Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 173967, 19 pages
doi:10.1155/2009/173967

Research Article

On the Use of the Correlation between Acoustic Descriptors for the Normal/Pathological Voices Discrimination

Thomas Dubuisson,[1] Thierry Dutoit,[1] Bernard Gosselin,[1] and Marc Remacle[2]

[1] *TCTS Lab, Faculté Polytechnique de Mons, 31 Boulevard Dolez, 7000 Mons, Belgium*
[2] *ORL-ORLO Lab, Université Catholique de Louvain, Avenue Dr Therasse, 8, 5530 Yvoir, Belgium*

Correspondence should be addressed to Thomas Dubuisson, thomas.dubuisson@umons.ac.be

Received 27 October 2008; Revised 25 February 2009; Accepted 23 April 2009

Recommended by Juan I. Godino-Llorente

This paper presents an analysis system aiming at discriminating between normal and pathological voices. Compared to literature of voice pathology assessment, it is characterised by two aspects. First the system is based on features inspired from voice pathology assessment and music information retrieval. Second the distinction between normal and pathological voices is simply based on the correlation between acoustic features, while more complex classifiers are common in literature. Based on the normal and pathological samples included the *MEEI* database, it has been found that using two features (spectral decrease and first spectral tristimulus in the Bark scale) and their correlation leads to correct classification rates of 94.7% for pathological voices and 89.5% for normal ones. The system also outputs a normal/pathological factor aiming at giving an indication to the clinician about the location of a subject according to the database.

1. Introduction

The acoustic evaluation of voice quality is an important tool for the assessment of pathological voices. This assessment may be performed following two different approaches: the perceptive judgement and the objective assessment. On the one hand, the perceptive judgement is used in the clinical domain and consists in qualifying and quantifying the voice pathologies by listening the production of a patient. This evaluation is performed by trained professionals who rate the speech samples on a grade scale (*GRBAS* scale [1]) according to their perception of voice disorder. This subjective evaluation suffers of the drawbacks to be highly dependent on the experience of the listener and on its inconsistency on judging pathological voice quality. On the other hand, the objective analysis consists in qualifying and quantifying the voice pathologies by acoustical, aerodynamic, and physiological measurement. It offers the advantages to be quantitative, cheaper, faster, and more comfortable for the patient than methods like the electroglottography (*EGG*) [2] or the imaging of the vocal folds (by stroboscopy [3] or more recently by high-speed camera [4]).

Many methods of acoustic evaluation of pathological voices have been proposed in literature. Among them, an important part consists of computing acoustic descriptors, using them in a classifier, and computing a classification performance from the outputs of this system. Interesting results are obtained but two drawbacks can be highlighted.

(i) Using a classifier like Neural Networks is equivalent to a kind of "black box," it is difficult to identify what really happens in it and, in the case of transformation of the input space, which feature or set of features discriminate well the normal and pathological samples.

(ii) The features used in literature are often linked to the clinical evaluation, while other acoustic features have been proposed in other domains of sound processing. Moreover, features like jitter or Harmonic-to-Noise Ratio (*HNR*) are based on the evaluation of fundamental period, which can be subject to controversy in speech analysis, even more in the case of pathological voices [5].

For these reasons, this paper presents an analysis system using only the information from the correlation between

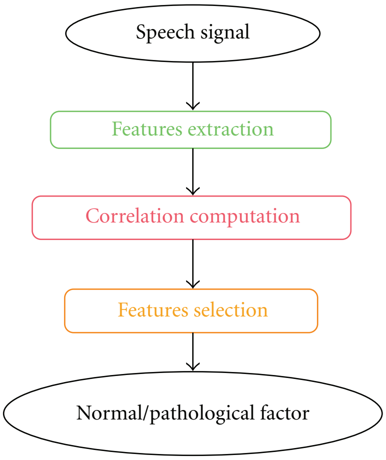

Figure 1: Structure of the analysis system.

acoustic descriptors in order to discriminate the normal and the pathological voices. These features come from both the clinical and sound analysis domains and have in common that none is based on the value of the fundamental frequency.

The analysis system is composed of three parts (see Figure 1).

(1) Feature extraction: this part consists in cutting the signal into frames, windowing them and computing descriptors (from both normal/pathological voices assessment and music analysis) in temporal and spectral domains. Only the value of descriptors corresponding to voiced parts of speech are considered. This part is described in Section 5.

(2) Correlation computation: the correlation between descriptors in voiced parts of speech is computed and stored into a matrix. This part is described in Section 6.

(3) Exploitation of the correlation: the elements of the correlation matrix are manipulated in order to discriminate between normal and pathological samples and to compute a final descriptor (normal/pathological factor). These manipulations are the results of a statistical study described in Section 7.

In a nutshell, the aims of this study are as follows.

(1) Giving an overview of the classic features and classifiers in normal/pathological voices discrimination. These two aspects are, respectively, described in Sections 2 and 3.

(2) Proposing features coming from other domain of sound analysis.

(3) Showing that simply using the correlation between features not based on fundamental frequency instead of a classic classifier allows to discriminate well between normal and pathological samples, extracted from the database described in Section 4.

2. Classic Features in Pathological Voice Assessment

The subject of this section is the overview of the classic features involved in pathological voice assessment. It is obviously not possible to include all the descriptors found in literature, only the most common are presented.

2.1. Fundamental Frequency. When working in speech processing domain, an obvious feature for researchers is the fundamental period, and its spectral equivalent, the fundamental frequency. This parameter is used in most of the studies, sometimes in conjunction with the *Mel*-Frequency Cepstral Coefficients (*MFCC*).

2.2. Mel -Frequency Cepstral Coefficients. *MFCCs* are one of the most widely-used way to represent the speech signal in domains like recognition or coding [6]. These coefficients are computed by weighting the Fourier Transform of the signal by a *MEL* filterbank (perceptive scale), then computing the cepstrum from this weighted spectrum and finally the Discrete Cosinus Transform (*DCT*) of this cepstrum.

Using the *MFCC* parameters provides three advantages.

(i) The human perception is taken into account by considering a perceptive scale of frequencies.

(ii) The *MFCC* parameters are uncorrelated thanks to the *DCT* operation. This may be an advantage if these parameters are used directly as input of a classification system. In this case, each parameter brings its own information, without link to other ones.

(iii) The spectral envelope is summarized into a limited set of parameters.

As *MFCC* coefficients are widely used in speech processing, some studies aim at adapting techniques of Automatic Speaker Recognition (*ASR*) to the pathological voice assessment. In [7, 8] the aim is to train a *GMM* classifier (see Section 3.1) able to determine the grade corresponding to a particular voice sample. 16 *MFCC* coefficients and their first derivative are computed by using a 24 *MEL* filterbank. In [9–11], 12 *MFCC* coefficients, with their first and second derivative, and the fundamental period are the inputs of an *HMM* classifier (see Section 3.2) trained in order to discriminate between normal and pathological samples. The distinction between these two classes is also the subject of [12], in which *MFCC* coefficients are used, among others, as inputs of a *SVM* classifier (see Section 3.3).

2.3. Acoustic Features from MDVP Software. The Multi-Dimensional Voice Program (*MDVP*) is a software produced by KayPentax Corp. [13]. When assessing the production of a subject, this system computes acoustic descriptors related to the perturbation of the fundamental frequency (or period) and to the amplitude of the signal (the whole definition of these descriptors is given in [14]).

As these features are considered as "classic" in the domain of speech pathology assessment, some authors use them in their classification system. Among these studies, some use the acoustic descriptors computed directly from *MDVP* software [10, 15], which is facilitated by the fact that these features are stored with speech samples in the *MEEI* database (see Section 4). However, one can object that there is little control on the computation of these acoustic descriptors. Some other studies use features inspired from those computed by *MDVP* software, meaning that their definitions is taken or inspired from [14]. For example, [16, 17] present a classification system of normal/pathological discrimination after transmission through a telephone channel in which input features are, among others, the perturbation of fundamental period and amplitude as defined in *MDVP* software.

2.4. Fundamental Frequency and Amplitude Perturbations (Jitter and Shimmer). Perhaps among the most famous acoustic descriptors in speech pathology assessment, jitter and shimmer are defined as the variation of the duration and the amplitude of the glottal cycle during the production of a sustained vowel.

MDVP software includes different ways of calculating jitter and shimmer, all of them being based on a classic estimation of fundamental frequency. Among these different implementations, *Perturbation Quotient* of fundamental frequency and amplitude are especially used as a measure of jitter and shimmer in a majority of studies [17–20]. An other descriptor derived from the *Perturbation Quotient* and showing more correlation with pathology is proposed in [21].

Most of the methods of jitter and shimmer computation rely on the assumption that periodicity exists in speech, which may be questionable in presence of pathology. That is why some methods propose alternative ways of computing jitter and shimmer than those based on the estimation of fundamental period. In [22] the salience of a sample (defined as the longest interval over this sample is maximum) is used to derive a duration quite close to the definition of glottal cycle length. Jitter and shimmer can easily be derived once this duration is available. One other interesting method relies on the modelisation of the power spectrum of speech into an harmonic part influenced by the jitter and a subharmonic one appearing because of jitter [5]. Jitter can be estimated by observing the behaviour of these two parts. Concerning the shimmer, the study proposed in [23] uses the waveform matching technique [24] to compute it. This study also proposes an interesting review of the different acceptations of the term *amplitude* in the definition of shimmer.

2.5. Spectral Balances. It is considered that the location of energy in spectral domain may be discriminant between the two populations. That is why descriptors are computed in limited frequency regions. Among these, the spectral balance is defined as the ratio between the spectral energy in two frequency bands.

Apart from *HNR*, de Krom defines in [25] 4 frequency bands ([60–400 Hz], [400–2000 Hz], [2000–5000 Hz], [5000–8000 Hz]) between which spectral balances are computed. The method exposed in [19] extends these frequency bands with the [8000–11025 Hz] band. Spectral balances between all possible pairs of bands and the whole spectrum are also considered. Other frequency bands (below 1 kHz, above 1 kHz, below 2 kHz, above 2 kHz) are proposed in [26] and are involved in the computation of two spectral tilt parameters, tuned to indicate the amount of noise without influence of jitter and shimmer.

2.6. Harmonics and Formant Level Variation. Considering the level in harmonic and formant regions is popular in speech pathology assessment because they reflect the presence of perturbation on speech signal (jitter, shimmer) and glottal source characteristics (spectral tilt, open quotient).

Concerning the harmonics, the first and second harmonic are most of the time considered. In [25, 27] the difference of amplitude between these harmonics and the relative level of the first one regarding to the level in the [400–2000 Hz] band are measured. The study [26] uses this measure in addition to the measure of the first harmonic level. Finally the authors of [19] choose to compute the level of the first harmonic and the relative level between the two first harmonics in the cepstral domain.

Concerning the formants, the level differences between the first harmonic and the strongest one in the first and third formant region are used in [26] and the level difference between the first and third formant is considered in [27]. The energy level in the region of the first, second and third formant are also selected in [28].

It must however be noticed that the measurement of these levels strongly relies on fundamental frequency estimation (which may be problematic in pathological cases) and formant detection.

2.7. Noise Features. As pathological speech is often perceived as noisy, researchers have been interested in measuring the harmonic and noise components of speech. This kind of measure is besides part of tools used in clinical domain (e.g., *MDVP* Software).

