# Are the unvoiced consonants relevant for dysphonia phenomenon observation?

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

# Are the unvoiced consonants relevant for dysphonia phenomenon observation?

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

[1]University of 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 speech signal for a better understanding of related phenomena while most of the studies have focused only on improving classification systems for diagnosis help purposes. This work is focused on an automatic and manual phonetic analysis, which highlights the potential and rather unexpected relevance of unvoiced consonants in the automatic classification task of dysphonia severity grades (based on the GRBAS scale).

**Index Terms**: Voice disorder, dysphonia characterization, automatic dysphonic voice classification, automatic and manual phonetic analysis, GRBAS scale

## 1. Introduction

In the medical domain, assessment of the pathological voice quality is a sensitive topic, involving multi-disciplinary domains: clinicians, phoneticians, and automatic speech processing specialists. Dysphonia is one of the medical terms to define phonation disorders. Currently, two main approaches are involved to deal with the phonation disorders. The first approach, the perceptual evaluation [1, 2], consists in qualifying and quantifying the vocal dysfunction by listening to the patient's speech production. It is the most used by the clinicians currently even if it is largely debated in the literature because of an intrinsic subjectivity, a lack of a universal scale, as well as a large intra and inter-variability in the human judgments. The second methodology, the objective measurement-based analysis [3, 4, 5, 6, 7, 8], consists in associating acoustic, aero-dynamical and/or physiological measures with an automatic classification system to provide a reproducible and objective decision. Even if the latter has been introduced as an alternative to the perceptual evaluation, the current performance of the automatic systems has not been sufficient yet to satisfy clinicians. This paper, related to the objective measurement-based analysis, aims at characterizing phenomena due to the dysphonia in the speech signal. Here, the authors propose to analyze the performance of the automatic classification system (according to the GRBAS scale) according to different phonetic classes, mixing voiced and unvoiced components. The goal of this study is to bring a better understanding of the dysphonia phenomena in the speech signal, helpful for the enhancement of the automatic classification systems, as well as for the medical domain.

## 2. Dysphonic voice corpus

Speakers involved in this study were, on the one hand, dysphonic women (aged 17 to 50) affected by nodules, polyps, oedema, cysts, ... and, on the other hand, control women (normal voice). They were recorded in the ENT department of the Timone University Hospital, Marseille, France. 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 subjects' voices are perceptually classified along the G parameter of the Hirano's GRBAS scale [9], where a normal voice is rated as grade G0, a slight dysphonia as G1, a moderate dysphonia as G2 and finally, a severe dysphonia as G3. 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.
For the following experiments, 80 voices equally balanced among the 4 grades (20 voices per each) were selected. Due to the small amount of data, cautions have been made to provide valid protocol and experimental results by applying leave_x_out technics. The latter consists in discarding *x* speakers' voices from the 80 voices, using the remaining data for training and the *x* speakers' voices for testing. This scheme is repeated until reaching a sufficient number of tests.

## 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 » (ALIZE/SpkDet [10]) and developed at the LIA laboratory. Three phases are necessary:
**Parameterization:** in this paper, 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 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 $g \in \{0, 1, 2, 3\}$. A GMM is a weighted sum of M multi-dimensional Gaussian distributions, each characterized by mean vector $\overline{x}$ (dimension *d*), covariance matrix $\Sigma$ (*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 ($\overline{x}$, $\Sigma$, $p$) thanks to the EM/ML algorithm (Expectation-Maximization/Maximum Likelihood).
Grade model $G_g$ is learned gathering all the voices evaluated perceptually as grade $g$. Two training phases are used here to cope with the lack of training data, as classically used in speaker recognition domain [11]: (1) training of a generic speech model estimated by the EM/ML algorithm on a large population of speakers; (2) training of the grade model, derived from the generic speech model by applying adaptation techniques (MAP, Maxi-

