Back-and-Forth Methodology for Objective Voice Quality Assessment: From/to Expert Knowledge to/from Automatic Classification of Dysphonia

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

Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 982102, 13 pages
doi:10.1155/2009/982102

Research Article

Back-and-Forth Methodology for Objective Voice Quality Assessment: From/to Expert Knowledge to/from Automatic Classification of Dysphonia

Corinne Fredouille,[1] Gilles Pouchoulin,[1] Alain Ghio,[2] Joana Revis,[2] Jean-François Bonastre,[1] and Antoine Giovanni[2]

[1] *Laboratoire Informatique d'Avignon (LIA), University of Avignon, 84911 Avignon, France*
[2] *LPL-CNRS, Aix-Marseille University, 13604 Aix-en-Provence, France*

Correspondence should be addressed to Corinne Fredouille, corinne.fredouille@univ-avignon.fr

Received 31 October 2008; Revised 1 April 2009; Accepted 10 June 2009

Recommended by Juan I. Godino-Llorente

This paper addresses voice disorder assessment. It proposes an original back-and-forth methodology involving an automatic classification system as well as knowledge of the human experts (machine learning experts, phoneticians, and pathologists). The goal of this methodology is to bring a better understanding of acoustic phenomena related to dysphonia. The automatic system was validated on a dysphonic corpus (80 female voices), rated according to the GRBAS perceptual scale by an expert jury. Firstly, focused on the frequency domain, the classification system showed the interest of 0–3000 Hz frequency band for the classification task based on the GRBAS scale. Later, an automatic phonemic analysis underlined the significance of consonants and more surprisingly of unvoiced consonants for the same classification task. Submitted to the human experts, these observations led to a manual analysis of unvoiced plosives, which highlighted a lengthening of VOT according to the dysphonia severity validated by a preliminary statistical analysis.

1. Introduction

Assessment of voice quality is a key point for establishing telecommunication standards as well as for medical area linked to speech and voice disorders. In the telecommunication field, voice quality assessment is mainly addressed at the perceptual level using the Mean Opinion Score (MOS) scale [1] standardized by the International Telecommunication Union (ITU). The evaluation of voice quality is done by a jury composed of nonspecialized listeners. Several algorithms were proposed in order to move from this human perception-based-measure to an automatic measure to reduce costs as well as to move from a subjective to an objective method. The most known algorithm is the Perceptual Evaluation of Speech Quality (PESQ) [2] also normalized by the ITU. The effectiveness of PESQ is measured by the correlation of the MOS measures obtained by a human jury and using the PESQ. If the PESQ (and its extensions) is well suited for the telecommunication field, it requires parallel audio records without and with noise disturbance to evaluate voice quality. This constraint is of course impossible to satisfy in the medical/pathological area. However, independently of this difference it is interesting to notice that the MOS/PESQ is estimated at the perceptual level and that there is no analytical description of information at the acoustic or phonetic field characterizing a given level of quality. In fact, the human subjective perception is used as a baseline (MOS) and an automatic approach (PESQ) is used to match some signal differences with the MOS scale.

In the field of voice disorder assessment, a general approach very similar to the one used in telecommunications is followed. Human experts are committed to evaluating quality of speech samples at the perceptual level, generally implying approaches based on the expertise of researchers and practitioners. Three main drawbacks of this scheme could be highlighted: assessment remains subjective, costly

(if an expert jury is involved), and not analytical, that is, the judgment may be global or not based on a standardized set of criteria. As opposed to the telecommunication area for which a standard scale (MOS) is proposed, only very few assessment scales [3–5] are to be found in the pathological field, which are generally accepted but not really considered as a standard due to the large diversity of pathological voice disorders and to the intrinsic difficulty to characterize some pathologies.

This paper describes the three points highlighted previously, by proposing a general approach based on both the human expertise and automatic voice classification approach. The proposed scheme allows to automate the voice quality estimate like PESQ, and to move from a subjective to an objective approach. Moreover, the most interesting part is to use the automatic approach in order to support the human expertise by highlighting some specific acoustic aspects of the addressed pathology or class of pathologies. Figure 1 presents the first part of the proposed scheme. The automatic voice classification system is fed by the pathological voice examples associated with the perceptual labels given by the human experts. A feedback loop is proposed to assess the ability of the automatic system in the classification task. Of course, several iterations involving inputs of machine learning experts are needed to obtain a satisfactory system.

Figure 2 illustrates the second part of the proposed scheme. Here, the automatic voice classification system is applied on a set of voice examples to produce analytical information. This information is given to the experts through a second feedback loop and associated with statistical measures and voice excerpts. It allows experts to listen to and/or to analyze manually small parts of a large speech database only in order to assess the interest of one information. Depending on the previous results, experts—with machine learning specialists—could change the feature selection and allow the system to output targeted information.

This scheme can be applied to any kind of pathology under a couple of main constraints: (1) enough expert knowledge is available (to seed the automatic classification system), (2) a good/large enough corpus is available.

In this paper, we focus on dysphonia—an impairment of the voice—for two main reasons. Firstly, dysphonia respects the constraints reported previously. Secondly, even if dysphonia is often considered as a "minor" trouble linked to an esthetic point of view, this pathology has a drastic impact on the patient's quality of life. An explanation of this subconsideration of dysphonia is that voice quality is generally described as a paralinguistic phenomenon with little impact on communication. However, the social relevance and economic impact of voice disorders are now obvious, especially for school teachers or other professionals who use their voice as a primary tool of trade. For instance, a recent study [6] has revealed that 10.5% of the teachers are clearly suffering from voice disorders, when several enquiries [7] show that voice is the primary tool for about 25% to 33% of the working population. In addition to medical and professional consequences, some voice disorders have also severe consequences regarding social activities and interaction with others. It is the reason why voice therapy is an important issue in a social, economical, clinical contexts, and among voice therapy activities, voice assessment is an important part of this clinical and scientific challenge.

A large set of methods can be used to assess voice disorders like discussion with the patient, endoscopic examination of the larynx, postural behavior of the patient [8], psychological and behavioral profile [9], auto-evaluation as Voice Handicap Index questionnaire [10], perceptual judgment [11], or instrumental assessment [12]. It is preferable to increase the fields of observation in order to take the multidimensional aspect of the spoken communication into account. Indeed, an assessment method taken individually appears as a reduced view of the voice disorder and provides only a part of the truth.

The perceptual dimension of voice is an essential aspect for the vocal evaluation as speech and voice are produced to be perceived. Evaluating voice without studying the impact on listeners amounts to lose its "raison d'être". Moreover, the majority of dysphonic speakers decides to consult a practitioner when their entourage hears changes in their vocal production on perceptive feelings only. In the same way, practitioners appreciate therapeutic results mainly listening to the patients' voice: auditory perception is the first and the most accessible method to evaluate vocal quality. Lastly, the human being and his/her perceptive system are powerful to decode speech [12]. However, the perceptual judgment remains a controversial method because of various drawbacks, notably its subjectivity [13, 14].

The multiparametric instrumental analysis represents an alternative solution to quantify vocal dysfunctions [15]. Methods can be based on acoustic measurements but also on aerodynamic parameters or electrophysiological signals. These measurements are carried out for vocal production with sensors designed to record and compute multiple parameters issued from the speech production. The majority of studies in this domain outlines the necessity of combining various complementary measures in order to take the multidimensional properties of vocal production into account [15–20]. In a recent study [21], we have applied a perceptual assessment (GRBAS scale [3]) on 449 voice samples including 391 patients with a voice disorder recorded in the ENT Department of the Timone University Hospital Center in Marseille (France). Concurrently, on the same cohort of patients, an instrumental voice analysis was carried out using the EVA workstation (SQLab-LPL, Aix-en-Provence, France). The subject was instructed to pronounce three consecutive sustained vowels and several consecutive/pa/. For more than 80% of this population, the grade proposed by perceptual and instrumental assessment was concordant, which was considered as an acceptable result by our practitioners for a clinical use. It is important to notice that the state of the art of such instrumental approaches allows only nonnatural, noncontinuous speech materials when studying the different phenomena on natural continuous speech is of a large interest.

