occlusives sourdes (70%) et les voyelles orales (60%). Par contre, les consonnes liquides et nasales obtiennent les plus faibles TCC avec 35% ;

• concernant le grade 2, la plupart des classes obtiennent des TCC plutôt faibles (en dessous de 50%). Seules les consonnes sonores dépassent ce seuil avec un TCC de 60%. Ces faibles TCC sont très proches du TCC de 50% obtenu sur l'ensemble des phonèmes ;

• pour le grade 3, la différence en terme de TCC entre la classe des consonnes (85%) et celle des voyelles (60%) est la plus grande en comparaison avec les autres grades. De plus, les consonnes obtiennent le meilleur TCC (90%) pour les fricatives sourdes et le plus faible (45%) avec les consonnes nasales. Un TCC intermédiaire de 70% est obtenu sur l'ensemble des phonèmes.

Concernant la colonne « Global », la classe des consonnes (notamment les consonnes sourdes) et la classe « Tous phonèmes » obtiennent des résultats très similaires (66.25% contre 65%). L'écart entre la classe vocalique (56.25%) et la classe consonantique (66.25%) est assez important. Finalement, les consonnes liquides et nasales affichent les plus faibles TCC (48.75% et 45% respectivement).

7.1.4 L'analyse comparative [0-8000]Hz vs [0-3000]Hz

L'analyse comparative va être présentée de manière ascendante c-à-d elle portera tout d'abord sur l'ensemble des classes de phonèmes, puis sur les consonnes et voyelles et enfin sur les consonnes.

► ***Ensemble des classes de phonèmes***

Comparant les performances du système automatique entre les deux plages de fréquences, [0-8000]Hz et [0-3000]Hz, il peut être observé sur les deux tableaux 7.2 ainsi que sur la figure 7.1 que :

• pour le grade 0, les valeurs de TCC sont améliorées sur l'ensemble des classes phonétiques en [0-3000]Hz, à l'exception des fricatives sourdes qui conservent leur TCC de 80% et des consonnes nasales pour lesquelles, le TCC baisse de 50% ([0-8000]Hz) à 30% ([0-3000]Hz). Les améliorations peuvent varier de 7% à 60% (consonnes liquides) en relatif, le meilleur TCC étant obtenu pour les occlusives sourdes (95 %) ;

• pour le grade 1, les valeurs TCC sont généralement plus faibles en [0-3000]Hz, notamment pour la classe vocalique avec seulement 30%. Les TCC des fricatives sourdes, occlusives sourdes et des voyelles orales, ont baissé par rapport à la bande [0-8000]Hz. Par contre, les consonnes liquides et sonores obtiennent les meilleurs TCC (50% et 70% resp. contre 35% et 45%), améliorant le TCC de la classe « tous phonèmes » (65% en [0-3000]Hz contre 55% en [0-8000]Hz) ;

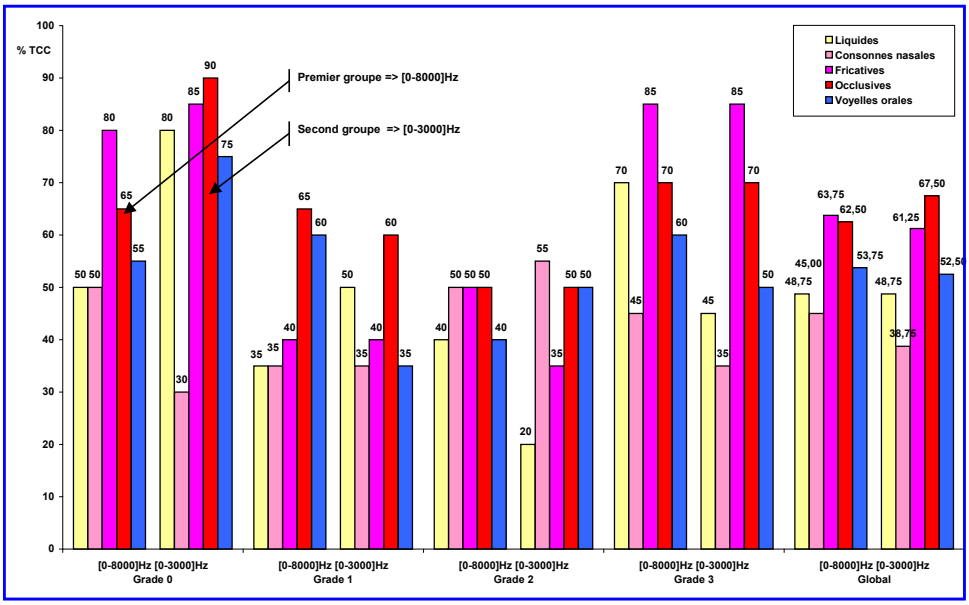


FIG. 7.1 – *Résultats de classification 4-G par classe phonétique en terme de % TCC (24LFSC) Comparaison entre la bande totale [0-8000]Hz et la bande restreinte [0-3000]Hz.*

• pour le grade 2, les classes des consonnes et des voyelles obtiennent un TCC satisfaisant de 70% en [0-3000]Hz, légèrement au-dessus du TCC de l'ensemble des phonèmes (65%) mais largement au-dessus des TCC obtenus sur la bande [0-8000]Hz (50% et 40% resp.). Par contre, des valeurs très faibles de TCC sont observées pour certaines classes de consonnes (sonores 30%, liquides 20%, fricatives sourdes 30%). Néanmoins, le TCC sur l'ensemble des phonèmes est bien meilleur en [0-3000]Hz avec 65% contre 50% en [0-8000]Hz ;

• pour le grade 3, la plupart des classes présentent une baisse de TCC en [0-3000]Hz, à l'exception des fricatives et occlusives qui conservent leur TCC (85% et 70% resp.). Malgré cela, 70% de TCC est atteint sur l'ensemble des phonèmes pour les deux bandes de fréquences.

L'analyse des résultats de la colonne « Global » nous amène aux mêmes observations qu'en [0-8000]Hz. La classe des consonnes atteint la même performance que celle obtenue sur l'ensemble des phonèmes en [0-3000]Hz, avec un TCC de 72.5%. Les résultats de la classe des voyelles restent plus faibles comparés à ceux de la classe des consonnes. Finalement, les consonnes liquides (48.75%) et nasales (38.75%) affichent les plus faibles TCC en [0-3000]Hz.

► ***Les consonnes et les voyelles***

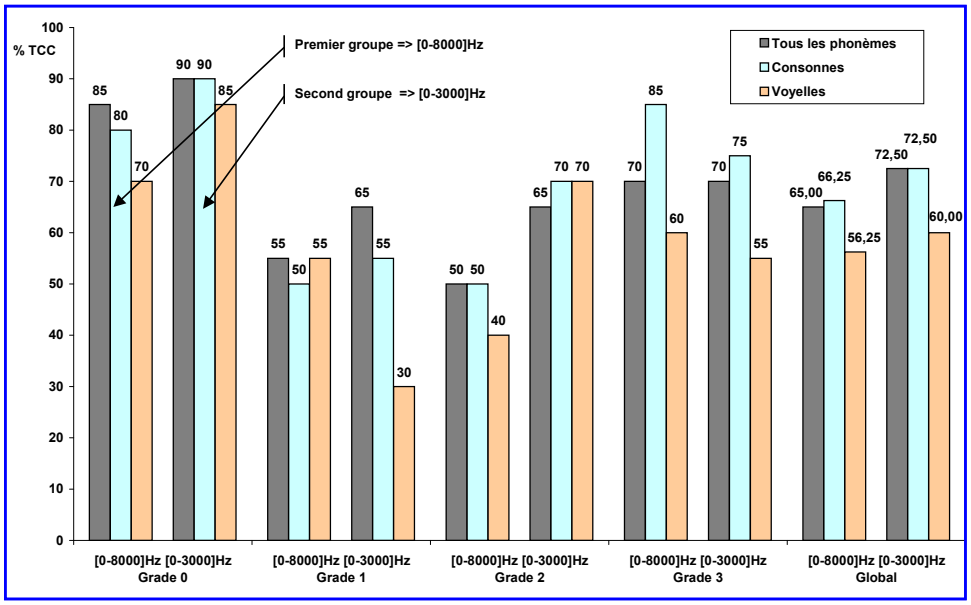


FIG. 7.2 – *Résultats de classification 4-G par classe phonétique en terme de % TCC (24LFSC) Comparaison entre la bande totale [0-8000]Hz et la bande restreinte [0-3000]Hz.*

La figure 7.2 présente pour chacun des grades, les résultats obtenus sur les deux bandes de fréquences ([0-8000]Hz et [0-3000]Hz) pour l'ensemble des phonèmes et les classes des consonnes et voyelles. Pour chacun des grades, on observe que :

• pour le grade 0, les valeurs sont améliorées sur la bande restreinte avec un TCC de 90.0 % pour l'ensemble des phonèmes et les consonnes, et 85.0 % pour les voyelles ;

• pour le grade 1, la classe des voyelles affiche une perte importante en [0-3000]Hz avec une baisse du TCC de 55.0 % à 30.0 % (soit -5 locuteurs). La classe des consonnes gagne un seul locuteur sur la bande restreinte avec un TCC de 55.0 %. Malgré cela, les dysphonies légères obtiennent un meilleur résultat sur la bande [0-3000]Hz en améliorant leur TCC de 55.0 % à 65.0 %. Ce comportement semble s'expliquer par un phénomène compensatoire lors de la prise de décision sur l'ensemble des phonèmes ;

• comme pour le grade 0, le grade 2 présente de meilleurs résultats sur la bande restreinte avec un score TCC de 70.0 % obtenu pour les voyelles et les consonnes, permettant d'atteindre 65.0 % sur l'ensemble des phonèmes ;

• sur l'ensemble des phonèmes, le grade 3 conserve un même TCC de 70.0 % pour les 2 bandes de fréquences. Cependant, on notera une baisse des performances en [0-3000]Hz pour les consonnes (de 85.0 % à 75.0 %) et les voyelles (de 60.0 % à 55.0 %).

Concernant la réduction à la plage de fréquences [0-3000]Hz, on observe :

1. Ensemble des phonèmes ⇒ amélioration sauf pour le grade G3 ;
2. Classe des consonnes ⇒ amélioration sauf pour le grade G3 ;
3. Classe des voyelles ⇒ amélioration sauf pour les grades G1 et G3.

Quelle que soit la bande de fréquences considérée, la classe des consonnes obtient de meilleurs résultats que la classe des voyelles. Cette observation est surprenante sachant que d'une part, les pathologies vocales étudiées sont directement liées à des altérations des sons laryngés (ce sont principalement des lésions des cordes vocales) ; d'autre part, la réduction à la plage de fréquences [0-3000]Hz aurait tendance à être plus favorable aux phonèmes voisés (principalement les sons vocaliques i.e. aptes à produire des voyelles).

Pour ces raisons, il nous aurait semblé logique que ce soit la classe des voyelles qui obtiennent de meilleurs performances que les consonnes.

► ***Les consonnes***

Nous nous sommes donc intéressés aux résultats obtenus sur la classe des consonnes en différenciant les sonores des sourdes comme illustré par la figure 7.3.

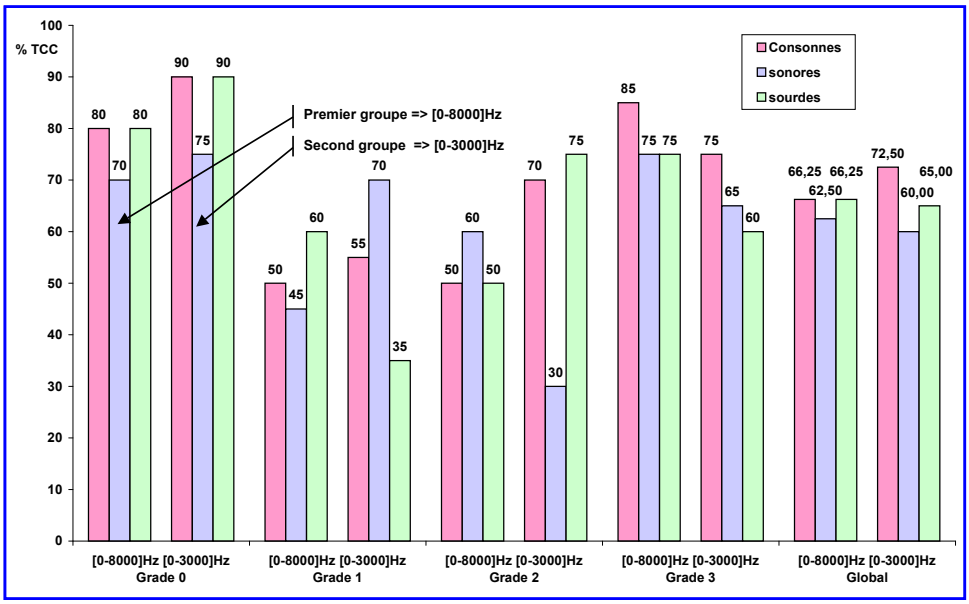


FIG. 7.3 – *Résultats de classification 4-G par classe phonétique en terme de % TCC (24LFSC) Comparaison entre la bande totale [0-8000]Hz et la bande restreinte [0-3000]Hz.*

Pour les mêmes raisons que celles évoquées précédemment, on s'attendrait à ce que les sonores soient plus discriminantes que les sourdes.

Pour chacun des grades, on observe que :

• pour le grade 0, les sourdes obtiennent le meilleur TCC de 90.0 % en [0-3000]Hz ;

• pour le grade 1, le comportement observé sur la bande totale s'inverse en [0-3000]Hz. Sur la bande [0-3000]Hz, on observe un gain important pour les sonores (70.0 % contre 45.0 %) et une baisse importante pour les sourdes (35.0 % contre 60.0 %). Malgré cela, la réduction de la bande de fréquences améliore sensiblement la valeur de la classe des consonnes avec un TCC de 55.0 %. Le comportement observé en [0-3000]Hz est celui attendu ;

• pour le grade 2, on observe un comportement inverse de celui observé pour le grade 1. Sur la bande [0-3000]Hz, les sourdes améliorent fortement leur score TCC, 75.0 % contre 50.0 %, tandis que les sonores accusent une forte baisse de leur TCC avec 30.0 % contre 60.0 %. Cela produit une amélioration conséquente sur la bande [0-3000]Hz de la classe des consonnes avec un score TCC de 70.0 % contre 50.0 % en bande totale ;

• pour le grade 3, les 3 scores TCC présentent une baisse en [0-3000]Hz comme observé précédemment.

Concernant la réduction à la plage de fréquences [0-3000]Hz, on observe :

1. Classe des consonnes ⇒ amélioration sauf pour le grade G3 ;
2. Classe des consonnes sonores ⇒ amélioration sauf pour les grades G2 et G3 ;
3. Classe des consonnes sourdes ⇒ amélioration sauf pour les grades G1 et G3.

7.1.5 Discussion

Les résultats présentés ci-dessus laissent apparaître que la classe consonantique semble être la plus pertinente pour la classification des voix dysphoniques quelle que soit la bande de fréquences considérée dans ce contexte expérimental. Cette observation soulève 3 questions.

Question 1
Cette observation est-elle conflictuelle au regard des évaluations perceptives et objectives qui s'appuient sur des voyelles tenues comme le /a/ ***?***

Le choix de ce support phonétique permet des conditions expérimentales indispensables pour évaluer la stabilité et le bruit du vibrateur en régime permanent (fluctuations à court terme telles que jitter, shimmer, HNR, ...)[1], sachant que l'instabilité vibratoire de la glotte est une cause essentielle des dysphonies.

[1]Variations cycle à cycle de la durée de cycle (jitter) et de l'amplitude (shimmer), harmonics-to-noise ratio (HNR)

En parole continue, il semble impossible d'extraire de tels indices de fluctuation car la succession très rapide des phonèmes ne permet pas d'avoir des parties stables par opposition à la voyelle tenue sur laquelle la partie stable peut atteindre plusieurs secondes. La parole rend compte des phénomènes phonémiques (coarticulation, transitions formantiques, attaques, interruptions vibratoires, ...) et prosodiques (rythme, intonation, mélodie, ...). La succession des phonèmes est extrêmement rapide, de l'ordre de quelques millisecondes, c'est pourquoi il est souvent considéré qu'il n'y a pas de partie stable (par opposition à la voyelle tenue sur plusieurs secondes).

Néanmoins, le support phonétique basé sur les voyelles tenues reste controversé dans la littérature car il tend à sous-estimer la dysphonie. Par ailleurs, certains phénomènes vocaux issus de la parole spontanée comme l'attaque sont reconnus comme pertinents dans l'évaluation des dysphonies [Revis et al., 2002].

Au regard de ces éléments et de la pertinence des consonnes mise en évidence dans cette étude, il semblerait par conséquent intéressant d'élargir le cadre de l'analyse phonétique menée ici, à l'étude de certains phénomènes vocaux transitoires comme le passage entre phonèmes « voisé ↷ non voisé » ou la séquence CV induisant la présence simultanée d'informations de la consonne et de la voyelle. Cette approche pourrait être, en fait, complémentaire aux méthodes d'évaluation basées sur les voyelles tenues.

Classes phonétiques	Grade 0	Grade 1	Grade 2	Grade 3
Consonne	Gain (+2)	Gain (+1)	Gain (+4)	Perte (-2)
Voyelle	Gain (+3)	Perte (-5)	Gain (+6)	Perte (-1)
Tous phonèmes	Gain (+1)	Gain (+2)	Gain (+3)	Stable

TAB. 7.5 – *Bilan de la réduction fréquentielle à la plage [0-3000]Hz en terme de Gain/Perte de locuteur(s) par classes phonétiques (24LFSC)*

Deuxièmement, la plage [0-3000]Hz a tendance à améliorer les TCC des grades 0 et 2 pour la classe des voyelles et celle des consonnes alors qu'elle pénalise les grades 1 et 3 lorsque les classes phonétiques sont considérées individuellement. Le tableau 7.5 présente le bilan par grade de l'analyse phonétique en terme de Gain/Perte de locuteur(s) entre la bande totale [0-8000]Hz et la bande restreinte [0-3000]Hz.

Le comportement en [0-3000]Hz du grade 1 (65% TCC sur l'ensemble des phonèmes) est particulièrement inattendu au regard des faibles TCC de la classe vocalique. Un phénomène compensatoire lors de la prise de décision sur l'ensemble des phonèmes semble ici être à l'origine de ce comportement. Il est intéressant de constater que l'observation des matrices de confusion en 7.4 sur l'ensemble des phonèmes en [0-3000]Hz montre une tendance à la sur-évaluation pour le grade 1 (6 voix en grade 2) et à la sous-évaluation pour le grade 2 (6 voix en grade 1).