Table 1: *Total duration in seconds per phonetic class and per grade as well as the number of phonemes (nb), their averaged duration ($\mu$) and associated standard deviation ($\sigma$).*

| Phonetic classes | Grades | | | | Info. per class | | |
|---|---|---|---|---|---|---|---|
| | G0 | G1 | G2 | G3 | nb | $\mu$ | $\sigma$ |
| Consonant | 135.1 | 139.2 | 149.8 | 167.3 | 6395 | 0.092 | 0.045 |
| . Voiced | 88.8 | 90.6 | 95.4 | 106.6 | 4719 | 0.081 | 0.039 |
| . Unvoiced | 46.3 | 48.7 | 54.5 | 60.7 | 1676 | 0.125 | 0.046 |
| Vowel | 103.6 | 98.8 | 103.5 | 109.8 | 5586 | 0.074 | 0.046 |
| All phonemes | 241.5 | 241.0 | 256.7 | 280.5 | 12140 | 0.084 | 0.046 |

mum a posteriori).
All GMM models are composed of 128 gaussian components with diagonal covariance matrices. It has to 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.
**Decision:** For dysphonic voice classification, decision is made by selecting the grade *g* of the model $G_g$ (among the four grade models) for which the largest similarity measure is computed given a test voice. 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 $i$ in the mixture.

## 4. Phoneme-based classification

The authors propose in this paper to observe the behaviour of the automatic dysphonic voice classification system following different phoneme classes. Thus, system performance is provided per phoneme class in order to evaluate how the dysphonia effects may impact on phonemes or phoneme classes according to the grades.

### 4.1. Automatic phoneme segmentation

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. In other words, phoneme boundaries of each expected words uttered by speakers are extracted automatically in a unsupervised way. This alignment was performed by the automatic alignment system developed at the LIA laboratory. This system is based on a Viterbi decoding algorithm, a text-restricted lexicon of words associated with their phonological variants and a set of 38 French phonemes. Each phoneme model relies on a three state HMM, initially trained on a French speech corpus, produced by a set of female speakers. Since the latter has no connection with the dysphonic corpus (described in section 2), classical unsupervised adaptation techniques are applied iteratively on phoneme models for the automatic phonetic alignment to enhance and refine phoneme boundaries.

### 4.2. Comparative phonetic analysis

The phonetic segmentation is coupled with the automatic dysphonic classification system for the decision step. 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 targeted phoneme class. Conversely, the grade models are learned on all the phonemic material available per grade in the corpus independently of the phonem class targeted.
In this paper, the authors focus on three main phoneme classes: both voiced and unvoiced consonants, and vowels. These individual phoneme classes are compared with the global set, joining all the phonemes available in the corpus. Table 1 provides duration information on these different phoneme classes per voice and grade. This automatic phonetic analysis of dysphonia is close to the "phonetic labelling" proposed in [2], in which a descriptive and perceptual study of pathological characteristics of different phonemes is presented.

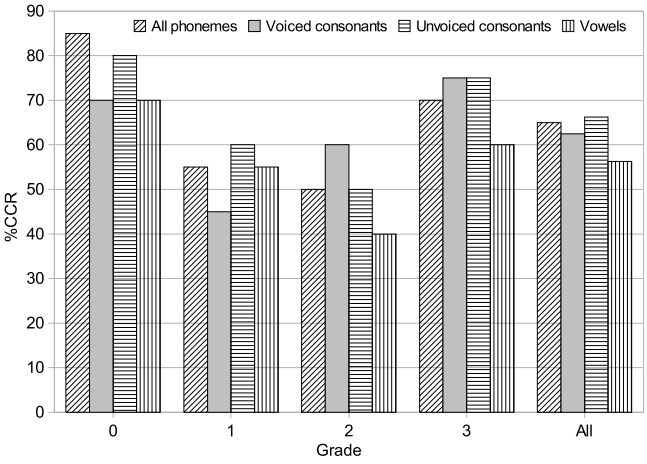


Figure 1: *Performance per grade in terms of Correct Classification Rate (CCR %) considering "All phonemes", both voiced and unvoiced consonant and vowel classes.*

