Dysphonic Voices and the 0-3000Hz Frequency Band

Gilles Pouchoulin, Corinne Fredouille, J.-F Bonastre, Alain Ghio, Antoine Giovanni

Dysphonic Voices and the 0-3000Hz Frequency Band

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

[1]Université d'Avignon, LIA, Avignon (France), [2]CNRS-LPL, Aix en Provence (France)
{gilles.pouchoulin,corinne.fredouille,jfb}@univ-avignon.fr,alain.ghio@lpl-aix.fr

Abstract

Concerned with pathological voice assessment, this paper aims at characterizing dysphonia in the frequency domain for a better understanding of related phenomena while most of the studies have focused only on improving classification systems for diagnosis help purposes. Based on a first study which demonstrates that the low frequencies ([0-3000]Hz) are more relevant for dysphonia discrimination compared with higher frequencies, the authors propose in this paper to pursue by analyzing the impact of the restricted frequency band ([0-3000]Hz) on the dysphonic voice discrimination from a phonetical and perceptual point of views. A discussion around the frequency band limitation of telephone channel is also proposed.
Index Terms: Voice disorder, dysphonia characterization, automatic dysphonic voice classification, frequency analysis

1. Introduction

Assessment of dysphonic voice quality is an important issue, resulting in a large amount of multidisciplinary research. Two main approaches can be considered. The first methodology, the perceptual evaluation [1, 2], consists in qualifying and quantifying the vocal dysfunction by listening to the patient's speech production. It is currently the most used by the clinicians. However, it is largely debated in the literature because of an intrinsic subjectivity, a lack of a universal scale, a large intra and inter-variability in the human judgments and finally a large cost in time and human resources when an expert jury is involved to reduce its subjectivity. The second methodology involved in the dysphonic voice assessment is the objective measurement-based analysis. This approach has been introduced as an alternative to the perceptual evaluation in order to cope with its drawbacks. In this context, acoustic, aero-dynamical and/or physiological measures are associated with an automatic classification system to provide a decision. Most of the studies proposed in the literature aims at improving automatic system performance since current systems are not sufficiently efficient from clinician point of view [3, 4]. Conversely, a few studies have been dedicated to the characterization of the dysphonia phenomena in the speech signal [5, 6]. This analysis should be useful for a better understanding of dysphonia impact in speech production or simply to enhance performance of the automatic classification by exploiting more relevant information.

As dysphonia is essentially relating to the vocal source, most of the studies have focused on parameters directly linked to this vibrator (FO stability, intensity, jitter, shimmer, harmonics to noise ratio, etc [7, 8, 9]). Other studies have been related on the global tone of the voice, assuming that the acoustic characteristics of dysphonia are uniformly distributed on the whole spectrum. Finally, information issued from long-term spectral analysis was also investigated many years ago, leading to different pathological voice classifications [10].

This paper pursues work reported in [11] in which the authors have investigated the characteristics of dysphonia in the frequency domain, especially by studying relating phenomena through a frequency subband analysis. In this context, the authors have shown that the [0-3000]Hz frequency subband tends to carry more relevant information for dysphonic voice discrimination, compared with higher frequency subbands as well as with the [0-8000]Hz full band. Here, the authors propose to extend investigation on the [0-3000]Hz frequency band by examining three different axes: (1) the analysis of the restricted frequency subband on the dysphonic voice discrimination from a phonetical point of view, (2) its effects on the perceptual judgment carried out by an expert jury, and finally (3) a parallel with the frequency band limitation involved by the telephone channel.

2. Dysphonic voice corpus

The corpus used in this study is composed of reading speech pronounced by both dysphonic subjects (affected by nodules, polyps, oedema, cysts, ...) and control group. The subjects' voices are classified according to the G parameter of the Hirano's GRBAS scale [12], 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 ORL department of the "Timone" University Hospital (Marseille - France). It is composed of 80 voices of females aged 17 to 50. 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, 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 one session only.

This corpus is used for all the experiments presented here. Due to its small size, cautions have been made to provide statistical significance of the results by applying specific methods like leave_x_out technics.