Finalement, le comportement du grade 3 sur [0-3000]Hz peut probablement être dû au filtrage de « parole bruitée » présente dans les hautes fréquences et caractéristique de certaines voix sévèrement dysphoniques. Malgré cela, la constance du TCC de l'en-

semble des phonèmes sur les deux bandes de fréquences, montre que l'information pertinente pour la discrimination du grade 3 vis à vis des autres grades est toujours présente en [0-3000]Hz.

Question 2
N'y-a-t-il pas un problème sur la qualité de la segmentation automatique ?

Pour Hammarberg (2000) : « les variations survenant dans la parole, comme l'attaque vocale, l'arrêt vibratoire, les cassures de la phonation, ..., sont des éléments cruciaux de la qualité de la voix ». Certains phénomènes vocaux issus de la parole spontanée, comme l'attaque vocale, sont reconnus comme pertinents pour l'évaluation des dysphonies. Donc au regard de ces éléments et de la pertinence des consonnes mise en évidence dans cette étude, et notamment sur les consonnes sourdes, nous nous sommes intéressés à la qualité de l'alignement phonétique. En effet, nous avons voulu vérifier si les frontières déterminées automatiquement sur les consonnes sourdes, n'incluaient pas de l'information provenant des phonèmes adjacents.

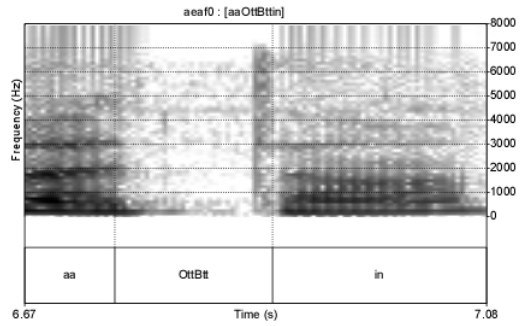


FIG. 7.4 – *Spectrogrammes sur la portion [aaOttBttin] du mot matin pour un locuteur de grade 0.*

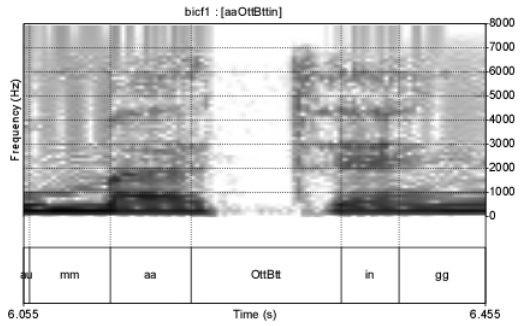


FIG. 7.5 – *Spectrogrammes sur la portion [aaOttBttin] du mot matin pour un locuteur de grade 1.*

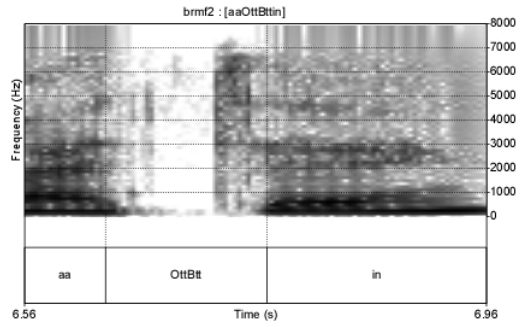


FIG. 7.6 – *Spectrogrammes sur la portion [aaOttBttin] du mot matin pour un locuteur de grade 2.*

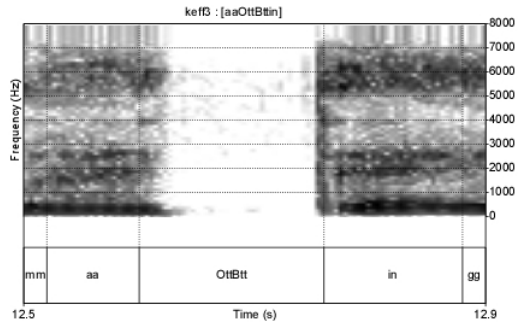


FIG. 7.7 – *Spectrogrammes sur la portion [aaOttBttin] du mot matin pour un locuteur de grade 3.*

A la vue de plusieurs spectrogrammes « large bande » (figures 7.4 à 7.7) et comme illustré par la figure 7.8, il apparaît que les consonnes sourdes dans les transitions de type « VC_{sourde} » incluent systématiquement[2] une partie de la finale de la voyelle précédente.

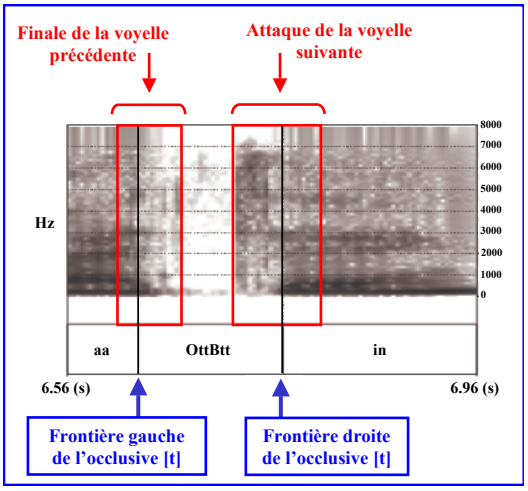


FIG. 7.8 – *Alignement automatique : frontières phonémiques de la séquence [aaOttBttin] du mot « matin ».*

Par contre, pour les transitions « $C_{sourde}V$ », cela paraît moins évident que la consonne sourde inclue une partie de l'attaque de la voyelle suivante. Cette constatation est intéressante car si elle se confirmait - ***si la portion finale d'une voyelle précédent un phonème non voisé ou l'attaque d'une voyelle suivant un phonème non voisé, sont des phénomènes transitoires caractéristiques de la dysphonie*** - cela montrerait, de façon inédite par la méthode, que la transition « VC_{sourde} » (et/ou « $C_{sourde}V$ ») est la partie la plus discriminante pour la catégorisation des dysphonies.

Il serait intéressant de déterminer manuellement les frontières réelles des consonnes sourdes sur l'ensemble du corpus dysphonique afin de quantifier l'attaque et la portion finale des voyelles incluses dans les consonnes sourdes. Il sera alors possible d'évaluer l'apport de l'attaque et/ou de la portion finale de la voyelle pour la discrimination des voix dysphoniques.

Question 3
La dysphonie est-elle uniquement un problème de source laryngée ?
Se manifeste-t-elle également au niveau supra-laryngé ?

Ces questions trouvent leur origine dans le comportement « plutôt inattendu » des paramètres cepstraux par rapport aux coefficients spectraux pour la reconnaissance des grades dysphoniques (chapitre 5). Par déconvolution spectrale, les premiers coefficients

[2] et nécessairement car la notion de frontière phonémique est un concept de travail mais pas une réalité physique

cepstraux sont censés ne contenir principalement que de l'information résonantielle du conduit vocal. Or, sachant que la dysphonie est directement liée à une perturbation de la vibration glottique, les deux interrogations soulevées trouvent ici toute leur justification. Cette question n'ayant pas encore été abordée à ce jour, elle rejoindra les perspectives de cette thèse.

Afin d'approfondir ces résultats observés sur les consonnes sourdes, nous avons entrepris une analyse sur les transitions entre les occlusives et les voyelles, et plus précisément sur le VOT qui sera décrit dans la section suivante 7.2.

7.2 L'étude du VOT

Souvent utilisé dans l'étude des dysarthries [Morris, 1989], le Voice Onset Time (VOT) correspond au temps d'établissement du voisement entre l'explosion d'une consonne occlusive et la mise en vibration des cordes vocales. Pour Özsancak et al. (2001), ce paramètre acoustique est un indice fiable de contrôle laryngé et de la coordination entre les organes articulateurs (supralaryngés) et le larynx. Il s'agit donc d'un mouvement délicat ; à un moment donné, il faut relâcher l'occlusion (apparition de l'explosion) tout en déclenchant la mise en vibration du larynx quelques milli-secondes après. Cela nécessite un geste très précis au niveau de la coordination des organes constricteurs et phonatoires car il s'agit à la fois de mouvements de langue/lèvres et du larynx

On sait depuis longtemps qu'un trouble neurologique peut entraîner des problèmes de coordination/contrôle de la motricité des mouvements. Cela peut se traduire par un rallongement ou une anticipation de l'ordre moteur. Ce phénomène n'a jamais été envisagé sur les dysphonies car le VOT relève de la coordination des organes. Mettre en évidence ce problème chez les dysphoniques signifie qu'il y a une perte du contrôle laryngé, une difficulté dans le geste phonatoire. Ne pouvant s'agir d'un problème neurologique, l'allongement de la mesure du VOT avec le grade peut signifier qu'au moment où la commande de mise en vibration des cordes vocales arrive, le larynx ne réagit pas immédiatement. Cette observation se rapproche de ce qui a été observé par Revis et al. (2000) sur de la parole isolée : «*l'attaque est porteuse d'une information sur la dysphonie*». Néanmoins cela n'a jamais été réellement montré sur de la parole articulée.

Dans l'analyse phonétique (section 7.1), le système automatique a mis en avant la pertinence des consonnes sourdes. Cette constatation est plutôt inattendue et surprenante sachant que le trouble de la dysphonie est directement lié à une perturbation de la vibration glottale. Une analyse manuelle a donc été entreprise afin de vérifier la qualité de l'alignement automatique. En effet, la première supposition s'orientait sur un défaut de positionnement de frontières phonémiques entre les consonnes sourdes et les phonèmes voisés adjacents. Cependant, cette intervention manuelle a plutôt souligné une qualité satisfaisante de la segmentation automatique. La deuxième supposition reposait alors sur un réel allongement de la partie «burst» de l'occlusive dans le contexte phonémique «$Occ_{sourde}V$» avec le degré de sévérité de la dysphonie. Un tel rallongement de durée permettrait la mise en évidence d'une dérégulation du larynx montrant ainsi un problème périphérique[3].

La question à laquelle nous allons répondre dans cette section est la suivante : «est-ce que l'allongement du VOT en fonction du grade est statistiquement significatif ?»

[3] problème physique et non neurologique

Résultats

Pour répondre à cette question, nous analyserons les mesures du VOT à l'aide du logiciel R [Gentleman & Ihaka, 1997] en ne prenant en compte que les séquences de phonèmes «$\text{Occlusive}_{\text{sourde}}\text{Phonème}_{\text{voisé}}$» du corpus CVD *i.e.* les occlusives sourdes ayant un contexte droit voisé. De plus, le système automatique présente la particularité d'utiliser deux symboles pour représenter les occlusives sourdes :

⟨ [Okk] + [Bkk] pour /k/ [Opp] + [Bpp] pour /p/ [Ott] + [Btt] pour /t/ ⟩

où le 1er symbole correspond à la tenue de l'occlusion et le 2^{e} symbole à l'explosion. Dans ces conditions, le VOT sera assimilé au 2^{e} symbole *i.e.* à la durée comprise entre l'explosion de la consonne occlusive et l'émergence de la 1re onde glottique du phonème voisé.

Le tableau 7.6 présente des statistiques descriptives calculées sur les mesures du VOT par grade (en secondes) extraites du corpus CVD.

Grade	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Total	Nb Phon.
0	0.03	0.05	0.07	0.06874	0.08	0.15	14.71	214
1	0.03	0.06	0.07	0.07134	0.09	0.19	15.48	217
2	0.03	0.06	0.08	0.08014	0.10	0.18	17.71	221
3	0.03	0.07	0.09	0.09263	0.11	0.33	19.73	213

TAB. 7.6 – *Statistiques descriptives sur les mesures du VOT par grade (en secondes)*

Au regard de ces différentes valeurs, la durée moyenne du VOT des occlusives sourdes semble *a priori* croître de manière monotone selon le niveau de sévérité de la dysphonie.

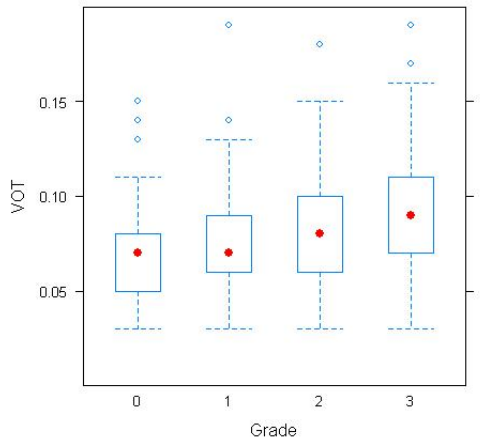


FIG. 7.9 – *Durée du VOT en fonction du grade.*

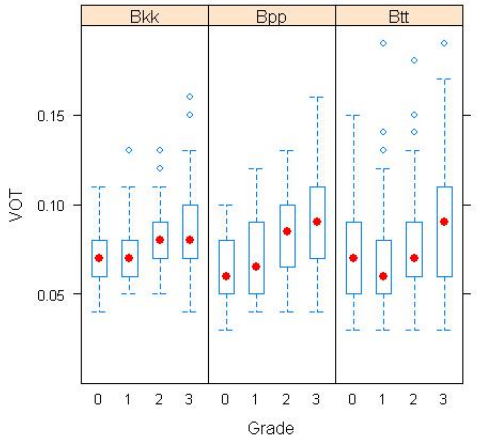


FIG. 7.10 – *Durée du VOT en fonction du grade par occlusive sourde.*

Sous la forme de « boîtes à moustaches », la figure 7.9 illustre les statistiques estimées sur le VOT en fonction des grades de dysphonie. Ce graphique reprend les caractéristiques de position affichées dans le tableau 7.6 - le minimum, le 1er quartile, la médiane, le 3^{e} quartile et le maximum - calculées sur les durées de VOT suivant les grades ; ces statistiques s'affinent par différenciation des occlusives sourdes dans la figure 7.10.

1re étape : effet du grade sur le VOT

Nous utiliserons l'analyse de la variance (ou Anova) pour étudier l'influence d'une ou de plusieurs variables indépendantes (le grade et le phonème) sur une variable quantitative (le VOT). La première étape va consister à vérifier si l'appartenance à un grade explique une partie des mesures du VOT *i.e.* quel effet produit le grade sur l'allongement de la durée du VOT. Pour cela, nous utiliserons un modèle linéaire avec comme variable quantitative la mesure du VOT et comme facteur le grade.

L'analyse de la variance est appliquée sur le modèle linéaire « VOT $\sim$ GRADE » :

```
Analysis of Variance Table

Response: VOT
           Df  Sum Sq Mean Sq F value    Pr(>F)
GRADE       3 0.07455 0.02485  26.361 2.561e-16 ***
Residuals 861 0.81170 0.00094
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sur la base des statistiques fournies ci-dessus, les grades présentent des différences hautement significatives dans la valeur du VOT (p-value[4] < 0.001).

Pour en savoir plus sur l'effet du grade sur l'allongement de la durée du VOT, on interroge le modèle linéaire « VOT $\sim$ GRADE » :

```
lm(formula = VOT ~ GRADE, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.062629 -0.020136 -0.002629  0.011262  0.237371

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0782099  0.0010441  74.908   <2e-16 ***
GRADE.L      0.0179940  0.0020987   8.574   <2e-16 ***
GRADE.Q      0.0049476  0.0020881   2.369   0.0180 *
GRADE.C     -0.0005606  0.0020775  -0.270   0.7873
---
```

[4]la valeur de p-value est la probabilité qu'un événement quelconque soit le simple fait du hasard ; de manière purement arbitraire, une valeur de p-value inférieure à 1 chance sur 20 est considérée comme statistiquement significative (soit p-value < 0.05)

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0307 on 861 degrees of freedom
Multiple R-squared: 0.08412,    Adjusted R-squared: 0.08093
F-statistic: 26.36 on 3 and 861 DF,  p-value: 2.561e-16
```

Le modèle linéaire « VOT ∼ GRADE » montre de manière très significative que le VOT répond linéairement au grade (p-value < 0.001). Par contre, un effet quadratique significatif apparaît aussi (p-value = 0.0180), effet qui pourrait être provoqué par le phonème [Btt] comme le laisserait supposer la figure 7.10.

2e étape : effet du grade suivant le phonème sur le VOT

Cela nous amène à étudier l'interaction entre le grade et le phonème, et à analyser leurs effets sur le VOT. A ce stade, il nous sera possible de savoir si les valeurs du VOT ont un comportement différent sur le grade en fonction du phonème.

Le modèle linéaire « VOT ∼ GRADE * PHONE » nous donne comme statistiques :

```
Call:
lm(formula = VOT ~ GRADE * PHONE, data = d)

Residuals:
      Min        1Q    Median        3Q       Max
-0.066283 -0.019145 -0.004833  0.013717  0.233717

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.0789876  0.0019651  40.196  < 2e-16 ***
GRADE.L           0.0127926  0.0039674   3.224  0.00131 **
GRADE.Q           0.0009880  0.0039301   0.251  0.80156
GRADE.C          -0.0011922  0.0038925  -0.306  0.75947
PHONEBpp         -0.0024221  0.0031155  -0.777  0.43712
PHONEBtt         -0.0006035  0.0024305  -0.248  0.80396
GRADE.L:PHONEBpp  0.0089613  0.0062453   1.435  0.15169
GRADE.Q:PHONEBpp -0.0038570  0.0062309  -0.619  0.53608
GRADE.C:PHONEBpp  0.0002446  0.0062165   0.039  0.96863
GRADE.L:PHONEBtt  0.0066074  0.0048999   1.348  0.17786
GRADE.Q:PHONEBtt  0.0087901  0.0048610   1.808  0.07091 .
GRADE.C:PHONEBtt  0.0011746  0.0048218   0.244  0.80759
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03067 on 853 degrees of freedom
Multiple R-squared: 0.09443,    Adjusted R-squared: 0.08275
F-statistic: 8.086 on 11 and 853 DF,  p-value: 1.545e-13
```

Les résultats présentés ci-dessus montrent que le facteur phonème ne produit aucun effet sur le VOT, à la différence du grade sur lequel le VOT répond linéairement (p-value = 0.00131). Il n'y a donc pas de comportement particulier des différentes occlusives

([Bkk], [Bpp], [Bpp]) par rapport à l'allongement de la durée du VOT. Néanmoins, on peut constater une légère composante quadratique entre le grade et le phonème [Btt] (p-value = 0.07091) qui confirmerait l'observation faite durant la 1re phase.

3e étape : effet du grade suivant le phonème sur le VOT avec prise en compte du sujet

Nous nous intéressons maintenant à modéliser les effets fixes qui peuvent être « noyés » dans les effets aléatoires comme la variabilité inter-individuelle, pouvant parfois être plus important que l'effet que l'on cherche à mesurer. Pour remédier à ces effets indésirables, l'utilisation d'un modèle mixte semble approprié car il permet une « normalisation » par sujet et une prise en compte de l'effet sujet (comportement spécifique de sujet).