This limitation encourages the authors to get interest to automatic speaker and speech recognition techniques for dysphonic voice characterization. Indeed, in addition to analytical instrumental assessment approaches, another kind of methods, mainly drawing upon both automatic

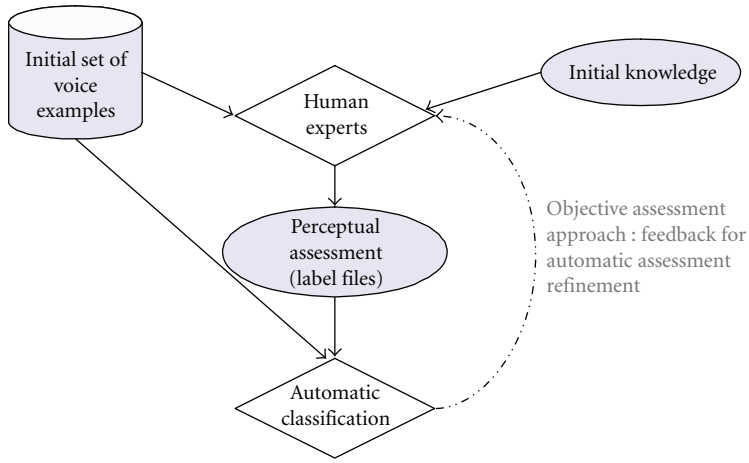

Figure 1: Objective assessment of voice quality based on an automatic classification approach and the human expertise.

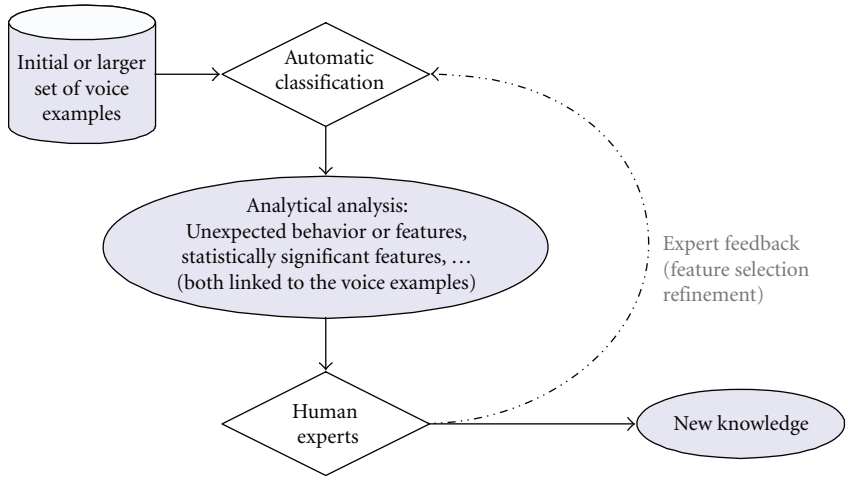

Figure 2: The "automatic system to human experts/knowledge" feedback loop.

speech processing and pattern recognition domains, has been proposed in literature. Mostly dedicated to voice disorder detection, these approaches rely on automatic acoustic analyses such as spectral [22, 23], cepstral [24–27], or multidimensional acoustic analysis inspired from the analytic instrumental assessment (F0, jitter, shimmer, or Harmonic Noise Ratio) [25, 28–30], combined with automatic classifiers based on Linear Discriminant Analysis (LDA) [28, 29], Hidden Markov Models (HMM) [24, 28], Gaussian Mixture Models (GMM) [24, 26], Support Vector Machines (SVM) [31], or Artificial Neural Networks (ANN) [23, 24, 30].

Compared with the analytic instrumental assessment methods described previously, originality and interest of these automatic classification-based-assessment approaches are: (1) the ability to analyze continuous speech near to natural elocution, (2) the ability to process a large set of data, authorizing studies on a large-scale and significant statistical results, (3) a simple and automatic acoustic analysis providing an easy-to-use and noninvasive tool for clinical use.

In this paper, we present a complete approach based on the "back-and-forth methodology" we have just presented. Section 2 is dedicated to the dysphonic voice classification system. Section 3 describes the experimental protocols as well as an experimental validation of the first part of the method: the objective assessment of dysphonic voices. The next section presents the core of our method, which aims to gather new knowledge about dysphonia. Finally, Section 5 concludes this paper and presents some future works.

2. Dysphonic Voice Classification System

The system presented below is part of the automatic system-based assessment approaches previously defined and is involved in the "back-and-forth" methodology. The principle retained here is to adapt a state of the art speaker recognition system to the dysphonic voice classification task. A speaker recognition system can be seen as a supervised classification process able to differentiate speech signals between classes. A class of signal generally belongs to a given speaker and is modeled using a set of examples from the latter. In

some cases, a composite class could be necessary (associated with several speakers) and modeled by grouping several classes modeled independently or by modeling a unique composite class on all the signals of speakers belonging to this class. Two adaptation levels are necessary to suit a speaker recognition system to the dysphonic voice classification task. Firstly, a class does not longer correspond to a given speaker but to a specific pathology or a severity grade of this pathology. The class is then modeled using data from a set of speakers affected by the corresponding pathology or severity grades. Obviously, voices used for training a pathological class cannot be included in the set of tested voices in order to differentiate pathology detection from speaker recognition. The second adjustment to apply to the speaker recognition system is the representation of speech data, which can be optimized for the voice disorder classification task.

The speaker recognition technique used in this study is based on the statistical Gaussian Mixture based modeling, which remains one of the state of the art alternative solutions for speaker recognition [32]. This approach consists in three phases:

(i) a parameterization phase;

(ii) a modeling phase;

(iii) a decision/classification phase.

2.1. Parameterization Phase. The parameterization phase is necessary to extract relevant information from the speech signal. Here, it is based on a short-term spectral analysis resulting on 24 frequency spectrum coefficients and performed as follows. The speech signal, sampled at 16 kHz, is first emphasized by applying a filter, which aims to enhance the high frequencies of the spectrum generally reduced for the speech production. This filter is defined as: $x(t) = x(t) - k \cdot x(t-1)$ with k fixed at 0.95 empirically. The speech signal is then windowed by using a 20 millisecond Hamming window, shifting at a 10 millisecond rate. The goal of the Hamming window is to reduce the side effects. The latter facilitates the application of a Fast Fourier Transform (FFT) locally on each window (512 points) and the computation of the FFT modulus leading to a power spectrum. This power spectrum is multiplied by a filterbank (series of passband frequency filters) in order to extract the envelope of the spectrum. Here, 24 triangular filters are used. According to experiments, they are either spaced linearly on the 8 kHz frequency band (referred to as LFSC standing for Linear Frequency Spectrum Coefficients in this paper), or spaced according to a MEL scale (referred to as MFSC standing for Mel Frequency Spectrum Coefficients), wellknown to be closer to the frequency scale of the human ear.

The feature vectors issued from this analysis, at a 10 millisecond rate, are finally normalized to fit a 0-mean and 1-variance distribution, coefficient by coefficient (means and variances are estimated on the non-silence signal portions). Classically, this normalization is employed to reduce the effect of recording channels and facilitates the following statistical process.

The LFSC/MFSC computation is done by using the (GPL) SPRO toolkit [33]. Finally, the feature vectors can be augmented by adding dynamic information representing the way these vectors vary in time. Here, first and second derivatives of static coefficients are considered (also named Δ and $\Delta\Delta$ coefficients) resulting in 72 coefficients.

2.2. Modeling Phase. Classically, this phase aims to estimate models of targeted classes like individual speakers in speaker recognition. In this paper, models represent either a set of pathological/normal voices or a set of voices related to a specific severity grade.

Modeling relies on Gaussian Mixture Models (GMM) and estimate techniques drawing upon the speaker recognition domain. In this context, a set of D-dimensional feature vectors, denoted by $X = x_1, \ldots, x_T$, is represented by a weighted sum of M multidimensional Gaussian distributions. Each distribution is defined by a $D \times 1$ mean vector μ_i, a $D \times D$ covariance matrix Σ_i, and a weight w_i of the distribution inside the mixture. The set of distributions and related parameters, also called the Gaussian Mixture Model, is denoted $\lambda = (w_i, \mu_i, \Sigma_i)$, $i = (1, \ldots, M)$. The modeling phase consists in estimating all these parameters according to training data.

