Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Biomedical Signal Processing and Control 1 (2006) 120–128

Biomedical Signal Processing and Control

www.elsevier.com/locate/bspc

Methodological issues in the development of automatic systems for voice pathology detection

Nicolás Sáenz-Lechón [a,*], Juan I. Godino-Llorente [a], Víctor Osma-Ruiz [a], Pedro Gómez-Vilda [b]

[a] *Department Ingeniería de Circuitos y Sistemas, Universidad Politécnica de Madrid, Spain*

[b] *Department Arquitectura y Tecnología de Sistemas Informáticos, Universidad Politécnica de Madrid, Spain*

Received 15 February 2006; received in revised form 29 May 2006; accepted 28 June 2006

Available online 7 September 2006

Abstract

This paper describes some methodological concerns to be considered when designing systems for automatic detection of voice pathology, in order to enable comparisons to be made with previous or future experiments.

The proposed methodology is built around the *Massachusetts Eye & Ear Infirmary* (*MEEI*) Voice Disorders Database, which to the present date is the only commercially available one. Discussion about key points on this database is included.

Any experiment should have a cross-validation strategy, and results should supply, along with the final confusion matrix, confidence intervals for all measures. Detector performance curves such as *detector error trade off* (*DET*) and *receiver operating characteristic* (*ROC*) plots are also considered.

An example of the methodology is provided, with an experiment based on short-term parameters and multi-layer perceptrons.

Keywords: Voice pathology detection; Pathological voice databases; MEEI Voice Disorders Database; Cross-validation; ROC and DET curves

1. Introduction

In the last years a large number of works have focused on the automatic detection and classification of voice pathologies, by means of acoustic analysis, parametric and non-parametric feature extraction, automatic pattern recognition or statistical methods. Table 1 lists several of these studies chronologically, and gives a good idea of the variety of approaches that can be found in existing literature. As can be seen, many research groups in speech technology have broached these problems, using their own databases and signal processing techniques. However, there is a lack of uniformity in their overall development that makes it very difficult to reach valid conclusions regarding the proposed methods. Focusing only on the databases, there are studies like [1] that employed just a few voice recordings, while others used several hundreds [2]. The type of the pathologies collected in these databases is broad, including all kinds of organic and functional lesions [3–5], although some studies concentrate only on laryngeal cancer [1,6–8]. Recording conditions are also varied, ranging from prepared rooms with professional equipments [9] to commercial sound cards under normal acoustic conditions [5,10]. Sampling frequencies vary from 8 kHz [1,6] up to 22 kHz [5]. The recorded material consists usually on sustained vowels, such as the Italian /a/ in [3], Japanese /e/ in [2] or the five Spanish vowels in [5,11]. In [7,9], electroglottographic data are also used along with the speech signals. Finally, most of the studies employed adult patients, with different distributions of male and female subjects, while [7] used exclusively male voices, and [4] employed children.

In 1999, Campbell and Reynolds wrote [12]: "*The use of standard speech corpora for development and evaluation is one of the major factors behind progress over the last 10 years in automatic speech processing research, particularly in speech and speaker recognition. Perhaps the main benefit of using standard corpora is that it allows researchers to compare performance of different techniques on common data, thus making it easier to determine which approaches are most*

[*] Corresponding author. Tel.: +34 913365527; fax: +34 913367829.
E-mail address: nicolas.saenz@upm.es (N. Sáenz-Lechón).