Voici les statistiques issues du modèle mixte « VOT ∼ GRADE * PHONE + (1 | LOC) » :

```
                Estimate MCMCmean HPD95lower HPD95upper  pMCMC Pr(>|t|)
(Intercept)       0.0787   0.0787     0.0754     0.0820 0.0001   0.0000
GRADE.L           0.0120   0.0120     0.0053     0.0185 0.0008   0.0004
GRADE.Q           0.0003   0.0002    -0.0067     0.0065 0.9496   0.9351
GRADE.C          -0.0016  -0.0016    -0.0082     0.0047 0.6166   0.6336
PHONEBpp         -0.0023  -0.0023    -0.0068     0.0020 0.3234   0.3089
PHONEBtt         -0.0046  -0.0046    -0.0083    -0.0012 0.0102   0.0096
GRADE.L:PHONEBpp  0.0093   0.0092     0.0001     0.0181 0.0390   0.0418
GRADE.Q:PHONEBpp -0.0035  -0.0035    -0.0122     0.0055 0.4330   0.4415
GRADE.C:PHONEBpp  0.0005   0.0005    -0.0083     0.0095 0.9158   0.9171
GRADE.L:PHONEBtt  0.0032   0.0031    -0.0039     0.0102 0.3772   0.3752
GRADE.Q:PHONEBtt  0.0061   0.0061    -0.0008     0.0131 0.0866   0.0884
GRADE.C:PHONEBtt -0.0010  -0.0009    -0.0081     0.0059 0.7914   0.7832

random
         MCMCmean HPD95lower HPD95upper
sigma    0.022225    0.02115    0.02340
LOC.(In) 0.007447    0.00535    0.01022
```

Au regard des statistiques, la composante linéaire du facteur grade sur les valeurs du VOT est très significative avec une p-value = 0.0008 (hautement significative).

De plus, on notera un effet linéaire significatif du grade suivant le phonème [Bpp] (p-value = 0.0390) ainsi qu'un léger effet quadratique du grade suivant le phonème [Btt] (p-value = 0.0866). Ces deux effets sont illustrés sur la figure 7.10.

La figure 7.11 trace les quartiles des résidus normalisés du modèle mixte en fonction des valeurs attendues selon une loi normale $\mathcal{N}(0\,;1)$.

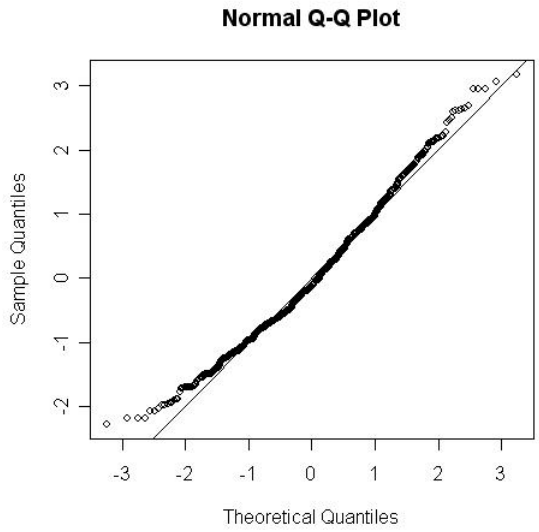


FIG. 7.11 – *Statistiques sur le modèle mixte « VOT ∼ GRADE * PHONE + (1 | LOC) ».*

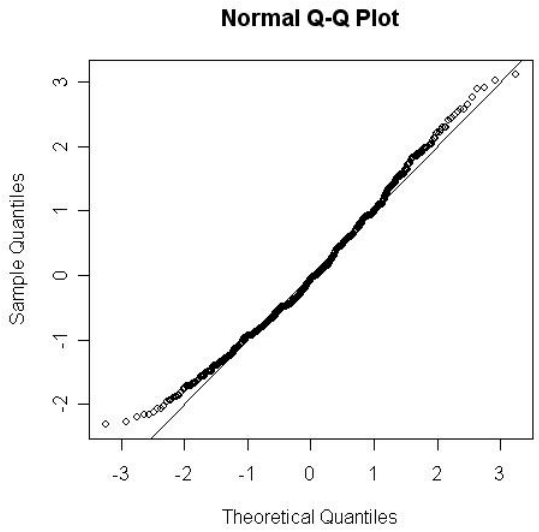


FIG. 7.12 – *Statistiques sur le modèle mixte « VOT ∼ GRADE + (1 | LOC) ».*

4e étape : effet du grade sur le VOT avec prise en compte du sujet

Du modèle mixte précédent, nous retirons l'effet phonème tout en gardant la prise en compte de l'effet sujet.

Voici les statistiques issues du modèle mixte « VOT ∼ GRADE + (1 | LOC) » :

```
            Estimate MCMCmean HPD95lower HPD95upper  pMCMC Pr(>|t|)
(Intercept)   0.0758   0.0759     0.0736     0.0781 0.0001   0.0000
GRADE.L       0.0154   0.0155     0.0108     0.0199 0.0001   0.0000
GRADE.Q       0.0028   0.0028    -0.0019     0.0072 0.2336   0.2256
GRADE.C      -0.0020  -0.0021    -0.0065     0.0026 0.3666   0.3770

random
         MCMCmean HPD95lower HPD95upper
sigma    0.022356   0.021210    0.02349
LOC.(In) 0.007499   0.005541    0.01009
```

L'effet linéaire du GRADE sur le VOT s'améliore encore (p-value = 0.0001, hautement significative).

La figure 7.12 trace les quartiles des résidus normalisés du modèle mixte.

Discussion

L'analyse statistique a montré que le grade de la dysphonie produit un effet linéaire et significatif sur l'allongement de la durée du VOT, sans que cela soit spécifique à une occlusive sourde bien précise. En d'autres termes, la durée du VOT est corrélée avec le grade de sévérité de la dysphonie et cela, quelle que soit la nature de l'occlusive sourde.

L'hypothèse que le VOT augmente avec le niveau de sévérité de la dysphonie, supposerait un problème de synchronisation des gestes articulatoires et phonatoires *i.e.* la difficulté de mettre en route le voisement chez les dysphoniques les plus sévères. En français, le VOT n'ayant pas vraiment de propriété linguistique[5], il peut alors rendre compte de la capacité du locuteur à coordonner ses organes constricteurs et phonatoires dans le cadre de la dysarthrie. Par contre dans le cadre de la dysphonie, l'allongement du VOT en fonction de la sévérité de la dysphonie peut être interprêté comme une difficulté bio-mécanique de démarrage de la vibration laryngée.

Au regard de la pertinence des consonnes sourdes relevée par l'analyse phonétique 7.1, le système automatique a fait preuve de sensibilité sur l'allongement du VOT. La paramétrisation utilisée pour l'analyse phonétique correspond à des coefficients spectraux (LFSC). On peut faire un lien entre l'allongement du VOT et l'inégalité de Heisenberg-Gabor (équation 7.1) dans laquelle $\Delta\tau$ est la durée de l'impulsion et $\Delta\omega$ la largeur de son spectre :

$$\Delta\tau.\Delta\omega \geq \frac{1}{4\pi} \tag{7.1}$$

Le fait de montrer la corrélation entre le VOT et le grade indique la présence d'évènements atypiques durant la production des occlusives sourdes chez les personnes dysphoniques. L'allongement d'un phénomène temporel implique une modification spectrale : une impulsion temporelle longue ou courte n'a pas la même représentation spectrale. Un phénomène mécanique très impulsif, très rapide fournira une représentation spectrale très étendue en fréquences (très aigu). A l'inverse, un phénomène lent, mou fournira moins de hautes fréquences (plus grave). Concernant la sensibilité du système automatique à l'allongement du VOT, il est tout à fait plausible d'émettre comme hypothèse que l'analyse spectrale ait permis de mettre en évidence un phénomène temporel connaissant le principe de la dualité temps-fréquence d'un signal. On peut donc faire l'hypothèse que, chez les dysphoniques, les mesures de VOT étant plus longues, le spectre contient probablement moins d'énergie dans les hautes fréquences.

Cependant, il faut souligner qu'il ne s'agit ici que d'une étude préliminaire sur laquelle plusieurs réserves doivent être faites, même si les résultats apparaissent comme significatifs d'un point de vue statistique. En effet, les mesures de VOT issues d'un alignement automatique doivent être validées par une analyse manuelle dans laquelle la qualité des frontières phonémiques des parties « burst » des occlusives sourdes sont à vérifier.

[5] à la différence de l'anglais

7.3 La méthode « Automatic Phonetic Labeling »

A travers la méthode « Phonetic Labeling » décrite en 1.5.3, [Revis, 2004; Revis et al., 2006] a étudié les caractéristiques pathologiques de l'ensemble des phonèmes constitutifs d'une phrase prononcée par des patients dysphoniques. Son principe, basé sur l'évaluation perceptive de chaque phonème, a montré l'importance de l'influence des contraintes phonétiques et linguistiques pour l'étude de la dysphonie.

A l'issue de l'application de la procédure dite « en entonnoir », les phonèmes sont étiquetés « normaux » ou « pathologiques ». Seuls les éléments jugés « pathologiques » sont alors caractérisés en fonction de 5 paramètres retenus :

- 2 paramètres dysphoniques : la raucité et le souffle ;
- 2 paramètres phonétiques : l'aspiration et le creak ;
- 1 paramètre mixte : le dévoisement.

Un étude présentée dans [Revis, 2004] a validé cette méthode en montrant sa haute reproductibilité en situation intra- comme inter-individuelle. Voici la phrase utilisée par les patients dysphoniques pour l'analyse phonétique avec la transcription phonétique :

« Il les perdait toutes de la même façon »
« i l e p ɛ r d e t u t ə d ø l a m ɛ m ə f a s ɔ̃ »

Dans ce travail, un principe similaire est reproduit, basé sur le système automatique. Ici, nommée « Automatic Phonetic Labeling », cette méthode repose sur une analyse phonétique automatique, telle que celle décrite en 7.1, appliquée sur la bande fréquentielle totale [0-8000]Hz, en interprétant chaque décision comme « normale » ou « pathologique » suivant le grade attribué par le système.

Menée avec la collaboration d'une orthophoniste [Azzarello, 2006], cette étude a été entreprise afin de pallier l'inconvénient majeur de la méthode « Phonetic Labeling ». En effet, malgré son haut niveau de fiabilité, le « Phonetic Labeling » ne peut être appliqué à la pratique clinique quotidienne « *du fait de la longueur des analyses, de la finesse des manipulations informatiques pour le séquençage des phonèmes, et de l'attention soutenue nécessaire à l'application de cette tâche* ». La confirmation des résultats par le système automatique permettrait à cette technique de sortir de son contexte exclusivement expérimental et d'offrir aux spécialistes de la voix, une meilleure compréhension du fonctionnement phonétique de la dysphonie de par son haut niveau de détail proposé.

Résultats

Les figures de 7.13 à 7.16 présentent les décisions du système pour chaque grade :

1. la répartition (en %) des vraisemblances par grade pour chaque phonème ;
2. le pourcentage de phonèmes jugés dans chaque grade ;

3. le pourcentage de phonèmes jugés « pathologiques ».

Les cellules de couleur rouge indique le grade (colonne correspondante) attribué par le système pour chaque phonème de la phrase étudiée.

	Grade 0			
	G0	G1	G2	G3
il	40	30	20	10
ll	10	60	25	15
ll	40	25	35	0
ei	75	25	0	0
opp	95	0	5	0
bpp	85	10	5	0
ai	60	30	10	0
rr	75	25	0	0
dd	75	5	20	0
ei	75	20	5	0
ott	80	10	5	5
btt	80	5	10	5
ou	80	5	10	5
ott	75	10	10	5
btt	75	15	10	10
eu	33	33	33	0
dd	55	25	15	5
eu	47	26	16	10
ll	25	60	15	0
aa	35	50	15	0
mm	30	50	20	0
ai	50	30	5	15
mm	55	30	10	5
eu	0	50	50	0
ff	85	15	0	0
aa	50	20	30	0
ss	85	10	5	0
on	70	20	5	5
% phon par grade	85	15	0	0
% phon patho	15			

FIG. 7.13 – *Décisions du système pour les locutrices G0.*

	Grade 1			
	G0	G1	G2	G3
il	35	20	30	15
ll	15	35	25	25
ll	25	20	55	0
ei	30	45	20	5
opp	5	55	40	0
bpp	5	40	45	10
ai	30	40	25	5
rr	25	50	20	5
dd	20	45	35	0
ei	10	50	30	10
ott	5	45	50	0
btt	10	40	25	25
ou	30	25	35	10
ott	20	30	30	20
btt	10	55	15	20
eu	17	33	17	33
dd	20	25	40	15
eu	5	78	11	5
aa	20	50	15	15
mm	20	45	35	0
ai	35	20	30	15
mm	30	30	30	10
eu	40	45	5	10
ff	0	0	0	100
aa	15	70	15	0
ss	15	15	55	15
on	20	45	35	0
% phon par grade	45	20	35	0
% phon patho	11	62	23	4
	89			

FIG. 7.14 – *Décisions du système pour les locutrices G1.*

Il faut noter que la transcription phonétique utilisée par le système automatique (détails en 2.1) dissocie les occlusives non-voisées (tableau 2.4) en 2 temps où le 1er temps correspond à la phase de l'occlusion (silence) et le 2^{e} temps à l'explosion (burst).

	Grade 2			
	G0	G1	G2	G3
il	10	30	20	40
ll	25	25	20	30
ll	20	35	20	25
ei	30	10	25	35
opp	0	55	35	10
bpp	0	35	45	20
ai	5	40	40	15
rr	10	50	20	20
dd	20	15	45	20
ei	15	35	30	20
ott	0	55	30	15
btt	0	15	70	15
ou	10	40	35	15
ott	0	35	45	20
btt	5	15	50	30
eu	25	50	25	0
dd	20	15	25	40
eu	22	39	28	11
ll	55	10	25	10
aa	5	30	50	15
mm	10	20	55	15
ai	10	25	55	10
mm	15	35	30	20
eu	33	0	0	66
ff	15	55	30	5
aa	10	10	55	25
ss	5	35	55	5
on	30	20	25	25
% phon par grade	7	37	37	19
% phon patho	93			

FIG. 7.15 – *Décisions du système pour les locutrices de G2.*

	Grade 3			
	G0	G1	G2	G3
il	0	10	45	45
ll	15	5	40	40
ll	10	10	30	50
ei	5	5	25	65
opp	10	10	20	60
bpp	5	15	15	65
ai	15	25	20	40
rr	5	30	30	35
dd	10	5	45	40
ei	0	5	40	50
ott	10	15	20	55
btt	0	0	30	70
ou	5	20	30	45
ott	5	30	20	45
btt	5	20	30	45
eu	0	0	40	60
dd	5	20	5	70
eu	8	8	16	67
ll	0	25	20	55
aa	0	10	10	80
mm	0	15	45	40
ai	0	25	20	55
mm	5	15	35	45
eu	0	0	33	66
ff	10	5	40	45
aa	0	5	45	50
ss	5	25	45	25
on	5	30	10	55
% phon par grade	0	0	12	88
% phon patho	100			

FIG. 7.16 – *Décisions du système pour les locutrices de G3.*

De plus, les phonèmes obtenant un même score pour plusieurs grades (situation indécise) n'ont pas été pris en compte dans l'analyse des résultats.

On peut observer que :

- les voix de grade 0 présentent 15 % de phonèmes « pathologiques » avec une majorité de 85 % jugés « non pathologiques » par le système ;
- le grade 1 obtient 89 % de phonèmes « pathologiques » avec une majorité de 62 % classés dans le même grade perceptif ;
- les voix de grade 2 atteignent le taux de 93 % de phonèmes « pathologiques » avec une majorité de 37 % obtenue pour les grades 1 et 2. Cela indique que le grade 2 n'est pas significativement marqué ;
- pour les voix de grade 3, la totalité des phonèmes sont jugés « pathologiques » par le système. Les grades 2 et 3 se partagent l'ensemble des décisions avec respectivement 12 % et 88 %.

Plus particulièrement pour les voix normales, on remarquera que :

- les phonèmes jugés majoritairement « pathologiques » par le système sont : /l/ de [i**l**], /lam/ de [**la** **m**ême] et /ə/ de [même**]** ;
- par contre, les phonèmes /l/ et /m/ de la séquence [**l**a **m**ême] sont au contraire jugés majoritairement comme « non pathologiques » pour les grades 1 et 2 ;
- le phonème /ə/ de [tout**e**] est jugé comme « non pathologique » à 33 % mais aussi comme « pathologique » à 33 % pour les grades 1 et 2.

Il faut tout de même préciser que sur les 20 sujets témoins (voix normales de grade 0), le phonème /ə/ du mot [tout**e**] n'a été prononcé que par 3 locutrices, et celui du mot [même] par seulement 2 sujets.

Discussion

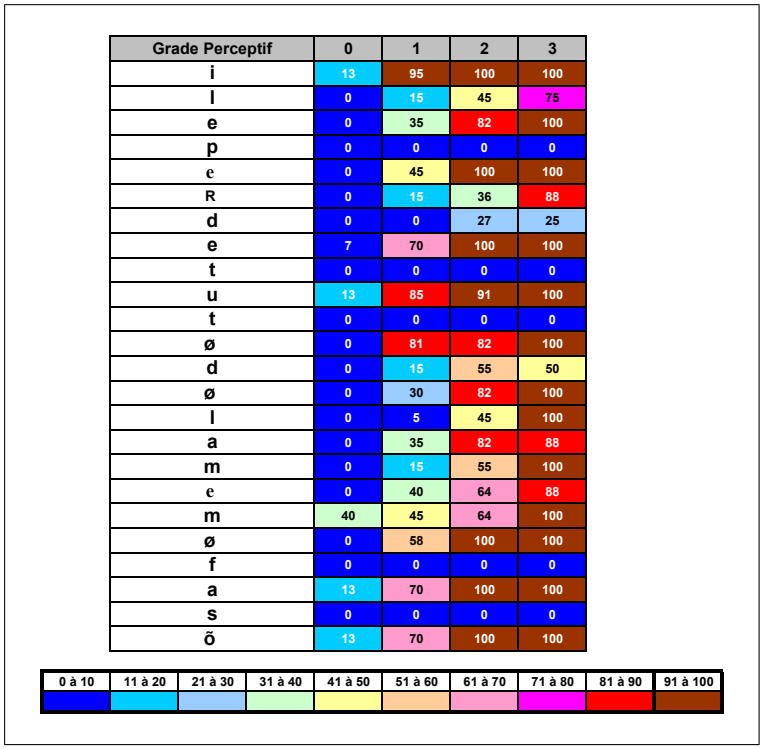
La figure 7.17 représente les pourcentages de phonèmes « pathologiques » par grade recencés dans une population de femmes à l'issue d'une étude de « Phonetic Labeling » menée par Revis (2004). Le corpus utilisé se composait de 54 voix de femmes : 15 sujets témoins, 20 dysphonies légères, 11 dysphonies moyennes et 8 dysphonies sévères. Les sujets étaient âgés de 19 à 83 ans (moyenne de 45.65 ans). La durée moyenne des échantillons analysés était de 1.99 secondes.