In speaker recognition, two-step modeling is typically applied for the model parameter estimate to improve their robustness, especially when a small amount of training data is available for some specific classes, as follows.

(i) Parameters of a GMM model are first estimated on a large amount of speech signal, issued from a generic population of speakers. This generic speech model, also called Universal Background Model (UBM), tends to represent the speaker-independent space of acoustic features. It is generally trained using the iterative Expectation-Maximization (EM) algorithm [34] associated with the Maximum Likelihood criterion (ML).

(ii) A speaker model is then derived from this UBM model by involving adaptation techniques like MAP [35] and the small amount of training data available for the given speaker. In practice, only the mean parameters are updated while covariance matrices and distribution weights remain generally unchanged, directly issued from the UBM model. The mean adaptation relies on a combining function involving mean values issued from both the UBM models and the speaker training data.

In this paper, the same scheme is applied due to the small amount of training data available for pathological and control speech (see Section 3.1 for more details on the corpus). The UBM model parameters are estimated on a French read-speech corpus composed of 76 female speech utterances of 2 minutes each. This female population is extracted from the BREF corpus [36], which is entirely separate from the dysphonic corpus and the targeted task. All the GMM models are composed of 128 Gaussian components with diagonal covariance matrices.

Regarding the dysphonic voice classification task, a GMM model will be estimated per class of information targeted (for instance, a GMM model per grade of dysphonia severity).

2.3. Decision/Test Phase. During this phase, a set of new feature vectors $Y = y_1, \ldots, y_T$ associated with an incoming speech signal is presented to the system and compared with one or several GMM models λ. This comparison consists in computing the averaged frame-based likelihood, denoted $L(Y \mid \lambda)$, as follows:

$$L(Y \mid \lambda) = \frac{1}{T}\sum_{t=1}^{T} L(y_t \mid \lambda) \tag{1}$$

with

$$\begin{aligned} L(y_t \mid \lambda) &= \sum_{i=1}^{M} w_i \cdot L_i(y_t), \\ L_i(y_t) &= \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \\ &\quad \times \exp\left\{-\frac{1}{2}(y_t - \mu_i)^T(\Sigma_i)^{-1}(y_t - \mu_i)\right\}, \end{aligned} \tag{2}$$

where $L_i(y_t)$ represents the likelihood of frame y_t according to the ith Gaussian distribution of the model $\lambda = (w_i, \mu_i, \Sigma_i)$, $i = (1, \ldots, M)$.

In the context of the dysphonic voice classification, the classification decision is made by selecting the GMM model λ, and consequently, the class of information associated with for which the largest likelihood measure is computed given the incoming speech signal Y.

3. Experimental Protocol

Results provided in the rest of the paper are expressed in terms of correct classification rates (named CCR).(In tables, the number of well-classified voices is also provided in brackets.) For indication, 95% confidence intervals are provided for the overall CCR only, given the small number of tests available from the corpus used and despite cautions taken by the authors (see Sections 3.1 and 3.2). Finally, it has to be pointed out that all these results are issued from the GMM classifier and have to be interpreted from a statistical viewpoint.

3.1. Corpus. The corpus, called DV, used in this study is composed of read speech pronounced by a set of dysphonic subjects, mostly affected by nodules, polyps, oedemas, and cysts as well as a control group. The subjects' voices are classified according to the G criterion of the Hirano's GRBAS scale [3], 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 choice of the G criterion was driven by two main reasons: (1) it refers to a global quality judgment as opposed to the other criteria (RBAS), which is more suitable regarding the type of parameterization used in this work, (2) like the R and B criteria, it is more robust to intra- and interlistener variability.

The corpus was supplied by the ENT Department of the Timone University Hospital Center in Marseille (France). 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), of which signal duration varies from 13.5 to 77.7 seconds (mean: 18.7 seconds). The 80 voices are equally balanced in the four GRBAS perceptual grades (20 voices per each), which were determined by a jury composed of three 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 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 over all the experiments by applying specific methods like the *leave-one-out* technique.

3.2. Leave-One-Out Technique. As shown in Section 2.2, training data used to learn models of pathological classes have to be separated from testing. In other words, speakers included in the training set should not be present in the testing set. As the DV corpus is relatively small (80 voices), it is not well suited to split it into two separate subsets. Consequently, some special protocols have been designed for different classification tasks (Task1 and Task2) in order to respect this constraint while providing more statistically significant results. These protocols rely on the *leave-one-out* technique, which consists in discarding a speaker, noted x, from the experimental set, in learning some models on the remaining data and in testing data of speaker x using these models. This scheme is repeated until reaching a sufficient number of tests.

3.3. Task1-Protocol P1. Task1 consists in determining whether a given voice is normal or dysphonic. Consequently, two different GMM models have to be estimated: the $\lambda_{\bar{d}}$ (normal) model trained on a subset of G_0 voices and λ_d (dysphonic) model trained on a voice subset equally-balanced in G_1, G_2 and G_3 grades.

In respect with the *leave-one-out* approach and this grade balancing, various voice subsets excluding the testing voice, composed of 18 voices each, are available to estimate both the normal and dysphonic GMM models. In the dysphonic case, these subsets are built randomly, including 6 voices per grade under the constraint that all the voices are used at least once.

For testing, each individual voice available in the DV corpus is first compared to all the normal voice models (from which it has been discarded if it is normal), resulting in an averaged normal voice likelihood $L(Y \mid \lambda_{\bar{d}})$, and secondly compared to all the dysphonic voice models (from which it has been discarded if it is dysphonic), resulting in an averaged dysphonic voice likelihood $L(Y \mid \lambda_d)$. The decision per individual voice relies on the maximum between the couple of likelihoods.

3.4. Task2-Protocol P2. Task2 consists in assessing a given voice according to the G criterion of the GRBAS scale. Four classes and corresponding models (one per grade, λ_{G_0}, λ_{G_1}, λ_{G_2} and λ_{G_3}) are in competition in the system. In this context,