1746-8094/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bspc.2006.06.003

Table 1
Summary of several research works on voice pathology detection, detailing the number of patients in the database (normal + pathologic), the acoustic features employed and the classification method

First author	DB	Features	Classifier	Best results (%)
Banci [3]	30 + 53	Pitch, noise, residual	Mahalanobis distance	91.6
Kasuya [2]	804 + 66	Perturbation, noise	Thresholds	–
Childers [9]	52 + 29	Linear prediction coefficients (LPC)	Vector quantization	82.6
Plante [4]	209 + 88	Cepstral, perturbation, noise	Thresholds	89
Wallen [6]	9 + 20	Perturbation, cepstral, LPC	Multi-layer perceptron	85.5
Gavidia [1]	10 + 24	Spectral	Hidden Markov models	92.8
Boyanov [10]	50 + 150	Perturbation, noise, energies	*K*-nearest neighbours, linear discriminant analysis, self-organized maps	96.5
Ritchings [7]	20 + 20	Spectral	Multi-layer perceptron	90.5
Tadeusiewicz [8]	?	Formants, perturbation	Multi-layer perceptron	84.75
Cheol-Woo [15]	33 + 33	Noise, perturbation	Multi-layer perceptron	91.6
Alonso [5,11]	100 + 68	High order statistics, bispectrum, chaos	Multi-layer perceptron	98.3

promising to pursue. In addition, standard corpora also can be used to measure current state-of-the-art performance in research areas for particular tasks and highlight deficiencies that require further research".

Unfortunately, in the area of voice quality assessment and automatic pathology detection there are no such standard speech corpora. As it is impossible to compare results when the experiments are performed with a private database, we have concentrated our study on works using the *Massachusetts Eye & Ear Infirmary* (*MEEI*) Voice Disorders Database, distributed by *Kay Elemetrics* [13], which is the only one that is commercially available and is rather extended. Table 2 lists several works with this database that will be discussed later on. But even when this database was employed in the state of the art, there were many differences in the way its files were chosen and handled. Furthermore, the experiments were carried out with such different criteria, that comparisons were fruitless.

Detection of voice pathology is closely related to a speaker verification task [14], where a candidate sample is compared against two different models (target and impostors versus normal and pathological speech). The system must provide a hard decision and a confidence score about which model the sample belongs to. We aim to develop a methodology that allows results from different classifiers and features to be compared. Thus we have adopted some methodological issues which are common in speaker verification.

The paper is organized as follows: Section 2 covers the *MEEI* database and discusses some of its particularities. Section 3 contains an overview of previous work on pathological voice detection using this database. Sections 4 and 5 present the proposed methodology and describe a simple experiment of detection based upon it. Finally, Section 6 presents discussion and conclusions.

2. MEEI voice disorders database

The *MEEI Voice Disorders Database* (*VDDb*) was delivered in 1994 [13]. It was compiled partly at the *MEEI Voice and*

Table 2
Summary of several research works with *MEEI VDDb*, detailing the number of patients used (normal + pathologic) and if there is an indication of which the files were, the features employed, the classification method if any and results obtained

First author	*MEEI VDDb*	Features	Classifier	Best results (%)
Qi [27]	0 + 48; Unknown	Harmonics-to-noise ratio	–	–
Cheol-Woo [28]	Unknown	Wavelet transform	–	–
Godino [33]	53 + 82; Unknown	Mel-frequency cepstral coefficients	Multi-layer perceptron, learning vector quantization	95
Wester [29]	36 + 607; Unknown	Harmonics-to-noise ratio	Hidden Markov models; linear discriminant analysis	65
Parsa [16,36]	53 + 173; Known	Noise	Linear discriminant analysis	98.7
Hadjitodorov [30]	53 + 638; Unknown	Perturbation, noise	Vector quantization, linear discriminant analysis	92.7
Dibazar [31]	All	Acoustic parameters given by MDVP, Mel-frequency cepstral coefficients	Hidden Markov models, Gaussian mixture models, multi-layer perceptron	98.3
Maguire [32]	58 + 573; Unknown	Perturbation, noise, Mel-frequency cepstral coefficients	Linear discriminant analysis	87.16
Moran [37]	58 + 573; Unknown	Perturbation, noise	Linear discriminant analysis	89.1
Marinaki [34]	21 + 42; Unknown	Linear prediction coefficients	Linear discriminant analysis, three nearest neighbours	85
Umapathy [35]	51 + 161; Unknown	Adaptive time–frequency transform	Linear discriminant analysis	93.4

Speech Lab. and partly at *Kay Elemetrics Corp*. It contains recordings of sustained phonation of vowel /ah/ (53 normal and 657 pathological files) and continuous speech (53 normal and 661 pathological). For this description we will mainly focus on the former ones.