En comparant ces taux de phonèmes « pathologiques » à ceux obtenus avec le système automatique, on constate que :

- les voix de grade 0 obtiennent 21 % contre 15 % ;
- les voix de grade 1 obtiennent 75 % contre 89 % ;
- les voix de grade 2 obtiennent 79 % contre 93 % ;
- les voix de grade 3 obtiennent 89 % contre 100 %.

Une première remarque concerne le système automatique qui juge « pathologiques » tous les phonèmes prononcés par les locutrices de grade 3. Mêmes les phonèmes non voisés qui ne mettent pas en vibration les cordes vocales sont jugés « pathologiques ».

Cette observation rejoint l'analyse phonétique automatique décrite en 7.1. Le lecteur pourra s'y référer pour plus de détails.



Grade Perceptif	0	1	2	3
i	13	95	100	100
l	0	15	45	75
e	0	35	82	100
p	0	0	0	0
e	0	45	100	100
R	0	15	36	88
d	0	0	27	25
e	7	70	100	100
t	0	0	0	0
u	13	85	91	100
t	0	0	0	0
ø	0	81	82	100
d	0	15	55	50
ø	0	30	82	100
l	0	5	45	100
a	0	35	82	88
m	0	15	55	100
e	0	40	64	88
m	40	45	64	100
ø	0	58	100	100
f	0	0	0	0
a	13	70	100	100
s	0	0	0	0
õ	13	70	100	100

FIG. 7.17 – *Pourcentage de phonèmes pathologiques par grade dans une population de femmes (Révis).*

Tout comme relevé en « Phonetic Labeling », le système tend à montrer que les voix normales ne sont pas exemptes d'occurrences pathologiques ; le taux de phonèmes atteints étant tout de même plus faible d'environ 28.5 %. Le système semble évaluer plus sévèrement les voix dysphoniques que les voix normales. Cependant, il serait légitime de s'interroger sur la variabilité et les précisions des jugements de l'analyse « Phonetic Labeling » qui a été effectuée par un seul juge. Une marge d'erreurs peut donc exister sur les décisions présentées. Il serait intéressant de citer les travaux de [Meunier & Floccia, 1999] qui montrent que le résultat de la perception peut varier en fonction de la taille de la fenêtre d'écoute.

De plus, il est intéressant de souligner que le système évalue les voix normales et les voix les plus dysphoniques dans leur grade perceptif de référence (85 % pour le grade 0 et 88 % pour le grade 3) ; les modèles statistiques correspondants semblent robustes et stables ; Parmi les 89 % de phonèmes « pathologiques » comptabilisés sur les voix légèrement dysphoniques, une majorité de 62 % sont évalués dans le grade 1. Cette tendance à la baisse s'accentue avec les dysphonies modérées pour lesquelles, les 93 % de phonèmes « pathologiques » se répartissent équitablement entre les grades 1 et 2 au taux de 37 % chacun. Cela montre la difficulté du système pour les grades 1 et 2 sur lesquels la plus grande confusion est généralement observée.

Cependant, en considérant uniquement les taux de phonèmes « pathologiques » pour chaque grade, il apparaît que le nombre d'occurrences est fortement corrélé avec le niveau de sévérité de la dysphonie établi par l'analyse perceptive « classique », à l'exception du grade 2. Relevée aussi par la méthode « Phonetic Labeling », cette constatation suggère que la perception semble fonctionner comme la somme d'événements dysphoniques intervenant au cours de la parole continue.

Par contre, aucune comparaison à un niveau phonémique ne peut être établie entre le système automatique et la méthode « Phonetic Labeling ». En effet, le système évalue comme « pathologiques » les phonèmes non voisés. Or, sachant que la dysphonie est un trouble directement lié à la perturbation de la vibration des cordes vocales, les consonnes sourdes ne devraient pas apparaître « pathologiques » comme cela est le cas pour la méthode « Phonetic Labeling ». Dans ce cas, les phonèmes non voisés ont été écartés systématiquement de l'étude considérés comme inintéressants dans ce contexte dysphonique. Il serait intéressant à la vue des résultats de l'analyse automatique de reproduire le « Phonetic Labeling » sur les consonnes non voisées sans *a priori* particulier sur la nature du trouble recherché. D'autre part, le « Phonetic Labeling » présentait des occurrences pathologiques sur les phonèmes /i/, /u/, /m/, /a/ et /õ/ pour les voix de grade 0. Or, le système ne trouve que le /m/ jugé « pathologique » pour les voix normales.

7.4 Conclusion

Dans ce chapitre, une analyse phonétique a permis d'étudier le comportement du système de classification de la voix pathologique selon différentes classes de phonèmes. Afin d'évaluer les manifestations de la dysphonie suivant les grades, les différents résultats obtenus par phonème ou classe de phonèmes, ont été comparés entre les 2 bandes de fréquences, [0-3000]Hz et [0-8000]Hz.

Quelle que soit la bande de fréquences considérée, l'analyse phonétique a fait ressortir comme principale observation, la pertinence de la classe des consonnes. On retiendra de cette étude les trois points suivants :

1. Le système automatique adapté au contexte pathologique permet d'analyser sur la parole continue certains phénomènes vocaux transitoires comme le passage entre phonèmes « non voisé/voisé » ou la séquence CV, pouvant être caractéristiques de la dysphonie ;
2. La pertinence relevée sur les consonnes nous a amené à contrôler la qualité de l'alignement phonétique en analysant plusieurs spectrogrammes, et plus particulièrement, ceux incluant les consonnes sourdes. Il apparaîtrait que dans les transitions de type « VC_{sourde} », les consonnes sourdes incluraient systématiquement une partie de la finale de la voyelle précédente. Cette constatation est intéressante car si elle se vérifiait, cela montrerait que la méthode automatique permet de mettre en évidence des phénomènes transitoires caractéristiques de la dysphonie ;
3. Toujours sur la pertinence de consonnes sourdes, une étude statistique préliminaire a montré une corrélation entre la durée du VOT et le grade de sévérité et cela, quelle que soit la nature de l'occlusive sourde.

La dernière étude de ce chapitre consistait en une étude comparative entre les méthodes « Phonetic Labeling » et « Automatic Phonetic Labeling ». Les principaux points relevés ont été :

1. les voix normales ne sont pas exemptes d'occurrences pathologiques ;
2. le système évalue les voix de grade 0 et 3 dans leur grade perceptif de référence (85 % et 88 % resp.) mais accuse de la difficulté pour les voix de grade 1 et 2 ;
3. il apparaît que le nombre d'occurrences « pathologiques » est fortement corrélé avec le jugement perceptif, à l'exception du grade 2.
« *La perception semble fonctionner comme la somme d'événements dysphoniques intervenant au cours de la parole continue* » ;
4. aucune corrélation à un niveau phonémique n'a pu être établie entre le système automatique et la méthode manuelle ;
5. le système évalue comme « pathologiques » des phonèmes non voisés.

Quatrième partie

Conclusion Générale et Perspectives

Conclusion Générale

Le travail réalisé dans ce document s'inscrit dans le domaine de la Reconnaissance Automatique du Locuteur (RAL) adapté au contexte de la voix pathologique. L'objectif de la RAL est d'identifier une personne à l'aide de sa voix en s'intéressant plus particulièrement aux informations extra-linguistiques véhiculées par le signal de parole (identité, caractéristiques physiques, particularités régionales, ...). Ainsi partant de l'hypothèse que la dysphonie peut être appréhendée comme n'importe quelle information extra-linguistique, un système RAL a été adapté à la reconnaissance des grades de l'échelle GRBAS pour observer les phénomènes pertinents de la dysphonie. Il est important de souligner que l'objectif des travaux présentés n'est pas d'optimiser les performances du système automatique en terme d'amélioration des scores de reconnaissance mais de caractériser les phénomènes liés à la dysphonie dans le signal de parole. L'objectif « infine » est d'apporter de nouvelles connaissances acoustiques à l'expertise humaine en vue d'affiner ou d'enrichir la compréhension de ces phénomènes.

Dans cette optique, trois axes de recherche de l'information pertinente sont explorés dans cette thèse.

L'étude paramétrique

Cette étude se décompose en deux parties :

[1] analyse comparative de plusieurs paramétrisations de coefficients statiques

Cette première étude expérimentale montre que la technique d'analyse par prédiction linéaire (LPC/LPCC) ne semble pas adaptée aux voix dysphoniques laissant supposer que le trouble vocal se manifeste acoustiquement de manière intermittente donc difficilement prédictible.

Les coefficients cepstraux MFCC obtiennent de bons résultats de manière « plutôt inattendue ». Par définition, la transformation cepstrale permet de dissocier l'influence de la source glottale (des cordes vocales) de celle du conduit vocal, en fournissant une représentation du signal de parole où les premiers coefficients sont censés ne contenir principalement que l'information résonantielle du conduit vocal. Or, sachant que la dysphonie se caractérise essentiellement par une perturbation de la source laryngée, il

est donc légitime de s'interroger sur les points suivants :

1. les coefficients cepstraux utilisés dans ce travail, ne contiennent-ils pas plus d'information sur la source glottique qu'ils ne sont censés contenir ?
2. la dysphonie se manifeste-elle également au niveau supra-laryngé ?

La dernière observation concerne les coefficients spectraux MFSC qui affichent la meilleure performance globale avec un score TCC de 73.75 % (59/80 locuteurs bien classés). La supériorité affichée de cette paramétrisation va se confirmer sur l'ensemble des expériences réalisées dans cette étude. Elle est probablement due à la particularité de l'analyse en banc de filtres d'estimer l'enveloppe spectrale du signal de parole, fournissant à la fois les informations du spectre de la source et de la réponse fréquentielle du conduit vocal. Au regard de la performance satisfaisante atteinte, une part de l'information pertinente liée à la perturbation laryngée a certainement été extraite durant l'analyse spectrale avant d'être caractérisée par les modèles de grade.

[2] analyse comparative de plusieurs paramétrisations augmentées de coefficients dynamiques

Quelle que soit la nature de l'information dynamique ajoutée aux coefficients statiques (Δ, $\Delta\Delta$, $\Delta\Delta\Delta$), les différentes paramétrisations améliorent leur performance de manière significative avec notamment comme meilleur score TCC global 78.75 % (63/80 locuteurs bien classés) pour les coefficients spectraux MFSC en configuration dynamique ($\Delta + \Delta\Delta$) contre 73.75 % avec les coefficients statiques uniquement.

Par contre, aucune nature d'information dynamique ne se distingue significativement des autres en terme d'amélioration des performances. L'espoir fondé sur les coefficients $\Delta\Delta\Delta$ associés aux variations temporelles de l'accélération qui génère des phénomènes d'instabilité par l'effort produit, ne s'est pas concrétisé dans les résultats globaux obtenus par les paramétrisations cepstrales et spectrales.

Concernant la prise en compte d'une fenêtre temporelle variable pour le calcul des coefficients dynamiques (5, 7 ou 9 trames), les résultats non probants dans ce travail tendraient à montrer que son augmentation n'apporte aucune information supplémentaire qui soit susceptible d'améliorer la discrimination des grades de la dysphonie ; le meilleur score TCC de 78.75 % étant obtenu dans un contexte de 5 trames pour les MFSC (TCC cité plus haut).

Malgré ces deux dernières observations, il est difficile d'en déduire que les caractéristiques des phénomènes liés à la dysphonie ne puissent transparaître dans les informations de nature dynamique. En effet, plusieurs suppositions peuvent être émises si l'on considère la dysphonie comme un phénomène irrégulier qui se superpose aux caractéristiques phonétiques et linguistiques du signal de parole :
◦ les coefficients delta, sont-ils capables de capturer cette information de nature intermittente ?

◦ s'ils sont capables de l'extraire, savent-ils la traiter sur des fenêtres temporelles de grande taille ?
◦ autre hypothèse : est-ce que les occurrences pathologiques ne sont-elles pas « *diluées* » dans les modèles statistiques à cause de leur manifestation irrégulière ?
◦ finalement dans la configuration paramétrique ($\Delta + \Delta\Delta + \Delta\Delta\Delta$), n'y-a-t-il pas un excédent de coefficients dynamiques pour que le système soit efficace ?

Plusieurs travaux consacrés à évaluer l'impact de la fenêtre temporelle sur l'efficacité des coefficients delta, ont montré que la taille optimale dépendait fortement de la tâche visée [Furui, 1981; Soong & Rosenberg, 1988; Bernasconi, 1990]. Par exemple, Furui (1981) a montré qu'une fenêtre de 90 milli-secondes (soit 9 trames d'une durée de 10 milli-secondes) semblait adéquate pour préserver les informations transitionnelles entre phonèmes. Tandis que Bernasconi (1990) considérait comme optimale une fenêtre de 135 milli-secondes (soit 9 trames de 15 milli-secondes) en VAL. A l'heure actuelle, la plupart des systèmes RAL utilisent les coefficients delta calculés sur une fenêtre de 5 trames. L'usage courant d'une telle taille de fenêtre laisserait supposer que les dérivées ne sont pas adaptées à une évaluation à long terme de l'information dynamique en reconnaissance du locuteur. De plus, Fredouille & Bonastre (1998) propose une manière originale d'exploiter l'information dynamique delta en concaténant des trames successives du signal.

En ce sens et malgré les résultats peu probants obtenus dans le contexte pathologique concernant la nature de l'information dynamique pertinente (delta ou autre) à prendre en compte ou l'empan temporel à considérer pour l'extraire, nous restons convaincus qu'il faille continuer à explorer cette voie de manière différente en raison de la particularité de la dysphonie de se manifester comme un phénomène intermittent, voire chaotique.

L'étude fréquentielle

Cette étude s'intéresse à la façon dont les caractéristiques acoustiques de la dysphonie sont réparties sur l'ensemble de l'espace fréquentiel. Pour cela, une architecture en sous-bandes de fréquences est associée au système automatique afin d'analyser la pertinence de certaines plages de fréquences pour la reconnaissance du degré de sévérité de la dysphonie.

Dans cette approche, seule la bande [0-3000]Hz constitue la plage de fréquences la plus intéressante, permettant une discrimination plus homogène sur l'ensemble des grades. La supériorité affichée de la bande restreinte [0-3000]Hz apparaît à travers l'ensemble de cette étude. Sachant que l'instabilité vibratoire de la glotte constitue une cause essentielle des dysphonies, la pertinence de la bande [0-3000]Hz ne paraît pas comme « inattendue » vu qu'elle constitue la plage fréquentielle la plus riche en informations caractéristiques de la source glottale comme la fréquence fondamentale, le taux de voisement, l'énergie, ... Ce résultat est d'ailleurs conforté par des méthodes objectives d'évaluation de la dysphonie qui reposent uniquement sur l'analyse des caractéris-

tiques spectrales du son laryngé telles que la fréquence fondamentale et les indices de fluctuations (jitter, shimmer, ...).

L'évaluation perceptive des signaux de parole du corpus des voix dysphoniques filtrés en [0-3000]Hz a été effectuée suivant le même protocole que celui mis en œuvre pour son évaluation initiale sur la bande totale [0-8000]Hz. L'analyse des jugements entre les deux évaluations perceptives, [0-3000]Hz et [0-8000]Hz, montre une concordance très élevée de 85 % qui représente un taux de reproductibilité nettement supérieur à celui habituellement atteint en intra ou inter-individuel (60 %). Cette performance conforte l'efficacité de la mise en place de protocoles d'analyse perceptive de la voix (jury d'experts, vote par consensus, ...) afin d'en améliorer la fiabilité. Les désaccords concernent principalement les voix normales ou avec une légère dysphonie (grade 1). Cela a tendance à montrer que le filtrage des basses fréquences (induisant un timbre vocal plus grave) des voix normales ou peu altérées, peut affecter les jugements d'un expert induisant une surestimation du trouble vocal. Par contre, le jury n'a pas été influencé par le filtrage pour les voix les plus dysphoniques. Enfin, aucune corrélation n'a pu être établie entre les deux évaluations en [0-3000]Hz, automatique et perceptive, sur les voix mal classées.

Finalement, l'analyse de la bande [300-3000]Hz a montré une baisse globale des performances du système, principalement pour les voix dysphoniques. Ce résultat tend à montrer que les fréquences inférieures à 300Hz contiennent de l'information utile à la caractérisation du trouble vocal. Avec notamment dans cette plage, la présence de la fréquence fondamentale (F0) dont la moyenne[6] pour les femmes dysphoniques diminue globalement avec le degré de sévérité du trouble vocal. Ainsi, la baisse globale des performances sur la bande [300-3000]Hz corrobore ce que disent certains spécialistes de la voix pour lesquels, une personne dysphonique paraît avoir une meilleure qualité de voix au téléphone qu'en situation conversationnelle directe c-à-d son trouble vocal étant amoindri par la bande téléphonique, elle paraît moins pathologique.

L'étude phonétique

Dans le troisième volet de cette thèse, nous nous sommes intéressés à l'étude phonétique. Ces travaux ont consisté à observer le comportement du système de classification des voix dysphoniques selon différentes classes de phonèmes. Ces comportements sont analysés sur les 2 bandes de fréquences, [0-3000]Hz et [0-8000]Hz, par phonème ou classe de phonèmes, afin d'évaluer les manifestations de la dysphonie selon les grades. La principale observation qui ressort de cette analyse phonétique concerne la pertinence de la classe des consonnes qui, quelle que soit la bande de fréquences considérée, atteint à elle seule le score global obtenu par l'ensemble des phonèmes.

La pertinence « peu attendue » des consonnes sourdes nous a amené à nous interroger sur la qualité de l'alignement phonétique. En effet, la dysphonie étant un trouble directement lié à la perturbation de la vibration des cordes vocales, seuls les phonèmes

[6] la F0 moyenne fournit une mesure globale de la hauteur tonale perçue (aiguë, grave, ...)

voisés devraient être affectés *a priori* par des occurrences pathologiques. Par conséquent, seule une analyse manuelle de l'alignement phonétique permettrait de vérifier si les frontières déterminées automatiquement sur les consonnes sourdes n'incluent pas de l'information provenant des phonèmes adjacents, notamment des portions de voisement. Dans ce sens, si de telles « erreurs de frontières phonémiques » [7] se confirmaient de manière constante, les limites de l'alignement automatique auront permis la mise en évidence de phénomènes transitoires caractéristiques de la dysphonie tels que « VC_{sourde} » et/ou « $C_{sourde}V$ ».

