Modélisation Statistique et Informations Pertinentes pour la Caractérisation des Voix Pathologiques (Dysphonies)

G. Pouchoulin[1], *C. Fredouille*[1], *J.-F. Bonastre*[1], *A. Ghio*[2], *M. Azzarello*[3]

[1]LIA, Avignon (France)
[2]LPL-CNRS, Aix en Provence (France)
[3]LAPC, Marseille (France)

ABSTRACT

This paper investigates the class of information relevant for the task of automatic classification of pathological voices. By using a GMM-based classification system (derived from the Automatic Speaker Recognition domain), the focus was made on three main classes of information : energetic, voiced, and phonetic information. Experiments made on a pathological corpus (dysphonia) have shown that phonetic information is particularly interesting in this context since it permits to refine the selection of the relevant information by looking at phonem- or phonem class-level (e.g. nasal vowels).

1. INTRODUCTION

Dans le domaine médical, l'évaluation de la qualité des voix pathologiques est un sujet sensible, au centre de nombreuses études dans des domaines multi-disciplinaires. Dans le cas des voix dysphoniques[1][12][8], sur lesquelles se concentre cet article, le dysfonctionnement vocal peut être évalué suivant deux approches, que sont l'analyse perceptive et l'analyse objective.

L'analyse perceptive ou auditive est la plus utilisée par le corps médical. Elle consiste à caractériser la qualité vocale par une simple écoute attentive ; Elle peut être effectuée par un jury d'experts afin d'augmenter la fiabilité de l'analyse en raison de sa subjectivité intrinsèque. Les inconvénients majeurs de cette approche sont le manque de fiabilité dû à différents facteurs comme la subjectivité des experts et la variabilité intra et inter-individuelle ainsi que le coût humain non négligeable dans le cas d'un jury d'écoute (réunions périodiques de plusieurs experts, durée des séances d'écoute, ...).

L'analyse objective ou instrumentale (comme le système EVA^{TM} [11], Evaluation Vocale Assistée - SQLab) s'appuie sur l'acquisition de nombreuses données quantitatives (mesures acoustiques, aérodynamiques et physiologiques, ...) au travers d'appareillages médicaux. Elle offre une approche complémentaire à l'examen laryngoscopique et à l'interrogatoire du patient effectués par les praticiens. Dans [13], 86 % de concordance entre l'analyse perceptive et objective sont atteints en utilisant 10 paramètres acoustiques et aérodynamiques (F0, intensité, jitter, coefficient de Lyapunov, rapport signal sur bruit, débit d'air buccal, pression sous-glottique estimée, étendue vocale, temps maximum de phonation). L'approche instrumentale offre donc des résultats très acceptables mais encore insuffisants pour son adoption dans une pratique clinique quotidienne. Par ailleurs, l'acquisition des mesures sur le /a/ tenu (généralement utilisée dans ce cas) reste controversée dans la littérature [9] car elle tend à sous-estimer la dysphonie. Des mesures effectuées sur de la parole spontanée ou conversationnelle permettraient de prendre en compte par exemple, les phénomènes vocaux de l'attaque, reconnus comme pertinents dans l'évaluation des dysphonies.

En résumé, à ce jour, aucune approche d'évaluation de la qualité de la voix ne semble répondre à l'attente des médecins et orthophonistes, même si l'analyse perceptive reste incontournable dans la pratique quotidienne. En outre, cette dernière reste la seule référence pour les méthodes objectives. En 2005, une étude préliminaire [4] a été menée par le LIA en vue d'adapter des techniques de reconnaissance automatique du locuteur (RAL) à l'évaluation de voix pathologiques (voix dysphoniques). Le but de cette étude était de proposer une méthode instrumentale mieux adaptée au suivi de la pathologie des patients : facilité et rapidité d'utilisation, non contraignante pour le patient et accessible pour les cliniciens. Comparée à des méthodes instrumentales classiques, les originalités de cette approche basée sur une modélisation statistique, reposent sur :

. sa capacité à analyser de la parole continue (et non des voyelles tenues) proche de l'élocution naturelle ;
. sa capacité à traiter de grands corpus, permettant de mener des études à grandes échelles et d'obtenir des informations statistiques significatives ;
. une analyse acoustique, simple et automatique, permettant une simplicité d'instrumentation et un faible coût humain.