Some noise measures proposed in literature can be highlighted:

(i) Harmonic to Noise Ratio (*HNR*): defined in [29] as the log ratio of the energy of the periodic and aperiodic components, different methods of *HNR* computation have been proposed in literature (a comparison between them is proposed in [29] in the context of voice quality analysis). Some methods are based on a model in which speech is assumed to be composed of a periodic component and an aperiodic component [25, 30] (notably by computing the cepstrum of speech signal) while other use the short-time autocorrelation function [31]. They all share the fact that they are based on the estimation of fundamental frequency.

When looking at studies of pathological speech assessment, *HNR* arises as a popular measure. It is sometimes computed for the whole frequency range [15] or more frequently in particular frequency bands because of the assumption that noise energy is located in different frequency regions in normal and pathological phonations. Indeed *HNR* in four frequency bands is used in [32] with spectral energy in critical bands for the discrimination between normal and pathological samples in the *MEEI* Database. The same measure is used in [33] to show that *HMM* is able to classify different voice qualities and in [16, 17] to discriminate normal and pathological samples after transmission through a telephone channel. Speech samples for both methods are also extracted from the *MEEI* Database.

(ii) Normalized Noise Energy (*NNE*): this measure is proposed in [34] and aims at quantifying the energy of the noise component from the spectrum of speech. In the computation of *NNE*, noise energy is obtained between the harmonics directly from spectrum while inside the harmonics noise energy is assumed to be the mean value of the adjacent minima in the spectrum. The authors point that there may be a problem of estimation when the harmonics are broadened (in case of jitter). This measure is used in [15] for the discrimination between normal and pathological samples in the *MEEI* database.

(iii) Glottal to Noise Excitation (*GNE*) ratio: this measure is proposed in [35] and aims at quantifying the amount of voice excitation by vocal-fold oscillations versus excitation by turbulent noise. This descriptor is computed as the maximum correlation between the Hilbert enveloppes of the inverse filtered speech wave, in different frequency bands.

GNE measure is compared to *HNR* and *NNE* in [35] in which the authors show its relevance for the measure of noise energy, even in presence of strong jitter and shimmer (in the case of synthetic speech signals). This work is continued in [18], in which *GNE* is compared to other features (*HNR*, *GNE*, jitter and shimmer from *MDVP* software) in the field of voice quality assessment (in the case of real speech signals). As for *HNR* and *NNE*, *GNE* is computed for different frequency bands. It turns from this study that, in pathological speech, *GNE* (measuring the additive noise due to air passing through the glottis in case of uncomplete closure), and jitter and shimmer (measuring the irregularity of vocal folds vibration) describe different voice aspects that often appear for this kind of voice.

GNE has also been proved to show significant difference between subjects with normal phonation or pathological phonation ([15] for various pathologies or [27] for the particular case of cancer).

3. Classifiers used in Normal/Pathological Voices Discrimination

This section aims at describing the different kinds of classifiers used in voice pathology assessment. Their structure and behavior are briefly presented, with a focus on the way they are adapted to this particular problem. This section is complementary to Section 2, since the features used as input are described in this latter.

3.1. Gaussian Mixture Model. Gaussian Mixture Modeling (*GMM*) is widely used in Automatic Speaker Recognition, where it acts as a supervised classification system able to differentiate speech samples into classes (two for speaker verification and *n* for speaker identification).

In [7] *GMM* is adapted from speaker identification so that a class does not longer belong to a given speaker but to one of the grades in *GRBAS* scale (from 0 (normal) to 3). Each class is thus learnt using samples whose pathology is associated to this grade. The normal and pathological corpus are part of a database developed by LAPEC (Hôpital de la Timonne, Marseille, France) and consists of 20 normal samples and 60 pathological samples whose grade has been assessed by experts. The building of the classification system consists into three phases.

(i) Parametrization: *MFCC* coefficients and their first derivatives are extracted from speech.

(ii) Model training: a generic *GMM* is estimated on a normal corpus and *GMMs* are derived from the generic one by adapting of the mean of all the gaussians (*MAP* technique). In case of normal/pathological discrimination, a normal and a pathologic *GMM* are adapted from the generic *GMM* while in the case of grade classification, each grade is represented by a *GMM* adapted from the generic *GMM*.

(iii) Classification: when a speech sample has to be classified, the likelihood between this sample and each *GMM* is estimated and the decision relies on the maximum between these likelihoods.

For the normal/pathological classification, 95% of normal subjects and 81.7% of pathological ones are correctly classified. For the grade classification, the same performance is obtained for the grade 0 (corresponding to the normal subjects) while a loss of performance is observed for the pathological ones, specially between adjacent grades. Although the results are judged promising by the authors, no particular attention is put on the choice of acoustic parameters. The same system is used in [8] to determine which kind of information is better suited to the classification of the four grades. Three levels are considered.

(i) Energy: only the information extracted from non-silence frames is considered.

(ii) Phonetic: the information is extracted from frames after automatic phonetic alignment.

(iii) Voiced: only the information extracted from voiced segments is considered.

Whatever the level of information, the same performance for grade 0 than in [7] is obtained. For other grades, it turns out that the information extracted from the phonetic level provides the best overall classification result (71% for the same database as in [7]). The authors pursue their work in [36], where the same system than in [7, 8] is applied but this time to parameters extracted from a cut of the frequency range [0–8000 Hz] into subbands. It turns out that the [0–3000 Hz] band seems to be more informative (in terms of discrimination between the four grades) that the full frequency range. Interesting results of this study are the proof that (1) an attention on acoustic features is important for the classification (2) the whole frequency range may not be as performant as a subband to discriminate normal and pathological voices.

GMMs are also used in [37] in order to discriminate between 41 normal samples and 111 pathological samples collected in a room of the ENT department of a hospital. This time, the features come from the *MDVP* analysis (see Section 2.3) and consists of *Jitt*, *RAP*, *Shim*, *APQ*, *HNR*, and *SPI*. The methodology for building the *GMM* is slightly different than, for example, in [7] because this time one *GMM* models all the pathological samples and the varying dimension is the number of Gaussians involved in the mixture. For the optimal number of gaussians, the correct classification rate is 92.9% for normal data and 98.6% for the pathological ones. As in [7], the authors point that it would be interesting to pay more attention to the choice of the acoustic features.

3.2. Hidden Markov Model. Hidden Markov Models (*HMMs*) are well known in speech processing, notably in speech recognition and more recently in speech synthesis.

In [9] *HMMs* are trained on 12 *MFCC* coefficients, their first and second derivatives and the pitch for the sustained vowels /a/ and the spoken utterances from the *MEEI* Database (see Section 4). The correct classification rate for the sustained vowels is 98.3% and 97.75% for the spoken utterances. The authors compare these performances to those obtained by using the results of *MDVP* analysis on the same corpus. It turns that using these features provides lower classification rate for the two kinds of production. This work is continued in [10], in which a discrimination between four degrees of a particular pathology (A-P Squeezing) are classified using an *HMM* structured as in [9]. It turns that the correct classification rate is higher when the degree of pathology is severe and that using an *HMM* with the same features than in [9] provides better classification rate than using the features from *MDVP* analysis. The classification of different pathologies is pursued in [11], where the authors aim at discriminating five pathologies (A-P Squeezing, hyper-function, ventricular compression, paralysis, gastric reflux) in the pathological corpus of vowels /a/ from the *MEEI* Database. The same features than in the two papers above are used as input of the *HMM*. In this case, five *HMMs* are trained, each one corresponding to a pathology versus the others. When a new sample is presented as input of the classification system, the assigned pathology is the one corresponding to the *HMM* that outputs the maximum score. The average correct classification rate is 71%. Although the results of discrimination between pathologies are encouraging, the authors point that extending this work to other pathologies would be difficult because of the sparseness of data in the *MEEI* database. In terms of features, these three papers show that (1) spectral enveloppe features and pitch tend to be more reliable than the features estimated in *MDVP* analysis and (2) using *HMM* in classification provides good results in the case of discrimination between normal and pathological voices, and between different kinds of pathologies.

3.3. Support Vector Machines. Support Vector Machines (*SVMs*) [38] are a well-known classifier used in problems of classification, regression, and novelty detection.

Some people use this classifier in discrimination between normal and pathological samples. For example, [12] proposes to use a set of features consisting of 11 *MFCC* coefficients, *HNR*, *NNE*, *GNE*, *Energy*, and their first derivatives. The classifier is trained on the vowels /a/ from the pathological corpus of *MEEI* Database (53 normal samples and 77 pathological samples) and the average correct classification rate is 95.12%. The author point that the cepstral and the noise features complement well and that the results are better than using an *MLP* classifier with the same inputs. This kind of classifier is also used in [39], in which input features are extracted from the wavelet transform of 30 normal and 60 pathological speech utterances (from a database designed in Republic Center of Hearing, Voice and Speech Pathologies, Minsk, Belarus). The correct classification rate is this time 97.5% for normal voices and 100% for pathological ones.

3.4. Linear Discriminant. Among the simplest classification systems, Linear Discriminant (*LD*) classifier aims at cutting the feature spaces under the hypothesis of Gaussian Distribution for features of each class. Under assumptions about the distributions, the decision boundaries are linear. When a new sample is presented as input of this classifier, its assigned class is the one for which the classifier outputs the highest probability.

The remote diagnosis tool presented in [16] uses as input features inspired from *MDVP* analysis (pitch and amplitude perturbations, *HNR* in different frequency bands) to discriminate normal and pathological samples when speech is transmitted through the telephonic channel. The database consists of all the samples from *MEEI* database after transmission through this channel. The authors use an *LD* classifier and obtain a correct classification rate of 89.10% for the original database and 74.15% for the telephone database. Although they point out that the results are promising, they admit that more samples are needed to increase the performance of the system and that the difficulty in accurately tracking the pitch in speech could severly limit the discriminant ability of pitch perturbation measures.

This work is pursued in [17] by using pitch and amplitude perturbation features to classify the pathological samples from the telephone database of [16] into different kinds of pathologies (neuromuscular, physical, and mixed). The *LD* classifier provides a correct classification rate of 87.27% for the neuromuscular corpus, 77.97% for the physical corpus and 61.08% for the mixed corpus. It also turns that, in the case of the database transmitted by telephone channel, *HNR* measures are not as relevant as others to discriminate between normal and pathological groups and between the different groups of pathologies.

3.5. K-Nearest Neighbours. The K-Nearest Neighbours (*KNNs*) classifier [38] is a system aiming at clustering a feature space into as many regions as classes, these regions being delimited by piecewise linear planes.

In [32] a modification of this classifier is used in order to classify 53 normal and 163 pathological samples extracted from the *MEEI* database. In this system, a new sample is not compared to its *K* nearest neighbours but to a vector which represents the mean of all the features vectors belonging to a class. Thus the class assigned to the new sample is the one corresponding to the closest mean vector to the new sample. As the method exposed in [32] considers *HNR* in 4 frequency bands and spectral energy in 21 critical bands as inputs, these two kinds of features are used to design classifiers between normal and pathological samples. For the first set of features, the obtained accuracy is 94.28% and for the second one 92.38%. Although the dimension of the first set of features is smaller and the obtained accuracy higher than for the second set, the authors point out that fundamental frequency is difficult to estimate for pathological voices, leading to erroneous estimation of harmonic and noise components. They also highlight that the computation load of *HNR* is higher than for spectral energy.