Toujours à propos de la performance « peu attendue » des consonnes sourdes, nous nous sommes intéressés aux mesures du VOT. L'étude a porté sur l'allongement de la durée du VOT dans un contexte « $C_{sourde}V$ » et sur l'hypothèse d'une corrélation avec le grade de sévérité. L'analyse statistique a montré que le grade de la dysphonie produisait un effet linéaire et significatif sur l'allongement de la durée du VOT et cela, quelle que soit la nature de l'occlusive sourde. L'augmentation du VOT avec le degré de sévérité supposerait un problème de synchronisation des gestes articulatoires et phonatoires. Elle soulignerait la difficulté croissante des sujets dysphoniques à déclencher le voisement au fur et à mesure que le degré de sévérité des troubles augmente. Cette corrélation entre grade et VOT ouvre de nouvelles perspectives de recherche.

Finalement, la méthode de classification automatique présentée dans cette thèse apparaît comme complémentaire aux méthodes objectives qui s'appuient sur les voyelles tenues comme le /a/ pour évaluer la stabilité et le bruit du vibrateur laryngé à travers des indices de fluctuations (jitter, shimmer, ...) difficiles (voire impossibles) à extraire en parole continue. En effet, la classification automatique permet d'analyser sur de la parole continue certains phénomènes vocaux transitoires comme le passage entre phonèmes « voisé $\curvearrowright$ non voisé » ou la séquence CV, pouvant être caractéristiques de la dysphonie.

$\backsim$ *o* $\backsim$

Hormis l'intérêt scientifique d'une telle étude, une collaboration « inter-laboratoire » a été des plus enrichissantes et des plus intéressantes. Elle a permis la rencontre et la coopération « inter-disciplinaire » (informaticiens, orthophonistes, phonéticiens, linguistes, médecins ORL, phoniatres, ...) favorisant l'accès à un domaine actuellement en pleine émergence, la « Phonétique Clinique ».

Aussi, de part la dimension « multi-disciplinaire » de ce travail de thèse suggérant un échange de connaissance et de savoir-faire, l'approche automatique, outre le fait d'améliorer les performances des systèmes, doit avoir comme principales vocations de :

1. « dégrossir » de grandes quantités d'informations pour ensuite pratiquer une analyse « locale », plus fine et manuelle ;
2. vérifier des hypothèses difficilement réalisables par des méthodes non automatiques c-à-d analytique ou manuelle ;

[7] en fait, la coda de la voyelle se superpose à l'occlusion dans le cas des occlusives

3. être confronté à des résultats expérimentaux « inattendus » grâce à la quantité et diversité des données traitées.

La méthodologie adoptée dans cette thèse - de ne pas chercher à optimiser les performances du système adapté mais plutôt de privilégier la recherche d'informations pertinentes pour la caractérisation de la dysphonie - s'est avérée comme étant la véritable attente des cliniciens. Le besoin des spécialistes de la voix est actuellement de mieux comprendre les phénomènes de la dysphonie. Cette première étape d'acquisition d'une meilleure connaissance des caractéristiques de la dysphonie pourra permettre par la suite d'améliorer la limite des 80 % atteinte par le système automatique, et ce d'autant plus que les « aller-retour » entre l'analyse automatique et l'analyse manuelle sont instaurés et fréquents.

Perspectives

Dans le cadre de la caractérisation de la voix dysphonique, plusieurs axes de perspectives peuvent être proposés :

[0] Validation des résultats sur un plus grand corpus de voix dysphoniques

Il est à noter que la taille réduite du corpus CVD (80 voix) peut influencer fortement la qualité des modèles des différents grades et, par conséquent, les résultats obtenus malgré les différents protocoles (technique de « leave_x_out ») mis en œuvre pour pallier le manque manifeste de données. L'ensemble des résultats obtenus par l'approche statistique doit donc être validé sur un plus grand corpus de voix dysphoniques afin de confirmer les observations relevées.

[1] Approfondissement sur la paramétrisation MFCC

L'étude paramétrique (chapitre 5) a montré que la paramétrisation de type MFCC obtenait de bonnes performances pour la reconnaissance des grades dysphoniques. Sachant qu'ils sont censés représenter essentiellement l'information du conduit vocal et que la dysphonie est directement liée à une perturbation de la vibration des cordes vocales, on peut se demander, d'une part, si les coefficients cepstraux ne contiennent pas plus d'information sur la source glottique, et d'autre part, si la dysphonie ne se manifeste pas également au niveau supra-laryngé. Considérée comme « l'état de l'art » en terme de paramétrisation pour la RAL, il serait intéressant d'approfondir les deux questions soulevées ci-dessus et notamment, d'étudier si l'information relative à la dysphonie est dépendante de la taille des vecteurs acoustiques de type MFCC.

[2] Pertinence des coefficients

Dans le système adapté à la classification des voix dysphoniques, l'utilisation des coefficients delta a montré qu'aucune nature d'information dynamique ne se distingue des autres en terme d'amélioration de performance (section 5.2). Il est pourtant difficile de considérer que les caractéristiques des phénomènes liés à la dysphonie ne puissent transparaître dans les informations de nature dynamique, sachant que le trouble vocal se manifeste comme un phénomène irrégulier, se superposant aux caractéristiques phonétiques et linguistiques du signal de parole. Partant de l'hypothèse que

l'inefficacité du système soit due à un excédent de coefficients dynamiques dans la configuration paramétrique ($\Delta + \Delta\Delta + \Delta\Delta\Delta$), nous proposons de réduire la dimension des vecteurs acoustiques en retirant les coefficients les moins pertinents. Pour cela, différentes techniques peuvent être utilisées :

- le retrait de certains coefficients dynamiques d'ordre élevé [Ljolje, 1994]
- la méthode « knock-out » [Sambur, 1975; Aha & Bankert, 1996]
- la « sélection ascendante » [Aha & Bankert, 1996; Charlet, 1997; Fredouille, 2000]
- une classification des coefficients selon un critère particulier [Bocchieri & Wilpon, 1993]
- une Analyse en Composantes Principales (ACP) pour décorréler les coefficients [Wang et al., 1993]
- une Analyse Linéaire Discriminante (ALD) par l'évaluation du F-Ratio de chaque coefficient [Tridgell et al., 1992]

Nous proposons d'étudier le comportement du système adapté au contexte pathologique à travers la réduction de la dimension de l'espace paramétrique par une ACP. Le principe est de décorréler les coefficients pour ensuite représenter leurs dispersions avec des matrices de covariances diagonales (matrices de vecteurs propres). L'exploration de cette voie nous paraît importante en raison de la particularité de la dysphonie de se manifester comme un phénomène intermittent voire chaotique.

[3] Les paramètres prosodiques

Les paramètres prosodiques (mélodie, intensité et durée) mettent en évidence le style d'élocution (débit, durée et fréquence des pauses, ...), ainsi que les caractéristiques de la source glottale (fréquence fondamentale, énergie, taux de voisement, ...). Ces paramètres s'avèrent cependant fragiles en pratique et ne permettent pas, à eux seuls, de discriminer les locuteurs. En conséquence, ils sont souvent associés aux paramètres de l'analyse spectrale (par exemple l'énergie). De plus, ils restent difficiles à extraire de manière automatique. Nous proposons d'extraire les valeurs de paramètres prosodiques (F0, intensité, ...), de les lisser avec un polynôme de Legendre avant de les concaténer aux vecteurs de paramètres. Cette technique d'approximation des contours de la fréquence fondamentale et de l'énergie a déjà été utilisée dans les domaines IAL et RAL [Qin et al., 2003; Dehak et al., 2007].

[4] Autres paramétrisations

Proposés comme une représentation spectrale améliorée [Hermansky, 1990], les coefficients par prédiction linéaire perceptive (PLP) sont caractérisés par leur capacité à modéliser étroitement la psychoacoustique de l'audition humaine. Il serait intéressant d'utiliser les coefficients PLP sur des voix pathologiques afin d'en apprécier leurs possibilités, ainsi que d'autres paramètres comme par exemple, les RASTA-PLP (RelAtive SpecTrAl Perceptual Linear Predictive) issus de la technique RASTA combinée à la méthode PLP ou encore, les coefficients issus de l'analyse en ondelettes.

[5] Analyse de phénomènes transitoires

L'analyse phonétique décrite en 7.1 a montré la pertinence « peu attendue » des consonnes et plus particulièrement des consonnes sourdes. Plusieurs spectrogrammes laissent apparaître que les consonnes sourdes dans les transitions de type « VC_{sourde} » incluent systématiquement une partie de la finale de la voyelle précédente. Dans le même sens, une étude préliminaire décrite en 7.2 a porté sur l'allongement de la durée du VOT dans un contexte « $C_{sourde}V$ » et sur l'hypothèse d'une corrélation avec le dégré de sévérité. L'analyse statistique a montré que le grade de la dysphonie produisait un effet linéaire et significatif sur l'allongement de la durée du VOT et cela, quelle que soit la nature de l'occlusive sourde.

Dans un permier temps, il serait intéressant d'analyser manuellement la qualité de l'alignement automatique en estimant les portions de voyelles incluses par erreur dans les segments des consonnes sourdes (et inversement) afin de déterminer, le cas échéant, des phénomènes transitoires pouvant être caractéristiques de la dysphonie. Dans un deuxième temps, l'analyse du comportement du système en faisant varier les frontières de certaines combinaisons de phonèmes, apparaissant comme pertinentes, permettrait d'affiner les régions phonémiques les plus caractéristiques de la dysphonie.

[6] Méthode « Automatic Phonetic Labeling »

La méthode « Phonetic Labeling » [Revis et al., 2006] reproduite automatiquement sur le système RAL (section 7.3) n'a pas apporté de résultat probant. Concernant le comportement des consonnes sourdes durant l'approche « Phonetic Labeling », Revis (2004) dit « *...que les phonèmes non-voisés ne sont pas affectés par les occurrences pathologiques dysphoniques (raucité, souffle, dévoisement) dans la mesure où la dysphonie apparaît exclusivement sur la vibration glottique. En revanche, ils peuvent être atteints spécifiquement par les critères phonétiques et particulièrement le paramètre d'aspiration ...* ».

Il serait donc intéressant de poursuivre l'approche « Automatic Phonetic Labeling » en ajoutant aux paramètres acoustiques des mesures « analytiques » sur la stabilité laryngée et sur le souffle de la voix. En effet, l'approche « Phonetic Labeling » semble être une voie intéressante pour élaborer une décision basée sur le critère du « vote majoritaire ». De plus, l'adaptation de cette méthode purement expérimentale à un système automatique permettrait d'offrir aux spécialistes de la voix une meilleure compréhension du fonctionnement phonétique de la dysphonie de par son haut niveau de détail proposé.

[7] Paradigme de décision

L'analyse phonétique (chapitre 7) a montré qu'un nouveau paradigme de décision pourrait être défini à partir de décisions locales prises sur des portions de parole spécifiques (phonème, classe de phonèmes, ...). Offrant des règles exhaustives et mutuellement exclusives, une structure en arbre de décision phonétique semble être une voie intéressante dans ce contexte pour améliorer les performances du système de classification des voix dysphoniques. Cependant, il faut préciser qu'en raison du peu de données disponibles du corpus CVD, l'élaboration d'un arbre de décision ne peut être entrepris à l'heure actuelle.

Par ailleurs, la conception d'un système de décision à partir de plusieurs experts (ou sources d'informations) peut être une manière de définir une stratégie de décision mieux adaptée à un contexte particulier. Sa mise en œuvre ici consisterait à considérer des sous-systèmes experts, chacun associé à une source d'information particulière (phonème, classe de phonèmes), comme des composantes du système de décision. La fusion des décisions issues des différents experts peut être réalisée par différentes approches comme des modèles statistiques (GMM, SVM, MLP, ...) ou encore, des opérations arithmétiques (addition, multiplication, moyenne, vote majoritaire, ...). Généralement, les systèmes multi-experts s'appuient sur une stratégie de pondération par des indices représentatifs du degré de confiance des différents experts [Rahman & FairHurst, 1998] qui constituent une connaissance *a priori* du pouvoir discriminant de chaque expert pour la tâche de reconnaissance.

[8] Apprentissage non supervisée

Malgré les imperfections du jugement auditif, l'approche statistique s'appuie tout de même sur l'analyse perceptive la considérant comme le «*Gold Standard*» pour l'évaluation de la dysphonie (faute d'un autre système de référence plus fiable). Dans ce sens, chaque modèle d'un grade donné étant entraîné avec les sujets de grade perceptif correspondant, peut comporter une part d'imprécision pouvant probablement être responsable de la limite actuelle atteinte par l'approche statistique. Face à cette difficulté, une première action envisageable serait de se libérer des imperfections du jugement perceptif en utilisant la technique d'apprentissage non supervisée pour les modèles de grade. Une seconde action serait alors de comparer les résultats de l'analyse perceptive, de méthodes instrumentales analytiques et de l'approche statistique en mode non supervisé afin d'étudier les points de désaccord et de concordance et ainsi, permettre que chaque approche soit indépendante les unes des autres.

Cinquième partie

Annexes

Les Dysarthries

Les dysarthries sont présentées suivant les six groupes définis dans la classification de Le Huche & Allali (2001b). Les morceaux de texte affichés «*en italique et entre parenthèse*» correspondent à des passages extraits de l'ouvrage du même auteur.

Les dysarthries paralytiques

« *Elles peuvent correspondre à des lésions centrales (syndrome pseudo-bulbaire), périphériques (atteintes de certains noyaux bulbaires), mixte (lésions dégénératives à la fois centrales et périphériques) ou par extension à un trouble de la jonction neuromusculaire.* »

Exemple : ***la maladie de Charcot (ou SLA)*** où la disparition progressive des neurones moteurs périphériques entraîne une atrophie musculaire progressive avec des troubles de la déglutition et de la parole. Les troubles de la parole contrastent fortement avec la conservation de l'intégrité des fonctions intellectuelles du malade (compréhension conservée, langage normal, ...) :

- rythme ralenti avec un débit métronomique ;
- articulation laborieuse et imprécise ;
- timbre poussif ;
- voix monotone, voire nasonnée ;
- ... .

Les dysarthries akinétiques

« *Elles concernent les altérations de la voix et de la parole dans le syndrome parkinsonien.* »

Sachant que ces troubles intéressent plus de la moitié des malades parkinsoniens, la dysarthrie peut alors être le 1er signe de la maladie.

Exemple : ***la maladie de Parkinson*** qui est caractérisée par trois éléments (1) le tremblement qui se manifeste au repos et disparait durant un mouvement volontaire ou dans le sommeil (2) l'hypertonie dite *plastique*[8] qui se traduit par un effet de *roue dentée* lors d'un mouvement musculaire (allongement de l'avant-bras par à-coups avec une résistance constante durant le mouvement d'extension) (3) l'akinésie provoquant la réduction générale du mouvement corporel (visage figé, inexpressif). Les troubles de la parole et de la voix chez le parkinsonien sont :

- réduction de l'activité verbale ;
- hauteur tonale souvent augmentée ;
- voix triste, monotone et chevrotante ;
- timbre souvent sourd et voilé ;
- difficulté de démarrage de la parole avec répétition des premières syllabes ;
- rythme rapide de la parole avec accélération progressive ;
- ... .

Les dysarthries dyskinétiques

« *Il s'agit d'une altération des noyaux gris centraux qui peut entraîner des mouvements anormaux susceptibles d'altérer plus ou moins gravement l'articulation de la parole.* »

Exemple : ***la chorée*** est due à une perte neuronale dans des zones cérébrales précises, notamment le striatum. Elle peut être acquise (d'origine infectieuse, toxique ou médicamenteuse) ou génétique (chorée de Huntington). Elle est caractérisée par des mouvements choréiques typiques (involontaires, imprévisibles, brusques, incessants, ...) associés à une hypotonie générale (baisse de la tonicité, de la force musculaire). La dysarthrie se caractérise par une articulation difficile de la parole : rythme entrecoupé de périodes de silence prolongées et de salves de syllabes prononcées brusquement avec une faible intensité, parfois la voix serrée.

[8] responsable du trouble postural caractéristique du sujet parkinsonien

Les dysarthries ataxiques

« *Elles regroupent les altérations de l'articulation de la parole et de la dysphonie, consécutives à une atteinte bilatérale du cervelet ou des voies cérébelleuses.* »

Exemple : ***le syndrome cérebelleux*** dont les caractéristiques principales sont (1) perturbation de l'amplitude du mouvement (2) incoordination de mouvements volontaires (3) retard au démarrage et à l'arrêt du mouvement (4) difficulté d'exécution rapide de mouvements répétitifs (5) baisse de la tonicité, de la force musculaire (hypotonie musculaire). La parole du dysarthrique cérébelleux est caractérisée par l'aspect irrégulier et explosif. Le débit de parole est relativement lent à cause des hésitations et des arrêts intempestifs. L'intensité est irrégulière avec de brusques montées de hauteur tonale dénotant la difficulté à synchroniser le souffle pulmonaire avec les gestes phonatoires et articulatoires. La dysarthrie cérébelleuse se caractérise aussi par son imprécision articulatoire lui attribuant une importante dimension dysphonique : assourdissement des consonnes sonores, attaques brutales, exagération du bruit d'explosion des occlusives, variation chaotique de la hauteur tonale avec des périodes de monotonie, ...

Les dysarthries apraxiques

« *Elles se manifestent par des troubles vocaux et articulatoires en rapport avec une lésion corticale en principe pariétale.* »

Selon la localisation de la lésion cérébrale, l'hémisphère dominant[9] ou l'hémisphère mineur[10], les conséquences sur la dysarthrie sont différentes. Dans le premier cas, on observe (1) une aphonie avec usage exclusif de la voix chuchotée (2) une difficulté à réaliser volontairement des actes élémentaires normalement effectués par divers gestes automatiques ou réflexes (comme tousser ou se racler la gorge) (3) un mauvais contrôle du souffle phonatoire (4) de la parole inintelligible due à la présence de souffle bruyant altérant l'articulation. Dans le second cas, les troubles concernent principalement un défaut de maîtrise de la hauteur tonale et du rythme, se traduisant par un discours haché au débit irrégulier.

Exemple : ***la dysprosodie après AVC***

[9] hémisphère cérébral responsable de l'expression et de la compréhension du langage

[10] hémisphère cérébral qui correspond au siège des fonctions de perception et d'orientation dans l'espace

Les dysarthries dystoniques

« *C'est un trouble moteur caractérisé par des contractions musculaires parasites soutenues et prolongées, déclenchées par l'incitation motrice volontaire et cessant en principe au repos.* »

La dysarthrie dystonique peut être « généralisée » à l'ensemble du corps comme dans *le spasme de torsion* ou alors être plus « localisée » à une région corporelle et n'affecter d'abord qu'un membre pour atteindre progressivement les quatre autres membres.