Le système conçu pour cette tâche particulière s'appuie sur l'approche à base de GMM, état de l'art pour la RAL. Il est issu des outils de RAL, disponibles en « version libre » (LIA_SpkDet et ALIZE) et développés au LIA.

Dans la continuité de ce travail, nous proposons ici une première étude sur l'extraction des informations utiles à la caractérisation des voix pathologiques. Nous nous intéresserons plus particulièrement à trois classes d'informations : informations « énergétiques », « phonétiques » et « voisées ».

Le corpus dysphonique utilisé dans cette étude ainsi que le système de classification automatique sont décrits en sections 2 et 3. La sélection des différentes classes d'information est détaillée en section 4, puis évaluée dans un contexte expérimental en section 5. Finalement, la section 6 fournit une conclusion à ce travail.

[1]dysfonctionnement anatomique se traduisant par une dégradation de la voix

2. EVALUATION DES VOIX DYSPHONIQUES

Dans cette étude, nous nous intéressons aux dysphonies fonctionnelles (nodules, polypes, oedèmes, kystes...) classées selon le paramètre G de l'échelle d'évaluation GRBAS de Hirano [6]. Le corpus mis à disposition par le LAPC (Hôpital de la Timone Marseille) est constitué de 80 échantillons de voix féminines correspondant à 20 sujets témoins et 60 patientes dysphoniques, âgées de 17 à 50 ans (moyenne de 32.2 ans). Chaque sujet a été enregistré sur la lecture d'un paragraphe de « La chèvre de Monsieur Seguin » d'Alphonse Daudet. Les enregistrements ont été évalués selon le grade global G de dysphonie de l'échelle GRBAS par un jury d'experts ; les décisions ont été prises par consensus afin d'en limiter la variabilité inter-auditeur et en une seule séance afin d'en limiter la variabilité intra-auditeur. L'ensemble du corpus étiqueté se présente donc de la manière suivante : 20 voix normales G0, 20 voix avec dysphonie légère G1, 20 voix avec dysphonie moyenne G2, 20 voix avec dysphonie sévère G3.

3. SYSTÈME DE CLASSIFICATION DES VOIX DYSPHONIQUES

Le principe retenu consiste en l'adaptation d'un système classique de RAL à la classification de voix dysphoniques. Trois phases sont nécessaires et sont décrites dans les sections suivantes.

3.1. Paramétrisation

Le signal de parole est paramétrisé comme suit : chaque signal est caractérisé par 16 coefficients cepstraux (MFCC) obtenus à partir de 24 coefficients de banc de filtres répartis sur une échelle de MEL (fenêtre de type Hamming de 20ms avec un pas de 10ms). Les dérivées premières des coefficients MFCC (Δ) sont ajoutées aux vecteurs de paramètres qui sont finalement normalisés pour obtenir une distribution de moyenne 0 et variance 1 (les moyennes et les variances sont estimées sur les portions jugées utiles du signal dont la sélection est décrite en section 4).

3.2. Modélisation

En RAL, l'état de l'art repose sur une modélisation statistique (GMM : Gaussian Mixture Model)[1]. Un GMM *X* est une somme pondérée de M distributions gaussiennes multidimensionnelles, chacune caractérisée par un vecteur moyen $\overline{x}$ de dimension *d*, une matrice de covariance Σ de dimension *d* x *d* et un poids p de la gaussienne au sein de la mixture. Durant la phase d'apprentissage, les paramètres ($\overline{x}$, Σ, p) sont estimés par l'algorithme EM/ML[2].
Classiquement, deux phases d'apprentissage sont nécessaires en RAL pour pallier le manque de données d'apprentissage disponibles pour chaque locuteur [1] :

. apprentissage d'un modèle générique (aussi appelé « modèle du monde ») estimé par l'algorithme EM/ML sur une grande quantité de données (population de locuteurs) ;
. apprentissage du modèle locuteur dérivé du modèle du monde par application des techniques d'adaptation (MAP, Maximum a Posteriori) [10].