3.6. Neural Networks. Artificial Neural Networks are a widely used classifier in various domains, in pattern classification and recognition or in speech recognition. Basically this type of classifier can be viewed as an interconnexion between simple small units, the neurons, designed to model to some extent the behaviour of human brain.

In [15] an *MLP* classifier is designed to discriminate between normal and pathological samples in the *MEEI* database. The network consists of a 26-neurons input layer (26 acoustic descriptors computed by *MDVP* software and stored in the database with the speech samples), one hidden layer and 1-neuron output layer (normal or pathological). The average correct classification rate is 94% when *HNR*, *VTI*, and *ShdB* are used as input features. The authors of [19] are also interested in the discrimination between normal and pathological samples in a database of 5 spanish sustained vowels (100 normal samples and 68 pathological samples). Each vowel is treated by a neural network which takes as input classic parameters and others extracted from the bicoherence. The decision from the 5 networks are then combined to decide if the input sample is healthy or not. The correct classification rate is 94.4% for the classic parameters and is increased of 4% when the others ones are added. A similar study is conducted in [20], in which the same classifier than in [19] is applied to two sets of features extracted from the database presented in the same paper. The two sets of features consist on classic parameters and classic parameters plus non-linear features inspired from the dynamic system theory (the correlation dimension and the largest Lyupanov exponent). The author shows that using this latter configuration of features leads to a correct classification rate of 93%.

4. Database

In the domain of pathological voices assessment, a widely-used database is the *MEEI* Disordered Voice Database, produced by KayPentax Corp. [14]. It has been chosen because a certain amount of studies [10–12, 15–17, 32, 33] use it in order to compare themselves to other methods and because it already provides a distinction between normal and pathological samples.

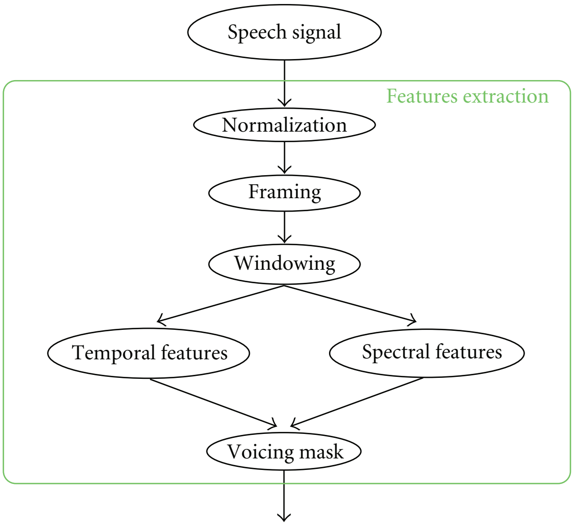

Figure 2: Details of the first part of the analysis system.

The database contains sustained vowels and reading text samples (12 seconds readings of the “Rainbow Passage”), from 53 subjects with normal voice and 657 subjects with a large panel of pathologies. The recordings are linked to informations about the subjects (age, gender, smoking or not) and to the results of the analysis by the *MDVP* software. The sampling frequency of the recordings is 25 kHz or 50 kHz, with only 25 kHz for the normal voices.

In this study, only the sustained vowels of the *MEEI* Database are considered. This group is split into a training and a test set, respectively, representing 65% and 35% of the whole database of sustained vowels.

(i) Training set: this set contains normal and pathological samples. The normal ones consist of 34 normal samples randomly chosen among the 53 samples of the database. The pathological ones consist of 427 samples randomly chosen among the 657 samples of the database.

(ii) Test set: this set contains the normal and pathological samples that are not part of the training set. It thus consists of 19 normal samples and 230 pathological ones.

The training set is used to find which information is the most discriminant for the discrimination between normal and pathological samples and the test set is used to assess the classification performance of this information.

In order to limit the computational load and to avoid an effect of the sampling frequency value on the discrimination between the two groups, all the voices are resampled at 16 kHz and quantified on 16 bits.

5. Feature Extraction

The first part of our analysis system consists in extracting features from speech signal (see Figure 2). This section thus aims at describing first the practical conditions of feature extraction, then the reasons of selecting features from the ones presented in Section 2 and from music sound analysis. The mathematical formulation of the selected descriptors and the implementation of voicing decision are finally presented.

5.1. Practical Conditions of Features Extraction. The computation of descriptors is preceded by several steps aiming at preparing the speech signal to be treated.

As explained in Section 4, the speech samples studied here (sustained vowels /a/) are part of the *MEEI* database. Their sampling frequency are various (25 or 50 kHz). It was then chosen resample them directly to 16 kHz.

In order to be as independent as possible to the recording conditions (e.g., tuning of the recording system), speech signal is normalized according to (1) ($x(n)$ stands for the samples of speech signal and N for its length):

$$x(n) = \frac{x(n) - (1/N)\sum_{i=1}^{N} x(i)}{\sqrt{(1/N)\sum_{i=1}^{N}\left(x(i) - (1/N)\sum_{i=1}^{N} x(i)\right)^2}}. \quad (1)$$

The descriptors considered in this paper are all local descriptors. They have thus to be computed for short time periods. Besides, in order to keep a good time resolution for the extraction of information, these time periods must overlap. As 30 milliseconds and 10 milliseconds are common values in speech processing community for, respectively, the duration and the delay between consecutive time periods, these two parameters are chosen in the analysis system. Besides, each time period is weighted by a window function (here a Hanning window), in order to avoid strong discontinuities at the boundaries.

5.2. Selecting Descriptors from Pathological Voices Assessment. Acoustic widely used descriptors in the domains of normal/pathological samples discrimination have been presented in Section 2. As already said in the introduction and stated in [5], the estimation of fundamental frequency may be doubtful in speech (especially in pathological speech). That is why it has been chosen not to consider descriptors relying on this measure (e.g., *HNR*, jitter, shimmer, harmonic level). Besides the results of some classification methods as [36] suggest that cutting the frequency range into frequency bands may be more informative than considering the whole frequency range. That is why, from the normal/pathological voice discrimination literature, the features related to spectral balances are considered in our system. These features are defined in Section 5.5.

5.3. Selecting Descriptors from Other Domains of Sound Analysis. Speech signal is itself a particular example of sound. Apart from the speech processing domain, many others are dedicated to the extraction of information from the sound. Among those, Music Information Retrieval (*MIR*) aims at extracting information from music in order to build classification system of music by, for example, artists. As this extraction is based on acoustic descriptors, it is interesting to highlight here some of them that could be used in voice pathologies assessment. These descriptors are part of the CUIDADO project [40], aiming at developing a new chain of applications through the use of audio/music content descriptors, and of the MPEG-7 [41] standard for multimedia content description. These features are not so far from speech processing: they are just complementary to the standard features exposed in Section 2.

All the considered features coming from the *MIR* domain (excepted the tristimuli) are not based on the estimation of fundamental frequency. However a modified definition of tristimuli is proposed in order to keep this measure in the feature vector. All these features are defined in Sections 5.4 and 5.5.

5.4. Temporal Domain. The features describing the temporal behaviour of speech signal are mathematically defined in this section. For the rest of the paper, $x(n)$ stands for a frame of speech signal and N for its length.

5.4.1. Temporal Energy. The energy of the frame (expressed in dB) is computed as

$$E_T = 10 \times \log_{10} \sum_{n=1}^{N} (x(n))^2. \quad (2)$$

5.4.2. Temporal Mean. The mean value of the frame is computed as

$$\mu_T = \frac{1}{N}\sum_{n=1}^{N} x(n). \quad (3)$$

5.4.3. Temporal Standard Deviation. The standard deviation of the frame is computed as

$$\sigma_T = \sqrt{\frac{1}{N}\sum_{n=1}^{N} (x(n) - \mu_T)^2}. \quad (4)$$

5.4.4. Temporal Zero Crossing. The zero-crossing rate [40, 42] aims at quantifying the frequency at which the signal crosses the zero-axis. This descriptor is notably used to indicate if a speech fragment is voiced or not [43]. For a given frame, the number of times that sign changes between a sample and the previous one (from positive to negative or the opposite) is computed. To convert this value in Hz, it is divided by the interval of time on which it is computed, 30 milliseconds in the present case.

5.5. Spectral Domain. The features describing the spectral behaviour of speech signal are mathematically defined in this section. For the rest of the paper, $X(k)$, $|X(k)|$, k, and N_{FFT} stand, respectively, for the Discrete Fourier Transform, its modulus, its bin index, and the number of frequency bins on which it is computed for the numeric sequence $x(n)$. N_{FFT} is set to 1024 in this study.

5.5.1. Spectral Delta. The delta value aims at quantifying the range of the amplitude spectrum. It is defined as

$$\mathrm{Delta}_S = \max_k(|X(k)|) - \min_k(|X(k)|). \tag{5}$$

5.5.2. Spectral Mean Value. The mean value of the amplitude spectrum is defined as

$$\mu_S = \frac{1}{N_{\mathrm{FFT}}}\sum_{k=1}^{N_{\mathrm{FFT}}}|X(k)|. \tag{6}$$

5.5.3. Spectral Median Value. The median value of the amplitude spectrum is defined as the amplitude value that divides all the values in two groups of same cardinality. The median is characterized by the fact that it is less influenced by extreme values than the mean value.

5.5.4. Spectral Standard Deviation. The standard deviation of the amplitude spectrum is defined as

$$\sigma_S = \sqrt{\frac{1}{N_{\mathrm{FFT}}}\sum_{k=1}^{N_{\mathrm{FFT}}}\left(|X(k)| - \mu_S\right)^2}. \tag{7}$$

5.5.5. Spectral Center of Gravity. The spectral center of gravity (also known as spectral centroid) is a very common feature in *MIR* domain [42, 44–47]. Perceptively connected to the perception of brightness, it indicates where the "center of mass" of the spectrum is. The spectral center of gravity of the amplitude spectrum is known as an economic spectral descriptor giving an estimation of the major location of spectral energy. It is computed as

$$\mathrm{COG} = \frac{\sum_{k=1}^{N_{\mathrm{FFT}}/2} k \times |X(k)|}{\sum_{k=1}^{N_{\mathrm{FFT}}/2}|X(k)|}. \tag{8}$$

The amplitude corresponding to this frequency is also stored as a feature.

5.5.6. Spectral Moments. As the power spectrum of a signal can be considered as the distribution of energy along frequency, one can describe this distribution by using descriptors from the theory of statistics. The spectral moments of the power spectrum [46] are well adapted to this description. The first four moments of the power spectrum [48] are considered in this study.

In order to compute them, the power spectral density and its total energy are computed as

$$\begin{gathered}\mathrm{PSD}(k) = \frac{1}{N_{\mathrm{FFT}}}|X(k)|^2, \\ E_S = \sum_{k=1}^{N_{\mathrm{FFT}}}\mathrm{PSD}(k).\end{gathered} \tag{9}$$

Then the four moments are computed as follows.