Elle peut être aussi « focalisée » et ainsi n'altérer que les muscles impliqués dans une même fonction comme le *symptôme de blépharospasme* pour le regard (les muscles périorbitaires et les muscles de la paupière), le *torticolis spasmodique* pour l'orientation de la tête (les muscles de la région cervicale), la *dystonie oro-mandibulaire* pour la fonction mimique (les muscles de la langue, de l'ouverture ou de la fermeture de la bouche), la *dystonie spasmodique* pour la fonction de la parole (les muscles laryngés), la *crampe de l'écrivain* pour l'écriture (les muscles de la main et de l'avant bras), ...

Exemple : ***la dystonie spasmodique*** concerne une atteinte dystonique des muscles laryngés correspondant à un dysfonctionnement des muscles vocaux et/ou respiratoires durant la phonation. On distingue deux formes de dystonies spasmodiques selon la configuration laryngée durant les spasmes (fermeture/ouverture) :

- *en adduction* des cordes vocales (la plus fréquente) : voix étranglée et irrégulière, souvent hachée en raison d'un contexte de serrage vocal interrompu par des relachements intempestifs. Une gêne respiratoire peut être entendue à cause des spasmes respiratoires ;
- *en abduction* des cordes vocales (la plus rare) : voix chuchotée, murmurée et de faible intensité, semblant parfois en suspens avant son émission irrégulière et saccadée au cours de laquelle des tremblements peuvent apparaître.

Le chant, le rire ou la « voix criée » peuvent améliorer (ou même faire disparaître) le trouble vocal de ces deux formes de dysphonies spasmodiques.

Les Dysphonies d'origine organique

Les principales dysphonies d'origine organique sont présentées selon leur étiologie.

La laryngite aiguë

Elle est consécutive à une simple infection rhyno-pharyngée comme un « coup de froid » avec l'apparition de picotements laryngés accompagnés d'une toux sèche. L'altération vocale évolue d'un timbre rauque, irrégulier vers une phonation pouvant devenir difficile et douloureuse, particulièrement lors de l'usage intensif et continu de la voix. Le larynx présente un aspect inflammatoire avec une muqueuse plus ou moins rouge qui peut s'étendre à l'ensemble du larynx. La muqueuse retrouve son aspect normal conjointement au rétablissement de la fonction vocale en quelques jours.

La laryngite chronique

Il s'agit d'une inflammation banale de la muqueuse laryngée qui peut être d'origine microbienne ou irritative. Il en existe plusieurs types comme oedémateuse, hypertrophique rouge, catarrhale, ... On notera principalement comme facteurs favorisants, l'abus d'alcool, de tabac ou le malmenage vocal. Le traitement dépend des différentes formes de lésions observées et de la multiplicité des facteurs étiologiques : antibiotiques, anti-inflammatoires, cures thermales, interventions chirurgicales et rééducations vocales en présence d'un comportement de forçage vocal.

Les laryngites spécifiques

Le Huche & Allali (2001b) les définissent « *comme des atteintes laryngées chroniques dues à un agent infectieux déterminé* ». Nous ne présenterons ici que deux types de ces laryngites :

• **la papillomatose laryngée** est une affection d'origine virale due à un papillomavirus. La prolifération tumorale peut être plus ou moins envahissante aux niveaux des

plis vocaux. La dysphonie se caractérise par une diminution de l'intensité vocale avec un timbre assourdi et une tonalité aggravée. Même à la suite de l'exérèse chirurgicale des papillomes, il existe une possibilité de récidive et de transformation maligne.

• **la tuberculose laryngée** est souvent révélatrice d'une tuberculose pulmonaire. Variable selon l'atteinte laryngée, la dysphonie se caractérise par une toux sèche avec des douleurs à la déglutition. Les cordes vocales peuvent être le siège d'infiltrations et d'ulcérations. Sous l'effet d'un traitement antibiotique, la dysphonie et la douleur à la déglutition peuvent régresser parfois rapidement. Ce n'est pas toujours le cas pour les lésions laryngées qui peuvent laisser des cicatrices responsables d'une dysphonie définitive.

Les traumatismes laryngés

On distinguera deux types de traumatismes laryngés suivant la voie, externe ou interne, par laquelle le larynx est exposé.

• **les traumatismes externes** correspondent à des atteintes laryngées qui surviennent au cours des accidents de la route ou sportifs pour les plus fréquents. Le choc au niveau du larynx peut provoquer oedème, hématome, contusion ou étirement des plis vocaux, ainsi que des fractures et luxations du larynx. Généralement, la dysphonie se caractérise par une aphonie complète avec une déperdition massive du souffle se traduisant par une voix chuchotée, très rauque ou sourde, avec un comportement de forçage vocal. Les troubles respiratoires et vocaux peuvent persister plusieurs années en raison des multiples interventions chirurgicales (explorations, réparations, ...) nécessaires et consécutives aux traumatismes occasionnés.

• **les traumatismes internes** correspondent principalement à des lésions laryngées provoquées à la suite d'interventions chirurgicales. On y retrouve les dysphonies résultant d'actes chirurgicaux de **la filière respiratoire** comme pour les sténoses (élargissement de l'espace glottique) ou la trachéotomie (création d'un orifice respiratoire sur la partie antérieure du cou), ou **après intubation** nécessaire durant l'anesthésie pour assurer la protection des voies respiratoires, mais aussi **en microchirurgie laryngée** pour des lésions bénignes (polypes, nodules, kystes, ...) pouvant provoquer des accidents tels que l'encoche cordale ou l'immobilisation du pli vocal.

Plus rares, les traumatismes laryngés internes regroupent les dysphonies déclenchées à la suite de brûlures laryngées (inhalation de vapeurs, de fumées chaudes, ...), de radiothérapie ou après un traumatisme vocal brutal et soudain (cri, ...).

Les paralysies laryngées

Le nerf laryngé inférieur (ou nerf récurrent) assure essentiellement l'innervation motrice des muscles intrinsèques du larynx. Fréquemment, une lésion du nerf récurrent entraîne un défaut de mobilité d'un pli vocal. Ainsi, on parlera de **paralysie récurrentielle unilatérale**. Les traumatismes occasionnés au nerf récurrent peuvent être provoqués par étirement, par échauffement ou par section à la suite d'interventions chirurgicales, par compression due au développement d'un nodule thyroïdien, ou par une atteinte virale comme la névrite.

La dysphonie se traduit par une voix bitonale et une déperdition importante du souffle respiratoire. L'altération vocale est telle que la voix est détimbrée avec de nombreuses désonorisations irrégulières. L'effort vocal constant rend la phonation fatigante et pénible. La rééducation vocale permet des progrès réguliers et très encourageants pour le patient. Le rééducateur apprendra au patient à obtenir l'affrontement des plis vocaux : par des manipulations latérocervicales, le pli vocal sain est amené, par hyperadduction, à franchir la ligne médiane et venir se porter vers le pli vocal paralysé. L'intervention chirurgicale pourra être envisagée en cas d'inefficacité de la rééducation vocale.

Plus rares, les **paralysies récurrentielles bilatérales** correspondent à un déficit de mobilité bilatérale des plis vocaux relevant essentiellement d'une atteinte neurogène (d'origine centrale ou périphérique).

Les anomalies laryngées congénitales

Il existe des dysphonies provoquées par des anomalies congénitales du larynx. Les troubles de la phonation peuvent alors être aussi bien acceptés par le patient que provoquer de graves difficultés aboutissant inéluctablement à des traitements rééducatifs et/ou chirurgicaux. De symptômes très variés, les dysphonies organiques congénitales peuvent être définies en trois catégories.

• **les anomalies de la structure laryngée** correspondent à des malformations rares. Citons **le diastème laryngé postérieur** (fente postérieure laryngée ou laryngotrachéale) provoquant des troubles de la déglutition et une phonation possible par la mise en place de mécanismes de compensation, ou **la palmure laryngée** (forme d'atrésie[11] laryngée mineure caractérisée par une fine membrane souvent localisée au niveau glottique et non gênante dans la mobilité des CV) provoquant des difficultés respiratoires et une phonation entrecoupée de râlements rauques.

• **les anomalies de la commande laryngée** englobent, entre autres, **les paralysies laryngées congénitales** pouvant être en rapport avec des affections neurologiques et se manifestant dans les premiers mois de la vie par des troubles respiratoires et de la dé-

[11] se définit comme un rétrécissement d'un conduit ou d'un orifice de l'organisme

glutition, **la maladie de Down** (**Trisomie 21**) ou **la maladie du cri du chat** (**Syndrome de Lejeune**, trouble génétique dû à la délétion du chromosome 5).

• **les pseudo-tumeurs bénignes** où se trouvent **le kyste congénital du pli vocal** dont l'altération vocale est assez similaire à celle du kyste muqueux par rétention, **le sulcus glottidis étroit** (kyste ouvert du pli vocal) caractérisé par une altération du timbre étouffé avec une tonalité élevée ou **la vergeture des plis vocaux** présentant une voix limitée dans le grave provoquant une certaine fatigabilité vocale.

Les Informations Dynamiques

Ajout des coefficients dynamiques en contexte de 7 trames

Les résultats obtenus à l'ajout des coefficients dynamiques calculés avec une taille de fenêtre de 7 trames, sont présentés dans le tableau 7.7.

Paramètres dynamiques	**16LFCC**	**16MFCC**	**24LFSC**	**24MFSC**
Coefficients statiques	*63.75 (51)*	*70.00 (56)*	*65.00 (52)*	*73.75 (59)*
Δ	68.75 (55)	71.25 (57)	65.00 (52)	**75.00 (60)**
$\Delta + \Delta\Delta$	66.25 (53)	75.00 (60)	71.25 (57)	**77.50 (62)**
$\Delta + \Delta\Delta + \Delta\Delta\Delta$	66.25 (53)	**73.75 (59)**	72.50 (58)	**73.75 (59)**

TAB. 7.7 – *Résultats globaux de la classification 4-G en terme de % TCC (nb/80).*
Ajout des paramètres dynamiques en contexte de 7 trames selon différentes paramétrisations.

Comparativement avec les résultats obtenus dans un contexte de 5 trames du tableau 5.2, l'augmentation du contexte temporel ne profite que pour les analyses à échelle linéaire :

- de 67.50 % ($\Delta + \Delta\Delta + \Delta\Delta\Delta$) $\rightsquigarrow$ 68.75 % (Δ) pour les paramétres LFCC (+ 1 locuteur) ;
- de 70.0 % (Δ) $\rightsquigarrow$ 72.50 % ($\Delta + \Delta\Delta + \Delta\Delta\Delta$) pour les paramétres LFSC (+ 2 locuteurs).

La paramétrisation de type MFCC n'améliore pas son meilleur TCC obtenu précédemment de 75.00 % avec l'ajout de ($\Delta + \Delta\Delta + \Delta\Delta\Delta$) et atteint ici avec l'ajout des coefficients dynamiques ($\Delta + \Delta\Delta$). *Même si nous sortons du cadre de cette thèse, nous pouvons souligner que l'intérêt ici est qu'en augmentant la taille de la fenêtre temporelle (de 5 à 7 trames), la performance reste la même avec un nombre de coefficients beaucoup plus réduit (de 64 à 48 coefficients), ce qui n'est pas négligeable en terme de temps de calcul et d'occupation mémoire.*

Pour les paramètres MFSC, l'aggrandissement de la fenêtre de 5 à 7 trames n'apporte aucune amélioration si l'on compare les scores TCC obtenus par nature de l'information dynamique :

- configuration Δ : TCC de 75.00 % conservé ;
- configuration $\Delta\Delta$: de 78.75 % $\rightsquigarrow$ 77.50 % (perte d'un locuteur) ;
- configuration $\Delta\Delta\Delta$: de 77.50 % $\rightsquigarrow$ 73.75 % (perte de trois locuteurs).

Ajout des coefficients dynamiques en contexte de 9 trames

Le tableau 7.8 présente les résultats obtenus à l'ajout des coefficients dynamiques calculés avec une taille de fenêtre de 9 trames.

Paramètres dynamiques	16LFCC	16MFCC	24LFSC	24MFSC
Coefficients statiques	*63.75 (51)*	*70.00 (56)*	*65.00 (52)*	*73.75 (59)*
Δ	66.25 (53)	72.50 (58)	63.75 (51)	**76.25 (61)**
Δ + ΔΔ	66.25 (53)	70.00 (56)	71.25 (57)	**76.25 (61)**
Δ + ΔΔ + ΔΔΔ	66.25 (53)	72.50 (58)	68.75 (55)	**75.00 (60)**

TAB. 7.8 – *Résultats globaux de la classification 4-G en terme de % TCC (nb/80). Ajout des paramètres dynamiques en contexte de 9 trames selon différentes paramétrisations.*

Il n'apparaît aucun bénéfice à utiliser un contexte temporel de cette dimension à la vue des résultats présentés. Aucune paramétrisation n'améliore son meilleur TTC atteint dans un contexte temporel plus réduit.

Synthèse sur l'ajout des coefficients dynamiques

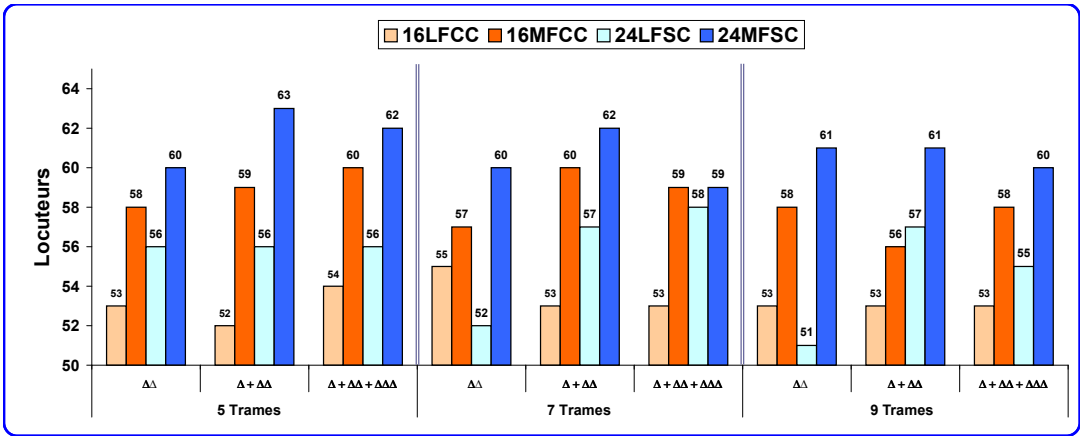


FIG. 7.18 – *Résultats de la classification 4-G selon différentes paramétrisations en terme de nombre de locuteurs correctement classés sur 80.*

La figure 7.18 synthétise les trois tableaux (5.2, 7.7, 7.8) présentés précédemment.

Pour chacune des différentes paramétrisations, les configurations[12] obtenant les meilleurs TCC globaux sont :

- **les paramètres LFCC** : 68.75 % (55) en contexte de 7 trames
- **les paramètres MFCC** : 75.00 % (60) en contexte de 5 et 7 trames
- **les paramètres LFSC** : 72.50 % (58) en contexte de 7 trames
- **les paramètres MFSC** : 78.75 % (63) en contexte de 5 trames

De plus, deux paramétrisations se distinguent nettement de part leurs performances :

- **les paramètres MFCC** : «l'état de l'art» en RAL ;
- **les paramètres MFSC** : meilleurs résultats pour des voix dysphoniques.

Pour résumé, le contexte de 5 trames est plutôt favorable aux analyses acoustiques à échelle Mel *i.e.* les paramétrisations MFCC/MFSC. Par contre, le contexte de 7 trames semble être bénéfique pour les coefficients cepstraux et les analyses acoustiques à échelle linéaire *i.e.* les paramétrisations LFCC/MFCC/LFSC. *Il sera tout de même préféré le contexte de 7 trames pour la paramétrisation MFCC en raison de la taille plus réduite du nombre de paramètres (48 vs. 64 coefficients).*

Il faut tout de même préciser que sur l'ensemble des résultats présentés, la variation de la fenêtre temporelle n'apporte pas une information significative, excepté peut-être pour les paramètres LFSC en contexte de 7 trames, et encore !

[12] taille de la fenêtre temporelle et nature de l'information dynamique ajoutée

Bibliographie Personnelle

Revue

C. Fredouille, G.Pouchoulin, A. Ghio, J. Revis, J.-F. Bonastre & A. Giovanni, 2009. Back-and-forth methodology for objective voice quality assessment : from/to expert knowledge to/from automatic classification of dysphonia. EURASIP Journal on Applied Signal Processing, special issue on « Analysis and signal processing of oesophageal and pathological voices », Mai 2009. (en soumission)

Conférences internationales

G. Pouchoulin, C. Fredouille, J.-F. Bonastre, A. Ghio & A. Giovanni, 2008. Dysphonic Voices and the 0-3000Hz Frequency Band. Interspeech'08, Brisbane, Australie, Septembre 2008.

J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve & J. Mason, 2008. ALIZE/SpkDet : a state-of-the-art open source software for speaker recognition. The Speaker Recognition Workshop, Afrique du Sud, Janvier 2008.

G. Pouchoulin, C. Fredouille, J.-F. Bonastre, A. Ghio & A. Giovanni, 2007. Frequency-Based Analysis for the Characterization of the Dysphonic Voices. PEVOC'07 : Pan European Voice Conference, Groningen, Pays-Bas, Août 2007. (sans acte)

G. Pouchoulin, C. Fredouille, J.-F. Bonastre, A. Ghio & A. Giovanni, 2007. Frequency Study for the Characterization of the Dysphonic Voices. Interspeech'07, Anvers, Belgique, Août 2007.

J.-F. Bonastre, C. Fredouille, A. Ghio, A. Giovanni, G. Pouchoulin, J. Revis, B. Teston & P. Yu, 2007. Complementary approaches for voice disorder assessment. Interspeech'07, Anvers, Belgique, Août 2007.

G. Pouchoulin, C. Fredouille, J.-F. Bonastre, A. Ghio & J. Revis, 2007. Characterization of the pathological voices (dysphonia) in the frequency space. International Congress on Phonetic Sciences, ICPhS'07, Saarbrucken, Allemagne, Août 2007.

C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni & A. Ghio, 2005. Application of automatic speaker recognition techniques to pathological voice assessment (dysphonia). Eurospeech'05, Lisbonne, Portugal, Septembre 2005.

Conférences nationales

G. Pouchoulin, C. Fredouille, J.-F. Bonastre, A. Ghio & A. Giovanni, 2008. Analyse Phonétique dans le Domaine Fréquentiel pour la Classification des Voix Dysphoniques. XXVIIeme Journées d'Etude sur la Parole, JEP'08, Avignon, France, 9-13 Juin 2008.