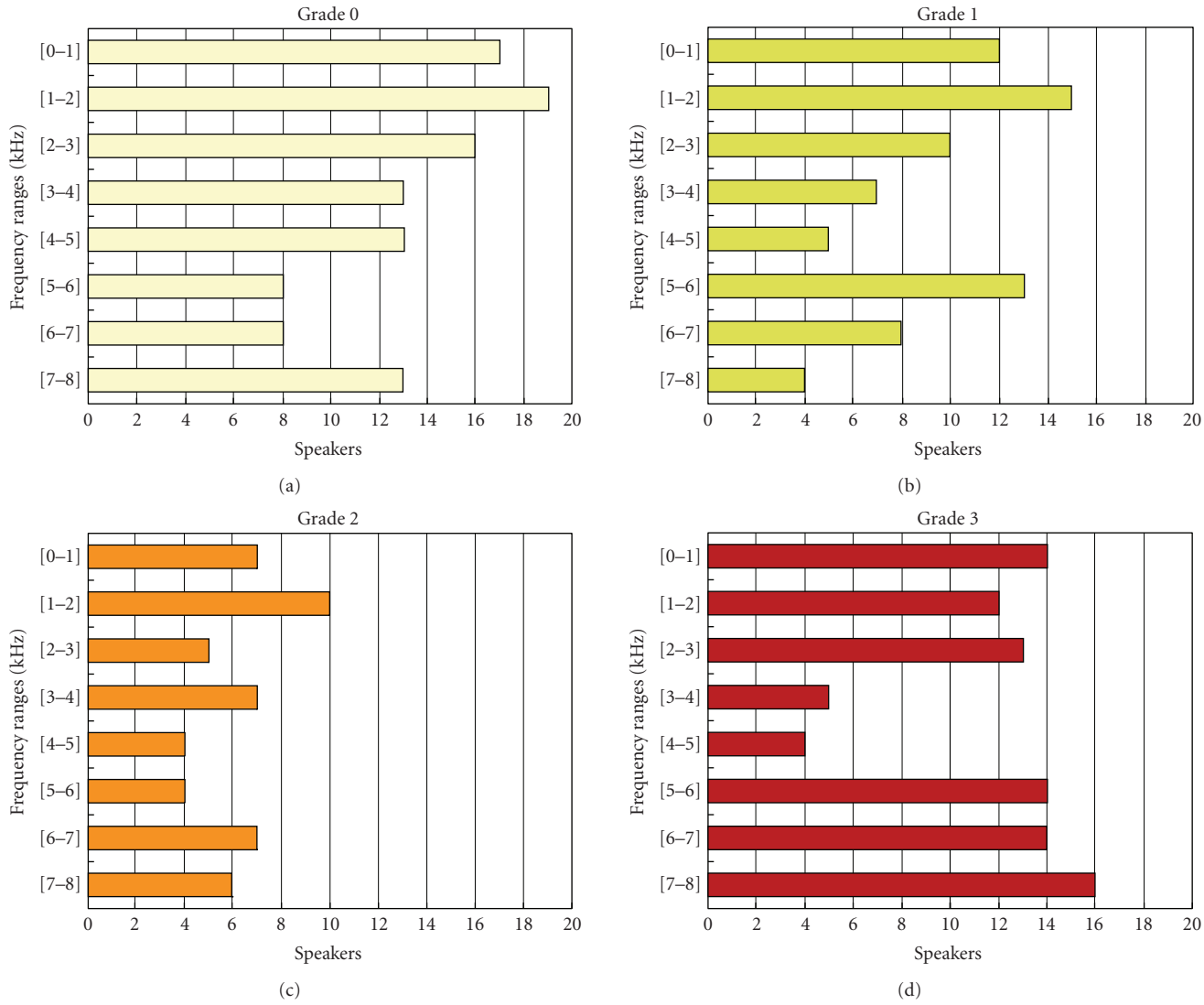

Figure 3: Number of voices correctly classified from the 4-G classification following 1000 Hz-width frequency subbands (24 LFSC).

the 20 normal voices (G_0 rated voices) and the 60 dysphonic ones (G_1, G_2 and G_3 rated voices) available in the DV corpus are used as follows.

(i) All the subsets of 19 voices among the G_0 set are used to estimate a model per each-$\lambda_{G_0}^{-Y}$-with Y the discarded voice. The same process is applied on the set G_1, G_2 and G_3. This results in 20 different models available per grade.

(ii) When voice Y, labeled perceptually as grade i, is tested, Y is first compared to model $\lambda_{G_i}^{-Y}$ leading to the likelihood computation: $L(Y \mid \lambda_{G_i}^{-Y})$. Then, an averaged likelihood is computed for all the other grades (different from i), by using the grade-dependent model sets (average on 20 likelihoods per grade).

(iii) The decision relies on the maximum of the four likelihoods.

3.5. Validation. To evaluate quality of the classification system, which next experimental results will depend on, Tables 1 and 2 provide its intrinsic performance through both protocols P1 and P2 respectively. In addition to the 24 LFSC-based parameterization, which will be used in the next sections, performance of a second system configuration, based on 72 MFSC is provided (see Section 2.1 for details of these different parameterizations). This second configuration aims to illustrate the potentiality of the automatic system when more complex and relevant information, like the joint use of static and dynamic features for instance, is extracted from the speech signal.

As expected, the 72 MFSC-based configuration shows the best classification performance independently of the protocols used, taking benefit of the more complex information. Focusing on the protocol P2, a large confusion can be observed on both grades 1 and 2 whatever the configurations used and the use of more relevant information involved in the 72 MFSC. This confirms the requirement of a better

Table 1: Performance of the normal and dysphonic voice classification (Task1) expressed in terms of correct classification rate (CCR in %) as well as the number of succeeded tests (in parentheses) on the *DV* corpus according to two parameterization configurations (24 LFSC and 72 MFSC). Confidence intervals (CIs) are provided for the overall scores.

	Correct classification rate (CCR in %) (Succeeded Test Nb)			
System	Normal	Dysphonic	Overall	±CI
24 LFSC	95.0 (19)	66.7 (40)	73.8 (59)	9.7
72 MFSC	95.0 (19)	91.7 (55)	92.5 (74)	5.8

understanding of the acoustic phenomena related to dysphonia and their different levels of severity.

4. Knowledge Gathering

The goal of this section is to describe how the automatic classification results allow to gather relevant knowledge for in-depth and refined the human expert analysis. In this way, the automatic system will be first joined to a frequency subband analysis. The aim of this subband-based analysis is to study how the acoustic characteristics of phenomena linked to dysphonia are spread out along different frequency bands depending on the severity level; in other words: "is a frequency subband more discriminant than another for dysphonic voice classification?" In a second step, this subband analysis will be coupled with a phonetic analysis to help for refining observations.

All the experiments have been conducted following the protocol P2. Despite its lower performance, the 24LFSC based-parameterization was preferred to the 72 MFSC for all the following experiments for two main reasons. Firstly, the use of linear filters is more straightforward in this context and facilitates the comparison between individual subbands. Secondly, the goal of the following experiments is to examine acoustic phenomena related to the dysphonia in the speech signal through the classification task instead of improving intrinsic performance of the automatic system.

4.1. Frequency Subband Analysis. 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, [37] shows that some subbands seem to be more relevant to characterize speakers than some others for the automatic speaker recognition 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 [38].

In this context, the full frequency band 0–8000 Hz is first split into individual fixed-width subbands (1000 Hz width), which the automatic classification system (described in Section 2) is applied on afterwards. According to performance observed on individual subbands, larger subbands are investigated.

4.1.1. 1000 Hz Subband Performance. In this first experiment, eight 1000 Hz-width subbands are processed individually through the classification system. Classification performance is presented per subband on Figure 3. Three main trends can be pointed out.

(i) Frequency bands between 0 and 3000 Hz get the best performance with an overall CCR varying from 55% to 70%.

(ii) Frequencies between 3000 and 5000 Hz exhibit the worse overall performance. Only the normal voices (grade 0) get a satisfactory score of 65% CCR, 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).

(iii) Frequencies upper than 5000 Hz provide better overall performance compared with 3000 to 5000 Hz 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).

Therefore, considering 1000 Hz-width frequency bands individually highlights (1) some difficulties to classify the grade 2 voices whatever the individual subband considered, (2) the ability of low frequencies to discriminate most of the voices, except for the grade 2 voices, (3) the "surprising" performance of the grade 3 voices on high frequencies, especially regarding the 7000–8000 Hz subband for which the speech amount is very low.

4.1.2. Joint Frequency Band Performance. This section focuses on the three frequency zones highlighted in the previous section. The automatic classification is now performed on the following frequency subbands: 0–3000 Hz, 3000–5400 Hz and 5400–8000 Hz, which aims to take benefit of the complementarity of the 1000 Hz-width subbands. Performance, reported in Table 3, shows that the behavior observed on the individual 1000 Hz-width subbands is emphasized here. Indeed, the 0–3000 Hz band (joining the first three 1000 Hz-width subbands) remains the most interesting frequency band, exhibiting an overall 71.25% CCR and achieving the best score for the grade 2 voices (65% CCR versus 50% for both full band and the best individual subband 1000–2000 Hz). Conversely, the 3000–5400 Hz band exhibits the lowest overall CCR (48.75%) compared with the other subbands. Confusion observed in the individual 1000 Hz-width subbands is still present, except for the grade 3 voices which tend to take benefit of the

Table 2: Performance of the 4-G classification (Task2) expressed in terms of correct classification rates (CCR in %) as well as the number of succeeded tests (in parentheses) on the *DV* corpus according to two parameterization configurations (24 LFSC and 72 MFSC). Confidence intervals (CI) are provided for the overall scores.