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 [13]) and developed at the LIA laboratory. Three phases are necessary (see [11]):

Parameterization: the pre-emphasized speech signal (0.95 value) is characterized by 24 spectrum coefficients issued from a filter-bank analysis (24 filters) applied on 20ms Hamming windowed frames at a 10ms frame rate. The filters are triangular and

Table 1: *Total duration in seconds per phonetic class and per grade - Information per phoneme class : count (nb) with duration mean (μ) and standard deviation (σ).*

Phonetic classes	Grades				Info. per class		
	G0	*G1*	*G2*	*G3*	*nb*	μ	σ
Consonant	**135.13**	**139.21**	**149.83**	**167.28**	**6395**	**0.092**	**0.045**
Liquid	34.56	34.01	36.04	43.03	2181	0.068	0.033
Nasal	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
Vowel	**103.58**	**98.77**	**103.46**	**109.79**	**5586**	**0.074**	**0.046**
Oral	84.37	80.45	85.22	93.66	4862	0.071	0.044
All phonemes	**241.51**	**240.96**	**256.66**	**280.52**	**12140**	**0.084**	**0.046**

equally spaced along the entire linear scale to yield Linear Frequency Spectrum Coefficients (LFSC). Parameters are normalized to match a 0-mean and 1-variance distribution.

Modelling: Gaussian Mixture Model (GMM)-based techniques are used to build a statistical model for each dysphonia severity grade, 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.
All GMM models are composed of 128 gaussian components with diagonal covariance matrices.

Decision: In the context of dysphonic voice classification, the classification decision is made by selecting the grade *g* of the model G_g (among the four grade models available) for which the largest similarity measure is computed given a test voice. Here, the similarity measure relies on a likelihood value as follows: $L(y_t|X) = \sum_{i=1}^{M} p_i \; L_i(y_t)$ 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 gaussian i.

4. [0-3000]Hz and phonetic analysis

In [11], the authors have studied how the acoustic characteristics of dysphonia are spread out on the overall frequency space by analyzing the performance of the automatic dysphonic voice classification (described in section 3) on different frequency subbands. The latter were obtained by filtering signal of the dysphonic corpus (defined in section 2) according to the following ranges: [0-3000]Hz, [3000-5400]Hz and [5400-8000]Hz. The classification tests applied on the different filtered corpora outline that the [0-3000]Hz frequency subband tends to be the most interesting zones (compared with the other frequency subbands as well as with the full band), leading to an homogeneous and better discrimination between voices.
To investigate further, the authors propose in this paper to observe the behaviour of the automatic dysphonic voice classification system following different phoneme classes. This behaviour will be analyzed according to the full frequency band and the [0-3000]Hz frequency subband. Indeed, performance of the classification system will be analyzed per phoneme class and per frequency range in order to evaluate how the dysphonia effects may impact on phonemes or phoneme classes in particular frequency bands according to the grades. This phonetic analysis is very close to the "phonetic labelling" proposed in [2], in which a descriptive and perceptual study of pathological characteristics of different phonemes is presented.

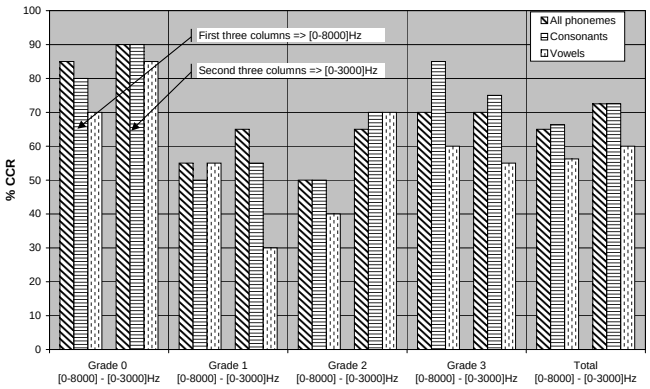

Figure 1: *Performance per grade in terms of Correct Classification Rate (CCR %) considering "All phonemes", consonant and vowel classes, for both the [0-8000]Hz (each first set of three columns) and [0-3000]Hz (each second set of three columns) frequency bands.*