The database also includes a spreadsheet with clinical and personal details from the subjects and the results of the acoustic analysis of the recordings, obtained with *Kay*'s *Multi-Dimensional Voice Program* (*MDVP*). The recordings were performed in matching acoustic conditions, using *Kay*'s *Computerized Speech Lab.* (*CSL*). Each subject was asked to produce a sustained phonation of vowel /ah/ at a comfortable pitch and loudness for at least 3 s. The process was repeated three times for each subject, and a speech pathologist chose the best sample for the database.

The usefulness of this database is clear and has been repeatedly tested in numerous research works since its development. Moreover, it is the most widespread and available of all the voice quality databases, but there are some key points that should be carefully taken into account when used for research purposes:

- Not all the pathological patients have corresponding recordings nor diagnoses, and there are some patients with more than one recording, from different visits to the clinic. Table 3 shows detailed information about the pathological subset of recordings of vowel /ah/.
- The files have different sampling frequencies. Normal and a small percentage of pathological files have 50 kHz, whereas most of the pathological ones have 25 kHz. In order to unify these frequencies, all the files should be down-sampled at least to 25 kHz before further processing.
- Normal and pathological voices were recorded at different locations (*Kay Elemetrics* and *MEEI Voice and Speech Lab.*, respectively), assumedly under the same acoustic conditions, but there is no guarantee that this fact has no influence in an automatic detection system. Normal subjects were not clinically evaluated, although according to [16], none of them had ‘‘complaints or history of voice disorders’’.
- The files are already edited to include only the stable part of the phonation. Several studies consider that the onset and offset parts of the phonation contain more acoustic information than stable parts [17]. The editing also makes it impossible to know the signal-to-noise ratio of the recordings.
- The normal files have an average length of 3 s for sustained vowels and 12 s for running speech, while pathological files have averages of around 1 and 9 s, respectively. These differences are possibly due to the fact that it is difficult for some pathological subjects to phonate for a long time. Hence, some of the *MDVP* measurements, provided with the spreadsheet, could be misleading, such as *SEG* (number of analyzed segments), *PER* (number of detected pitch periods), etc. Common sense dictates that when training automatic models, it has to be ensured that the length is not used as a parameter to discriminate between classes.
- There is only one phonation per patient and visit. Sometimes it is useful to have available several samples of the same vowel to model intra-speaker variability or samples of different vowels [18,19].
- There is a heterogeneous number of pathologies in the database, with almost 200 different diagnoses, probably because they were included as they were captured in the clinical practice. There are a lot of files labelled with several diagnoses, pertaining sometimes to different categories of voice disorders (e.g. physical and neuromuscular). According to [20], the only mutually exclusive possible categorization is at the highest level (i.e. ‘‘normal’’ and ‘‘pathological’’).
- There is a limited number of normal recordings in comparison with the number of pathological ones. This is a problem for training supervised pattern recognition systems, which work best with large amounts of data which are well balanced between the different classes.
- There is no perceptual subjective evaluation of the recordings, such as GRBAS [21] or others [22–24], which would be very useful for research purposes. This would require a similar number of recordings of each perceptual rank.
- There are no video recordings (stroboscopy, endoscopy). The importance of this kind of material is highlighted in [25]. Moreover, there are no electroglottographic (*EGG*) data with the voice registers. *EGG* signals have proven to be an important complement for acoustic analysis and detection of pathology [9,26].
- The database has been on the market for more than 10 years now and has been extensively used for pathology detection. Therefore, the newly developed algorithms and parameters could have been adapted to these particular data. The results obtained with this database should be contrasted with new databases to take this possibility into account.