	Correct classification rate (CCR in %) (Succeeded Test Nb)					
System	Grade 0	Grade 1	Grade 2	Grade 3	Overall	±CI
24 LFSC	85.0 (17)	55.0 (11)	50.0 (10)	70.0 (14)	65.0 (52)	10.5
72 MFSC	95.0 (19)	65.0 (13)	70.0 (14)	85.0 (17)	78.8 (63)	9

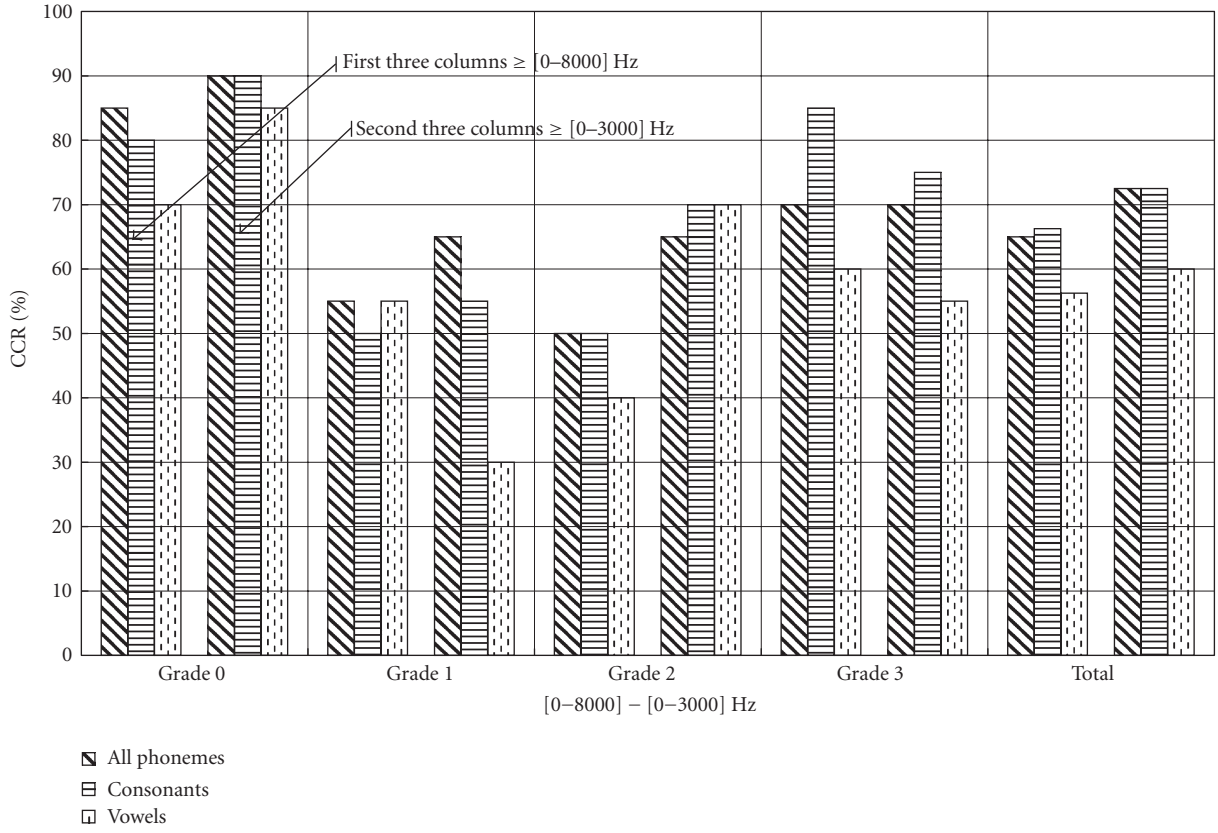

Figure 4: Performance per grade in terms of correct classification rate (CCR %) considering "All phonemes", consonant and vowel classes, for the 0–8000 Hz (each first set of three columns) and 0–3000 Hz (each second set of three columns) frequency bands.

complementarity of the individual subbands (65% CCR versus 25% and 20% for the corresponding individual 1000 Hz-width subbands). Finally, the 5400–8000 Hz band, related to the residual zone of fricative and plosive consonants, provides reasonable performance for the normal (65% CCR) and severe dysphonic voices (70% CCR). Regarding speech information carried by this band, CCR of the grade 3 voices 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 compared with other grades.

4.2. Frequency Band-Based Phonetic Analysis. To help in understanding and interpreting the behavior of the automatic classification system in the 0–3000 Hz frequency band, the authors propose to pursue the classification system observation through a frequency band-based phonetic analysis. In this way, performance of the classification system will be analyzed per phoneme class and per frequency range (0–8000 Hz and 0–3000 Hz) in order to evaluate which impact may have the dysphonia effects on phonemes or phoneme classes in particular frequency bands according to the grades. This phonetic analysis is close to the "phonetic labeling" proposed in [39], in which a descriptive and perceptual study of pathological characteristics of different phonemes is proposed. To perform this frequency band-based phonetic analysis, a phonetic segmentation is necessary for each speech signal available in the *DV* corpus. This segmentation was extracted automatically by realizing an automatic text-constrained phonetic alignment. The latter was yielded by the LIA alignment system, based on a Viterbi decoding

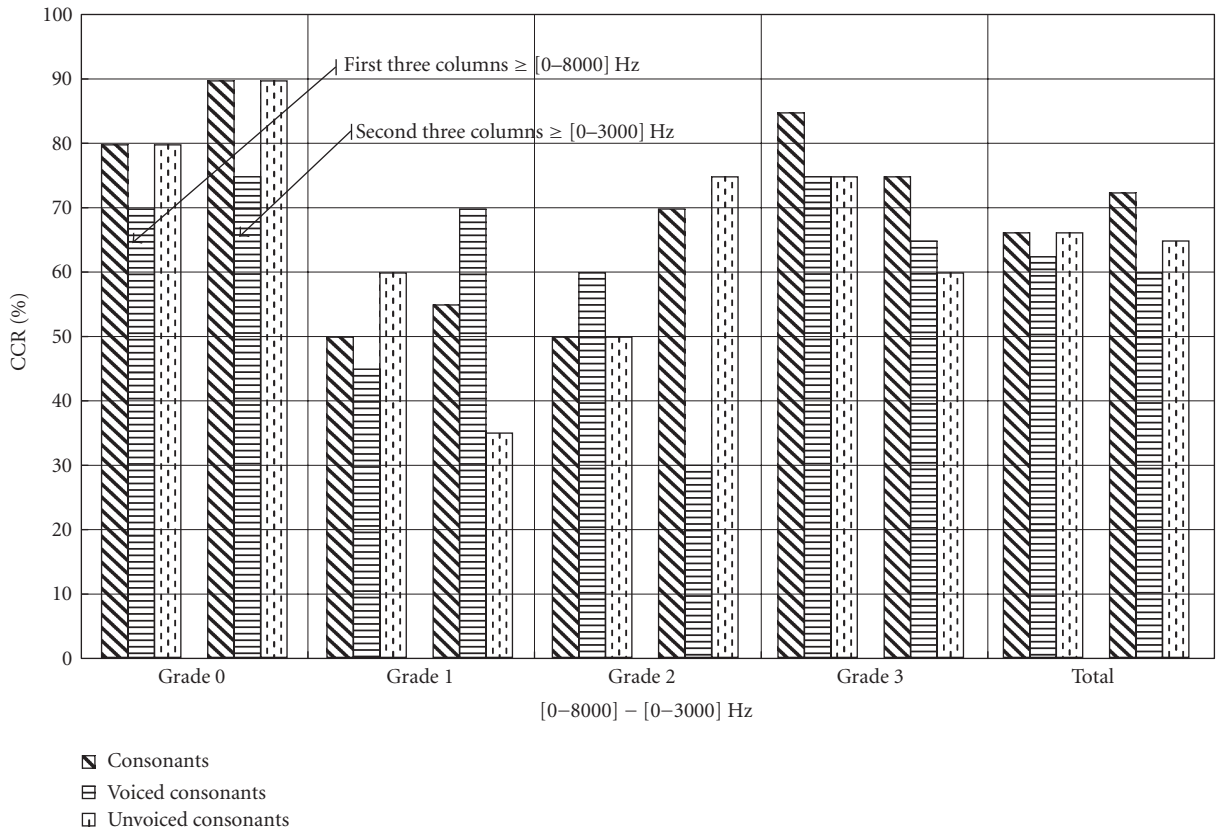

Figure 5: Performance per grade in terms of correct classification rate (CCR %) considering voiced and unvoiced consonant classes, for the 0–8000 Hz (each first set of three columns) and 0–3000 Hz (each second set of three columns) frequency bands.