Figure 1 presents performance of the automatic dysphonic voice classification system depending on the different phoneme classes: all phoneme set, both voiced and unvoiced consonants, and vowels. Performance is expressed in terms of Correct Classification Rates (named $CCR$ in the rest of the paper).
From this figure, it can be observed that:
• best performance (85% CCR) is obtained for grade 0 voices considering all the phoneme while other phoneme classes reach between 70 and 80% CCR. Performance of grade 3 voices is rather close (from 60 to 75% CCR) while more confusion is observed on grade 1 and 2 voices (CCR varying from 40 to 60%);
• individual classes perform in the similar range of performance as considering all the phoneme set. Indeed, while CCR variation for all the phonemes is between 50 to 85%, CCR of voiced consonants vary from 45 to 75% depending on the grades, CCR of unvoiced consonants from 50 to 80%, and CCR of vowels from 40 to 70%.
• unvoiced consonant class obtains quite high CCR, compared with all the phoneme set. It outperforms the vowel class whatever the grades observed. It obtains higher CCR than the voiced consonant class for grades 0 and 1, and a similar and lower CCR for grades 3 and 2 respectively.

*Discussion*
The performance observed on the unvoiced consonant class is particularly surprising. Indeed, focus is generally made on voiced components in the literature while considering dysphonia since they are directly affected by this kind of pathology related to the glottic source. Sustained vowels are, for instance, extensively used in the literature by perceptual or objective approaches, since they facilite the measurement of parameters directly linked

to the vocal source.
In this context, two hypotheses may be proposed to explain the behaviour of the automatic classification system on the unvoiced consonant class:
• Hyp 1: phoneme boundaries are determined by the automatic text-constraint alignment system. As reported in section 4, the latter has not been developped to process pathological speech specifically. Even if unsupervised techniques have been applied to enhance the quality of the phoneme models, voice alteration due to dysphonia may decrease the system performance, degrading the quality of phoneme boundaries. Mislabelling in the phonetic alignment could involve, for instance in the case of vowel-consonant ($VC$) or consonant-vowel ($CV$) contexts, that parts of vowels were merged with unvoiced components. In this case, previous observations would be biased.
• Hyp 2: dysphonia may have effects on the unvoiced consonant production, especially in the $VC$ or $CV$ contexts. Moreover, it can be assumed that this effect is gradual according to the dysphonia severity since the classification system provides satisfactory discrimination between grades.

# 5. Validation of the phonetic segmentation

This section aims at verifying the first hypothesis outlined in the previous section.

## 5.1. Manual phonetic segmentation

In order to be able to evaluate the performance of the automatic text-constrained alignment system regarding the unvoiced consonants, manual verification is necessary. The focus has been made on 18 unvoiced consonants among the 20 ones uttered by each speaker in the corpus[1]. The 18 unvoiced consonants are split into 8 fricatives (6 /s/, 1 /f/, 1 /ch/), and 10 plosives (2 /p/, 5 /t/, 3 /k/). A human expert listened to all the speech signal in the corpus and marked them with different information, distinguishing in some cases unvoiced fricatives from plosives:
• formant end of each vowel preceding an unvoiced consonant noted $FEV$ (both fricatives and plosives);
• voicing end of each vowel preceding an unvoiced consonant noted $VEV$ (both fricatives and plosives);
• start of noise for the fricative consonants noted $SNO$;
• end of noise for the fricative consonants noted $ENO$;
• plosive burst;
• voicing start of each vowel following an unvoiced consonant noted $VSV$ (both fricatives and plosives);
• formant start of each vowel following an unvoiced consonant noted $FSV$ (both fricatives and plosives);

This manual labelling was performed in a blind way, using the Praat software [12]. Listening, signal visualization, spectrograms, and F0 measurements issued from Praat were utilized. No information about the boundaries determined by the automatic system was made available for the human expert to avoid any influence. Only rough location of unvoiced consonants in the speech signal was provided to save time.
Globally, 1440 unvoiced consonants have been processed in the overall corpus, corresponding to 7840 markers potentially. Nevertheless, the human expert was not able to fix the overall set of markers, especially due to the voice quality degradation. For instance, in a few contexts, vowels have been totally devoiced, making unavailable information, and corresponding markers.

