CHARACTERIZATION OF THE PATHOLOGICAL VOICES (DYSPHONIA) IN THE FREQUENCY SPACE

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

[1]LIA, Avignon (France), [2]LPL-CNRS, Aix en Provence (France), [3]LAPEC, Marseille (France)
(gilles.pouchoulin,corinne.fredouille,jfb)@univ-avignon.fr,alain.ghio@lpl.univ-aix.fr

ABSTRACT

This paper is related to dysphonic voice assessment. It aims at characterizing dysphonia in the frequency domain. In this context, a GMM-based automatic classification system is coupled with a frequency subband architecture in order to investigate which frequency bands are relevant for dysphonia characterization.
Through various experiments, the low frequencies [0-3000]Hz tend to be more interesting for dysphonia discrimination compared with higher frequencies.

Keywords: dysphonia, pathological voice and speech, automatic speaker recognition

1. INTRODUCTION

In the medical domain, assessment of the pathological voice quality is a sensitive topic, involving multi-disciplinary domains. Concerning the dysphonic voices[14][11], which this article focuses on, vocal dysfunction can be assessed following two approaches: the perceptual judgment and the objective measurement-based analysis.
The perceptual judgment is the most widely used by clinicians. This method consists in qualifying and quantifying the vocal dysfunction by listening to patients' speech production; It can be performed by an expert jury to increase the reliability of the analysis because of its intrinsec subjectivity. The objective measurement-based analysis (such as the EVA^{TM} [12] system, Computerised Vocal Assessment - SQLab) consists in acquiring numerous quantitative data (like acoustic, aerodynamic and physiological measures) through simple computer-based systems or more complex medical equipments. It offers an additional approach to the clinical examination of the larynx and to the questioning of patients by clinicians.
A few studies have been dedicated to the acoustic analysis of dysphonia effects on the speech signal[13][7][9]. Indeed, if an expert is able to assess a dysphonic voice according to a quality scale like the Hirano's GRBAS scale [5], it is more difficult for him/her to bring acoustic justification for his/her choice.
As dysphonia is essentially related to the vocal source, most of the studies have focused on parameters directly linked to this vibrator (FO stability, intensity, harmonics to noise ratio, ...). Other studies are related on the global timbre of the voice, assuming that the acoustic characteristics of dysphonia are distributed uniformly on the whole spectrum. One of the originality of our study is to investigate the characteristics of dysphonia in the frequency domain, especially by studying relating phenomena through a subband analysis. The second originality is to rely on an automatic system dedicated to the dysphonic voice classification and derived from the Automatic Speaker Recognition technology[4]. This system will be coupled with a subband architecture, which should permit to analyse the relevance of different frequency subbands for the characterization of the dysphonic voices.

2. CORPUS

The corpus used in this study is composed of speech excerpts pronounced by both dysphonic subjects (affected by nodules, polyps, oedema, cysts, ...) and control group. The subjects' voices are classified according to G parameter of the Hirano's GRBAS scale [5], where a normal voice is rated as grade 0, a slight dysphonia as 1, a moderate dysphonia as 2 and finally, a severe dysphonia as 3. The corpus was supplied by the Experimental and Clinical Audio-Phonology Laboratory (LAPEC - Hospital La Timone - Marseille). It is composed of 80 voices of females aged 17 to 50 (mean: 32.2). The speech material is obtained by reading the same short text (French), which signal duration varies from 13.5 to 77.7 seconds (mean: 18.7s). The 80 voices are equally balanced among the 4 grades (20 voices per each). These perceptual grades were determined by a jury composed of 3 expert listeners. This perceptual judgment was carried out by consensus between the different jury members as it is the usual way to assess voice quality by our therapist partners. The judgment was done during only one session.
This corpus is used for all the experiments presented in this paper. Due to its small size, cautions have been made to provide statistical significance of the results over all the experiments by applying specific methods like, for instance, leave_x_out technics [4].

3. BASELINE CLASSIFICATION SYSTEM

The baseline system is derived from a classical speaker recognition (ASR) system adapted to dysphonic voice classification. The ASR system is based on the state-of-the-art GMM modelling. It relies on the ASR toolkit, available in « open source » (LIA_SpkDet and ALIZE [3]) and developed at the LIA laboratory.
Three phases are necessary and are described in the following sections.