Dans le contexte pathologique, un modèle ne correspond plus à un locuteur donné 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 est appris en utilisant l'ensemble des locuteurs de même grade. 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 (mise en oeuvre de la technique leave_x_out).
Les modèles GMM représentant les grades pathologiques sont établis comme suit :

. un modèle GMM générique est d'abord estimé par l'algorithme EM/ML sur un corpus français composé de 76 enregistrements de 2 mn chacun de voix exclusivement féminines. Cette population est extraite du corpus BREF [7] entièrement disjoint du corpus dysphonique
. les modèles de grade sont ensuite dérivés du modèle GMM générique suivant la technique d'adaptation MAP [10]. Seules les moyennes sont adaptées.

Tous les modèles GMM se composent de 128 composantes gaussiennes à matrices de covariance diagonales.

3.3. Classification

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 : $L(y_t|X) = \sum_{i=1}^{M} p_i \; L_i(y_t)$
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.
La **décision** correspondra au grade *g* du modèle G_g sur lequel la plus grande vraisemblance est obtenue. Cette définition de la décision est proche de celle de la tâche d'identification automatique du locuteur. On dira que le système a classé la voix du locuteur *Y* dans le grade *g*.

4. SÉLECTION DES INFORMATIONS PERTINENTES

Dans ce papier, nous nous intéressons aux informations pertinentes pour la caractérisation des voix dysphoniques. Trois niveaux d'extraction de l'information utile sont étudiés :

. « information énergétique » : le signal de parole est nettoyé des trames de silence (système de détection « parole/non parole » du LIA basé sur une modélisation statistique de l'énergie) ;
. « information phonétique » : extraite d'un alignement phonétique automatique contraint par le texte (système d'alignement du LIA basé sur un décodage Viterbi, à partir d'un lexique de mots et leurs variantes phonologiques - 38 phonèmes du français) ;
. « information voisée » : extraction des sons de parole voisés par analyse de la fréquence fondamentale F0 (logiciel PRAAT [2] sur l'intervalle de fréquence [75,600] Hz).

Une fois extraite, l'information jugée utile est utilisée lors des phases de normalisation des signaux de parole, de modélisation des grades et de prise de décision.
Les durées moyennes des différentes classes d'informations extraites du corpus dysphonique, décrites ci-dessus, sont données dans la figure 1.

5. EXPÉRIENCES

Les expériences ont été réalisées après « adaptation » du système de RAL du LIA. Ce système (appelé LIA_SpkDet) repose entièrement sur la plateforme libre ALIZE [3] conçue et réalisée dans le cadre du programme

[2] Expectation-Maximization/Maximum Likelihood

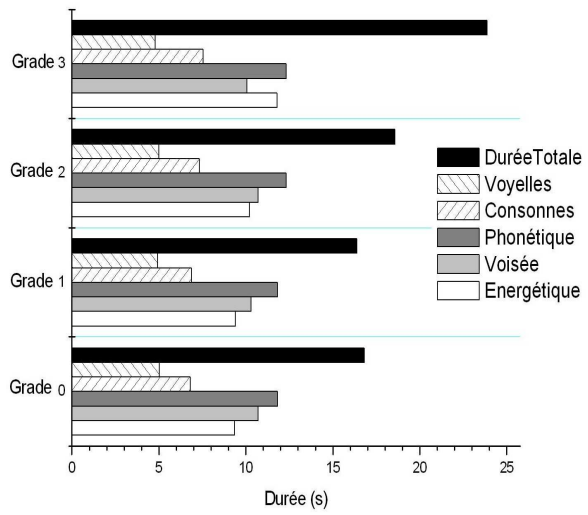

FIG. 1: Durées moyennes par grade de différentes classes d'informations extraites du corpus dysphonique.