4.1. Experimental protocol

To perform dysphonic voice classification tests according to different phoneme classes, a phonetic segmentation is necessary for each speech signal of the corpus. This segmentation was extracted automatically by realizing an automatic text-constrained phonetic alignment. This alignment was performed by using the LIA alignment system, based on a Viterbi decoding algorithm, a text-restricted lexicon of words associated with their phonological variants and a set of 38 French phonemes.
The phonetic segmentation is coupled with the automatic dysphonic classification system for the decision step only i.e this segmentation is not used for both the parameterization and training phases. In the latter case, all the grade models are learned on all the phonemic material available per grade in the corpus. Indeed, for the classification tests and decision making, the similarity measure (see section 3) between the test voice and the grade models is computed on the restricted set of segments associated with a given phoneme class. Table 1 provides the targeted phoneme classes available through the dysphonic voice corpus as well as information on their durations.
Results provided in this section are expressed in terms of Correct Classification Rates (named CCR in the rest of the paper).

4.2. Comparative phonetic analysis [0-8000] vs [0-3000]Hz

This section presents performance of the automatic dysphonic voice classification system depending on the frequency bands: [0-8000]Hz and [0-3000]Hz frequency bands, and on different phoneme classes: on the one hand, all phoneme set, consonant and vowel classes illustrated in figure 1 and more specific consonant and vowel classes like liquid, nasal, fricative, plosive and oral vowels illustrated in figure 2 on the other hand. From these figures, it can be observed that:
• CCR is improved for the global consonant class (fig. 1) on the [0-3000]Hz for most of the cases (e.g. from 50 to 70% for the grade 2), except for the grade 3 (from 85 to 75% CCR). Nevertheless, the behaviour of the individual consonant classes (fig. 2) is rather different, with a CCR improvement observed uniquely for the liquids on both the grades 0 and 1, for the nasal consonant on the grade 2, for the fricative on the grade 1 only, and for the plosive on both the grade 0 and 2. Regarding the grade 3, the reduction of the frequency band affects CCR of most of the consonant classes, except CCR of both plosive and fricative consonants, which remains unchanged.

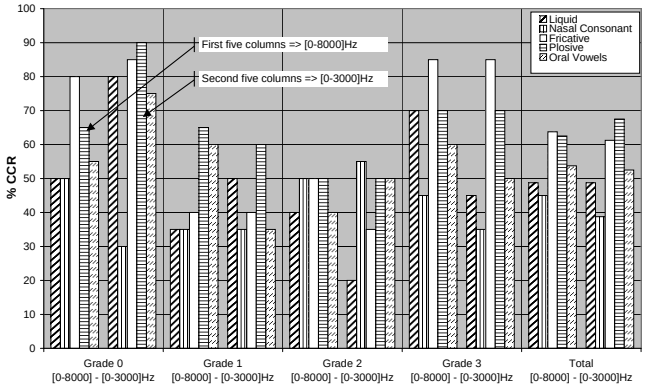

Figure 2: *Performance per grade in terms of Correct Classification Rate (CCR %) considering specific consonant and vowel classes, for both the [0-8000]Hz (each first set of three columns) and [0-3000]Hz (each second set of three columns) frequency bands.*

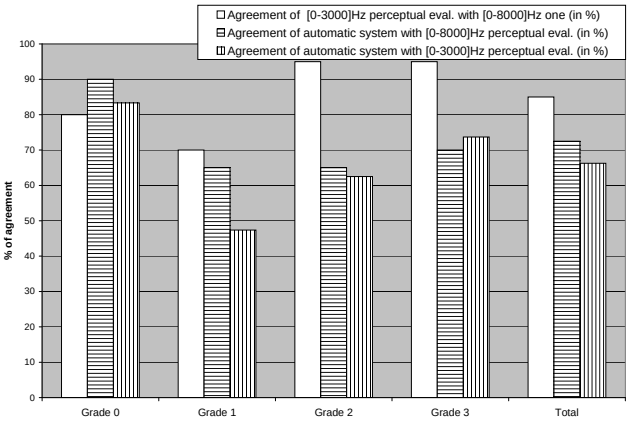

Figure 3: *Agreement rate (in %) between both [0-3000]Hz and [0-8000]Hz perceptual evaluation and between automatic grade classification and perceptual evaluations.*