(1) The first moment is equivalent to the spectral center of gravity but computed this time on the *PSD*:

$$M_1 = \frac{2}{E_S}\sum_{k=1}^{N_{\mathrm{FFT}}/2} k \times \mathrm{PSD}(k). \tag{10}$$

(2) The second moment expresses the spread of the spectrum around its first moment:

$$M_2 = \frac{2}{E_S}\sum_{k=1}^{N_{\mathrm{FFT}}/2}(k - M_1)^2 \times \mathrm{PSD}(k). \tag{11}$$

(3) The third moment is defined as

$$M_3 = \frac{2}{E_S}\sum_{k=1}^{N_{\mathrm{FFT}}/2}(k - M_1)^3 \times \mathrm{PSD}(k). \tag{12}$$

As itself, the third moment is not stored as a feature because it is used to compute the skewness [49], defining the orientation of the *PSD* around its first moment. If it is positive, the *PSD* is more oriented to the right and to the left if is negative. The skewness is computed as

$$\mathrm{Skewness} = \frac{M_3}{{M_2}^{3/2}}. \tag{13}$$

(4) The fourth moment is defined as

$$M_4 = \frac{2}{E_S}\sum_{k=1}^{N_{\mathrm{FFT}}/2}(k - M_1)^4 \times \mathrm{PSD}(k). \tag{14}$$

As itself, the fourth moment is not stored as a feature because it is used to compute the kurtosis [49], defining the acuity of the *PSD* around it first moment. A Gaussian distribution having a kurtosis equal to 3, a distribution with a higher kurtosis is more acute than a Gaussian one while a distribution with a lower kurtosis is more flat than a gaussian distribution. The kurtosis is computed as

$$\mathrm{Kurtosis} = \frac{M_4}{M_2^2}. \tag{15}$$

5.5.7. Spectral Decrease. The spectral decrease [40] aims at quantifying the amount of decrease of the amplitude spectrum. Coming from perceptive studies, it is supposed to be more correlated with human perception. This descriptor is computed as

$$\mathrm{Decrease} = \frac{\sum_{k=2}^{N_{\mathrm{FFT}}/2}\left((|X(k)| - |X(1)|)/(k-1)\right)}{\sum_{k=2}^{N_{\mathrm{FFT}}/2}|X(k)|}. \tag{16}$$

5.5.8. Spectral Slope. The spectral slope [46, 50] is an other representation of the amount of decreasing of the amplitude spectrum. It is computed by linear regression of

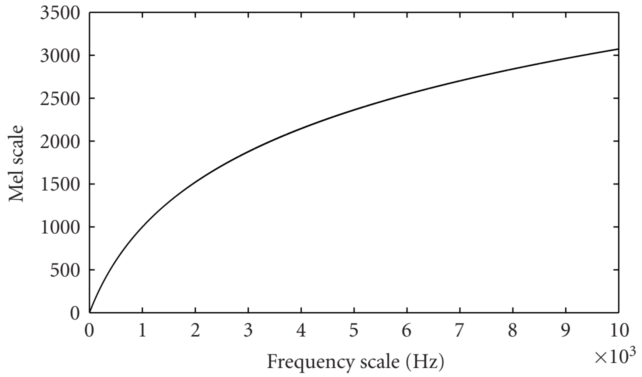

FIGURE 3: Relation between *Hz* and *MEL* scales.

the spectrum. In this formulation, the amplitude spectrum is approximated

$$\hat{X}(k) = S \times k + \text{constant}, \tag{17}$$

and the slope is computed

$$S = \frac{\left[(N_{\text{FFT}}/2)\sum_{k=1}^{N_{\text{FFT}}/2} k|X(k)|\right] - \left[\sum_{k=1}^{N_{\text{FFT}}/2} k \times \sum_{k=1}^{N_{\text{FFT}}/2} |X(k)|\right]}{\left[\sum_{k=1}^{N_{\text{FFT}}/2} |X(k)|\right]\left[(N_{\text{FFT}}/2)\sum_{k=1}^{N_{\text{FFT}}/2} k^2 - \left(\sum_{k=1}^{N_{\text{FFT}}/2} k\right)^2\right]}. \tag{18}$$

5.5.9. Spectral Roll-Off. The spectral roll-off [42, 46] is the frequency so that 95% of the energy is located below this point. k_c is computed by solving the equality

$$\sum_{k=1}^{k_c} |X(k)|^2 = 0.95 \times \sum_{k=1}^{N_{\text{FFT}}/2} |X(k)|^2. \tag{19}$$

5.5.10. Perceptive Scales. As already said in Section 2.2, perceptive behaviour of human hearing can be approximated by non linear scale of frequencies. Among those, one may cite the *MEL* scale and the *Bark* scale.

The *MEL* scale consists in a non linear division of frequency range, guided by perceptive considerations. Proposed in [51], this perceptive scale of pitches is defined so as a constant variation in the *MEL* scale is *perceived* as constant in the *Hz* scale. One particular link between these scales is that 1000 Hz corresponds to 1000 mels. The relation between the Hz scale and *MEL* scale is presented in Figure 3 and obeys

$$m = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right). \tag{20}$$

Based on this scale, a filterbank is designed and consists on 24 triangular-shaped filters whose center frequency is linearly distributed in the *MEL* scale and whose bandwith increases with central frequency.

The *Bark* scale [52] divides the frequency range into critical bands. This division is defined so as two sinusoids located in a critical band and whose amplitude is the same are perceived in the same way while their perceived intensity is different if they are located in different bands. The relation between *Hz* and *Bark* scale obeys

$$\text{Bark} = 13 \times \arctan\left(\frac{f}{1315.8}\right) + 3.5 \times \arctan\left(\frac{f}{7518}\right). \tag{21}$$

Based on this scale, the critical bands are implemented by using 24 rectangular filters whose center frequency is linearly distributed in the *Bark* scale and whose bandwith increases with central frequency. The *Bark* scale is used to compute the loudness and derived measures [46] with the sharpness and the spread (see Section 5.5.12).

5.5.11. Spectral Tristimuli. The tristimuli [46] are proposed in [53] as a timbre equivalent to the color attributes of vision. These are defined as energy ratio between the fundamental frequency and its harmonics. As it is decided to use in this study descriptors not based on fundamental frequency estimation, the implementation of tristimuli is modified by using frequency bands from the *MEL* and *Bark* scales instead of the harmonics. The 3 tristimuli are defined as in (22), in which $k_{\text{Band}_{[1]}}$ stands for the *FFT* bins corresponding to the frequency range defined by the first *MEL* or *Bark* frequency band and $k_{\text{Band}_{[1,\ldots,24]}}$ for the bins corresponding to the frequency range defined from the first to the 24th *MEL* or *Bark* frequency bands:

$$\begin{aligned} T_1 &= \frac{\sum_{k_{\text{Band}_{[1]}}} |X(k)|}{\sum_{k_{\text{Band}_{[1,\ldots,24]}}} |X(k)|}, \\ T_2 &= \frac{\sum_{k_{\text{Band}_{[2,3,4]}}} |X(k)|}{\sum_{k_{\text{Band}_{[1,\ldots,24]}}} |X(k)|}, \\ T_3 &= \frac{\sum_{k_{\text{Band}_{[5,\ldots,24]}}} |X(k)|}{\sum_{k_{\text{Band}_{[1,\ldots,24]}}} |X(k)|}. \end{aligned} \tag{22}$$

5.5.12. Spectral Loudness. The specific loudness [46] is the loudness associated to each *Bark* band and is defined as in (23), where z is the index of the *Bark* band (z standing for values from 1 to 24) and $k_{\text{Band}_{[z]}}$ the *FFT* bins corresponding to the frequencies included in the zth critical band:

$$\text{Loudness}(z) = \left(\sum_{k_{\text{Band}_{[z]}}} |X(k)|\right)^{0.23}. \tag{23}$$

The total loudness is defined as the sum of the specific loudness:

$$\text{Loudness}_{\text{Total}} = \sum_{z=1}^{24} \text{Loudness}(z). \tag{24}$$

For each band, a relative loudness is defined

$$\text{Loudness}_{\text{Relative}}(z) = \frac{\text{Loudness}(z)}{\text{Loudness}_{\text{Total}}}. \tag{25}$$

Based on the *Bark* scale and the loudness, a perceptive equivalent to the spectral center of gravity is computed as

$$A = 0.11 \times \frac{\sum_{z=1}^{24} z g(z) \text{Loudness}(z)}{\text{Loudness}_{\text{Total}}}, \tag{26}$$

where $g(z)$ is defined

$$g(z) = \begin{cases} 1, & \text{if } z < 15, \\ 0.066 \times e^{0.171 \times z}, & \text{if } z \geq 15. \end{cases} \tag{27}$$

Finally the *Spread* measures the distance from the larget specific loudness to the total loudness:

$$\text{Spread} = \left(\frac{\text{Loudness}_{\text{Total}} - \max_{[z]} [\text{Loudness}(z)]}{\text{Loudness}_{\text{Total}}} \right)^2 . \tag{28}$$

5.5.13. Spectral Balances. As defined in [19], 5 frequency bands are considered:

(1) L_0: [60–400 Hz],

(2) L_1: [400–2000 Hz],

(3) L_2: [2000–5000 Hz],

(4) L_3: [5000–8000 Hz],

(5) L_T: [60–8000 Hz].

The energy in each of these bands is computed as in (29), where L_i stands for the ith frequency band and k_{L_i} for the *FFT* bins corresponding to the frequencies included in the L_i frequency band (i stands for the values $0, 1, 2, 3, T$):

$$E_{L_i} = 10 \times \log_{10} \sum_{k_{L_i}} \text{PSD}(k). \tag{29}$$

The energy ratio between two of these frequency bands is defined as in (30) (i and j stand for the values $0, 1, 2, 3, T$):

$$E_{L_{i,j}} = 10 \times \log_{10} \frac{\sum_{k_{L_i}} \text{PSD}(k)}{\sum_{k_{L_i}} \text{PSD}(k)}. \tag{30}$$

The Soft Phonation Index is defined in the same way than in (30), but for the [0–1000 Hz] and [0–8000 Hz] frequency bands.

5.5.14. Spectral Flux. The spectral flux [42, 46, 47] is a descriptor aiming at quantifying the variation of the spectrum along time. It is particularly useful when particular event (such as voice onsets [54]) must be detected. This temporal variation is computed from the normalized cross-correlation between two successive amplitude spectra:

$$\text{SF}(t) = 1 - \frac{\sum_{k=1}^{N_{\text{FFT}}/2} |X(t-1,k)| \times |X(t,k)|}{\sqrt{\sum_{k=1}^{N_{\text{FFT}}/2} |X(t-1,k)|^2} \times \sqrt{\sum_{k=1}^{N_{\text{FFT}}/2} |X(t,k)|^2}}. \tag{31}$$

5.6. Voicing Decision. As it has been chosen to compute the correlation between features only for voiced parts of speech, a voicing detection algorithm dedicated to this purpose has been developed. The different steps of this algorithm are as follows.

(1) Prior estimation of fundamental period: a lot of methods are proposed in literature, but the YIN algorithm [55] has emerged since recent years in the speech processing and *MIR* communities. This algorithm provides a prior estimation of fundamental period, necessary for the following step.

(2) Computation of the local cross-correlation [49]: the cross-correlation function (see (32) for two sequences $y(n)$ and $z(n)$ whose length is N) is the major element to determine if the speech segment is voiced or not:

$$R_{yz}(m) = \frac{1}{N - |m|} \sum_{n=0}^{N-m-1} y(n+m) z(n). \tag{32}$$

Every 30 milliseconds, the corresponding estimation of fundamental period is considered and two frames are extracted from speech signal: one fundamental period on the left of the current instant of analysis and one fundamental period on the right of this instant. The cross-correlation between these two frames is then computed according to (32).