[1] 2 unvoiced consonants have been discarded here since they appear in particular consonantic groups: /ks/ and /tr/.

| Configuration | Count |
|---|---|
| $VEV > START$ | **1014** |
| $VEV \cong START$ | **67** |
| $VEV < START$ | **107** |
| $END < VSV$ | **899** |
| $END \cong VSV$ | **199** |
| $END > VSV$ | **147** |

Table 2: Comparison between the unvoiced consonant boundaries ($START$ and $END$) issued from the automatic text-constrained alignment system and markers fixed by a human expert ($VEV$ represents the voicing end of the vowel preceding an unvoiced consonant and $VSV$ the voicing start of the vowel following an unvoiced consonant). Possible configurations and their counts.

## 5.2. Manual vs automatic phonetic segmentation

The boundaries of unvoiced consonants, issued from the automatic text-constrained alignment system and noted $START$ and $END$ are compared with both the $VEV$ and $VSV$ markers, in terms of time location. Assuming that boundaries of unvoiced consonants are synchronized with the end and start of vowel voicing, this comparison aims at pointing out whether parts of vowels are included in the unvoiced consonants according to the following configurations:
• $VEV > START$ : the unvoiced consonant includes the final part of the preceding vowel;
• $VEV \cong START$ : the start boundary of the unvoiced consonant is considered as correct;
• $VEV < START$ : the beginning part of the unvoiced consonant is missing;
• $END < VSV$ : the final part of the unvoiced consonant is missing;
• $END \cong VSV$ : the final boundary of the unvoiced consonant is considered as correct;
• $END > VSV$ : the unvoiced consonant includes the beginning part of the following vowel.

Table 2, providing counts relating to each configuration, shows that the automatic text-constrained alignment system used to misplace the start of the unvoiced consonants rather systematically (1014 times on 1188 measures), including the final part of the preceding vowel. In contrast, the beginning part of the following vowel is rather excluded (included in 147 times only on 1245 measures) whereas the final part of the unvoiced consonant is most of the times missing (899 times on 1245 measures).

Therefore, these results could confirm the first hypothesis, which assumes that unvoiced consonants could be mislabelled by the automatic system. Next section will verify whether this mislabelling should have an impact on the classification system performance reached by the unvoiced consonants.

## 5.3. Experimental validation

Experiment reported in section 4.2 is reproduced here, focusing on unvoiced consonants only. Considering manual markers, six different case studies are conducted here:
• $Auto1$ and $Auto2$ : classification system is based on the automatic boundaries of unvoiced consonants for which manual markers of the preceding ($Auto1$) or the following ($Auto2$) vowels were available. These experiments provide system performance computed on the same number of unvoiced conso-

| Study cases | Grade 0 % CCR (nb/20) | Grade 1 % CCR (nb/20) | Grade 2 % CCR (nb/20) | Grade 3 % CCR (nb/20) | Total % CCR (nb/80) |
|---|---|---|---|---|---|
| $Auto1$ | 75 (15) | 55 (11) | 60 (12) | 80 (16) | 67.5 (54) |
| $VEV1$ | 75 (15) | 55 (11) | 60 (12) | 80 (16) | 67.5 (54) |
| $VEV2$ | 65 (13) | 55 (11) | 60 (12) | 80 (16) | 65 (52) |
| $Auto2$ | 75 (15) | 55 (11) | 50 (10) | 80 (16) | 65 (52) |
| $VSV1$ | 75 (15) | 55 (11) | 55 (11) | 85 (17) | 67.5 (54) |
| $VSV2$ | 75 (15) | 55 (11) | 55 (11) | 85 (17) | 67.5 (54) |
| $VEV - VSV$ | 65 (13) | 55 (11) | 60 (12) | 85 (17) | 66.25 (53) |