	Grade 0	Grade 1	Grade 2	Grade 3	Total
Information	**% succès** (nb/20)	**% succès** (nb/20)	**% succès** (nb/20)	**% succès** (nb/20)	**% succès** (nb/80)
Energétique	95,0 (19)	70,0 (14)	50,0 (10)	60,0 (12)	68,75 (55)
Voisement	95,0 (19)	65,0 (13)	50,0 (10)	75,0 (15)	71,25 (57)
Phonétique	95,0 (19)	60,0 (12)	55,0 (11)	75,0 (15)	71,25 (57)

TAB. 1: Résultats de la classification 4-Grades suivant différentes classes d'informations extraites

Technolangue. LIA_SpkDet est également distribué en logiciel libre.

5.1. Protocole expérimental

Il s'agit de classer une voix suivant les 4 niveaux du grade général de l'échelle GRBAS. Par conséquent, 4 modèles de grade G_g sont à estimer avec $g \in \{0, 1, 2, 3\}$.
Lors de la phase de test, la mise en oeuvre de la technique leave_x_out (en vue de séparer les données de test et d'apprentissage) permet de comparer chaque voix y_t de grade g avec :
- 1 modèle G_g appris à partir des 19 voix restantes de grade g ($y_t \notin$ aux 19 voix) ;
- 3 x 20 modèles $G_{\overline{g}}$ appris chacun à partir de 19 voix de grade $\overline{g}$ avec $\overline{g} \in \{0, 1, 2, 3\} - \{g\}$.

A l'issu 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.
Note : Le même nombre de voix (19) est utilisé pour l'ensemble des modèles de grade.

5.2. Résultats

Classes d'informations utiles
Le tableau 1 donne les résultats des différentes classes d'informations sélectionnées en vue de la caractérisation des voix pathologiques. Les expériences relatives aux informations « voisées » et « phonétiques » obtiennent le meilleur résultat (71,25 % de réussite). Quelque soit la classe d'information, on peut remarquer que le grade 0 est le mieux reconnu (95,0 %) et que la confusion provient

Classification	Grade 0	Grade 1	Grade 2	Grade 3
Locuteurs de Grade 0	**19**	1	0	0
Locuteurs de Grade 1	2	**12**	4	2
Locuteurs de Grade 2	2	5	**11**	2
Locuteurs de Grade 3	0	1	4	**15**

TAB. 2: Matrice de confusion - Phonétique

Classification	Grade 0	Grade 1	Grade 2	Grade 3
Locuteurs de Grade 0	**19**	1	0	0
Locuteurs de Grade 1	3	**13**	2	2
Locuteurs de Grade 2	3	5	**10**	2
Locuteurs de Grade 3	0	1	4	**15**

TAB. 3: Matrice de confusion - Voisement

principalement des grades 1 et 2 (voir matrices de confusion Tab. 2 et 3). Concernant les informations « énergétiques », à durée équivalente avec les informations « phonétiques », la classification en grade 3 est dégradée, pouvant montrer une faiblesse du détecteur parole/non parole sur des voix dysphoniques très sévères. Les informations « voisées » obtiennent un taux de réussite équivalent aux informations « phonétiques » alors que leurs durées sont plus courtes (figure 1) de 10 à 20 % suivant les grades. Cette observation tend à démontrer que les informations non voisées présentes dans la classe des informations « phonétiques » sont moins pertinentes dans ce contexte.
Note : Tous les résultats fournis dans ce papier sont issus du classifieur GMM et doivent être interprétés d'un point de vue statistique.

Analyse phonétique
A partir de l'expérience sur les informations « phonétiques », une analyse a été réalisée sur la pertinence des différentes classes de phonèmes dans la classification des voix pathologiques. Il est à noter que cette analyse porte uniquement sur le pouvoir décisionnel de ces différentes classes (cette catégorisation n'est pas utilisée lors des phases de normalisation des paramètres ni d'apprentissage pour lesquelles l'ensemble des informations « phonétiques » ont été utilisées). Le tableau 4 présente une première analyse des décisions suivant une catégorisation détaillée « Consonnes/Voyelles ». Il est intéressant de remarquer :
. la pertinence des différentes classes pour le grade 0 (85% en moyenne de bonne classification) ;
. des différences marquées entre les classes pour le grade 3 (de 75 à 85% pour les voyelles non nasales, les semi-consonnes et les occlusives contre de 45 à 60% pour les voyelles nasales, les consonnes nasales, liquides et fricatives) ;
. pour les grades 1 et 2, les consonnes occlusives et les voyelles nasales apportent peu d'information (de 35% à 45% uniquement de bonne classification). En revanche, les semi-consonnes, liquides et fricatives réagissent plus favorablement au grade 2 (de 60% à 65% de bonne classification) ; les voyelles non nasales et consonnes nasales plus favorablement au grade 1 (60%) ;
. les semi-consonnes, malgré leurs courtes durées (de 0.39s à 0.44s), obtiennent un taux global de réussite de 67.5% contre 71.25% sur la totalité des informations « phonétiques ».