Table 3
Pathological recordings of sustained vowel /ah/ in *MEEI VDDb*

	No. of visits	No. of patients
Spreadsheet entries	720	617
Audio recordings	657	566
Files without diagnosis	306	253
Files with diagnosis ‘‘normal’’	6	6
Remainder files	345	307

3. Pathological voice detection with MEEI VDDb

This section presents an overview of previous works found in the literature using *MEEI VDDb*. The objective here is to concentrate on the way they handle the database, how they design the experiments and evaluate their results.

In [27], Qi and Hillman employed 48 voices from *MEEI* to test an algorithm to compute a harmonics-to-noise ratio (*HNR*) in the spectral domain. They used some of the original files, prior to being edited, not detailed and not publicly available.

In 1998, Cheol-Woo and Dae-Hyun [28] proposed two novelty measurements, based on the wavelet transform, and compared their discriminative power against some of the available *MDVP* features, though they do not state which files were employed.

In 1998, Wester [29] compared linear regression techniques and Hidden Markov Models (*HMM*) to detect voice pathologies. She employed 36 normal and 607 pathological voices from the running speech files. A number of *HNR*-based features were extracted by acoustic analysis every 10 ms. 80% of the data were used to train the system and the rest were for testing it. The word ''sunlight'' was segmented from each file, and perceptually evaluated by two expert listeners into three scales: roughness, breathiness, and general degree of deviance. Results were favourable to *HMM*s yielding best results close to 65% of correct classification rate.

Parsa and Jamieson in 2000 [16] approached the detection task based on six different noise measurements. They employed 53 normal and 173 pathological voices, enumerated in an appendix. All the files were down-sampled to 25 kHz, chosen to have a diagnosis and similar age distributions between both groups. Only the first second of each file was used. Discrimination results were obtained building the histograms of the two classes and *receiver operating characteristic* (*ROC*) curves were employed to compare them, yielding a best accuracy of 98.7% using a spectral flatness ratio (*SFR*) measure.

Hadjitodorov and Mitev in 2002 [30] describe a system for acoustic analysis of voice, which also allows the automatic detection of pathology, using jitter, shimmer and noise measurements. Classification is achieved by means of linear discriminant analysis (*LDA*) and nearest neighbours (*NN*) clustering. They employed 106 normal (''two phonations by each non-pathological speaker'') and 638 pathological files. The accuracy of the system was 96.1%.

Dibazar et al. [31] presented some of the best results in pathology detection with this database. They used all the files in the database, along with all *MDVP* parameters, short-term *Mel-Frequency Cepstral Coefficients* (*MFCC*) and fundamental frequency. They classified the voices with *HMM*s, to achieve a best accuracy of 98.3%. However, they do not give many methodological details due to the great amount of experiments carried out.

Maguire et al. [32] propose a pathology detector, based on sustained phonation, combining long-term acoustic, spectral and *cepstral* parameters. They used 58 normal and 573 pathological voices. The classifier was a *LDA* with a 10-fold cross-validation strategy. They achieved an 87.16% accuracy with a subset of the *MDVP* parameters (shimmer and noise features).

Godino et al. have several papers using this database. In [33] they employed 53 normal files and 82 pathological files, the latter chosen randomly among the whole database. All files were down-sampled to 25 kHz. The files were short-term parameterised using *MFCC*s and their derivatives, and the detector system was based on neural networks (multi-layer perceptrons, *MLP* and learning vector quantifiers, *LVQ*). The training test was composed with 70% of the files from each class. Results were presented with confusion matrices, providing confidence intervals for the measurements, yielding a best accuracy of around 95% with *LVQ*.

Moran and co-workers [20] presented a telephone system for detecting voice pathologies, with the same data and classifying scheme as [32], down-sampling the recordings to 10 kHz. They used 36 short-term parameters based on jitter, shimmer and noise measures. The system yielded 89.1% accuracy for the original data and 74.15% for simulated telephone data.