3.1. Parameterization

The speech signal is parameterized as follows: the signal (pre-emphasized with 0.95 value) is characterized by 24 spectrum coefficients issued from a filter-bank analysis (24 filters) applied on 20ms Hamming windowed

Table 1: Description of the frequency subbands

SB_n	Coefficients of the filter-bank	Band ranges (Hz)
SB_1	1-4	0-1600
SB_2	5-8	1280-2880
SB_3	9-12	2560-4160
SB_4	13-16	3840-5440
SB_5	17-20	5120-6720
SB_6	21-24	6400-8000

frames at a 10ms frame rate. The filters are triangular and either equally spaced along the entire linear scale to yield Linear Frequency Spectrum Coefficients (LFSC) or distributed along a MEL scale (close to the hearing perception) to yield MEL Frequency Spectrum Coefficients (MFSC). Here, given F_{ci} the central frequency of the ith filter, the boundaries of this filter are fixed to the central frequencies of the two adjacent filters [F_{ci-1},F_{ci+1}]. Outside that range, magnitude is totally attenuated.
The first and second derivatives of the LFSC/MFSC coefficients are added (Δ and $\Delta\Delta$) to the parameters in order to catch short-term dynamic information. Finally, parameters are normalized to match a 0-mean and 1-variance distribution (mean and variance are estimated on speech signal only, after discarding non-speech signal).

3.2. Modelling

In ASR, state-of-the-art systems rely on the statistical modelling: Gaussian Mixture Model (GMM)[2]. A GMM is a weighted sum of M multi-dimensional Gaussian distributions, each characterized by mean vector $\overline{x}$ (dimension *d*), covariance matrix Σ (*d* x *d*) and weight p of the Gaussian component within the mixture (diagonal covariance matrices are used in this work). A GMM model is built on a training data set by estimating the parameters ($\overline{x}$, Σ, p) thanks to the EM/ML algorithm (Expectation-Maximization/Maximum Likelihood).
Classically, two training phases are necessary to cope with the frequent lack of training data available for a speaker [2]: (1) training of a generic speech model estimated by the EM/ML algorithm on a large population of speakers; (2) training of the speaker model, derived from the generic speech model by applying adaptation techniques (MAP, Maximum a posteriori).
In the pathological context, a model doesn't correspond anymore to a speaker but to a dysphonia severity level. It will be named **grade model** G_g with $g \in \{0, 1, 2, 3\}$. Grade model G_g is learned gathering all the voices evaluated as grade g. It can be noted that all the voices used for the grade model training are excluded from the test trials in order to differentiate the detection of the pathology from the speaker recognition.

3.3. Classification and decision

In ASR domain, a test trial consists in computing a similarity measure between a test signal and the GMM model of a given speaker, following: $L(y_t|X) = \sum_{i=1}^{M} p_i \; L_i(y_t)$ where $L_i(y_t)$ is the likelihood of signal y_t given gaussian i, M the number of gaussians and p_i the weight of the gaussian i.
In this paper, the **decision** will be made by selecting grade *g* of model G_g for which the largest likelihood is measured given a test voice.

Table 2: Comparison between LFSC and MFSC - Results of the 4-G classification in [0-8000]Hz in terms of % CCR

	Grade 0	**Grade 1**	**Grade 2**	**Grade 3**	**Total**
Parameter [0-8000Hz]	**% CCR** (nb/20)	**% CCR** (nb/20)	**% CCR** (nb/20)	**% CCR** (nb/20)	**% CCR** (nb/80)
24LFSC + 24Δ + 24$\Delta\Delta$	**95.0 (19)**	50.0 (10)	55.0 (11)	**75.0 (15)**	68.75 (55)
24MFSC + 24Δ + 24$\Delta\Delta$	**95.0 (19)**	**60.0 (12)**	**75.0 (15)**	**75.0 (15)**	**76.25 (61)**

4. MULTIBAND APPROACH

4.1. Objective

The multiband approach consists in cutting the frequency domain in subbands processed independently. The main motivation of this approach resides in the assumption that the quality of frequency information can be dependent on the band of frequencies considered. For example, [1] shows that some subbands seem to be more relevant to characterize speakers than some others for the ASR task. In the same way, the multiband approach has been used for the automatic speech recognition task in adverse conditions, since subbands may be affected differently by noise [8].
In this paper, the multiband approach is used in order to study how the acoustic characteristics of dysphonia are spread out along different frequency bands depending on the severity level: « is a frequency subband more discriminant than another for dysphonic voice classification ? » Therefore, each subband will be processed as the full band by the classification system.