Un premier approfondissement de cette étude montre qu'au sein d'une même classe de phonèmes, les compor-

	Grade 0	Grade 1	Grade 2	Grade 3	Total
Classe phonét.	**% succès** (nb/20) dur. moy.	**% succès** (nb/20) dur. moy.	**% succès** (nb/20) dur. moy.	**% succès** (nb/20) dur. moy.	**% succès** (nb/80)
Voyelle	**95,0 (19)**	**60,0 (12)**	**35,0 (7)**	**70,0 (14)**	**65,00 (52)**
Voyelle non nasale	95,0 (19) 4,14 (s)	60,0 (12) 4,09 (s)	45,0 (9) 4,15 (s)	75,0 (15) 4,12 (s)	68,75 (55)
Voyelle nasale	95,0 (19) 0,89 (s)	40,0 (8) 0,85 (s)	35,0 (7) 0,85 (s)	45,0 (9) 0,66 (s)	53,75 (43)
Consonne	90,0 (18)	50,0 (10)	60,0 (12)	65,0 (13)	66,25 (53)
Semi -consonne	90,0 (18) 0,39 (s)	35,0 (7) 0,43 (s)	60,0 (12) 0,43 (s)	85,0 (17) 0,44 (s)	67,50 (54)
Consonne liquide	80,0 (16) 1,68 (s)	45,0 (9) 1,66 (s)	60,0 (12) 1,72 (s)	60,0 (12) 1,86 (s)	61,25 (49)
Consonne nasale	75,0 (15) 1,40 (s)	60,0 (12) 1,41 (s)	50,0 (10) 1,57 (s)	50,0 (10) 1,46 (s)	58,75 (47)
Consonne fricative	90,0 (18) 1,53 (s)	40,0 (8) 1,56 (s)	65,0 (13) 1,64 (s)	45,0 (9) 1,77 (s)	60,00 (48)
Consonne occlusive	85,0 (17) 2,01 (s)	45,0 (9) 2,06 (s)	45,0 (9) 2,18 (s)	85,0 (17) 2,25 (s)	65,00 (52)

TAB. 4: Analyse par classe phonétique : % de réussite et durée moyenne par classe et par grade

Phonèmes	Grade 0	Grade 1	Grade 2	Grade 3
[ɑ̃]	**68,3 (41/60)**	**28,3 (17/60)**	**41,7 (25/60)**	**36,7 (22/60)**
[ɛ̃]	**47,5 (19/40)**	**35,0 (14/40)**	**35,0 (14/40)**	**25,0 (10/40)**
[ɔ̃]	**85,0 (51/60)**	**28,3 (17/60)**	**28,3 (17/60)**	**51,7 (31/60)**
[œ̃]	**75,0 (15/20)**	**30,0 (6/20)**	**25,0 (5/20)**	**70,0 (14/20)**

TAB. 5: Résultats de la classification 4-Grades suivant les voyelles nasales (décision au niveau du phonème)

tements peuvent être très différents suivant les grades. Par exemple, le tableau 5 présente les résultats de classification des voyelles nasales.
Note : Dans ce cas, la décision est prise au niveau du phonème uniquement.
On peut observer, par exemple, pour le grade 3 des taux de réussite variant de 25% pour le phonème [ɛ̃] à 70% pour le phonème [œ̃].
Ce même comportement a été observé sur d'autres classes de phonèmes.

