Frequency Study for the Characterization of the Dysphonic Voices

Gilles Pouchoulin, Corinne Fredouille, Jean-François Bonastre, Alain Ghio, Antoine Giovanni

▶ **To cite this version:**

Gilles Pouchoulin, Corinne Fredouille, Jean-François Bonastre, Alain Ghio, Antoine Giovanni. Frequency Study for the Characterization of the Dysphonic Voices. INTERSPEECH 2007, Aug 2007, Antwerp, Belgium. pp.1198-1201. hal-00173730

HAL Id: hal-00173730
https://hal.science/hal-00173730

Submitted on 20 Sep 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Frequency Study for the Characterization of the Dysphonic Voices

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

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

Abstract

Concerned with pathological voice assessment, this paper aims at characterizing dysphonia in the frequency domain for a better understanding of relating phenomena while most of the studies have focused only on improving classification systems for diagnosis help purposes. In this context, a GMM-based automatic classification system is applied on different frequency ranges in order to investigate which ones are relevant for dysphonia characterization. Experiment results demonstrate that the low frequencies [0-3000]Hz are more relevant for dysphonia discrimination compared with higher frequencies.

Index Terms: dysphonia characterization, pathological voice and speech, automatic speaker recognition

1. Context

Many studies have focused on the objective measurement-analysis for dysphonic voice assessment, proposed as an alternative to the perceptual evaluation [13] (the most widely used by clinicians). In most cases, these studies describe classification systems, acoustic, physiological and/or aerodynamical analysis in order to improve voice classification performance and to help clinicians making their decision [15][8][14]. A few studies have been dedicated to the analysis of dysphonia effects on the speech signal [5][10][16]. Indeed, if an expert is able to assess a dysphonic voice according to a quality scale like the Hirano's GRBAS scale [6], 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 this paper is to investigate the characteristics of dysphonia in the frequency domain, especially by studying relating phenomena through a frequency 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 applied on different frequency subbands, which should permit to analyse the relevance of the latter for the characterization of the dysphonic voices. This paper pursues work reported in [12] in which a simpler subband architecture (directly built from the spectrum coefficients) was utilized.

The paper is organized as follows: the dysphonic voice corpus used in this paper is first presented in section 2, followed by the baseline classification system in section 3 as well as the subband-based analysis in section 4. Section 5 is dedicated to the experiments and result discussion. Finally, section 6 draws some conclusions and proposes some perspectives.

2. Dysphonic speakers

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 the G parameter of the Hirano's GRBAS scale [6], 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 perceptual grades (20 voices per each), which 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, considering the G parameter of the GRBAS scale uniquely. The judgment was done during one session only.

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 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 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). 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.
All GMM models are composed of 128 gaussian components with diagonal covariance matrices.

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.
The **decision** is made by selecting grade g of model G_g for which the largest likelihood is measured given a test voice.

4. Subband-based analysis

4.1. Objective

The subband-based analysis consists in cutting the frequency domain in subbands processed independently. The main motivation of this approach resides in the assumption that the relevance 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, subband architecture-based approaches have been used for the automatic speech recognition task in adverse conditions, since subbands may be affected differently by noise [9].
In this paper, the subband-based analysis 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 ? ».

4.2. Subband description

In this paper, the full frequency band [0-8000]Hz is split into individual variable width subbands (e.g. 1000Hz width) on which the classification system (described in section 3) is applied afterwards. The linear scale is preferred to the MEL scale in order to keep homogeneous spectral analysis over the subbands.
It has to be noted that the subband-based analysis involved in this paper is different from a subband architecture (as used in [1][9][12]) since the overall classification system is directly applied on a given frequency range instead of utilizing information extracted from the parameterization.

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. Subband-based analysis

In this first experiment, eight 1000Hz-width subbands are processed invidually through the classification system. Classification performance is presented per subband: Table 1 compares performance of the individual subbands and the full band (CCR) while table 2 provides confusion matrices per subband. From these different results, three main trends can be observed:
• Frequency bands between 0 and 3000Hz get the best performance with an overall CCR varying from 55% to 70%. In detail, the subband [0-1000]Hz exhibits 70% CCR for grade 3 voices (similarly to the full band), the subband [1000-2000]Hz gets 95% CCR for grade 0 voices, outperforming the full band rate (85% CCR) (see table 3 for the full band confusion matrix), and 75% CCR for grade 1 voices (vs 55% CCR for the full band). It also provides the best performance for the grade 2 voices with 50% CCR. Furthermore, it can be noted that the classification errors are distributed on the adjacent grades in most cases (e.g. On subband [0-1000]Hz, classification errors for the grade 2 are reported on grades 1 and 3 with 8 and 5 errors respectively).
• Frequencies between 3000 and 5000Hz exhibit the worse overall performance. The normal voices (grade 0) get a satisfactory score of 65% CCR only, despite a loss of performance compared with the full band (85% CCR). On the other side, a strong confusion can be observed for the dysphonic voices, leading to very low scores (20% CCR).
• Frequencies upper than 5000Hz provide better overall performance compared with the previous subbands even though most of the classification errors are scattered over the grades, still demonstrating a large confusion. On the contrary, it can be observed that severe dysphonic voices (grade 3) are well classified in both subbands between 5000 and 7000 Hz (70% CCR) and [7000-8000]Hz (80% CCR, best score).