4.2. Subband description

In this paper, the full frequency band [0-8000]Hz is split into six subbands (SB_1, SB_2, ..., SB_6), as described in table 1. In practice, each subband is composed of 4 successive linear coefficients issued from the 24 filter-bank spectral analysis defined in section 3.1. Here, the linear scale is preferred to the MEL scale in order to keep homogeneous spectral analysis over the different subbands.

5. EXPERIMENTS

Results provided in this section are either expressed in terms of correct classification rates (named CCR in the rest of the paper), the number of well-classified voices is also provided in brackets, or presented in confusion matrix form (a confusion matrix provides the error number and the type of confusion between the response given by the system - noted TGx in the paper - and the perceptual reference - noted RGx. The matrix diagonal provides the number of correct matches).
Note: all the results, presented in next sections, are issued from the GMM classifier and have to be interpreted from a statistical viewpoint.

5.1. Baseline system

In this first experiment, the effect of a MEL scale associated with the spectrum analysis-based parameterization is investigated on the baseline dysphonic voice classification system. Table 2 gives performance of the classification system according to the LFSC (Linear Frequency Spectrum Coefficients) and MFSC (MEL Fre-

Table 3: Confusion matrices of the 4-G classification following frequency subbands (LFSC parameters)

SB_1

	RG0	RG1	RG2	RG3
TG0	**19**	1	0	0
TG1	5	**14**	1	0
TG2	4	10	**5**	1
TG3	0	6	4	**10**

SB_2

	RG0	RG1	RG2	RG3
TG0	**17**	3	0	0
TG1	9	**10**	1	0
TG2	6	5	**8**	1
TG3	2	5	9	**4**

SB_3

	RG0	RG1	RG2	RG3
TG0	**10**	4	2	4
TG1	1	**3**	6	10
TG2	1	4	**3**	12
TG3	0	2	2	**16**

SB_4

	RG0	RG1	RG2	RG3
TG0	**6**	7	3	4
TG1	2	**5**	4	9
TG2	4	2	**1**	13
TG3	1	2	2	**15**

SB_5

	RG0	RG1	RG2	RG3
TG0	**5**	6	9	0
TG1	3	**11**	4	2
TG2	1	5	**11**	3
TG3	0	1	5	**14**

SB_6

	RG0	RG1	RG2	RG3
TG0	**6**	9	4	1
TG1	6	**7**	5	2
TG2	3	5	**6**	6
TG3	0	2	3	**15**

quency Spectrum Coefficients) parameters. It can be observed that the use of the MEL scale, through the MFSC, allows to decrease the classification errors on both grades 1 and 2 and reaches 76.25% CCR.
Since the MFSC based-parameterization provides a better resolution in the low frequencies, it would be interesting to observe the behaviour of the system in different frequency subbands, as done in the next sections.

5.2. Subband Analysis

Following the multiband protocol described in section 4.2., confusion matrices given in table 3 compare the performance of the classification system according to the 6 frequency subbands SB_1, SB_2, ..., SB_6. Three main trends may be observed from these confusion matrices:
• SB_1 and SB_2 underestimate dysphonia since 25 errors over 32 are located in grade 0 or 1 for subband SB_1 and 27 errors over 41 for subband SB_2. Consequently, both normal or slightly dysphonic voices (Grades 0 and 1) tend to be well classified in subbands SB_1 and SB_2; Grade 0 voices get 95% CCR in subband SB_1, outperforming the full band rate (85% CCR) (see table 5 for the full band confusion matrix) similarly to grade 1 voices getting 70% CCR (vs 55% CCR);
• SB_3 and SB_4 overestimate dysphonia since 34 errors over 48 are located in grade 2 or 3 for the subband SB_3 and 33 errors over 53 for the subband SB_4. Consequently, severe dysphonic voices (grade 3) tend to be well classified in both subbands SB_3 (80% CCR) and SB_4 (75% CCR);
• SB_5 and SB_6 do not show particular tendancy (over or under-estimate). Conversely, most of the classification errors are scattered over the grades. Nevertheless, it can be observed that grade 2 reaches its better performance on subband SB_5 and that severe dysphonic voices (grade 3) are still well classified in both subbands SB_5 (70% CCR) and SB_6 (75% CCR).