(3) Thresholding of the cross-correlation: by observing the evolution of the maximum of the cross-correlation function (let us call it *MaxXC*) and according to [43], it has been observed that this descriptor, correctly thresholded, provides a preliminary discrimination between voiced and unvoiced frames. The most satisfying value for the threshold is 0.02. A voiced mask is defined for the whole speech signal:

$$\text{Voiced Mask} = \begin{cases} 1, & \text{if } MaxXC \geq 0.02, \\ 0, & \text{if } MaxXC < 0.02. \end{cases} \tag{33}$$

(4) Correction of the voicing mask: although the results of the previous step are already satisfying, some mistakes remain, as in other problems in which a threshold has to be applied. A typical mistake is an isolated voiced frame among unvoiced ones. To overcome these detection errors, a second-order moving average filter has been applied on the voiced mask. Thus, for a given frame, if the output of the filter is lower than 1, it is tagged as unvoiced and voiced otherwise.

Once the voiced mask is available, each evolution of the features presented above is multiplied by the mask in order to keep only the value of features in voiced parts of speech.

6. Correlation Computation

As presented in Section 5, a total of 87 features are considered in this study. They were originally intended to be inputs of

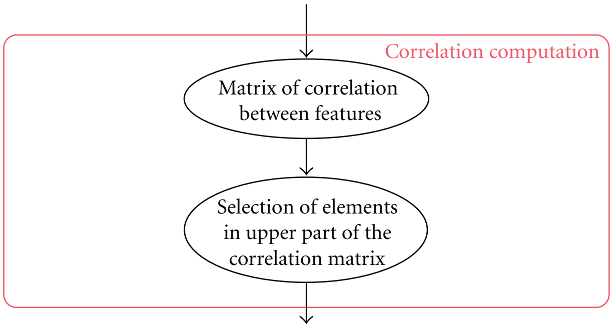
Correlation computation
Matrix of correlation between features
Selection of elements in upper part of the correlation matrix

Figure 4: Details of the second part of the analysis system.

a classification system. In order to eliminate the redundant information in the features, the correlation between features is first computed. This operation is included in the second part of the analysis system (see Figure 4).

6.1. Definition of the Correlation. The Pearson correlation coefficient [49] is computed as (for two numeric sequences x and y whose length is N)

$$R_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x}) \times (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2} \times \sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}}, \tag{34}$$

where

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{35}$$

The values of the correlation coefficient are constricted into the $[-1, 1]$ interval, $|R_{xy}| = 1$ corresponding to perfectly correlated sequences and $R_{xy} = 0$ to perfectly uncorrelated sequences.

In the case of multiple sequences of features, the correlation is computed between each pair of sequences and the overall correlation matrix is computed as

$$M(p,q) = \frac{\sum_{i=1}^{N}\left(x_{i,p} - \overline{x}\right) \times \left(y_{i,q} - \overline{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_{i,p} - \overline{x}\right)^2} \times \sqrt{\sum_{i=1}^{N}\left(y_{i,q} - \overline{y}\right)^2}}, \tag{36}$$

where p and q (constricted into the $[1, 87]$ interval) identify two sequences of features and where $\overline{x}$ and $\overline{y}$ are computed for each sequence. The correlation matrix for a normal subject and a pathological sample (from the database described in Section 4) are presented in Figures 5 and 6.

When looking at those matrices, one can see that their structures are quite different, this fact being confirmed for other samples in the database. That is why it was decided to exploit the information from the correlation matrix rather than the features themselves in order to see if significant differences could be found between normal and pathological samples.

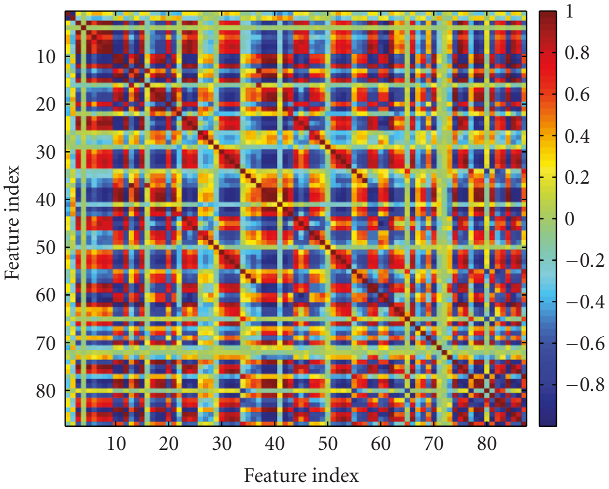

Figure 5: Correlation matrix for a normal sample.

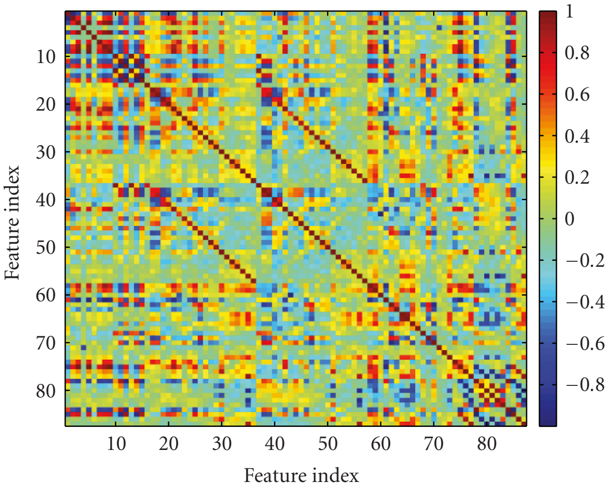

Figure 6: Correlation matrix for a pathological sample.

6.2. Exploitation of the Correlation Matrix. As shown in Figures 5 and 6, the correlation matrix for the normal sample contains more elements close to 1 (in absolute values) than the one for the pathological sample. An information could be extracted from the correlation matrix by considering the elements of the upper part of this matrix and by considering each of these elements as a feature itself. As correlation matrix is symetric and its diagonal consisting on elements equals to 1 by definition, the number of elements in the upper part of the correlation matrix is

$$\frac{N_{\text{Descriptors}} \times \left(N_{\text{Descriptors}} - 1\right)}{2}, \tag{37}$$

where $N_{\text{Descriptors}}$ stands for the number of acoustic descriptors. In the present case, as 87 descriptors are considered as inputs of the system, there are 3741 elements to consider in the correlation matrix.

In order to find a statistical discriminant factor between normal and pathological samples, the correlation matrix is

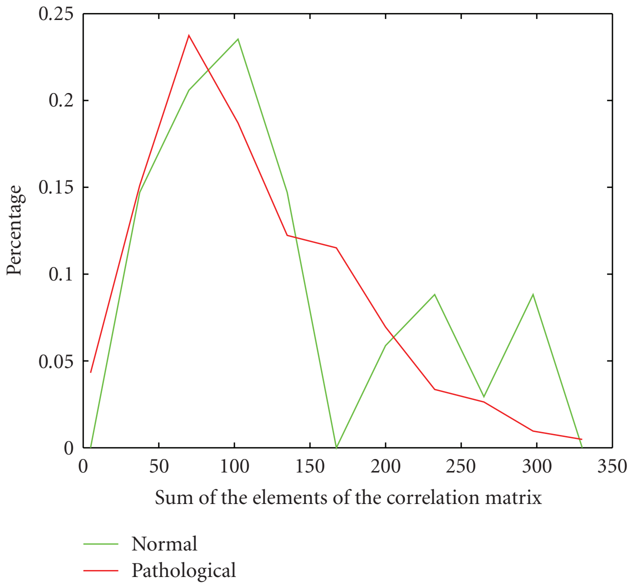

FIGURE 7: Distribution of the sum of the elements of the correlation matrix for normal samples (green) and pathological samples (red) included in the training set.

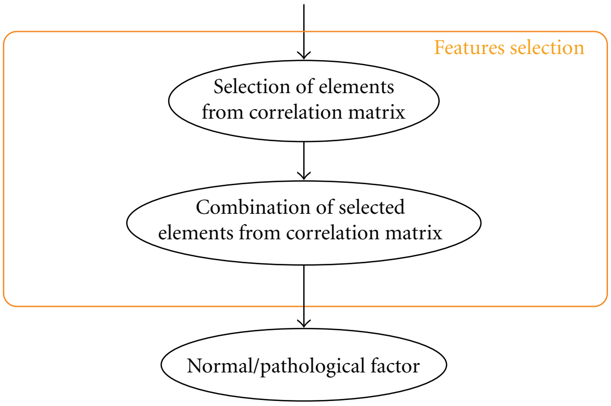
Features selection
Selection of elements from correlation matrix
Combination of selected elements from correlation matrix
Normal/pathological factor

FIGURE 8: Details of the third part of the analysis system.

computed for all samples of the training set and, to assess the classification performance of this factor, the correlation matrix has also been computed for the samples in the test set.

A first attempt to the extraction of information from the correlation matrix is the sum of the elements in its upper part. The distribution of this sum for the samples of the training set is shown in Figure 7. The normal distribution is in green and the pathological one in red. The y-axis is graduated in percentages of samples that are associated to a particular value of the sum (x-axis). When looking at these distributions, one can see that each one is large and that they strong overlap, leading to the conclusion that the sum value is not able to separate the normal and pathologic samples. One can make the hypothesis that some features bring confusion in the sum operation. That is why it was decided to select only few of them in order to see if the separation between normal and pathological samples is better in this case. The operation of selecting features is presented in the next section.

7. Feature Selection

As each element of the correlation matrix can be considered as a feature, the analysis system computes now 3741 features. As seen in the previous section, the sum of these features does not allow to separate well the distribution of normal and pathological samples in the training set. It is thus necessary to select among the 3741 features the few ones that discriminate best between the two populations. Some methods are proposed in literature, each of them is briefly described here. The reasons of choosing one of them are then presented and the selected method is finally applied to the present problem. The selection and combination of the elements of the correlation matrix are included in the third part of the analysis system (see Figure 8).

7.1. Methods for Feature Selection

7.1.1. Principal Component Analysis. Principal Component Analysis (*PCA*) is a well-known method for the preprocessing of features in a classification system [38, 56]. It is used to linearly transform the features in order to find the best way to represent them (in terms of least square error). If X represents the normalized features matrix, the new features matrix Z is obtained by

$$Z = U^T X, \tag{38}$$

where U is the linear transformation matrix. One can show that the matrix U that leads to the best final representation consists of the eigen vectors of the autocorrelation matrix XX^T. The dispersion of features around each new axis of representation is given by the eigen value associated to this axis. A reduction of features dimensionality is possible by selecting the axis of representation associated to the highest eigen values. It must be however emphasized that the transformation defined in (38) is not based on a class labelling but only on the features. Besides, *PCA* consists of computing a linear combination of original features, leading to a difficult physical interpretation of the new ones.

7.1.2. Generalized Fisher Criterion. The generalized Fisher criterion [56] is a class separation criterion based on the features and the class labelling. It is based on the ratio between two matrices.

(i) Within-class covariance matrix: quantifies the amount of inner features dispersion for all the classes.