6. CONCLUSION

Dans ce papier, nous proposons une analyse de l'information pertinente, véhiculée par le signal de parole, pour la caractérisation des voix pathologiques. Trois niveaux d'informations ont été testés : informations « énergétiques » , « phonétiques » et « voisées » . Cette étude a été menée en utilisant un système de classification de voix pathologiques dérivé du domaine de la RAL et basé sur une approche statistique (GMM).
D'un point de vue expérimental, les informations « phonétiques » semblent les plus intéressantes dans cette étude, au sens où elles permettent d'affiner la sélection de l'information utile. En effet, l'étude des différentes classes phonétiques a montré l'influence de certains phonèmes ou classes de phonèmes dans la tâche de classification des voix pathologiques (68.75% de réussite pour les voyelles non nasales contre 53.75% pour les voyelles nasales). Néanmoins, nous avons montré également que des comportements différents peuvent intervenir au sein d'une même classe. Par ailleurs, il est intéressant de constater que certaines classes de phonèmes sont plus discriminantes que d'autres pour les grades 1 et 2, qui restent problématiques dans le cadre d'une décision globale (totalité de l'information présente). Il est à noter cependant que la petite taille du corpus (80 voix dont 60 dysphoniques) est à prendre en compte dans la validité de ces résultats ainsi que les caractéristiques intrinsèques du système de classification utilisé.
Il serait à présent intéressant 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.

REMERCIEMENTS

Les auteurs tiennent à remercier le Laboratoire Audio-Phonologie Clinique (LAPC - Hôpital La Timone-Marseille) d'avoir mis à leur disposition le corpus dysphonique utilisé dans cette étude.

RÉFÉRENCES

[1] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, and D. A. Reynolds. A tutorial on text-independent speaker verification. In *EURASIP Journal on Applied Signal Processing*, volume 39, pages 430–451, 2004.

[2] P. Boersma and D. Weenink. Praat : doing phonetics by computer. http ://www.praat.org/.

[3] J.-F. Bonastre, F. Wils, and S. Meignier. Alize, a free toolkit for speaker recognition. In *ICASSP-05, Philadelphia, USA*, volume 39, pages 430–451, 2005.

[4] C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni, and A. Ghio. Application of automatic speaker recognition techniques to pathological voice assessment (dysphonia). In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 05)*, 2005.

[5] G. Gravier. Spro : a free speech signal processing toolkit. http ://www.irisa.fr/metiss/guig/spro/.

[6] M. Hirano. Psycho-acoustic evaluation of voice : Grbas scale for evaluating the hoarse voice. In *Clinical Examination of voice, Springer Verlag*, 1981.

[7] L. Lamel, J. Gauvain, and L. Eskénazi. Bref, a large vocabulary spoken corpus for french. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 99)*, 1991.

[8] J. Revis. L'analyse perceptive des dysphonies : approche phonétique de l'évaluation vocale. In *Phd thesis, Université de la Méditerranée*, 2004.

[9] J. Revis, A. Giovanni, FL. Wuyts, and J.M. Triglia. Comparaison of different voice samples for perceptual analysis. In *Folia Phoniatr Logop*, pages 108–116, 1999.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models, digital signal processing (dsp). In *a review journal - Special issue on NIST 1999 speaker recognition workshop 10 (1-3)*, pages 19–41, 2000.

[11] B. Teston and B. Galindo. A diagnosis of rehabilitation aid workstation for speech and voice pathologies. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech 95)*, pages 1883–1886, 1995.

[12] F. L. Wuyts, M. S. De Bodt, G. Molenberghs, M. Remacle, L. Heylen, B. Millet, K. Van Lierde, J. Raes, and P. H. Van de Heyning. The dysphonia severity index : an objective measure of vocal quality based on a multiparameter approach. In *Journal of Speech, Language, and Hearing Research 43*, pages 796–809, 2000.

[13] P. Yu, M. Ouakine, J. Revis, and A. Giovanni. Objective voice analysis for dysphonic patients : a multiparametric protocol including acoustic and aerodynamic measurements. In *Journal Voice 15*, pages 529–542, 2001.