Marinaki et al. [34] implemented a system to distinguish between 21 normal speakers, 21 patients with vocal fold paralysis and 21 patients with vocal fold edema, with similar distributions of age and gender. These patients had also other pathologies. They use short-term linear predictive coding (*LPC*) parameters, principal components analysis (*PCA*) and *LDA* to classify the voices. Results yielded an accuracy of nearly 85% and were presented via *ROC* curves.

Umapathy et al. [35] presented a detection system using 51 normal and 161 pathological continuous speech samples. They extracted five new acoustic parameters based on adaptive time–frequency transformations (*ATFT*) and employed *LDA* and leave-one-out cross-validation to discriminate between classes. Results were presented by means of *ROC* curves, areas under the curves and confusion matrices. Best results achieve a correct classification rate of 93.4%.

Although all these works represent novel contributions to automatic detection of voice disorders and to voice quality assessment and they share the same database, their achievements and conclusions are not easily comparable, due to a lack of uniformity when computing and presenting the results. There does not normally exist an adequate description of the files used or the reasons for using them. Many works use pathological files without a diagnosis or make a subset of the database without taking into consideration the distribution of features such as gender, age, or origin. Another important point, which has rarely been addressed until recently, is the reliability of the results. These only provide a single measure of the performance of a detector, but they do not take into account what the answer of the system would be when facing unknown data. This can only be achieved through cross-validation and confidence intervals.

4. Methodology

Bearing in mind all of the considerations presented in the previous sections, our goal is to discuss a series of key points for designing experiments to automatically distinguish pathological voices from normal ones, and to set up a methodology that could allow comparisons between different experiments, in order to outline the benefits of each approach.

The first thing to establish should be the database. A good decision is to use *MEEI VDDb*, due to its availability.

If any other database is available, it could be a good choice to repeat the experiments with both databases, in order to test the robustness or independence of the algorithms to the database. We have considered only a subset of all the available files, 53 normal and 173 pathological voices, according to Parsa and Jamieson [16]. This paper included a list with the names of the

Table 4
Gender and age distribution of the recordings in the chosen subset of *MEEI* database, adapted from [16]

	Subjects		Margin (years)		Mean (years)		Standard deviation (years)	
	Male	Female	Male	Female	Male	Female	Male	Female
Normal	21	32	26–58	22–52	38.8	34.2	8.49	7.87
Pathologic	70	103	26–58	21–51	41.7	37.6	9.38	8.19

Table 5
Typical aspect of a confusion matrix

Detector's decision	Actual diagnosis	
	Pathological	Normal
Pathological	tp	fp
Normal	fn	tn

tp, fp, fn and tn stand for true positive, false positive, false negative and true negative rates, respectively. See text for definitions.

pathological files, so it is easy to reuse their work.[1] The splitting was accomplished to assure that all the files have a diagnosis (though very often more than one), and gender and age characteristics are uniformly distributed between the two classes (Table 4).

After this, the files in the database should be arranged into at least two sets, one for training and one for testing and validating the results. Typical possible sizes for these sets are around 70 and 30% of the files, respectively. When the feature extraction is performed on a short-term basis, it is important to avoid mixing segments from the same file in both sets, which could affect the results [38].

Once the system is trained or the models have been computed, the test set is employed to estimate the performance of the detector. The final results are presented through confusion matrices (Table 5), where we define the next measures: *true positive* rate (*tp*), also called *sensitivity*, is the ratio between pathological files correctly classified and the total number of pathological voices. *False negative* rate (*fn*) is the ratio between pathological files wrongly classified and the total number of pathological files. *True negative* rate (*tn*), sometimes called *specificity*, is the ratio between normal files correctly classified and the total number of normal files. *False positive* rate (*fp*) is the ratio between normal files wrongly classified and the total number of normal files. The final accuracy of the system is the ratio between all the hits obtained by the system and the total number of files. If the parameterization is performed on a short-term basis, then these measures can also be calculated on a segment or frame basis, besides of on a file basis.