(ii) Between-class covariance matrix: quantifies the features dispersion around the general mean for all the classes.

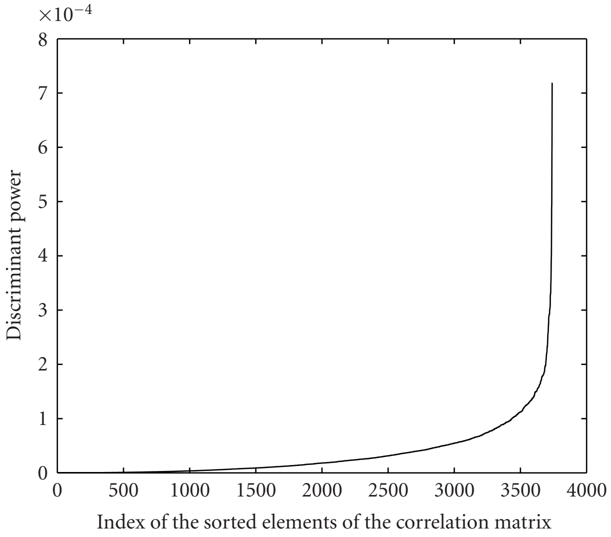

Figure 9: Discriminant power of the 3741 correlations for the samples of the training set.

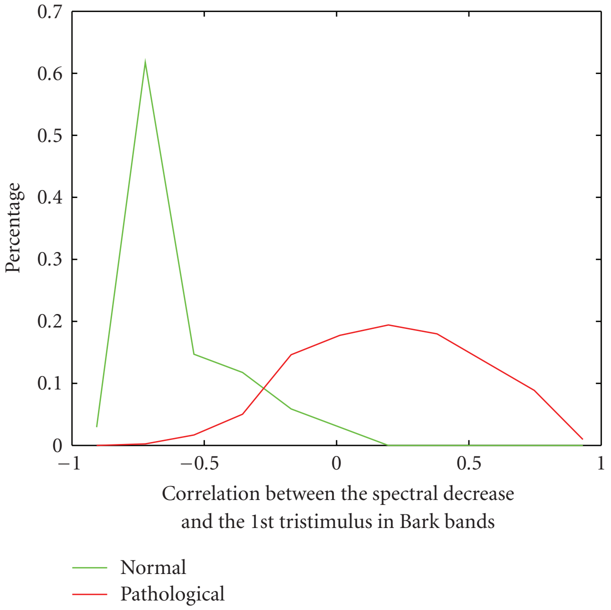

Figure 10: Distribution of the first discriminant correlation for the samples of the training set (green; normal class; red: pathological class).

For a given feature k, only the diagonal elements of the two matrices defined above are considered and its discriminant power between C classes is defined as

$$D_k = \frac{\sum_{c=1}^{C} p(\omega_c)(\mu_{ck} - \mu_k)^2}{\sum_{c=1}^{C} p(\omega_c)\sigma_{ck}^2}, \qquad (39)$$

where $p(\omega_c)$ stands for the percentage of representation of class c in the database, μ_{ck} for the mean of the feature k in the class c, μ_k the mean of feature k for all the classes, and σ_{ck} for the standard deviation of feature k in the class c. A feature selection is possible by selecting the features associated to the highest values of discriminant power. Comparing to *PCA*, it has the advantage to be based on class labelling and to conserve the meaning of the features.

7.1.3. Fisher Discriminant Analysis. The Fisher discriminant analysis [56, 57] is a procedure allowing to change the representation system of features and to select among them in one operation. For a C classes problem, this method consists of finding $C - 1$ linear discriminant functions, these functions maximizing the ratio between the within-class covariances and the between-class covariances. One can prove that these functions are the eigen vectors of a particular matrix. This method allows to reduce the dimensionality of a problem although this dimensionality is fixed by the number of involved classes. Besides, as *PCA*, the new features are the result of linear combination of the original ones, leading to a difficult physical interpretation after transformation.

7.2. Application of Feature Selection. It has been chosen in this study to apply the generalized Fisher criterion in order to keep the choice of the final dimensionality (contrary to the linear discriminant analysis) and the physical meaning of the features (contrary to *PCA* and linear discriminant analysis).

Figure 9 shows the discriminant power of the 3741 correlations, sorted in ascending order, for the samples of the training set. One can see that, for some correlations, the discriminant power is higher than for others. It has been chosen to study two cases here: the case in which only the correlation associated to the highest discriminant power is kept and the case in which only the two ones associated to the highest discriminant powers are kept.

7.2.1. One Correlation Case. The selected correlation is the one between the first spectral tristimulus in the Bark scale (see Section 5.5.11) and the spectral decrease (see Section 5.5.7). The distribution of this correlation for the two classes of the training set is shown in Figure 10. The normal samples are characterized by the fact that there is a high concentration for values around −0.75. This means that the evolution of the spectral decrease and the first spectral tristimulus are fairly strong linked for a large majority of samples, although this link is not absolute (because the correlation is not −1 but −0.75). The pathological samples are characterized by a larger dispersion of the correlation value, meaning that for some samples the two characteristics are fairly slightly linked and for others no link exists at all. Compared to Figure 7, the two classes are much more separated. It may thus be possible to split the normal and pathological samples of the training set by thresholding the most discriminant correlation.

In order to have an overview of the classification performances of this thresholding for the training set, a Receiver Operating Curve (*ROC*) is built by computing the False Positive Rate (*FPR*) and True Positive Rate (*TPR*) for thresholds uniformly distributed between the lower and

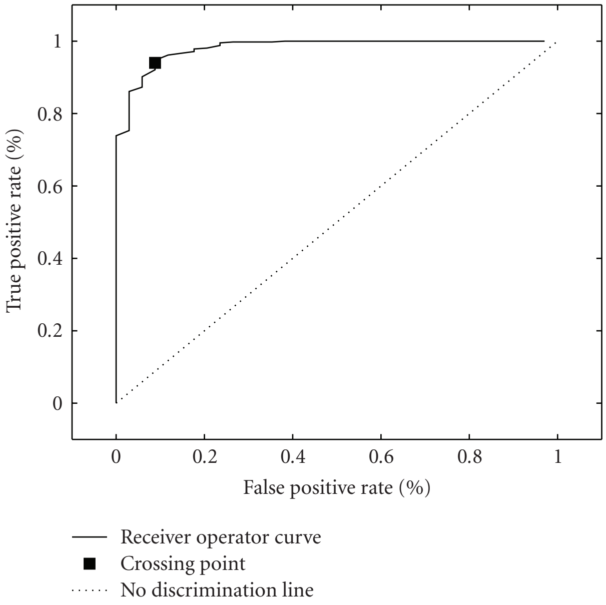

Figure 11: *ROC* for the "One Correlation Case."

upper limits of the correlation. These numbers are computed for each threshold value as follows.

(1) For each sample of the training set, computing an automatic labelling by assigning the class *Normal* if the correlation is lower than the threshold and assigning the class *Pathological* otherwise.

(2) Computing the confusion matrix based on the confrontation between the manual class labelling (in other terms if a sample has been manually tagged *Normal* or *Pathological*) and on the automatic class labelling. The elements of the confusion matrix are defined as in Table 1 where *TP* stands for *True Positive*, *FP* for *False Positive*, *FN* for *False Negative* and *TN* for *True Negative*. These values may be normalized according to the cardinality of the *Normal* class (called *#Normal*) and the cardinality of the *Pathological* class (called *#Pathological*). *TPR* and *FPR* are therefore obtained by dividing, respectively, *TP* and *FP* by *#Pathological* and *#Normal*. One may also define the accuracy (*Acc*), measuring how well a binary classifier correctly identifies or excludes a condition and defined as

$$\text{Acc} = \frac{TP + TN}{\#Normal + \#Pathological}. \tag{40}$$

The *ROC* for the "One Correlation Case" is shown in Figure 11. In this curve, the point $(0, 0)$ corresponds to the case in which all the normal samples are correctly classified but all the pathologic ones are misclassified and the point $(1, 1)$ corresponds to the opposite situation. One may also cite the ideal point $(0, 1)$ corresponding to the perfect classification of both normal and pathological samples. The more the *ROC* is close to this point the best the classifier is. Between the points $(0, 0)$ and $(1, 1)$,the choice of a particular threshold depends on the objective. If one wants to avoid errors on *Normal* class identification, the corresponding threshold will lead to low *FPR* (but also to low *TPR*). On the contrary if it is important to avoid mistakes on *Pathological* class identification, the corresponding threshold will lead to high *TPR* (but also high *FPR*).

Table 1: Confusion matrix.

	Manual pathological	Manual Normal
Auto pathological	TP	FP
Auto normal	FN	TN

Table 2: Confusion matrix for the one correlation case (Training set).

	Manual pathological	Manual normal
Auto pathological	0.947	0.088
Auto normal	0.053	0.912

Table 3: Confusion matrix for the one correlation case (Test set).

	Manual pathological	Manual normal
Auto pathological	0.947	0.105
Auto normal	0.053	0.895

Table 4: Mean confusion matrix for the 10 training sets (One correlation case).

	Manual pathological	Manual normal
Auto pathological	0.943	0.109
Auto normal	0.057	0.891

A particular point is highlighted in the *ROC* (black square), corresponding to the threshold located at the crossing point of the two distributions in Figure 10 (threshold = -0.3). For this threshold, the confusion matrix is shown in Table 2 (*Acc* = 0.9446). These first results are already satisfying.

Now that the most discriminant correlation has been chosen and its classification performance assessed on the training set, this performance has to be evaluated for samples that are not part of the training set, here samples forming the test set. When applying the threshold of the correlation on samples of the test set, one obtains the confusion matrix shown in Table 3 (*Acc* = 0.9426). One can see that the performance is in the same order as for the training set although the chosen correlation and its threshold lead to lower classification performance for the normal samples. It must be emphasized here that the normal samples are much less represented than the pathological ones in the *MEEI* Database, and thus in training and test sets. Consequently a misclassification of a normal sample leads to a higher variation of classification performance than a misclassification of a pathological sample because *#Normal* is much lower than *#Pathological*. That is why the results of classification should be interpreted while keeping in mind this difference.

Table 5: Mean confusion matrix for the 10 test sets (One correlation case).

	Manual pathological	Manual normal
Auto pathological	0.955	0.074
Auto normal	0.045	0.926

Table 6: Accuracy for the 10 pairs of training and test sets (One correlation case).

Number	Training set	Test set
1	0.946	0.947
2	0.942	0.947
3	0.942	0.947
4	0.940	0.951
5	0.940	0.951
6	0.938	0.955
7	0.929	0.971
8	0.938	0.955
9	0.940	0.963
10	0.942	0.945

In order to validate the fact that the chosen correlation and its thresholding are the most appropriate for the distinction between normal and pathological samples, 10 training sets and 10 test sets (different from the training and test sets defined in Section 4) have been randomly formed from the samples of the *MEEI* Database. This has been done in the proportion described in Section 4. For each training set, the Fisher analysis has been performed. It turned out that the most discriminant correlation is always the correlation between the first spectral tristimulus in the Bark scale and the spectral decrease. Moreover, it appeared that the same threshold than the one corresponding to the crossing point of the distributions in Figure 10 could be appropriate for the classification task. Therefore this threshold has been applied on the chosen correlation in the 10 training sets and 10 test sets and the associated confusion matrices have been computed. Tables 4, 5, and 6 show, respectively, the mean confusion matrix for the 10 training sets, the mean confusion matrix for the 10 test sets, and the accuracy for the 10 pairs of sets.