These observations show a different behaviour of

Table 4: Confusion Matrices of the 4-G classification joining two adjacent subbands (LFSC parameters)

$SB_1 + SB_2$

	RG0	RG1	RG2	RG3
TG0	**19**	1	0	0
TG1	4	**16**	0	0
TG2	3	8	**9**	0
TG3	0	4	4	**12**

$SB_3 + SB_4$

	RG0	RG1	RG2	RG3
TG0	**12**	6	1	1
TG1	5	**12**	2	1
TG2	6	7	**4**	3
TG3	0	3	5	**12**

$SB_5 + SB_6$

	RG0	RG1	RG2	RG3
TG0	**11**	6	3	0
TG1	9	**7**	3	1
TG2	4	6	**7**	3
TG3	0	2	5	**13**

the classification system according to the subbands. In order to refine these results, a second experiment was conducted merging two adjacent subbands together, as shown in confusion matrices (table 4). Here, it can be observed that:
• the joint used of SB_1-SB_2 permits to avoid some confusion errors (CCR improvement for all the grades compared with individual subband CCR) while still under-estimating dysphonia;
• the joint used of subbands SB_3-SB_4 leads to an unexpected behaviour, for which boundaries between adjacent grades are confusing (6 confusion errors for grade 0 mis-classified in grade 1, 5 confusion errors for grade 1 mis-classified in grade 0 and 5 confusion errors for grade 3 in grade 2). The over-estimate trend observed on the subbands SB_3 and SB_4 individually has been softened with this joint used. Particularly, confusion errors related to grade 2 occur in lower grades (7 with grade 1 and 6 with grade 0).
• the joint used of SB_5-SB_6 presents a similar behaviour to the individual subbands.

5.3. Restricted frequency bands

The previous observations highlight three frequency zones, interesting for further analysis. In this sense, a complete parameterization is applied on three targeted frequency bands: [0-3000]Hz (mainly formant zone), [3000-6400]Hz (mainly fricative and plosive zone), and [6400-8000]Hz (residual zone of fricatives and plosives), based on 24 LFSC (see section 3.1. for computation details). Here, the 24 triangular filter banks are spread out along the targeted frequency band, leading to a better resolution in the spectrum analysis.
Classification confusion errors per restricted frequency band are provided in table 5. It can be pointed out that the underestimate observed with the joint used of subbands SB_1 and SB_2 is softened in the restricted [0-3000]Hz frequency band, leading to better classification rates in intermediate grades (grades 1 and 2) compared with the [0-8000]Hz (full band) based-system. An overall 71.25% CCR is reached (compared to 65% CCR on [0-8000]Hz). The [3000-6400]Hz and [6400-8000]Hz restricted frequency bands show a behaviour similar to the one observed with joint subbands SB_3-SB_4 and SB_5-SB_6 respectively, with an increase of correct classification for

Table 5: Confusion matrices of the 4-G classification following different frequency ranges (24LFSC)

[0-3000]Hz

	RG0	RG1	RG2	RG3
TG0	**18**	1	1	0
TG1	1	**13**	6	0
TG2	0	6	**13**	1
TG3	0	2	5	**13**

[3000-6400]Hz

	RG0	RG1	RG2	RG3
TG0	**13**	6	1	0
TG1	6	**9**	3	2
TG2	3	4	**10**	3
TG3	1	1	4	**14**

[6400-8000]Hz

	RG0	RG1	RG2	RG3
TG0	**12**	2	5	1
TG1	10	**4**	4	2
TG2	9	2	**1**	8
TG3	1	1	3	**15**

[0-8000]Hz (Full Band)

	RG0	RG1	RG2	RG3
TG0	**17**	2	1	0
TG1	2	**11**	5	2
TG2	2	6	**10**	2
TG3	0	1	5	**14**

grade 3 (especially in [6400-8000]Hz).

5.4. Synthesis

The frequency analysis has shown that the classification of dysphonic voices may differ according to the subband considered. The high frequencies (over 3000Hz) seem to be relevant for severe dysphonic voices (grade 3) only. The central frequencies [3000-6400]Hz do not outline discriminant information (regarding the confusion errors), useful for the classification sheme. Finally, low frequencies ([0-3000]Hz) tend to be the most pertinent (compared to the others) since classification performance is homogeneous and satisfactory along the different grades. The latter proposition is emphasized through table 6, in which the complete system (described in section 3. and which performance is illustrated in table 2) is applied on the restricted [0-3000]Hz frequency band (24 spectrum coefficients plus first and second derivative coefficients). Here, classication performance is improved over all the grades compared with the full frequency band ([0-8000]Hz) (80% CCR against 76.25% for the MFSC coefficients). The performance gain classically brought by using the derivative coefficients (Δ and $\Delta\Delta$) is still observed here.