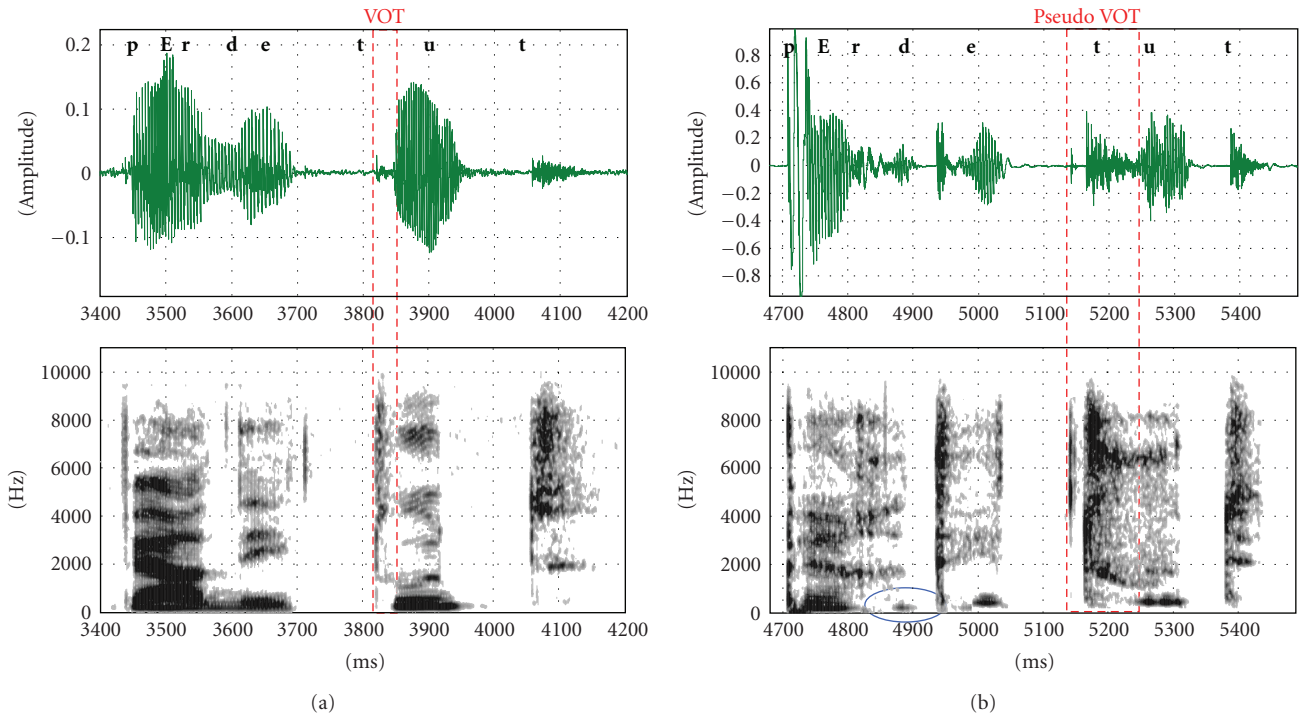

Figure 6: Oscillogram and spectrogram for the extract "[· · ·] perdait toutes [· · ·]"/pErdetut/; grade 0 on the left (normal voice), grade 3 on the right (severe dysphonia). On the right, we can note (1) the abnormal extension of the voice onset time (VOT) on the unvoiced plosive/t/. The consonant is quasi transformed in fricative (spirantisation), (2) the quasi unvoiced consonant/d/ transformed in/t/.

and graph-search algorithms [40], a text-restricted lexicon of words associated with their phonological variants, and a set of 38 French phonemes. It is worth noting that the phonetic segmentation is coupled with the automatic dysphonic classification system for the decision step only. Indeed, for the classification tests and decision making, the averaged frame based likelihood (see Section 2.3) between the incoming voice and the grade models is computed on the restricted set of segments associated with a targeted phoneme class. Conversely, grade models are learned on all the phonemic material available per grade in the *DV* corpus independently of the phoneme class targeted. Table 4 provides duration information of the phoneme classes targeted.

Figure 4 compares performance of the overall phoneme set (denoted "All phonemes" in the figure) with consonant and vowel classes according to the 0–8000 Hz and 0–3000 Hz frequency bands. Figure 5 focuses on consonant performance comparing voiced and unvoiced classes.

Comparing vowel and consonant classes (Figure 4) on the 0–8000 Hz band uniquely, it can be observed that consonants outperform slightly vowels for the grade 0 (80 against 70% CCR), for the grade 2 (50 against 40% CCR), and for the grade 3 (85 against 60% CCR). Best performance is reached by the vowel class on the grade 1 uniquely even if the difference is slight with the consonant class (55 against 50% CCR). Therefore, consonants tend to be more efficient in this context for discriminating dysphonia severity grades than vowels, although the former contains voiced and unvoiced components. In this way, Figure 5 shows that the unvoiced consonants outperform the voiced ones for both the grade 0 (80 against 70% CCR) and the grade 1 (60 against 45% CCR) as opposed to the grade 2 for which voiced consonants obtain 60% CCR against 55% for unvoiced consonants. Both of them reach 75% CCR for the grade 3.

Considering now the 0–3000 Hz frequency band, Figure 4 shows that consonants outperform vowels on both the grades 1 (55 against 30% CCR) and 3 (75 against 55% CCR) while reaching similar performance on grades 0 (90 against 85% CCR) and 2 (70% CCR for both). Thus, similar behavior can be observed on the grades 0 and 3 by comparing the frequency bands, performance of the consonant and vowel classes becomes equal for the grade 2 on the 0–3000 Hz band. Only the behavior of the consonant and vowel classes for the grade 1 is quite different since performance of the vowels decreases largely on the 0–3000 Hz. This tends to indicate that confusion with other grades is largely higher for the grade 1 considering the first formants of vowels only (present in the 0–3000 Hz). Regarding now Figure 5, the behavior of the unvoiced consonant is rather similar for grades 0 and 3 by comparing both the frequency bands. Inversely, the behavior of the grades 1 and 2 is quite different since the unvoiced consonants reach 35% CCR for the grade 1 on the 0–3000 Hz against 60% CCR on the 0–8000 Hz frequency band and 75% against 50% CCR for the grade 2. Therefore, the 0–3000 Hz frequency band seems to increase the confusion of grade 1 with the other grades considering the unvoiced consonants only.

4.3. Discussion. The progress in the experiments based on the automatic classification system reported previously, from the subbands to the phonetic analysis, tends to underline the relevancy of the unvoiced consonant in the discrimination of the GRBAS grades of dysphonia. This observation is rather unexpected regarding the definition of dysphonia. For that matter, studies reported in literature generally focus on voiced components because they are directly affected by pathologies related to the glottic source. For instance, sustained vowels are extensively associated with perceptual or objective approaches in literature since they make the assessment or measurement of parameters directly linked to the vocal source easier. The relatively high performance of the unvoiced consonants exhibited in this paper tends to highlight that these components can be of interest for assessing severity grade of dysphonia similarly to the voiced components. An interesting assumption for this observation would be that the consequences of dysphonia on the vowel production may impact the production of the unvoiced consonants as well, considering Vowel-Consonant (VC) or Consonant-Vowel (CV) contexts.

4.4. From Automatic Classification to Expertise: Preliminary Results on Prolonged Voice Onset Time. Previous sections have raised different interesting observations, requiring further analysis by human experts. As dysphonia is a laryngeal disorder, the quite good performance reached on the unvoiced consonants by the automatic classifier was rather unexpected. It is the reason why data were manually analyzed, focusing first on the unvoiced plosives (Figure 6). By verifying the automatic boundaries of the plosive, a lengthening of the voice onset time according to the dysphonia severity has been highlighted.