G. Pouchoulin, C. Fredouille, J.-F. Bonastre, A. Ghio & A. Giovanni, 2007. Analyse fréquentielle pour la caractérisation des voix dysphoniques. Deuxièmes Journées Phonétique Clinique (JPC2), Grenoble, France, 13-14 Décembre 2007. (sans acte)

G. Pouchoulin, C. Fredouille, J.-F. Bonastre, A. Ghio, M. Azzarello & A. Giovanni, 2006. Modélisation Statistique et Informations Pertinentes pour la Caractérisation des Voix Pathologiques (Dysphonies). XXVIeme Journées d'Etude sur la Parole, JEP'06, Dinar, France, 12-15 Mai 2006.

Travaux interdisciplinaires

A. Ghio, G. Pouchoulin, A. Giovanni, C. Fredouille, B. Teston, J. Révis, J.-F. Bonastre, D. Robert, P. Yu, M. Ouaknine, M.-D. Guarella, C. Spezza, T. Legou & A. Marchal, 2007. Approches complémentaires pour l'évaluation des dysphonies : bilan méthodologique et perspectives. Travaux Interdisciplinaires du Laboratoire Parole et Langage, vol. 26, p. 33-74.

A. Ghio, B. Teston, F. Viallet, L. Jankowski, A. Purson, D. Duez, J. Locco, T. Legou, S. Pinto, A. Marchal, A. Giovanni, D. Robert, J. Révis, C. Fredouille, J.-F. Bonastre & G.Pouchoulin, 2006. Corpus de « parole pathologique ». État d'avancement et enjeux méthodologiques au LPL. Travaux Interdisciplinaires du Laboratoire Parole et Langage, vol. 25, p. 109-126.

Liste des acronymes

ACP	Analyse en Composantes Principales
ALD	Analyse Linéaire Discriminante
API	Alphabet Phonétique International
AVC	Accident Vasculaire Cérébral
CMS	Cepstral Mean Subtraction
CV	Corde Vocale
CVD	Corpus des Voix Dysphoniques
DCT	Discrete Cosine Transform
EM	Expectation-Maximization
EP	Evaluation Perspective
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IAL	Identification Automatique du Locuteur
IC	Intervalle de Confiance
IDCT	Inverse Discrete Cosine Transform
INRIA	Institut National de Recherche en Informatique et en Automatique
LFSC	Linear Frequency Spectrum Coefficients
LFCC	Linear Frequency Cepstral Coefficients
LIA	Laboratoire d'Informatique d'Avignon
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
LPC	Linear Predictive Coefficients
LPCC	Linear Predictive Cepstral Coefficients
MAP	Maximum A Posteriori
MFSC	Mel Frequency Spectrum Coefficients
MFCC	Mel Frequency Cepstral Coefficients
ML	Maximum Likelihood
NIST	National Institute of Standards and Technologies
QV	Qualité de la Voix
ORL	Ortho-Rhino-Laryngologie
PLP	Perceptual Linear Predictive
PV	Plis Vocaux
RAL	Reconnaissance Automatique du Locuteur
RAP	Reconnaissance Automatique de la Parole
SLA	Sclérose Latérale Amyotrophique
SVM	Support Vector Machine
TAP	Traitement Automatique de la Parole
TC	Traumatisme Cranien
TCC	Taux Correct de Classification
TCD	Transformée en Cosinus Discrète
TCDI	Transformée en Cosinus Discrète Inverse
TFD	Transformée de Fourier Discrète
UBM	Universal Background Model
VAL	Vérification Automatique du Locuteur

Liste des illustrations

Liste des tableaux

Bibliographie

[Aha & Bankert, 1996] D. W. Aha & R. L. Bankert, 1996. A comparative evaluation of sequential feature selection algorithms. *Artificial Intelligence and Statistics*.

[Alonso et al., 2005] J. Alonso, F. Diaz, C. Travieso, & M. Ferrer, 2005. Using nonlinear features for voice disorder detection. Dans les actes de *NOLISP'05*, Barcelona, Spain, 94–106.

[Alpan et al., 2008] A. Alpan, Y. Maryn, F. Grenez, A. Kacha, & J. Schoentgen, 2008. Multi-band and multi-cue analyses of disordered connected speech. Dans les actes de *Interspeech'08*, Brisbane, Australia.

[Anders et al., 1988] L. Anders, H. Hollien, P. Hurme, A. Sonninnen, & J. Wendler, 1988. Perceptual evaluation of hoarseness by several classes of listeners. *Folia Phoniatrica et Logopaedica 40*, 91–100.

[Auckenthaler et al., 2000] R. Auckenthaler, M. Carey, & H. Lloyd-Thomas, 2000. Score normalization for text-independent speaker verification system. *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop 10(1-3)*, 42–54.

[Auckenthaler & Mason, 1997] R. Auckenthaler & J. S. Mason, 1997. Equalizing subband error rates in speaker recognition. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech'97)*, Rhodes (Greece), 2303–2306.

[Azzarello, 2006] M. Azzarello, 2006. Analyse phonétique de la dysphonie : Application des méthodes statistiques issues de la reconnaissance automatique du locuteur. Mémoire de Master, Université de Provence, Laboratoire d'Audio Phonologie Expérimentale Clinique, CHU Timone, Marseille, France.

[Bernasconi, 1990] C. Bernasconi, 1990. On instantaneous and transitional spectral information for text-dependent speaker verification. *Speech Communication 9(2)*, 129–139.

[Besacier, 1998] L. Besacier, 1998. *Un Modèle Parallèle pour la Reconnaissance Automatique du Locuteur*. Thèse de Doctorat, Académie d'Aix-Marseille, Université d'Avignon et des Pays de Vaucluse.

[Besacier & Bonastre, 1998] L. Besacier & J.-F. Bonastre, 1998. Frame pruning for speaker recognition. Dans les actes de *International Conference on Acoustics Speech and Signal Processing (ICASSP'98)*, Seatle (USA).

[Besacier et al., 2000] L. Besacier, J.-F. Bonastre, & C. Fredouille, 2000. Localization and selection of speaker specific information with statistical modelling. *Speech Communication 31*, 89–106.

[Bimbot et al., 2004] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, & D. A. Reynolds, 2004. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing 39*, 430–451.

[Bocchieri & Wilpon, 1993] E. L. Bocchieri & J. G. Wilpon, 1993. Discriminative feature selection for speech recognition. *Computer Speech and Language 7(3)*, 229–246.

[Boite & Kunt, 1987] R. Boite & M. Kunt, 1987. *Traitement de la parole. Presses* (Polytechniques Romandes ed.).

[Bonastre et al., 2005] J.-F. Bonastre, F. Wils, & S. Meignier, 2005. Alize, a free toolkit for speaker recognition. Dans les actes de *International Conference on Acoustics Speech and Signal Processing (ICASSP'05)*, Philadelphia, USA.

[Bourlard & Dupont, 1996] H. Bourlard & S. Dupont, 1996. A new ASR approach based on independent processing and combination of partial frequency bands. Philadelphia (USA). International Conference on Spoken Language Processing (ICSLP'96).

[Briffa, 2004] C. Briffa, 2004. Analyse de la dysphonie : Application des méthodes statistiques issues de la reconnaissance automatique du locuteur. Mémoire de Master, Université de Provence, Laboratoire d'Audio Phonologie Expérimentale Clinique, CHU Timone, Marseille, France.

[Bürki et al., 2008] A. Bürki, C. Gendrot, G. Gravier, G. Linarès, & C. Fougeron, 2008. Alignement automatique et analyse phonétique : comparaison de différents systèmes pour l'analyse du schwa. *TAL*.

[Charlet, 1997] D. Charlet, 1997. *Authentification vocale par téléphone en mode dépendant du texte*. Thèse de doctorat.

[Chen et al., 2007] W. Chen, C. Peng, X. Zhu, B. Wan, & D. Wei, 2007. Svm-based identification of pathological voices. Dans les actes de *29th Annual International Conference of the IEEE*, 3786–3789.

[Crevier-Buchman et al., 1993a] L. Crevier-Buchman, M.-C. Monfrais-Pfauwadel, D. Begue, L. Lauga-Houdoyer, O. Laccourreye, & D. Brasnu, 1993a. Acoustic evaluation and use of computers. *Revue de laryngologie, d'otologie et de rhinologie 114*, 311–314.

[Crevier-Buchman et al., 1993b] L. Crevier-Buchman, M.-C. Monfrais-Pfauwadel, O. Laccourreye, V. Jouffre, D. Brasnu, & H. Laccourreye, 1993b. La laryngostroboscopie. *Annales d'oto-laryngologie et de chirurgie cervico-faciale 110*, 355–357.

[Darley et al., 1969a] F. L. Darley, A. E. Aronson, & J. R. Brown, 1969a. Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research 12*, 462–496.

[Darley et al., 1969b] F. L. Darley, A. E. Aronson, & J. R. Brown, 1969b. Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research 12*, 246–269.

[Darley et al., 1975] F. L. Darley, A. E. Aronson, & J. R. Brown, 1975. *Motor speech disorders*. Philadelphia : W. B. Saunders and Co.

[Davis & Mermelstein, 1980] S. Davis & P. Mermelstein, 1980. Comparison of parametric representation for monosyllabic word recognition in continuous spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing 4*, 357–366.

[De Bodt et al., 1996] M. S. De Bodt, P. H. Van de Heyning, F. L. Wuyts, & L. Lambrechts, 1996. The perceptual evaluation of voice disorders. *Acta Oto-rhino-laryngolica Belg. 50(4)*, 283–291.

[De Bodt et al., 1997] M. S. De Bodt, F. L. Wuyts, P. H. Van de Heyning, & C. Croux, 1997. Test-retest study of the GRBAS scale : influence of experience and professional background on perceptual ratings of voice quality. *Journal of Voice 1*, 74–80.

[De Krom, 1993] G. De Krom, 1993. A cepstrum based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research 36*, 254–266.

[De Krom, 1994] G. De Krom, 1994. Consistency and reliability of voice quality ratings for different types of speech fragments. *Journal of Speech and Hearing Research 37(5)*, 985–1000.

[De Krom, 1995] G. De Krom, 1995. Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of Speech and Hearing Research 38(4)*, 794–811.

[Dehak et al., 2007] N. Dehak, P. Dumouchel, & P. Kenny, 2007. Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing 15 (7)*, 2095–2103.

[Dejonckere, 1986] P. Dejonckere, 1986. Acoustic analysis of voice production. production trial from a clinical perspective. *Acta Oto-rhino-laryngolica Belg. 40*, 377–385.

[Dejonckere, 1996] P. Dejonckere, 1996. Electroglottography : A useful method in voice investigation. *Voice Update. Excerpta Medica. Elsevier*, 29–33.

[Dejonckere et al., 2001] P. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. Van De Heyning, M. Remacle, & V. Woisard, 2001. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques : Guidelines elaborated by the committee on phoniatrics of the european laryngological society (ELS). *European archives of oto-rhino-laryngology 258(2)*, 77–82.

[Dejonckere et al., 1993] P. Dejonckere, C. Obbens, G. De Moor, & G. Wieneke, 1993. Perceptual evaluation of dysphonia : reliability and relevance. *Folia Phoniatrica et Logopaedica 45*, 76–83.

[Dejonckere & Villarosa, 1986] P. Dejonckere & D. Villarosa, 1986. Analyse spectrale moyennée de la voix. Comparaison de voix normales et de voix altérées par différentes catégories de pathologies laryngées. *Acta Oto-rhino-laryngolica Belg. 40*, 426–435.

[Deller et al., 1999] J. R. Deller, J. H. L. Hansen, & J. G. Proakis, 1999. Discrete-time processing of speech signals. *IEEE Press 2*, 636–775.

[Dempster et al., 1977] A. P. Dempster, N. M. Laird, & D. B. Rubin, 1977. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Acoustical Society of America 39*, 1–38.

[Dibazar et al., 2002] A. A. Dibazar, S. Narayanan, & T. W. Berger, 2002. Feature analysis for automatic detection of pathological speech. Dans les actes de *Engineering Medicine and Biology Symposium'02*, Volume 1, 182–183.

[Fant, 1960] C. G. Fant, 1960. *The Acoustic Theory of Speech Production. With Calculations based on X-Ray Studies of Russian Articulations.*

[Ferragne & Pellegrino, 2006] E. Ferragne & F. Pellegrino, 2006. Les systèmes vocaliques des dialectes de l'anglais britannique. Dans les actes de *Actes XXVIème Journées d'études sur la Parole*, Dinard, France, 411–414.

[Fredouille, 2000] C. Fredouille, 2000. *Approche Statistique pour la Reconnaissance Automatique du Locuteur : Informations Dynamiques et Normalisation Bayesienne des Vraisemblances*. Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse.

[Fredouille & Bonastre, 1998] C. Fredouille & J.-F. Bonastre, 1998. Use of dynamic information with second order statistical methods in speaker identification. Dans les actes de *Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, Avignon (France), 50–54.

[Fredouille et al., 2005] C. Fredouille, G. Pouchoulin, J.-F. Bonastre., M. Azzarello, A. Giovanni, & A. Ghio, 2005. Application of automatic speaker recognition techniques to pathological voice assessment (dysphonia). Dans les actes de *Interspeech'05*, Lisbon (Portugal).

[Furui, 1981] S. Furui, 1981. Cepstral analysis techniques for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing 29*, 254–272.

[Furui, 1986] S. Furui, 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. Dans les actes de *International Conference on Acoustics Speech and Signal Processing (ICASSP'86)*, Volume 34, 52–59.

[Gauvain & Lee, 1994] J. L. Gauvain & C. H. Lee, 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing 22*, 291–298.

[Gavidia-Ceballos & Hansen, 1996] L. Gavidia-Ceballos & J. Hansen, 1996. Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection. *EEE Transaction on Biomedical Engineering 43 (4)*, 373–383.

[Gentleman & Ihaka, 1997] R. Gentleman & R. Ihaka, 1997. The r project for statistical computing.

[Gerratt et al., 1993] B. R. Gerratt, J. Kreiman, N. Antonnanzas-Barroso, & G. S. Berke, 1993. Comparing internal and external standards in voice quality judgments. *Journal of Speech and Hearing Research 36*, 14–20.

[Giovanni et al., 2006] A. Giovanni, C. Assaiante, A. Galmiche, M. Vaugoyeau, M. Ouaknine, & F. Le Huche, 2006. Forçage vocal et posture : études expérimentales chez le sujet sain. *Revue de laryngologie, d'otologie et de rhinologie 127(5)*, 285–291.

[Giovanni et al., 1995] A. Giovanni, N. Estublier, D. Robert, B. Teston, M. Zanaret, & M. Cannoni, 1995. Evaluation vocale objective des dysphonies par la mesure simultanée de paramètres acoustiques et aérodynamiques à l'aide de l'appareillage EVA. *Annales d'oto-laryngologie et de chirurgie cervico-faciale 112(3)*, 85–90.

[Giovanni et al., 2000] A. Giovanni, C. Heim, D. Demolin, & J. M. Triglia, 2000. Estimated subglottic pressure in normal and dysphonic subjects. *The Annals of Otology, Rhinology and Laryngology 109*, 500–504.

[Giovanni et al., 1999] A. Giovanni, M. Ouaknine, & J. M. Triglia, 1999. Determination of largest lyapounov esponents of vocal signal : Application to unilateral laryngeal paralysis. *Journal of Voice 13 (3)*, 341–354.

[Godino-Llorente et al., 2006] J. I. Godino-Llorente, P. Gomez-Vilda, & M. Blanco-Velasco, 2006. Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters. *EEE Transaction on Biomedical Engineering 53(10)*, 1943–1953.

[Gravier, 2003] G. Gravier, 2003. Spro : a free speech signal processing toolkit (version 4.0.1). http ://gforge.inria.fr/projects/spro.

[Hajaiej et al., 2006] Z. Hajaiej, K. Ouni, & N. Ellouze, 2006. Paramétrisation de la parole basée sur une modélisation des filtres cochléaires : Application au RAP. Dans les actes de *Actes XXVIème Journées d'études sur la Parole*, Dinard, France.

[Hammarberg, 1986] B. Hammarberg, 1986. Perceptual and acoustic analysis of dysphonia. *Dept. of Logopedics and Phoniatrics, Karolinska Institutet*.

[Hammarberg, 2000] B. Hammarberg, 2000. Voice research and clinical needs. *Folia Phoniatrica et Logopaedica 52*, 93–102.

[Hammarberg et al., 1980] B. Hammarberg, B. Tritzell, J. Gauffin, J. Sundberg, & L. Wedin, 1980. Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngol 90*, 441–451.

[Harris, 1978] F. J. Harris, 1978. On the use of windows for harmonic analysis with discrete fourier transform. *Proc. IEEE 66*, 51–83.

[Haton et al., 2006] J.-P. Haton, C. Cerisara, D. Fohr, Y. Laprie, & K. Smaïli, 2006. *Reconnaissance Automatique de la Parole : du signal à son interprétation*. UniverSciences.

[Hermansky, 1990] H. Hermansky, 1990. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America 87*, 1738–1752.

[Hermansky et al., 1996] H. Hermansky, S. Tibrewala, & M. P. M., 1996. Towards ASR on partially corrupted speech. Dans les actes de *International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia (USA).

[Hernandez-Espinosa et al., 2000] C. Hernandez-Espinosa, M. Fernandez-Redondo, P. Gomez-Vilda, J. I. Godino-Llorente, & S. Aguilera-Navarro, 2000. Diagnosis of vocal and voice disorders by speech signal. Dans les actes de *Neural Networks, IEEE-INNS-ENNS International Joint Conference*, Volume 4, 253–258.

[Hirano, 1974] M. Hirano, 1974. Morphological structure of the vocal cord as a vibrator and its variations. *Folia Phoniatrica (Basel) 26*, 89–94.

[Hirano, 1981] M. Hirano, 1981. Psycho-acoustic evaluation of voice : GRBAS scale for evaluating the hoarse voice. *Clinical Examination of voice, Springer Verlag*.

[Hirano, 1989] M. Hirano, 1989. Objective evaluation of the human voice : Clinical aspects. *Folia Phoniatrica et Logopaedica 41*, 89–144.

[Hirano et al., 1968] M. Hirano, Y. Koike, & H. Von Leden, 1968. Maximum phonation time and air usage during phonation. *Folia Phoniatrica et Logopaedica 20*, 185–201.

[Hiraoka et al., 1984] N. Hiraoka, Y. Kitazoe, & H. Ueta, 1984. Harmonic-intensity analysis of normal and hoarse voices. *Journal of the Acoustical Society of America 76*, 1648–1651.