All these results confirm that the chosen correlation and its associated threshold perform well in the task of discriminating between normal and pathological samples.

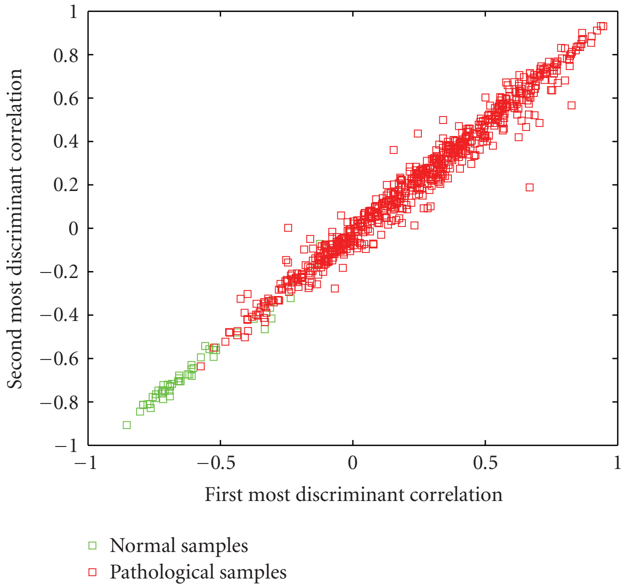

Figure 12: Location of the normal and pathological samples of the training set in the "two most discriminative correlations" representation space (Normal: green; Pathological: red).

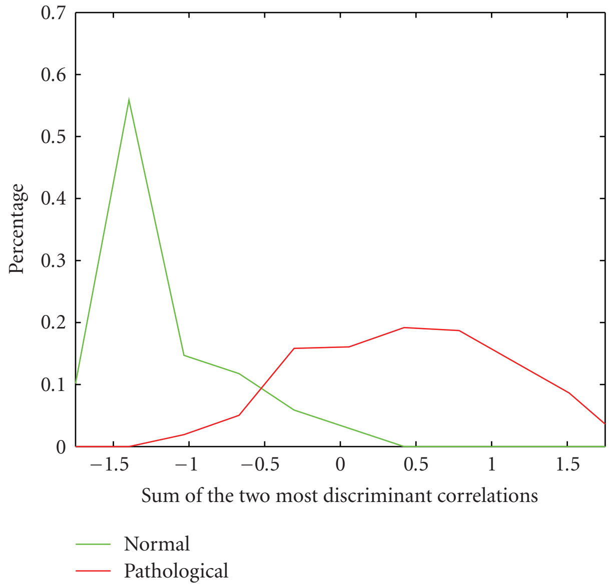

Figure 13: Distribution of the sum of the two most discriminative correlations for the two corpora of the training set (Normal: green; Pathological: red).

7.2.2. Two Correlations Case. In this case the two correlations associated to the highest discriminant power are selected. The first one is the correlation between the first spectral tristimulus in the Bark scale and the spectral decrease and the second one is the correlation between the relative loudness in the first Bark band (see Section 5.5.12) and the spectral decrease. For these two correlations, the location of the normal and pathological samples of the training set is shown in Figure 12. Based on this distribution, it has been chosen to evaluate to what extent the sum of the two correlations could separate the normal and pathological samples.

The distribution of this sum for the samples of the training set is represented in Figure 13. As for the "One Correlation Case," it may be possible to split the two populations by thresholding the sum of the two correlations. The *ROC* for thresholds uniformy distributed between lower and upper limits of this sum is computed in the same way as for the previous case and is shown in Figure 14. The same

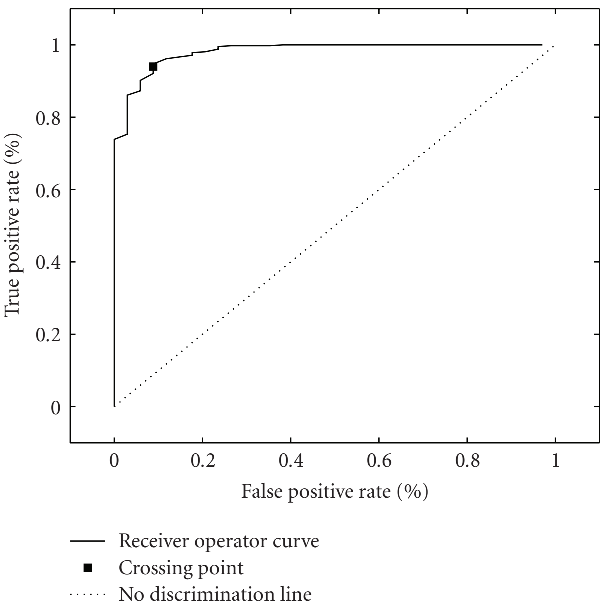

Figure 14: *ROC* for the "Two Correlations Case."

Table 7: Confusion matrix for the two correlations case (Training set).

	Manual pathological	Manual normal
Auto pathological	0.935	0.088
Auto normal	0.065	0.912

Table 8: Confusion matrix for the two correlations case (Test set).

	Manual pathological	Manual normal
Auto pathological	0.938	0.105
Auto normal	0.062	0.895

Table 9: Mean confusion matrix for the 10 training sets (Two correlations case).

	Manual pathological	Manual normal
Auto pathological	0.930	0.106
Auto normal	0.070	0.894

Table 10: Mean confusion matrix for the 10 test sets (Two correlations case).

	Manual pathological	Manual normal
Auto pathological	0.945	0.074
Auto normal	0.055	0.926

Table 11: Accuracy for the 10 pairs of training and test sets (Two correlations case).

Number	Training set	Test set
1	0.933	0.934
2	0.933	0.934
3	0.931	0.939
4	0.925	0.951
5	0.931	0.938
6	0.927	0.947
7	0.920	0.960
8	0.927	0.947
9	0.929	0.943
10	0.929	0.943

remarks that in the first case can be made about the meaning of the curve. Moreover, when comparing the two curves, one may observe that, for a same value of *TPR*, the second configuration is better that the first one for *TPR* values below 0.92 and the contrary for higher values.

A particular point is highlighted in the *ROC* (black square), corresponding to the threshold located at the crossing point of the two distributions in Figure 13 (threshold $= -0.5$). For this threshold, the confusion matrix is shown in Table 7 ($Acc = 0.9335$). One may remark that, comparing to Table 2, the correct classification rate is lower for the pathological class (correct classification decreased by 1.2%) but remains unchanged for the normal class. The same threshold has been applied on the sum of the two correlations for the samples in the test set. The results are shown in Table 8 ($Acc = 0.9394$). One can see that the performance is in the same order as for the training set although the chosen correlation and its threshold again lead to slightly lower classification performance for the pathological samples and unchanged classification performance for the normal ones.

The 10 training and test sets of validation defined in the "One Correlation Case" have been used to assess the validity of the "Two Correlations Case" approach. For each training set, the Fisher analysis has been performed and it turned out that the two most discriminant correlations are always the correlation between the first spectral tristimulus in the Bark scale and the spectral decrease and the correlation between the relative loudness in the first Bark band and the spectral decrease. Moreover, it appeared that the same threshold than the one corresponding to the crossing point of the distributions in Figure 13 could be appropriate for the classification task. Therefore this threshold has been applied on the sum of the chosen correlations for the 10 pairs of training and test sets and the associated confusion matrices have been computed. Tables 9, 10, and 11 show, respectively, the mean confusion matrix for the 10 training sets, the mean confusion matrix for the 10 test sets, and the accuracy for the 10 pairs of sets.

Concerning the mean confusion matrices, the classification performance for the pathological samples is in both cases lower than in the "One Correlation Case" (see Table 5). When looking at the accuracies, one can see that they are lower than the ones in the "One Correlation Case" for all the pairs of sets (see Table 6).

7.3. Discussion. The application of a feature selection on the correlations between acoustic descriptors has proved its ability to separate the normal and pathological samples in the *MEEI* database. When comparing the "One Correlation

Case" and the "Two Correlations Case," one may say that the first one is better than the second one. This decision is supported by different considerations. Firstly the *ROC* of the first configuration is better than the *ROC* of the second one for *TPR* higher than 0.92. Although it corresponds to higher *FPR*, this result is better because the *FPR* is more sensitive to misclassification than *TPR* (because *#Normal* is lower than *#Pathological*). Secondly, when comparing the confusion matrices, it has been found that the second configuration leads to lower classification performance than the first one for the pathological samples and unchanged performances for the normal ones. It is of significant importance since it is more important to detect all the pathological samples than all the normal ones. Thirdly, when comparing the accuracies, they are always lower for the second configuration than for the first one. The accuracy depending on *#Normal* and *#Pathological*, a lower number of correctly classified pathological samples induces a smaller accuracy than a lower number of correctly classified normal samples. The second configuration is thus characterized, by means of accuracy, by a higher number of misclassified pathological samples than the first configuration. Fourthly, the first configuration has the advantage to keep more facilities of interpretation than the second configuration (only one correlation instead of the sum of two correlations). Finally, the first configuration only requires the computation of two spectral features and one correlation while the second one requires the computation of three spectral features and two correlations.

Some interpretations can be given about the features selected in the first configuration. As shown in Figure 10, the normal samples are characterized by correlation values concentrated around -0.75. That means that the evolution of the spectral decrease and the first spectral tristimulus are fairly strong linked for a large majority of the samples, although this link is not perfect (because the correlation is not -1 but -0.75). The pathological samples are characterized by a larger dispersion of the correlation value, meaning that for some samples the two features are fairly slightly linked and for others no link exists at all. Although the trend is clearer for normal samples than for the pathological ones, one must keep in mind that the number of normal samples is much lower than the number of pathological ones in the database.

Concerning the speech utterances used in this work, it is interesting to discuss about the sense of jointly assess sustained and continuous speech samples since these two kinds of samples are included in the *MEEI* database. On the one hand, the sustained vowel offers the advantage to be acquired in relatively stable conditions, meaning that the characteristics of the source and the vocal tract are quite stable. This enables computing features and especially their perturbation in an easier way than in the case of continuous speech. The correlation between features is also easier to understand and to interpret. Besides, analyzing the sustained vowels also allows the computation and the interpretation of features to be relatively less influenced than continuous speech by intonation, stress, or phonetic context. On the other hand, continuous speech reflects more the dynamics of speech production since the characteristics of source and vocal tract are no longer stable. This production includes onset, terminations, variation of pitch and amplitude, and voice breaks. According to clinicians, this kind of information is also informative about the presence of pathology and more representative of the every-day life of a patient than the sustained vowel. Assessing jointly sustained vowels and continuous speech seems to make sense because these two kinds of productions describe different (but complementary) conditions: the sustained vowel is more relative to stable conditions while continuous speech is more relative to dynamic conditions.

Apart from the discussion above, it must be emphasized that the output of the analysis system presented in this paper is a normal/pathological factor (see the overview of the system in Section 1). When a new subject is presented at the analysis system, this output could be the value of the most discriminant correlation and the position of the subject according to the distribution of this correlation in the test database. The aim would be not to provide an unilateral decision about the presence of pathology or not but to provide an indication to the clinician, who remains the person who has the final appreciation.