• CCR is slightly improved for the global vowel class on the [0-3000]Hz. However, the analysis per grade reveals balanced behaviours between the 0 and 2 grades, for which a quite significant improvement is observed, and the 1 and 3 grades. Results per grade reported for the oral vowels follow the same scheme.
• CCR of the consonant class is higher in [0-3000]Hz than the vowel consonants in most of the cases (equal for the grade 2). Specific consonant and vowel classes tend to exhibit similar behaviour, notably while comparing CCR of plosive and fricative consonants with oral vowels. Nasal consonants seem to be the least relevant phoneme class here.
These different observations permit to draw some assumptions about the discrimination of dysphonic voices (according to the GRBAS scale): first, vowel formants (variably located in the [0-3000Hz] frequency band) may not carry sufficient information to discriminate grade 1 voices, contrary to the grades 0 and 2. Secondly, consonants seem to bring, amazingly, more useful information for dysphonic voice discrimination in this context.

5. [0-3000]Hz and perceptual judgment

If the baseline automatic classification system (described in section 3) tends to be positively sensitive to the frequency band reduction (from [0-8000] to [0-3000]Hz), the question which can be raised is: "Which effects can this reduction have on perceptual judgment carried out by an expert jury ?". To be able to bring some responses to this question, the [0-3000]Hz filtered corpus of dysphonic voices (described in section 2) was perceptually analyzed following the same rules as the original corpus (evaluated on the full frequency band). In this way, the perceptual judgment was carried out by consensus involving the same expert jury as previously for one session only.

Experimental Results

The analysis of the perceptual evaluation done on the [0-3000]Hz filtered corpus of dysphonic voices permits to draw up different comparisons involving the perceptual evaluation done on the original corpus as well as the automatic classification system. In this way, figure 3 presents agreement rates (1) between both the perceptual assessment conditions, (2) between the automatic classification system and the [0-8000]Hz perceptual judgment, and (3) between the automatic classification system and the [0-3000]Hz perceptual judgment.

Table 2: *Confusion matrix of the [0-3000]Hz filtered dysphonic corpus perceptual judgment against the original perceptual judgment ([0-8000]Hz dysphonic corpus).*

	G0	**G1**	**G2**	**G3**
RG0	**16**	4	0	0
RG1	2	**14**	4	0
RG2	0	1	**19**	0
RG3	0	0	1	**19**

The case (1) reveals 85% of overall agreement of the [0-3000]Hz perceptual judgment with the [0-8000]Hz one, exhibiting very high agreement rates (95%) for the grades 2 and 3, high agreement rate (80%) for the grade 0 and moderate agreement rate (70%) for the grade 1. If the overall agreement rate of 85% could be due to the intra-listener variability accross sessions, it is interesting to highlight that disagreements mainly occur for normal and slight dysphonic voices as illustrated by the confusion matrix given in table 2. This tends to show that the reduction of the frequency band, resulting in more low-pitched voices, may affect the expert jury's judgments, inducing an overestimate of dysphonia level for normal and slight dysphonic voices.
Regarding the cases (2) and (3), the automatic classification obtains higher agreement rates with the [0-8000]Hz perceptual judgment. Only the grade 3 agreement rate is slightly more favorable for the [0-3000]Hz perceptual judgment (73.7% vs 70%), due to a unique grade 3 voice reclassified in grade 2 in the latter. It is interesting to remark that the agreement rate for the grade 1 is quite low for the [0-3000]Hz perceptual judgment with only 47% against 65% for the [0-8000]Hz perceptual judgment. This behaviour is strongly correlated to observations made in the case (1) regarding the low agreement rate for the grade 1. Indeed, a large part of dysphonic voices rated as 1 for the [0-8000]Hz perceptual judgment and well classified by the automatic system has been rated differently for the [0-3000]Hz perceptual judgment (4 voices in grade 0 and 6 in grade 2). Finally, it seems meaningful that the agreement rates are globally higher between the automatic system and the [0-8000]Hz perceptual judgment, since the latter was used as reference for the grade model training.