6. CONCLUSION

In this paper, the authors propose to study dysphonic voice classification, according to the GRBAS scale, in the frequency domain. Indeed, the main idea is to observe how the acoustic characteristics of dysphonia are spread out along different frequency subbands. In this context, an automatic dysphonic voice classification is used, coupled with a multiband approach in order to evaluate the effect of frequency bands on the classification paradigm. The subband analysis outlines that low frequencies tend to be the most interesting zones for an homogeneous discrimination between grades. Additional experiments, involving a more complex parameterization (MFSC plus Δ and $\Delta\Delta$), show that the use of the restricted frequency band [0-3000]Hz (compared with the [0-8000]Hz full band) provides a very good compromise for the classification over all the grades.
In further work, this study will be coupled with a phonetic analysis [10] in order to evaluate how the dysphonia

Table 6: Comparison between LFSC and MFSC - Results of the 4-G classification in [0-3000]Hz, in terms of % CCR

	Grade 0	Grade 1	Grade 2	Grade 3	Total
Parameter [0-3000Hz]	**% CCR** (nb/20)	**% CCR** (nb/20)	**% CCR** (nb/20)	**% CCR** (nb/20)	**% CCR** (nb/80)
24LFSC + 24Δ + 24$\Delta\Delta$	**95.0 (19)**	**75.0 (15)**	50.0 (10)	**85.0 (17)**	76.25 (61)
24MFSC + 24Δ + 24$\Delta\Delta$	**95.0 (19)**	70.0 (14)	**70.0 (14)**	**85.0 (17)**	**80.00 (64)**

effects may impact on phonemes or phoneme classes in particular subbands according to the grades. Moreover, it will be interesting to compare the results presented in this paper with a perceptual evaluation of dysphonic voices performed by an expert jury within restricted frequency bands. On the other side, the results reported in this paper are issued from statistical observations. For instance, even if a subband appears as discriminant (e.g. [6400-8000]Hz for the grade 3), relevancy may be due to either a presence of signal information or a lack of energy, compared with the other bands. These two alternatives can draw very different interpretations. Therefore, results outlined in this paper have to be validated in the future from a physio-pathological or clinical analysis. The authors will first investigate some results in laryngology [6], which could bring some explanations to the observed behaviours.

7. REFERENCES

[1] Besacier, L., et al. 2000. Localization and selection of speaker specific information with statistical modelling. *Speech Communication*, Vol. 31, 89–106.

[2] Bimbot, F., et al. 2004. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, Vol. 39, 430–451.

[3] Bonastre, J.-F., et al. 2005. *ALIZE, a free toolkit for speaker recognition*. ICASSP-05, Philadelphia, USA.

[4] Fredouille, C., et al. 2005. *Application of Automatic Speaker Recognition techniques to pathological voice assessment (dysphonia)*. Proc. of Eurospeech'05.

[5] Hirano, M. 1981. Psycho-acoustic evaluation of voice : GRBAS Scale for evaluating the hoarse voice. *Clinical Examination of voice, Springer Verlag*

[6] Honda, K., et al. 2004. *Resonance Characteristics of Hypopharyngeal Cavities*. International Conference on Voice Physiology and Biomechanics, Marseille, France.

[7] Maguire, C., et al. 2003. *Identification of voice pathology using automated speech analysis*. Third International Workshop on Models and Analysis of Vocal Emission for Biomedical Applications, Florence, Italy.

[8] McCowan, I. A., Sridharan, S. 2001. Multi-Channel Sub-Band Speech Recognition. *EURASIP Journal on Applied Signal Processing*, Vol. 1, 45–52.

[9] Kacha, A., Grenez, F., Schoentgen, J., Benmahammed, K. 2005. *Dysphonic speech analysis using generalized variogram*. In Proc. ICSLP'05, Vol. 1, 917–920.

[10] Pouchoulin, G., et al. 2006. *Modélisation Statistique et Informations Pertinentes pour la Caractérisation des Voix Pathologiques (Dysphonies)*. XXVIèmes JEP'06, Dinard, France.

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

[12] Teston, B., Galindo, B. 1995. *A diagnosis of rehabilitation aid workstation for speech and voice pathologies*. Proc. of Eurospeech, 1883–1886.

[13] Wester, M. 1998. *Automatic classification of voice quality: Comparing regression models and hidden Markov models*. VOICEDATA98, December, 92–97, Utrecht.

[14] Wuyts, F. L., et al. 2000. The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach. *Journal of Speech, Language, and Hearing Research 43*, 796–809.