Finally, figure 1, which summarizes results (number of correct classification per grade and per individual subband) highlights (1) the difficulties to classify grade 2 voices whatever the individual subband considered, (2) the ability of low frequencies to discriminate most of the voices, except grade 2 voices, (3) the "surprising" performance of grade 3 voices on high frequencies despite the low rate of speech in this zone.

5.2. Joint frequency bands

This section focuses on the three frequency zones highlighted in the previous section. Here, the classification scheme is

Figure 1: Voices correctly classified from the 4-G classification following different frequency subbands (LFSC parameters)

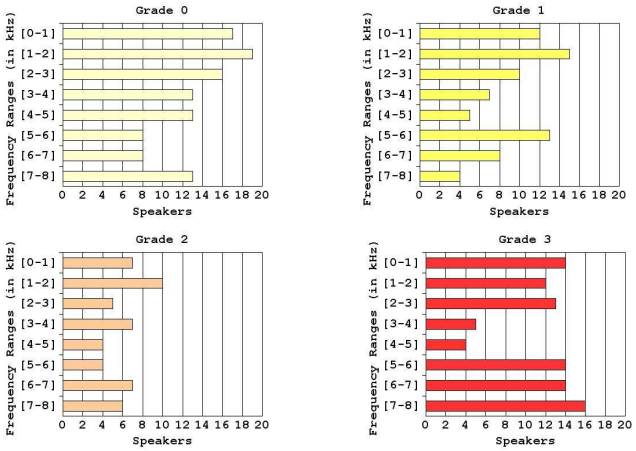

Table 1: 24LFSC - Results of the 4-G classification following different frequency subbands in terms of % CCR

	Grade 0	Grade 1	Grade 2	Grade 3	Total
24LFSC	**% CCR** (nb/20)	**% CCR** (nb/20)	**% CCR** (nb/20)	**% CCR** (nb/20)	**% CCR** (nb/80)
Full Band	85.0 (17)	55.0 (11)	**50.0 (10)**	70.0 (14)	65.00 (52)
[0-1000]Hz	**85.0 (17)**	**60.0 (12)**	35.0 (7)	**70.0 (14)**	62.50 (50)
[1000-2000]Hz	**95.0 (19)**	**75.0 (15)**	**50.0 (10)**	60.0 (12)	**70.00 (56)**
[2000-3000]Hz	80.0 (16)	50.0 (10)	25.0 (5)	65.0 (13)	55.00 (44)
[3000-4000]Hz	65.0 (13)	35.0 (7)	35.0 (7)	25.0 (5)	40.00 (32)
[4000-5000]Hz	65.0 (13)	25.0 (5)	20.0 (4)	20.0 (4)	32.50 (26)
[5000-6000]Hz	40.0 (8)	**65.0 (13)**	20.0 (4)	**70.0 (14)**	48.75 (39)
[6000-7000]Hz	40.0 (8)	40.0 (8)	35.0 (7)	**70.0 (14)**	46.25 (37)
[7000-8000]Hz	65.0 (13)	20.0 (4)	30.0 (6)	**80.0 (16)**	48.75 (39)
[0-3000]Hz	**90.0 (18)**	**65.0 (13)**	**65.0 (13)**	65.0 (13)	**71.25 (57)**
[3000-5400]Hz	65.0 (13)	40.0 (8)	25.0 (5)	65.0 (13)	48.75 (39)
[5400-8000]Hz	65.0 (13)	35.0 (7)	45.0 (9)	**70.0 (14)**	53.75 (43)

performed on the following frequency subbands: [0-3000]Hz, [3000-5400]Hz and [5400-8000]Hz. This experiment aims at taking benefit of the complementarity of individual subbands. Tables 1 and 3 report CCR and the confusion matrices per frequency band respectively, on which it can be pointed out that:

• the [0-3000]Hz band, mainly covering the formant zone, is the most interesting frequency band. First, an overall 71.25% CCR is reached (compared with 65% CCR on [0-8000]Hz and 70% CCR on [1000-2000]Hz). Secondly, grade 2 voices reach their best CCR (65% CCR vs 50% for both the full band and the best individual subband [1000-2000]Hz); Finally, the joint use of the individual subbands results in classification performance more homogeneous and satisfactory along the different grades, especially regarding the grade 2 voices.

• the [3000-5400]Hz band, mainly related to the fricative and plosive zone, gets the lowest overall CCR (48.75%) compared with the other bands. Confusion observed in the individual subbands is still present, except for the grade 3 voices, which tend to take benefit of the complementarity of the individual subbands (65% CCR vs 25% and 20%).