Table 3: *Confusion matrix of the automatic classification system on both the [0-3000]Hz and [300-3000]Hz filtered dysphonic voices.*

[0-3000]Hz

	G0	**G1**	**G2**	**G3**
TG0	**18**	1	1	0
TG1	1	**13**	6	0
TG2	0	6	**13**	1
TG3	0	2	4	**14**

[300-3000]Hz

	G0	**G1**	**G2**	**G3**
TG0	**17**	2	1	0
TG1	3	**11**	5	1
TG2	0	7	**11**	2
TG3	0	2	7	**11**

6. [0-3000]Hz and telephone band

One of the issues of speech transmitted through the telephone channel is the restriction of the frequency band to [300-3400]Hz. It is well-known that this limitation strongly disturbs automatic systems related to speech (e.g. automatic speech and speaker recognition).
Considering in this paper the [0-3000]Hz frequency band, it makes sense to examine the disturbance of the telephone band restriction to the automatic dysphonic voice classification. Consequently, signals related to the dysphonic corpus have been filtered according to the [300-3000]Hz frequency band and processed in the same way as described in section 3.

Experimental Results

Table 3 provides the confusion matrices issued from the automatic classification of the dysphonic voices for both the [0-3000]Hz and [300-3000]Hz frequency bands. Here, this classification has been compared with the original perceptual judgment (full band corpus). As expected, it can be observed that all the grades are affected by discarding the low frequency subband [0-300]Hz, resulting in an absolute overall CCR loss of 10% (from 72.5 to 62.5% CCR). Indeed, confusion with adjacent grades is drastically increased, notably for the grade 3.
By extrapolating the disagreement between the perceptual judgments carried out on both the [0-3000]Hz and [0-8000]Hz dysphonic corpora, it might be assumed that the telephone band may affect the perceptual judgment similarly.
Finally, considering the other issues of the telephone channel such as the amplitude signal distorsion and noise, a more significant CCR decrease may be expected in real conditions.

7. Conclusion

This paper aims at exploring the characterization of dysphonia voices in the frequency domain. It pursues a first study, in which the authors have shown that the [0-3000]Hz subband tends to be the most relevant zone for the automatic discrimination of dysphonic voices in the proposed context. First, focus is made on the effects of this frequency band reduction from a phonetic viewpoint. This study has underlined that consonant classes tend to be more relevant than vowel classes for the automatic dysphonic voice discrimination task. Moreover, this analysis has revealed a better discrimination of 0 and 2 rated voices observed on the vowel classes compared with the 1 and 3 ones. Secondly, perceptual judgments involving the dysphonic voice corpus, in its original form as well as filtered in the [0-3000]Hz frequency band have been compared. This comparison has shown that expert jury has been affected by the limitation of the frequency band, notably for 0 and 1 rated voices, inducing an overestimate of dysphonia level in this case. Finally, a parallel between the [0-3000]Hz frequency band and the band limitation of telephone channel has been proposed. As expected, automatic classification rates have been affected when considering the [300-3000]Hz frequency band. Making connection with the perceptual judgment study, it can be assumed that the telephone band may also affect the perceptual judgment provided by an human expert similarly, especially in real conditions for which additional signal damages have to be taken into account like the signal amplitude distorsion and noise.

8. References

[1] P.H. Dejonckere et al., "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)*, vol. 258, pp. 77–82, 2001.

[2] J. Revis et al., "Phonetic labeling of dysphonia: a new perspective in perceptual voice analysis," *7th International Conference Advances in Quantitative Laryngology, Voice and Speech Research*, October 2006.

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

[4] N. Saenz-Lechon et al., "Methodological issues in the development of automatic systems for voice pathology detection," *Journal of Biomedical Signal Processing and Control, Elsevier*, 2006.

[5] J.I. Godino-Llorente et al., "Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters," *EEE Trans. on Biomedical Engineering*, vol. 53(3), pp. 1943–1953, 2006.

[6] A. Kacha et al., "Dysphonic speech analysis using generalized variogram," *In Proc. ICSLP'05*, vol. 1, pp. 917–920, 2005.

[7] J. Schoentgen et al., "Acoustic analysis of dysphonic voices: descriptors and methods," in *LARYNX'97*, 1997, pp. 37–46.

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

[9] P. Yu et al., "Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements," in *Journal Voice 15*, 2001, pp. 529–542.

[10] N. Yanagihara, "Significance of harmonic changes and noise components in hoarseness," *Journal Speech, Hear, Res*, vol. 10, pp. 531–541, 1967.

[11] G. Pouchoulin et al., "Frequency study for the characterization of the dysphonic voices," *Interspeech'07, Antwerp, Belgium*, August 2007.

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

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