Voice onset time (VOT) is the duration between the release of a plosive and the beginning of the vocal fold vibration. This duration can be indicative of the speaker capacity to coordinate his/her articulatory and phonatory organs. For instance, during the production of the sequence /pa/, the speaker must control the relaxation of the lips, which creates a burst followed by the vibration of the vocal cords to produce the vowel. However, a deregulation can involve a lengthening or a shortening of this duration because of some peripheral biomechanical constraints or if motor control for the laryngeal vibration is delayed or anticipated compared to the gesture of opening of the consonant. This deregulation could also appear if the speaker does not have a well-tuned pneumophonatory control, for instance in dysphonia without laryngeal pathology. Abnormal VOT has been studied for second-language learning [41], aphasia or apraxia of speech [42], dysarthria [43], stuttering speech [44], dysphagia [45], spasmodic dysphonia [46]. To confirm VOT lengthening observation, VOT was measured from 865 unvoiced plosives (161 /p/, 244 /k/, 460 /t/), present on the French text uttered by the 80 female speakers of the *DV* corpus. For the statistical analysis, the "R" software v.2.6.2—a language and environment for statistical computing (http://www.R-project.org)—was used, associated with a linear mixed model [47]. The latter is a powerful model class used for the analysis of grouped

Table 3: Performance of the 4-G classification following joint frequency subbands in terms of correct classification rates (% CCR)—24 LFSC. Confidence intervals (CI) are provided for the overall scores.

	Correct classification rate (CCR in %) (Succeeded test Nb)					
	Grade 0	Grade 1	Grade 2	Grade 3	Overall	±CI
Full Band	85.0 (17)	55.0 (11)	50.0 (10)	**70.0 (14)**	65.00 (52)	10.5
0–3000 Hz	**90.0 (18)**	**65.0 (13)**	**65.0 (13)**	65.0 (13)	**71.25 (57)**	10
3000–5400 Hz	65.0 (13)	40.0 (8)	25.0 (5)	65.0 (13)	48.75 (39)	11
5400–8000 Hz	65.0 (13)	35.0 (7)	45.0 (9)	**70.0 (14)**	53.75 (43)	11

Table 4: Total duration in seconds per phonetic class and per grade as well as the number of phonemes (nb), their averaged duration (μ) and associated 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

data such as the repeated observations of a speaker available in this study. A key feature of mixed models is that they allow to address multiple sources of variation by introducing some random effects in addition to fixed effects; in other words, they permit to take both within- and between-subject variation into account in this context. The model VOT = f(GRADE) was studied here where GRADE is a 4-level ordered factor as regressor and SPEAKER is a random effect (intercept only). The statistical analysis, depicted in Figure 7, shows that the linear component of the GRADE factor is significant ($P = .0001$) with no significant quadratic or cubic effect. The main result obtained is that VOT is increasing with the dysphonia level in a significant way. This result can be explained by the difficulty to initiate the vocal cord vibration correctly, encountered by dysphonic speakers. It confirmed the phenomenon observed manually on the attack of sustained vowels by [48], which showed the importance of the vowel onset to identify the dysphonia severity perceptually. Of course, this study needs to be pursued by observing other kinds of data. To conclude this set of experiments, we can point out the interest of the automatic system to highlight features which cannot be a priori observed or expected by the human experts, especially on continuous speech.

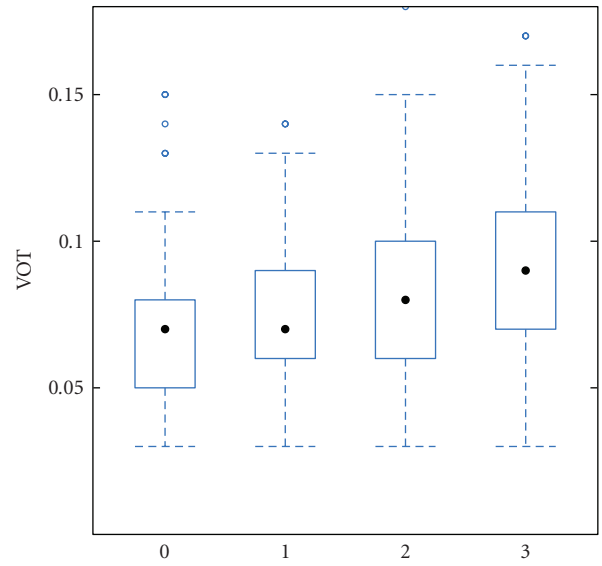

Figure 7: VOT (in seconds) according to the dysphonia severity grade (0: normal, 3:severe dysphonia).

5. Conclusion

The work presented in this paper aims to show that machine learning approaches could help the human experts to analyze more deeply acoustic features linked to voice disorders. Initially, the described approach relies on the human expertise, necessary for feeding an automatic voice classification system. More precisely, the latter requires a set of voice samples associated with some meaningful labels provided by the experts like pathologies and related perceptual grades. After several tuning and validation steps involving human experts (machine learning experts, phoneticians and pathologists) the voice classification system is able to model initial knowledge related to the labels. In a second phase, the automatic classification system is used to determine relevant information available in different parts of the input

speech recordings or exploited through different kinds of acoustical features. This second phase aims to determine some new hypotheses on voice disorders (more precisely, new hypotheses on the features characterizing voice disorders). Therefore, a new feedback loop between the automatic classification system and the human experts is mandatory. To assess the general methodology, this approach was experimented through an automatic classification system of dysphonia severity grades (following the *G* criterion of the GRBAS scale). This paper presents in details the automatic classification system, the class of original information it allows to highlight, as well as the first preliminary study related to the VOT carried out by the human experts.

Experiments based on this automatic classification system and conducted on a dysphonic corpus led to different interesting observations. First, the 0–3000 Hz frequency band achieved the best performance compared to [3000–5400] and 5400–8000 Hz, with an overall correct classification rate of about 71% (to be compared to 48% and 53% respectively for the other subbands) for the task of severity grade classification. When the system was used to rank the part of useful speech in terms of phonemic content, it was observed that consonants outperform vowels and, more surprisingly, that unvoiced consonants appeared to be very relevant for the classification task. Submitted to the human experts, these results led to a manual observation and analysis of unvoiced consonants. Focused on unvoiced plosives, this analysis highlighted a lengthening of the voice onset time (VOT) according to the dysphonia severity. This observation was confirmed by a statistical analysis performed on 865 unvoiced plosives issued from the dysphonic corpus. This phenomenon can be intuitively explained by the difficulty in initiating the vocal cord vibration correctly encountered by dysphonic speakers. However, from the author's knowledge, it has never been discussed from a scientific point of view. Even if this preliminary study on the VOT has to be pursued by observing, for instance, other classes of unvoiced consonants, the approach proposed in this paper has shown the potentiality of the back-and-forth loop between the automatic dysphonic voice classification system and the human experts. It should to drive the latter towards a better understanding of the acoustic phenomena related to voice disorders in the speech signal. In addition to the validation of the VOT lengthening according to the dysphonia severity levels, future work will be dedicated to bring human expertise on the potentiality of the unvoiced components for discriminating dysphonia severity grades. First studies will examine more complex phonemic contexts like Consonant-Vowel (CV) or Vowel-Consonant (VC) in order to determine if vowel alterations due to dysphonia may have impacts on the adjacent unvoiced consonants. Once validated, this new knowledge will be analyzed by the machine learning experts for its potential integration in the automatic classification system.

Acknowledgment

This research was partially supported by COST Action 2103 "Advanced Voice Function Assessment".

References

[1] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[2] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," International Telecommunication Union, Geneva, Switzerland, February 2001.

[3] M. Hirano, "Psycho-acoustic evaluation of voice: GRBAS scale for evaluating the hoarse voice," Clinical Examination of voice, Springer, 1981.

[4] B. Hammarberg, *Perceptual and Acoustic Analysis of Dysphonia*, Department of Logopedics and Phoniatrics, Karolinska Institutet, 1986.

[5] M. S. De Bodt, P. H. Van de Heyning, F. L. Wuyts, and L. Lambrechts, "The perceptual evaluation of voice disorders," *Acta Oto-Rhino-Laryngologica Belgica*, vol. 50, no. 4, pp. 283–291, 1996.

[6] D. Morsomme, S. Russel, J. Jamart, M. Remacle, and I. Verduyckt, "Evaluation subjective de la voix (VHI) chez 723 enseignants en région bruxelloise," in *Actes du Congrès de la Société Française de Phoniatrie*, Paris, France, 2008.

[7] INSERM, *La Voix. Ses Troubles Chez Les Enseignants*, INSERM, 2006.

[8] J. D. Hoit, "Influence of body position on breathing and its implications for the evaluation and treatment of speech and voice disorders," *Journal of Voice*, vol. 9, no. 4, pp. 341–347, 1995.