• the [5400-8000]Hz band, related to the residual zone of fricatives and plosives, provides reasonable performance for the normal (65% CCR) and severe dysphonic voices (70% CCR). Regarding the speech information carried by this band, grade 3 voice CCR may be explained by the resulting noise of the « veiled » (or « blown ») features of severe dysphonic voices. In contrary, it is more difficult to explain the behavior of the normal voices in this band, except by a « discriminant lack of information ».

Table 2: Confusion matrices of the 4-G classification following different frequency subbands (LFSC parameters)

[0-1000]Hz

	RG0	RG1	RG2	RG3
TG0	**17**	3	0	0
TG1	2	**12**	5	1
TG2	0	8	**7**	5
TG3	2	2	2	**14**

[1000-2000]Hz

	RG0	RG1	RG2	RG3
TG0	**19**	1	0	0
TG1	2	**15**	1	2
TG2	1	3	**10**	6
TG3	0	1	7	**12**

[2000-3000]Hz

	RG0	RG1	RG2	RG3
TG0	**16**	3	1	0
TG1	5	**10**	3	2
TG2	5	3	**5**	7
TG3	0	0	7	**13**

[3000-4000]Hz

	RG0	RG1	RG2	RG3
TG0	**13**	5	2	0
TG1	7	**7**	4	2
TG2	4	8	**7**	1
TG3	0	9	6	**5**

[4000-5000]Hz

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

[5000-6000]Hz

	RG0	RG1	RG2	RG3
TG0	**8**	7	4	1
TG1	3	**13**	3	1
TG2	4	7	**4**	5
TG3	1	1	4	**14**

[6000-7000]Hz

	RG0	RG1	RG2	RG3
TG0	**8**	4	6	2
TG1	5	**8**	3	4
TG2	4	4	**7**	5
TG3	0	2	4	**14**

[7000-8000]Hz

	RG0	RG1	RG2	RG3
TG0	**13**	2	3	2
TG1	9	**4**	6	1
TG2	5	3	**6**	6
TG3	1	1	2	**16**

5.3. Application to the complete system

According to the frequency analysis performed above, the complete classification system is applied on the best frequency band: [0-3000]Hz and compared with the full band [0-8000]Hz-based system. 24 spectrum coefficients plus first and second derivative coefficients are utilized for classification, involving either a Linear or MEL scale.

Table 4 gives performance in terms of CCR for each configuration. Here, we can observe that [0-3000]Hz band permits to improve classification performance over all the grades compared with the full frequency band ([0-8000]Hz) using both MEL and Linear scales. The best performance is reached by the MFSC coefficients, which, coupled with the derivative coefficients, gets 80% CCR (against 76.25% on the full band). The performance gain classically brought by using the derivative coefficients (Δ and $\Delta\Delta$) is still observed here.

Table 3: 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**

[3000-5400]Hz

	RG0	RG1	RG2	RG3
TG0	**13**	6	1	0
TG1	8	**8**	2	2
TG2	7	6	**5**	2
TG3	2	1	4	**13**

[5400-8000]Hz

	RG0	RG1	RG2	RG3
TG0	**13**	4	3	0
TG1	8	**7**	4	1
TG2	5	3	**9**	3
TG3	0	1	5	**14**

[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**

Table 4: Comparison between LFSC and MFSC - Results of the 4-G classification acoording to frequency bands [0-8000]Hz and [0-3000]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 + 24ΔΔ	**95.0 (19)**	50.0 (10)	55.0 (11)	**75.0 (15)**	68.75 (55)
24MFSC + 24Δ + 24ΔΔ	**95.0 (19)**	**60.0 (12)**	**75.0 (15)**	**75.0 (15)**	**76.25 (61)**
Parameter [0-3000Hz]	**% CCR** (nb/20)	**% CCR** (nb/20)	**% CCR** (nb/20)	**% CCR** (nb/20)	**% CCR** (nb/80)
24LFSC + 24Δ + 24ΔΔ	**95.0 (19)**	**75.0 (15)**	50.0 (10)	**85.0 (17)**	76.25 (61)
24MFSC + 24Δ + 24ΔΔ	**95.0 (19)**	70.0 (14)	**70.0 (14)**	**85.0 (17)**	**80.00 (64)**

6. Conclusion

In this paper, the authors propose to study how the acoustic characteristics of dysphonia are spread out along the frequency domain by analyzing the performance of an automatic dysphonic voice classification on different frequency ranges. This subband analysis outlines that low frequencies tend to be the most interesting zones, leading to an homogeneous discrimination between voices. Additional experiments, involving a more complex parameterization (MFSC plus Δ and ΔΔ), 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 [11] in order to evaluate how the dysphonia 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. [5400-8000]Hz for the grade 3), relevance 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 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 [7], which could bring some explanations to the observed behaviors.

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] Godino-Llorente J.I., et al. 2006. Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters. *EEE Trans. on Biomedical Engineering, Vol53(3), pp.1943-1953*.

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

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

[8] 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.

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

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

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

[12] Pouchoulin, G., et al. 2007. *Characterization of pathological voices (dysphonia) in the frequency space*. ICPHS'07, August 2007, Saarbrücken.

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

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

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

[16] 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.