To assess the generalization capabilities of the system, it is important to adopt a cross-validation scheme [39, Chapter 9]. A simple one is to repeat each experiment N times, with a different test set, randomly chosen from the whole set of files, or the *K-fold* cross-validation, where the dataset is randomly split in K different subsets and the experiment is repeated K times, using each time a different subset for testing the performance. When the number of folds is equal to the number of available files F, then this method is known as *leave-one-out* cross-validation. The experiment is repeated F times, and each time the system is trained with $F-1$ files, leaving the remaining file for testing. This method is computationally expensive. After the cross-validation, the final results are averaged across these repetitions, and confidence intervals can be computed using the standard deviation of the measures.

When we use short-term parameters, accuracies for both frames and files should be presented.

Statistics also provide other simple indicators of the generalization error for linear models, working under certain conditions of the sample. These statistics can be also considered as rough estimations of the generalization error for non-linear models if the database is large enough, although their adaptation to the non-linear model framework (e.g. neural networks) is not always possible.

Eq. (1) represents a statistic used in speech technology to measure the generalization error [40]. Testing with N patterns and obtaining an accuracy p, the confidence interval for this measure is:

$$CI = \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{N}} \tag{1}$$

where the value z is obtained from a standard normal distribution as a function of the required confidence level α. Using $\alpha = 0.05$ (95% of confidence), z is 1.96.

The calculation requires statistical independence among the patterns, so although the training patterns would be short-time features, N should be the number of speech files available (and p should be calculated on a file basis). This is a conservative approach, but matches well with the assumption that there must be expected more inter than intra-speaker variability. Regarding the amount of data needed for this kind of measurements, it is useful to consider the "*rule of 30*" found in [41]: "*To be 90% confident that the true error rate is within ±30% of the observed error rate, there must be at least 30 errors*". This rule could lead to the need of huge amounts of data for an experiment to be considered meaningful. An alternative is to assume independency among the patterns and not to take assertions about the statistical significance of the results too seriously.

During the system testing, a score representing the likelihood of the input vector for belonging to the desired class (i.e. pathological voice) is given. This score has to be compared to a threshold value in order to compute the confusion matrix. If we move this threshold we obtain a set of possible operating points for the system, which can be represented through a *detector error trade off* plot [42], widely used in speaker verification. In this curve, the false positives (or *false acceptances*) are plotted against the false negatives (*false rejections*), for different threshold values (Fig. 1, left). Another

[1] In fact, Parsa and Jamieson affirm that they use 175 pathological files in their work, but in the final list they give two repeated file names. Data in Table 4 have been adapted to reflect just the 173 registered files.

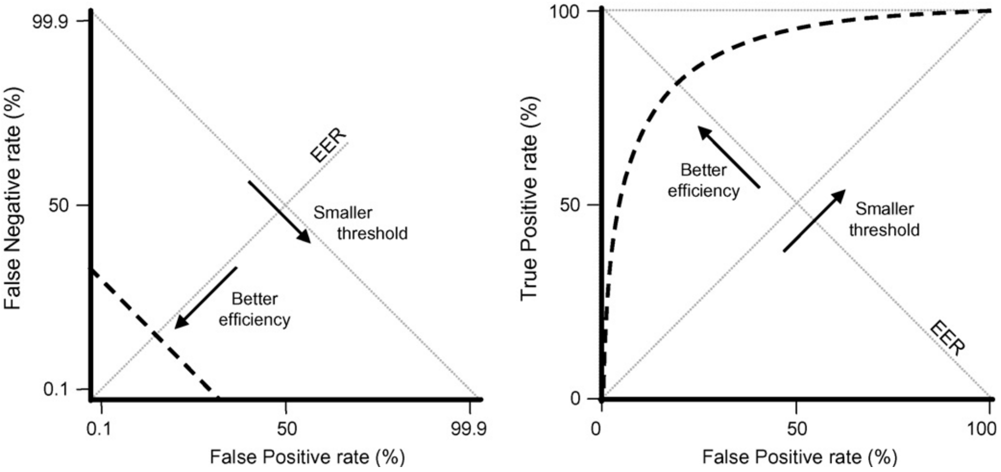

Fig. 1. Schematic depiction of DET (left) and ROC (right) curves.