[9] N. Roy, D. M. Bless, and D. Heisey, "Personality and voice disorders: a multitrait-multidisorder analysis," *Journal of Voice*, vol. 14, no. 4, pp. 521–548, 2000.

[10] B. H. Jacobson, A. Johnson, C. Grywalski, et al., "The Voice Handicap Index (VHI): development and Validation," *American Journal of Speech-Language Pathology*, vol. 6, no. 3, pp. 66–69, 1997.

[11] P. H. Dejonckere, C. Obbens, G. M. de Moor, and G. H. Wieneke, "Perceptual evaluation of dysphonia: reliability and relevance," *Clinical Linguistics and Phonetics*, vol. 45, no. 2, pp. 76–83, 1993.

[12] P. H. Dejonckere, P. Bradley, P. Clemente, et al., "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques: guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS)," *European Archives of Oto-Rhino-Laryngology*, vol. 258, no. 2, pp. 77–82, 2001.

[13] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality: review, tutorial, and a framework for future research," *Journal of Speech and Hearing Research*, vol. 36, no. 1, pp. 21–40, 1993.

[14] L. Anders, H. Hollien, P. Hurme, A. Sonninnen, and J. Wendler, "Perceptual evaluation of hoarseness by several classes of listeners," *Clinical Linguistics and Phonetics*, vol. 40, pp. 91–100, 1988.

[15] F. L. Wuyts, M. S. De Bodt, G. Molenberghs, et al., "The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 3, pp. 796–809, 2000.

[16] J. F. Piccirillo, C. Painter, D. Fuller, and J. M. Fredrickson, "Multivariate analysis of objective vocal function," *The Annals*

of Otology, Rhinology and Laryngology, vol. 107, no. 2, pp. 107–112, 1998.

[17] A. Giovanni, V. Molines, B. Teston, and N. Nguyen, "L'évaluation objective de la dysphonie: une méthode multiparamétrique," in *Proceedings of the International Congress of Phonetic Sciences (ICPhS '91)*, pp. 274–277, Aix-en-Provence, France, 1991.

[18] B. Teston and B. Galindo, "A diagnosis of rehabilitation aid workstation for speech and voice pathologies," in *Proceedings of European Conference on Speech Communication and Technology (Eurospeech '95)*, pp. 1883–1886, Madrid, Spain, 1995.

[19] A. Giovanni, D. Robert, N. Estublier, B. Teston, M. Zanaret, and M. Cannoni, "Objective evaluation of dysphonia: preliminary results of a device allowing simultaneous acoustic and aerodynamic measurements," *Clinical Linguistics and Phonetics*, vol. 48, no. 4, pp. 175–185, 1996.

[20] A. Ghio and B. Teston, "Evaluation of the acoustic and aerodynamic constraints of a pneumotachograph for speech and voice studies," in *Proceedings of the International Conference on Voice Physiology and Biomechanics*, pp. 55–58, 2004.

[21] P. Yu, R. Garrel, R. Nicollas, M. Ouaknine, and A. Giovanni, "Objective voice analysis in dysphonic patients. New data including non linear measurements," *Clinical Linguistics and Phonetics*, vol. 59, pp. 20–30, 2007.

[22] L. Gavidia-Ceballos and J. H. L. Hansen, "Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 4, pp. 373–383, 1996.

[23] R. T. Ritchings, G. V. Conroy, M. McGillion, et al., "A neural network based approach to objective voice quality assessment," in *Proceedings of the 18th International Conference on Expert System (ES '98)*, pp. 198–209, Cambridge, UK, 1998.

[24] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proceedings of the Engineering Medicine and Biology Symposium*, vol. 1, pp. 182–183, 2002.

[25] C. Maguire, P. de Chazal, R. B. Reilly, and P. Lacy, "Identification of voice pathology using automated speech analysis," in *Proceedings of the 3rd International Workshop on Models and Analysis of Vocal Emission for Biomedical Applications*, Florence, Italy, December 2003.

[26] C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni, and A. Ghio, "Application of automatic speaker recognition techniques to pathological voice assessment (dysphonia)," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech '05)*, pp. 149–152, Lisboa, Portugal, 2005.

[27] J. I. Godino-Llorente, P. Gómez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 10, pp. 1943–1953, 2006.

[28] M. Wester, "Automatic classification of voice quality: comparing regression models and hidden Markov models," in *Proceedings of the Symposium on Databases in Voice Quality Research and Education (VOICEDATA '98)*, pp. 92–97, Utrecht, December 1998.

[29] V. Parsa and D. G. Jamieson, "Acoustic discrimination of pathological voice: sustained vowels versus continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 2, pp. 327–339, 2001.

[30] J. B. Alonso, F. Diaz, C. M. Travieso, and M. A. Ferrer, "Using nonlinear features for voice disorder detection," in *Proceedings of the International Conference on Non-Linear Speech Processing (NOLISP '05)*, pp. 94–106, Barcelona, Spain, April 2005.

[31] W. Chen, C. Peng, X. Zhu, B. Wan, and D. Wei, "SVM-based identification of pathological voices," in *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2007, pp. 3786–3789, 2007.

[32] F. Bimbot, J.-F. Bonastre, C. Fredouille, et al., "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 4, pp. 430–451, 2004.

[33] G. Gravier, "SPRO: a free speech signal processing toolkit (version 4.0.1)," 2003, http://gforge.inria.fr/projects/spro.

[34] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *Journal of the Acoustical Society of America*, vol. 39, pp. 1–38, 1977.

[35] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing: A Review Journal*, vol. 10, no. 1–3, pp. 19–41, 2000.

[36] L. F. Lamel, J. L. Gauvain, and M. Eskénazi, "BREF, a large vocabulary spoken corpus for french," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '91)*, pp. 505–508, Genoa, Italy, 1991.

[37] L. Besacier, J.-F. Bonastre, and C. Fredouille, "Localization and selection of speaker specific information with statistical modelling," *Speech Communication*, vol. 31, pp. 89–106, 2000.

[38] I. A. McCowan and S. Sridharan, "Multi-channel sub-band speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2001, no. 1, pp. 45–52, 2001.

[39] J. Revis, A. Ghio, and A. Giovanni, "Phonetic labeling of dysphonia: a new perspective in perceptual voice analysis," in *Proceedings of the 7th International Conference Advances in Quantitative Laryngology, Voice and Speech Research*, October 2006.

[40] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993.

[41] J. Alba-Salas, "Voice Onset Time and foreign accent detection: are L2 learners better than monolinguals?" *Revista Alicantina de Estudios Ingleses*, vol. 17, November 2004.

[42] P. Auzou, C. Özsancak, R. J. Morris, M. Jan, F. Eustache, and D. Hannequin, "Voice onset time in aphasia, apraxia of speech and dysarthria: a review," *Clinical Linguistics and Phonetics*, vol. 14, no. 2, pp. 131–150, 2000.

[43] R. J. Morris, "VOT and dysarthria: a descriptive study," *Journal of Communication Disorders*, vol. 22, no. 1, pp. 23–33, 1989.

[44] L. Jäncke, "Variability and duration of voice onset time and phonation in stuttering and nonstuttering adults," *Journal of Fluency Disorders*, vol. 19, no. 1, pp. 21–37, 1994.

[45] J. Ryalls, K. Gustafson, and C. Santini, "Preliminary investigation of voice onset time production in persons with dysphagia," *Dysphagia*, vol. 14, no. 3, pp. 169–175, 1999.

[46] J. D. Edgar, C. M. Sapienza, K. Bidus, and C. L. Ludlow, "Acoustic measures of symptoms in abductor spasmodic dysphonia," *Journal of Voice*, vol. 15, no. 3, pp. 362–372, 2001.

[47] J. C. Pinheiro and D. M. Bates, *Mixed-Effects Models in S and S-PLUS (Statistics and Computing)*, Springer, New York, NY, USA, 2000.

[48] J. Revis, A. Giovanni, and J.-M. Triglia, "Influence de l'attaque sur l'analyse perceptive des dysphonies," *Clinical Linguistics and Phonetics*, vol. 54, no. 1, pp. 19–25, 2002.