Table 3: Performance of the automatic classification system considering the unvoiced consonants only. Performance is reported for different study cases considering automatic segmentation before and after corrections according to the manual markers.

| Study cases Duration in s | Grade 0 | Grade 1 | Grade 2 | Grade 3 | Total |
|---|---|---|---|---|---|
| $Auto1$ | 37.09 | 39.25 | 44.05 | 37.37 | 157.76 |
| $VEV1$ | 38.24 | 40.23 | 44.96 | 38.8 | 162.23 |
| $VEV2$ | 27.41 | 33.33 | 37.91 | 33.92 | 132.57 |
| $Auto2$ | 37.69 | 41.10 | 44.51 | 38.68 | 161.98 |
| $VSV1$ | 42.60 | 46.88 | 50.04 | 45.30 | 184.82 |
| $VSV2$ | 42.22 | 46.46 | 49.46 | 44.57 | 182.71 |
| $VEV - VSV$ | 29.59 | 37.11 | 39.68 | 31.57 | 137.95 |

Table 4: Duration of the unvoiced consonants considering automatic segmentation before and after correction according to the manual markers.

nants as the rest of study cases;
• $VEV1$ : similar to $Auto1$ except that segmentation of unvoiced consonants for which the beginning part is missing (107 cases) is corrected according to the manual markers;
• $VEV2$ : similar to $VEV1$ except that segmentation of unvoiced consonants which include the final part of the preceding vowels ($VEV > START$: 1014 cases) is corrected;
• $VSV1$ : similar to $Auto2$ except that segmentation of unvoiced consonants for which the final part is missing (899 cases) is corrected;
• $VEV2$ : similar to $VSV1$ except that segmentation of unvoiced consonants which include the beginning part of the following vowels ($END > VSV$: 147 cases) is corrected;
• $VEV - VSV$: segmentation of unvoiced consonants is directly issued from the manual markers : $VEV$ and $VSV$.

Tables 3 and 4 report classification system performance in terms of CCR% and global durations of unvoiced consonants related to each case study respectively. Very few variation in terms of system classification performance can be observed comparing the different case studies.

Correcting the segmentation of the unvoiced consonants by discarding all the mislabelled components (ie the vowel components) permits to reach very similar performance to the one obtained in section 4.2. This enables to invalidate the first hypothesis raised previously, highlighting the second hypothesis : dysphonia may have impact on the unvoiced consonant production, especially in the $VC$ or $CV$ contexts.

# 6. Conclusion

This paper aims at studying the characterization of dysphonia voices through an automatic classification system coupled with a phonetic analysis. Comparing system performance according to vowel, voiced and unvoiced consonant classes shows that unvoiced consonants outperform vowels and reach relatively high classification performance. The first hypothesis, raised by the authors to explain this quite surprising behaviour and directly linked to the quality of the automatic phonetic segmentation has been invalidated thanks to a manual correction done by a human expert. Indeed, additional experiments have still demonstrated the relevance of unvoiced consonants in the classification task. The second hypothesis, raised by the authors, assuming that dysphonia may impact on the unvoiced consonant production, especially in the vowel-consonant $VC$ or consonant-vowel $CV$ contexts is therefore emphasized. Further work will consist in examining unvoiced consonant components of the speech corpus manually with the help of phonetician experts.

# 7. 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 ELS*, vol. 258, pp. 77–82, 2001.

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

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

[4] 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.

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

[6] 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.

[7] 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.

[8] G. Pouchoulin et al., "Dysphonic voices and the 0-3000hz frequency band," *Interspeech'08, Brisbane, Australia*, 2008.

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

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

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

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