8. Conclusion

A classification scheme between normal and pathological voices has been presented in this paper. When applied on speech samples extracted from the *MEEI* database, this system provides a correct classification rate of 94.7% for pathological samples and 89.5% for normal samples. Regarding to litterature, these results are slightly below those offered by methods basing on this database but our method is unique in several aspects: the considered features are not based on the estimation on the fundamental period, they come from both the normal/pathologic voice assessment and Music Information Retrieval domains, the correlation between selected features is used to discriminate normal and pathological samples instead of using a complex classifier. Besides, a potential use of our system is the computation of a normal/pathological factor, aiming at giving an indication to the clinician about the location of a subject according to the database.

Among the future works, the test of this classification system on larger databases is planned in order to see if using correlation remains powerful for the discrimination between the two populations. Using the mutual information for estimating the link between features will also been investigated since it has not been considered in this study. Finally some features provided by the source-tract separation of speech could be integrated in the system in order to see if they are relevant for classification purposes.

Acknowledgments

The authors would like to thank the Walloon Region, Belgium, for its support (Grant WALEO II ECLIPSE 516009), the Université Catholique de Louvain (ORL-ORLO Laboratory) for the availability of the database and the COST Action 2103 "Advanced Voice Function Assessment." This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its author(s).

References

[1] M. Hirano, *Psycho-Acoustic Evaluation of Voice: GRBAS Scale for Evaluating the Hoarse Voice*, Springer, Berlin, Germany, 1981.

[2] K. Marasek, "An attempt to classify lx signals," in *Proceedings of the 4th European Conference on Speech Communication and Technology (EuroSpeech '95)*, Madrid, Spain, September 1995.

[3] D. Deliyski, "High-speed videoendoscopy: recent progress and clinical prospects," in *Proceedings of the 7th International Conference on Advances in Quantitative Laryngology Voice and Speech Research (AQL '06)*, vol. 7, pp. 1–12, University of Groningen, 2006.

[4] J. Demeyer and B. Gosselin, "Glottis segmentation with a highspeedglottography: a new approach," in *Proceedings of Liege Image Days*, Liege, Belgium, March 2008.

[5] M. Vasilakis and Y. Stilyanou, "A mathematical model for accurate measurement of jitter," in *Proceedings of 5th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA '07)*, Firenze, Italy, December 2007.

[6] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, Springer, Berlin, Germany, 2008.

[7] 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 (EuroSpeech '05)*, pp. 149–152, Lisbon, Portugal, September 2005.

[8] G. Pouchoulin, C. Fredouille, J. Bonastre, A. Ghio, M. Azzarello, and A. Giovanni, "Modélisation statistique et informations pertinentes pour la caractérisation des voix pathologiques (dysphonies)," in *Proceedings of JEP (Journée d'Etudes sur la Parole)*, 2006.

[9] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proceedings of the 2nd Joint Engineering in Medicine and Biology, 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society (BMES/EMBS '02)*, vol. 1, pp. 182–183, Houston, Tex, USA, October 2002.

[10] A. Dibazar and S. Narayanan, "A system for automatic detection of pathological speech," in *Proceedings of the 36th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, Calif, USA, November 2002.

[11] A. Dibazar, T. Berger, and S. Narayanan, "Pathological voice assessment," in *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '06)*, New York, NY, USA, August 2006.

[12] J. Godino-Llorente, P. Gómez-Vilda, N. Sáenz-Lechón, M. Blanco-Velasco, F. Cruz-Roldán, and M. A. Ferrer, "Discriminative methods for the detection of voice disorders," in *Proceedings of International Conference on Non-Linear Speech Processing (NOLISP '05)*, Barcelona, Spain, April 2005.

[13] K. E. Corp, "Multi-dimensional voice program (mdvp) [computer program]," Tech. Rep., Kay Elemetrics Corp., 2008.

[14] K. E. Corp, "Disordered voice database model (version 1.03)," Tech. Rep., Massachussets Voice Eye and Ear Infirmary Voice and Speech Lab, 1994.

[15] J. I. Godino-Llorente, S. Aguilera-Navarro, C. Hernandez-Espinosa, M. Fernandez-Redondo, and P. Gomez-Vilda, "On the selection of meaningful speech parameters used by a pathologic/non pathologic voice register classifier," in *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH '99)*, Budapest, Hungary, September 1999.

[16] R. B. Reilly, R. Moran, and P. Lacy, "Voice pathology assessment based on a dialogue system and speech analysis," in *Proceedings of the AAAI Symposium on Dialogue Systems for Health Communication*, pp. 104–109, 2004.

[17] R. J. Moran, R. B. Reilly, P. de Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 468–477, 2006.

[18] D. Michaelis, M. Fröhlich, and H. W. Strube, "Selection and combination of acoustic features for description of pathologic voices," *Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1628–1639, 1998.

[19] J. B. Alonso, J. de Leon, I. Alonso, and M. A. Ferrer, "Automatic detection of pathologies in the voice by HOS based parameters," *EURASIP Journal on Advances in Signal Processing*, vol. 2001, no. 4, pp. 275–284, 2001.

[20] J. Alonso, F. D. de Maria, C. Travieso, and M. Ferrer, "Using nonlinear features for voice disorder detection," in *Proceedings of International Conference on Non-Linear Speech Processing (NOLISP '05)*, Barcelona, Spain, April 2005.

[21] H. Kasuya, Y. Endo, and S. Saliu, "Novel acoustic measurements of jitter and shimmer characteristics from pathological voice," in *Proceedings of 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, Geneva, Switzerland, September 2003.

[22] A. Alpan, F. Grenez, and J. Schoentgen, "Estimation of vocal noise and cycle duration jitter in connected speech," in *Proceedings of MAVEBA*, 2007.

[23] C. Ferrer, M. E. Hernandez-Diaz, and E. Gonzalez, "Using waveform matching techniques in the measurement of shimmer in voiced signals," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH '07)*, vol. 4, pp. 2436–2439, Antwerp, Belgium, August 2007.

[24] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Transactions on Signal Processing*, vol. 39, no. 1, pp. 40–48, 1991.

[25] G. de Krom, "Spectral correlates of breathiness and roughness for different types of vowel fragments," in *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP '94)*, Yokohama, Japan, September 1994.

[26] E. O'Leidhin and P. Murphy, "Analysis of spectral measures for voiced speech with varying noise and perturbation levels," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, Philadelphia, Pa, USA, March 2005.

[27] M. Fröhlich, D. Michaelis, and H. Strube, "Acoustic breathiness measures in the description ofpathologic voices," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, Seattle, Wash, USA, May 1998.

[28] W. Wszolek, R. Tadeusiewicz, A. Izworski, and T. Wszolek, "Automated understanding of selected voice tract pathologies based on the speech signal analysis," in *Proceedings of the 23rd*

Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '01), vol. 2, pp. 1719–1722, Istanbul, Turkey, October 2001.

[29] F. Severin, B. Bozkurt, and T. Dutoit, "Hnr extraction in voiced speech oriented towards voice quality analysis," in *Proceedings of 13th European Signal Processing Conference (EUSIPCO '05)*, Antalya, Turkey, September 2005.

[30] C. d'Alessandro, B. Yegnanarayana, and V. Darsinos, "Decomposition of speech signals into deterministic and stochastic components," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '95)*, vol. 1, pp. 760–763, Detroit, Mich, USA, May 1995.

[31] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences (IFA '93)*, Amsterdam, The Netherlands, 1993.

[32] K. Shama, A. Krishna, and N. U. Cholayya, "Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 85286, 9 pages, 2007.

[33] M. Wester, "Automatic classification of voice quality: comparing regression models and hidden markov models," in *Proceedings of Symposium on Databases in Voice Quality Research and Education (VOICEDATA '98)*, 1998.

[34] H. Kasuya, S. Ogawa, and Y. Kikuchi, "Adaptive comb filtering method as applied to acoustic analyses of pathological voice," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86)*, pp. 669–672, Tokyo, Japan, April 1986.

[35] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio—a new measure for describing pathological voices," *Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[36] G. Pouchoulin, C. Fredouille, J.-E. Bonastre, A. Ghio, and A. Giovanni, "Frequency study for the characterization of the dysphonic voices," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH '07)*, vol. 3, pp. 1789–1792, Antwerp, Belgium, August 2007.

[37] J. Wang and C. Jo, "Performance of gaussian mixture models as a classifier for pathological voice," in *Proceedings of the 11th Australian International Conference on Speech Science and Technology*, 2006.

[38] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, NY, USA, 2006.

[39] P. Kukharchik, I. Kheidorov, E. Bovbel, and D. Ladeev, "Image and signal processing," in *Speech Signal Processing Based on Wavelets and SVM for Vocal Tract Pathology Detection*, Lecture Notes in Computer Science, pp. 192–199, Springer, Berlin, Germany, 2008.

[40] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," Tech. Rep., IRCAM, 2004.

[41] MPEG-7, "Information technology - multimedia content description interface—part 4: audio," Tech. Rep. ISO/IEC JTC 1/SC 29, ISO/IEC FDIS 15938-4, 2002.

[42] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, vol. 2, pp. 1331–1334, Munich, Germany, April 1997.

[43] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 1, pp. 333–336, Orlando, Fla, USA, May 2002.

[44] K. Martin and Y. Kim, "2pmu9. musical instrument identification: a pattern-recognition approach," in *Proceedings of the 136th Meeting of The Acoustical Society of America*, 1998.

[45] E. Wold, T. Blum, D. Keislar, and J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE Transactions on Multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[46] G. Peeters and X. Rodet, "Hierarchical gaussian tree with inertial ratio maximization for the classification of large musical instrument databases," in *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx '03)*, London, UK, September 2003.

[47] N. Misdariis, B. Smith, D. Pressnitzer, P. Susini, and S. McAdams, "Validation of a multidimensionnal distance model for perceptual dissimilarities among musical timbres," *Journal of the Acoustical Society of America*, vol. 103, 1998.

[48] C. E. Pearson, "Probability and Statistics," in *Handbook of Applied Mathematics*, pp. 1185–1196, Van Nostrand Reinhold, New York, NY, USA, 1974.

[49] G. Korn and T. Korn, *Mathematical Handbook for Scientists and Engineers*, chapter 18, McGraw-Hill, New York, NY, USA, 1967.

[50] X. Serra and J. Bonada, "Sound transformation based on the sms high level attributes," in *Proceedings of the International Conference on Digital Audio Effects (DAFx '98)*, 1998.

[51] S. Stevens and J. V. E. Newman, "A scale for the measurement of the psychological magnitude of pitch," *Journal of the Acoustical Society of America*, vol. 8, pp. 185–190, 1937.

[52] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *Journal of the Acoustical Society of America*, 1961.

[53] H. Pollard and E. Jansson, "A tristimulus method for the specification of musical timbres," *Acustica*, vol. 51, pp. 162–171, 1982.

[54] K. Jensen, "Multiple scale music segmentation using rythm, timbre and harmony," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 73205, 11 pages, 2007.

[55] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[56] K. Fukunaga, *Introduction to Statistical Pattern Recognition* , Academic Press, New York, NY, USA, 2nd edition, 1990.

[57] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, NY, USA, 1973.