[Holmgren, 1967] C. L. Holmgren, 1967. Physical and psychological correlates of speaker recognition. *Journal of Speech and Hearing Research 10*, 57–66.

[Isshiki et al., 1969] N. Isshiki, H. Okamura, M. Tanabe, & M. Morimoto, 1969. Differential diagnosis of hoarseness. *Folia Phoniatrica et Logopaedica 21*, 9–19.

[Kacha et al., 2005] A. Kacha, F. Grenez, J. Schoentgen, & K. Benmahammed, 2005. Dysphonic speech analysis using generalized variogram. Volume 1, 917–920.

[Kajarekar, 2005] S. S. Kajarekar, 2005. Four weightings and a fusion : A cepstral-svm system for speaker recognition. Dans les actes de *IEEE Speech Recognition and Understanding Workshop*, San Juan, Puerto Rico, 17–22.

[Kasuya et al., 1986] H. Kasuya, S. Ogawa, K. Mashima, & S. Ebihara, 1986. Normalised noise energy as an acoustic measure to evaluate pathologic voice. *Journal of the Acoustical Society of America 80*, 1329–1334.

[Kay Elemetrics Corporation, 1995] Kay Elemetrics Corporation, 1995. The disordered voice database version 1.03.

[Kitzing, 1986] P. Kitzing, 1986. Ltas criteria pertinent to the measurement of voice quality. *Phonetics 14*, 477–482.

[Kreiman et al., 1993] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, & G. S. Berke, 1993. Perceptual evaluation of voice quality : review, tutorial and a framework for future research. *Journal of Speech and Hearing Research 36*, 21–40.

[Kreiman et al., 1992] J. Kreiman, B. R. Gerratt, K. Precoda, & G. S. Berke, 1992. Individual differences in voice quality perception. *Journal of Speech and Hearing Research 35*, 512–520.

[Lamel & Gauvain, 1994] L. F. Lamel & J.-L. Gauvain, 1994. Language identification using phone-based acoustic likelihoods. Dans les actes de *International Conference on Acoustics Speech and Signal Processing (ICASSP'94)*, Adelaide, Australia, 293–296.

[Lamel et al., 1991] L. F. Lamel, J. L. Gauvain, & M. Eskénazi, 1991. BREF, a large vocabulary spoken corpus for french. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech'91)*, Genoa, Italy, 505–508.

[Laver, 1980] J. Laver, 1980. *The Phonetic Description of Voice Quality*. Cambridge.

[Laver et al., 1985] J. Laver, S. Wirz, J. Mackenzie, & S. M. Hiller, 1985. Vocal profile analysis in the description of voice quality. Dans Lawrence (Ed.), *Transactions of the 14th Symposium on the Care of the Professional Voice*, 184–192.

[Le Huche & Allali, 2001a] F. Le Huche & A. Allali, 2001a. *la Voix : Pathologie vocale d'origine fonctionnelle*, Volume 2.

[Le Huche & Allali, 2001b] F. Le Huche & A. Allali, 2001b. *la Voix : Pathologie vocale d'origine organique*, Volume 3.

[Lee et al., 2007] J.-Y. Lee, S. Jeong, & M. Hahn, 2007. Classification of pathological and normal voice based on linear discriminant analysis. *Computer Science 4432*, 382–390.

[Linarès et al., 2007] G. Linarès, D. Massonié, P. Nocéra, & C. Lévy, 2007. A scalable system for embedded large vocabulary continuous speech recognition. *Lecture Notes in Computer Science 4629 LNAI*, 302–308.

[Ljolje, 1994] A. Ljolje, 1994. The importance of cepstral parameter correlations in speech recognition. *Computer Speech and Language 8*, 223–232.

[Lorch & Whurr, 2003] M. Lorch & R. Whurr, 2003. A cross-linguistic study of vocal pathology : Perceptual features of spasmodic dysphonia in French-speaking subjects. *Journal of Multilingual Communication Disorders 1*, 35–52.

[Maguire et al., 2003] C. Maguire, P. de Chazal, R. B. Reilly, & P. Lacy, 2003. Identification of voice pathology using automated speech analysis. Dans les actes de *Third International Workshop on Models and Analysis of Vocal Emission for Biomedical Applications*, Florence, Italy.

[Makhoul, 1975] J. Makhoul, 1975. Linear prediction : A tutorial review. *IEEE-PROC 63(4)*, 561–580.

[Markel & Gray, 1976] J. D. Markel & A. H. Gray, 1976. Linear prediction of speech. *Springer Verlag*.

[McCowan & Sridharan, 2001] I. A. McCowan & S. Sridharan, 2001. Multi-channel sub-band speech recognition. *EURASIP Journal on Applied Signal Processing 1*, 45–52.

[MEEI Database, 2002] MEEI Database, 2002. Disorder database model 4337. *Massachusetts Eye and Ear Infirmary Voice and Speech Lab, Boston, MA*.

[Meunier & Floccia, 1999] C. Meunier & C. Floccia, 1999. Syllabe ou mot : quelle unité permet permet d'identifier les catégories phonétiques ? *Actes des 2ièmes Journées d'Etudes Linguistiques "Syllabes"*, 87–92.

[Miet, 2001] G. Miet, 2001. *Towards wideband speech by narrowband speech bandwidth extension : magic effect or wideband recovery ?* Thèse de Doctorat, University of Maine.

[Millet & Dejonckere, 1998] P. Millet & P. Dejonckere, 1998. What determines the differences in perceptual rating of dysphonia between experienced raters ? *Folia Phoniatrica et Logopaedica 50 (6)*, 305–310.

[Misra et al., 2003] H. Misra, H. Bourlard, & V. Tyagi, 2003. New entropy based combination rules in HMM/ANN multi-stream ASR. Dans les actes de *International Conference on Acoustics Speech and Signal Processing (ICASSP'03)*, Hong Kong.

[Moreau, 1995] N. Moreau, 1995. *Technique de compression des signaux*. Masson, collection technique et scientifique des communications.

[Morris, 1989] J. R. Morris, 1989. V.O.T. and dysarthria : a descriptive study. *Journal of communication disorders 22 (1)*, 23–33.

[Moses, 1954] P. J. Moses, 1954. *The voice of neurosis* (Grune and Stratton ed.). New York.

[Nocera, 1992] P. Nocera, 1992. *Utilisation conjointe de réseaux neuronaux et de connaissances explicites pour le décodage acoustico-phonétique*. Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse.

[Openshaw & Mason, 1994] J. Openshaw & J. Mason, 1994. On the limitations of cepstral features in noise. Dans les actes de *International Conference on Acoustics Speech and Signal Processing (ICASSP'94)*, Volume 2.

[Osgood et al., 1957] C. E. Osgood, G. J. Suci, & P. H. Tannenbaum, 1957. *The Measurement of Meaning*. Urbana.

[Özsancak et al., 2001] C. Özsancak, P. Auzou, M. Jan, & D. Hannequin, 2001. Measurement of voice onset time in dysarthric patients : Methodological considerations. *Folia Phoniatrica et Logopaedica 53*, 48–57.

[Parsa & Jamieson, 2001] V. Parsa & D. G. Jamieson, 2001. Acoustic discrimination of pathological voice : sustained vowels versus continuous speech. *Journal of Speech, Language, and Hearing Research 14*, 327–339.

[Pelecanos & Sridharan, 2001] J. Pelecanos & S. Sridharan, 2001. Feature warping for robust speaker verification. Dans les actes de *2001 : a Speaker Odyssey. The Speaker Recognition Workshop*, Crete, Greece, 213–218.

[Piccirillo et al., 1998a] J. Piccirillo, P. Colin, F. Dennis, & J. Frederickson, 1998a. Assessment of two objective voice function indices. *The Annals of Otology, Rhinology and Laryngology 107(5)*, 396–400.

[Piccirillo et al., 1998b] J. Piccirillo, P. Colin, F. Dennis, & J. Frederickson, 1998b. Multivariate analysis of objective vocal function. *The Annals of Otology, Rhinology and Laryngology 107*, 107–112.

[Picone, 1993] J. W. Picone, 1993. Signal modeling techniques in speech recognition. *IEEE 81(9)*, 1215–1247.

[Pouchoulin et al., 2007] G. Pouchoulin, C. Fredouille, J.-F. Bonastre, A. Ghio, & A. Giovanni, 2007. Frequency study for the characterization of the dysphonic voices. Dans les actes de *Interspeech'07*, Antwerp, Belgium.

[Qi et al., 1995] Y. Qi, B. Weinberg, N. Bi, & W. K. Hess, 1995. Minimizing the effect of period determination of the computation of amplitude perturbation in voice. Dans les actes de *Workshop on acoustic voice analysis*, Iowa City. IA : National Center for voice and speech.

[Qin et al., 2003] Y. Qin, S. Vaseghi, D. Rentzos, H. Ching-Hsiang, & E. Turajlic, 2003. Analysis of acoustic correlates of british, australian and american accents. *Automatic Speech Recognition and Understanding, IEEE, ASRU 2003*, 345–350.

[Rabiner, 1989] L. R. Rabiner, 1989. A tutorial on hidden markov models and selected applications in speech recognition. *IEEE Transactions on Speech and Audio Processing 77(2)*, 257–285.

[Rabiner & Schafer, 1978] L. R. Rabiner & R. W. Schafer, 1978. Digital processing of speech signals. *Prentice-Hall, Signal Processing Series*.

[Rabinov et al., 1995] R. Rabinov, J. Kreiman, B. R. Gerratt, & S. Bielamowicz, 1995. Comparing reliability of perceptual ratings of roughness and acoustic measures of jitter. *Journal of Speech and Hearing Research 38*, 26–32.

[Rahman & FairHurst, 1998] A. Rahman & M. FairHurst, 1998. A novel confidence-based framework for multiple expert decision fusion. Dans les actes de *BMVC'98*, Southampton University, Royaume-Uni.

[Revis, 2004] J. Revis, 2004. *L'analyse perceptive des dysphonies : approche phonétique de l'évaluation vocale*. Thèse de Doctorat, Université de la Méditerranée.

[Revis et al., 2000] J. Revis, S. Barberis, A. Giovanni, & J. M. Triglia, 2000. Définition d'une mesure temporelle de l'attaque vocale. *Revue de laryngologie, d'otologie et de rhinologie*.

[Revis et al., 2006] J. Revis, A. Ghio, & A. Giovanni, 2006. Phonetic labeling of dysphonia : a new perspective in perceptual voice analysis. Dans les actes de *7th International Conference Advances in Quantitative Laryngology, Voice and Speech Research*.

[Revis et al., 2002] J. Revis, A. Giovanni, & J. M. Triglia, 2002. Influence de l'attaque sur l'analyse perceptive des dysphonies. *Folia Phoniatrica et Logopaedica 54*, 19–25.

[Revis et al., 1999] J. Revis, A. Giovanni, F. L. Wuyts, & J. M. Triglia, 1999. Comparison of different voice samples for perceptual analysis. *Folia Phoniatrica et Logopaedica*, 108–116.

[Reynolds, 1992] D. A. Reynolds, 1992. *A gaussian mixture modeling approach to text-independent speaker identification*. Thèse de Doctorat, Georgia Institute of Technology, USA.

[Reynolds, 1994] D. A. Reynolds, 1994. Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing 2(4)*, 639–643.

[Reynolds, 1995] D. A. Reynolds, 1995. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication 171–2*, 91–108.

[Reynolds, 1997] D. A. Reynolds, 1997. Comparison of background normalization methods for text-independent speaker verification. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech'97)*, Volume 19.

[Reynolds, 2003] D. A. Reynolds, 2003. Channel robust speaker verification via feature mapping. Dans les actes de *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume 2, Hong Kong, 53–56.

[Reynolds et al., 2000] D. A. Reynolds, T. F. Quatieri, & R. B. Dunn, 2000. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop 10 (1-3)*, 19–41.

[Ritchings et al., 2002] R. T. Ritchings, M. McGillion, & C. J. Moore, 2002. Pathological voice quality assessment using artificial neural networks. *Medical engineering and physics 24(7-8)*, 561–564.

[Roy et al., 2000] N. Roy, D. M. Bless, & D. Heisey, 2000. Personality and voice disorders : a multitrait-multidisorder analysis. *Journal of Voice 14*, 521–548.

[Rumelhart & McClelland, 1986] D. E. Rumelhart & J. L. McClelland, 1986. *Parallel Distributed Processing*, Volume 1. MIT Press.

[Saenz-Lechon et al., 2006] N. Saenz-Lechon, J. I. Godino-Llorente, V. Osma-Ruiza, & P. Gomez-Vilda, 2006. Methodological issues in the development of automatic

systems for voice pathology detection. *Journal of Biomedical Signal Processing and Control, Elsevier*.

[Sambur, 1975] M. R. Sambur, 1975. Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing 23(2)*, 176–182.

[Scheffer, 2006] N. Scheffer, 2006. *Structuration de l'espace acoustique par le modèle générique pour la vérification du locuteur*. Thèse de Doctorat, thèse de Doctorat de l'Université d'Avignon.

[Schoentgen & Bucella, 1997] J. Schoentgen & F. Bucella, 1997. Acoustic analysis of dysphonic voices : descriptors and methods. Dans les actes de *LARYNX'97*, 37–46.

[Schutte & Seidner, 1983] H. K. Schutte & W. Seidner, 1983. Recommendation by the Union of European Phoniatricians (UEP) : standardizing voice area measurement/phonetography. *Folia Phoniatrica et Logopaedica 35*, 286–288.

[Smitheran & Hixon, 1981] J. Smitheran & T. A. Hixon, 1981. A clinical method for estimating laryngeal airway resistance during vowel production. *Journal of Speech and Hearing Disorders 46*, 138–148.

[Soong & Rosenberg, 1988] F. K. Soong & A. E. Rosenberg, 1988. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing 36(6)*, 871–879.

[Teston, 2004] B. Teston, 2004. *L'évaluation instrumentale des dysphonies. Etat actuel et perspectives*. ISBN 2-914513-62-3. In Giovanni A. (ed) Le bilan d'une dysphonie. Marseille : Solal, 105–169.

[Teston & Galindo, 1995] B. Teston & B. Galindo, 1995. A diagnosis of rehabilitation aid workstation for speech and voice pathologies. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech'95)*, Madrid, Spain, 1883–1886.

[Tridgell et al., 1992] A. Tridgell, B. Millar, & K.-A. Do, 1992. Alternative pre-processing techniques for discrete hidden markov model phoneme recognition. Dans les actes de *International Conference on Spoken Language Processing (ICSLP'92)*, 631–634.

[Tubach, 1989] J. Tubach, 1989. *La parole et son traitement automatique* (Masson, collection technique et scientifique des télécommunications ed.).

[Vapnik, 1998] V. Vapnik, 1998. *Statistical learning theory*.

[Viterbi, 1967] A. J. Viterbi, 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory 13(2)*, 260–269.

[Voiers, 1964] W. Voiers, 1964. Perceptual bases of speaker identity. *Journal of the Acoustical Society of America 36(6)*, 1065–1073.

[Wang & Jo, 2006] J. Wang & C. Jo, 2006. Performance of gaussian mixture models as a classifier for pathological voice. Dans les actes de *11th Autralian International Conference on Speech Science and Technology*.

[Wang et al., 1993] X. Wang, L. ten Bosch, & L. Pols, 1993. Impact of dimensionality and correlation of observation vectors in hmm-based speech recognition. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech'93)*, Berlin, Germany, 1583–1586.

[Wester, 1998] M. Wester, 1998. Automatic classification of voice quality : Comparing regression models and hidden markov models. Dans les actes de *VOICEDATA98, Symposium on Databases in Voice Quality Research and Education*, Utrecht, 92–97.

[Wirz & Beck, 1995] S. Wirz & J. M. Beck, 1995. *Assessment of voice quality : The vocal profiles analysis scheme*. Whurr, London : S. Wirz.

[Woisard et al., 2004] V. Woisard, S. Bodin, & M. Puech, 2004. The Voice Handicap Index : impact of the translation in French on the validation. *Revue de laryngologie, d'otologie et de rhinologie 125(5)*, 307–312.

[Wolfe et al., 1995] V. Wolfe, J. Fitch, & R. Cornell, 1995. Acoustic prediction of severity in commonly occurring voice problems. *Journal of Speech, Language, and Hearing Research 38(2)*, 273–279.

[Wolfe et al., 2000] V. Wolfe, D. Martin, & C. Pamer, 2000. Perception of dysphonic voice quality by naive listeners. *Journal of Speech, Language, and Hearing Research 43*, 697–705.

[Wuyts et al., 2000] F. L. Wuyts, M. S. De Bodt, G. Molenberghs, M. Remacle, L. Heylen, B. Millet, K. Van Lierde, J. Raes, & P. H. Van de Heyning, 2000. The dysphonia severity index : an objective measure of vocal quality based on a multiparameter approach. *Journal of Speech, Language, and Hearing Research 43(3)*, 796–809.

[Yanagihara, 1967] N. Yanagihara, 1967. Significance of harmonic changes and noise components in hoarseness. *Journal of Speech, Language, and Hearing Research 10*, 531–541.

[Yang et al., 2005] C. Yang, F. K. Soong, & T. Lee, 2005. Static and dynamic spectral features : Their noise robustness and optimal weights for ASR. Dans les actes de *European Conference on Speech Communication and Technology*, Philadelphia, U.S.A.

[Yi & Loizou, 2008] H. Yi & P. C. Loizou, 2008. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing 16 (1)*, 229–238.

[Yu et al., 2007] P. Yu, R. Garrel, R. Nicollas, M. Ouaknine, & A. Giovanni, 2007. Objective voice analysis in dysphonic patients. New data including non linear measurements. *Folia Phoniatrica et Logopaedica 59*, 20–30.

[Yu et al., 2001] P. Yu, M. Ouakine, J. Revis, & A. Giovanni, 2001. Objective voice analysis for dysphonic patients : a multiparametric protocol including acoustic and aerodynamic measurements. *Journal of Voice 15*, 529–542.

[Yu et al., 2000] P. Yu, M. Ouaknine, & A. Giovanni, 2000. Intérêt clinique du calcul des coefficients de lyapunov pour l'analyse objective des dysphonies. *Revue de laryngologie, d'otologie et de rhinologie 121(5)*, 301–305.

[Yu et al., 2002] P. Yu, J. Revis, F. L. Wuyts, M. Zanaret, & A. Giovanni, 2002. Correlations of instrumental voice evaluation with perceptual analysis using a modified visual analogic scale. *Folia Phoniatrica et Logopaedica 54(6)*, 274–281.

[Yumoto & Gould, 1982] E. Yumoto & W. Gould, 1982. Harmonics to noise ratio as an index of the degree of hoartheness. *Journal of the Acoustical Society of America 71(6)*, 1544–1550.