choice is to represent the false positives in terms of the true positives in a *receiver operating characteristic* (Fig. 1, right), which is very common in medical decision systems [43]. There are numerical measurements that convert the whole curve into a single figure, to compare different curves from different systems. The most usual is the *area under ROC curve* (*AUC*) [44]. If the performance of a system is poor, the *AUC* will be close to 0.5 and if the performance is perfect, the area will be 1. This area also allows the calculation of other statistics to determine if there exist statistical significances between different curves drawn from the same data. As presented in [45], another way to obtain a measurement of the confidence interval is to calculate the standard error of the *AUC*.

In a *detector error trade off* (*DET*) plot, the better the detector, the closer the curve will get to the bottom-left corner. In a *ROC* plot, the curve gets nearer to the upper-left corner as the efficiency of the system improves. Another important difference between both types of curves is that the *DET* is plotted in a normal deviate scale; therefore, if the data distributions from the two classes are close to normal, the curve will tend to be linear. This is typically the case when data come from cross-validation averaging of different folds.

5. An example detector

The goal of the following experiment is not to improve the results of previous works in the state of the art, but to illustrate the proposed methodology with a brief example. We have designed an automatic system following [33]: the voices from the database are processed in a short-term basis and used to train a neural network that models the two classes of voices.

The voices are selected from *MEEI VDDb* as detailed in the previous section (53 normal speakers, 173 pathological patients from [16]). Each file is segmented into frames, using 40 ms Hamming windows, with a 50% of overlapping between consecutive frames. This size ensures that each frame contains at least one pitch period. The frames are analyzed in order to detect silence of unvoiced segments, which are removed.

From the remaining frames, a certain number of *Mel-Frequency Cepstral Coefficients* are extracted. For this example we have chosen 18 coefficients. *MFCC*s have been calculated following a non-parametric approach, based on the human auditory perception system. The term ''*mel*'' refers to a frequency scale related to the human perceptual auditory system. The mapping between the real frequency scale (Hz) and the perceived frequency scale (mels) is approximately linear below 1 kHz and logarithmic for higher frequencies. This matches with the idea that a trained speech therapist is able, most of the times, to detect the presence of a disorder by just listening to the speech.

The detector is a basic feedforward *multi-layer perceptron* (*MLP*) with three layers [46, Chapter 6]. The input layer is made of as many inputs as *MFCC* parameters, the hidden layer has 12 neurons and the output layer has two nodes. These two outputs are employed to obtain a logarithmic likelihood ratio or *score* from every input pattern. Supervised learning is carried out by the *backpropagation* algorithm with delta rule and momentum. The activation functions on all nodes are logistic. The connection weights are initialized with random values drawn from a Gaussian distribution of zero mean and a standard deviation inverse to the number of weights of each neuron. The training is performed on-line, that is, the weights are updated immediately after each example is presented to the net. For this example, 40 iterations of the training algorithm were performed.

The database is split into two subsets: a training set with the 70% of the normal and pathological files and a test set with the remainder 30%. The normal recordings are approximately three times longer than the pathological recordings, so they produce more short-term frames per file than the latter. This is compensated by the bigger number of pathological files in the database (173–53). The data in the training set are normalized to be in the range [0, 1] and shuffled randomly. The test set is normalized according to the normalization values used for the training set.

The experiment was repeated 10 times, each time building different training and test sets randomly. The scores produced by the neural network are used to calculate the curves that

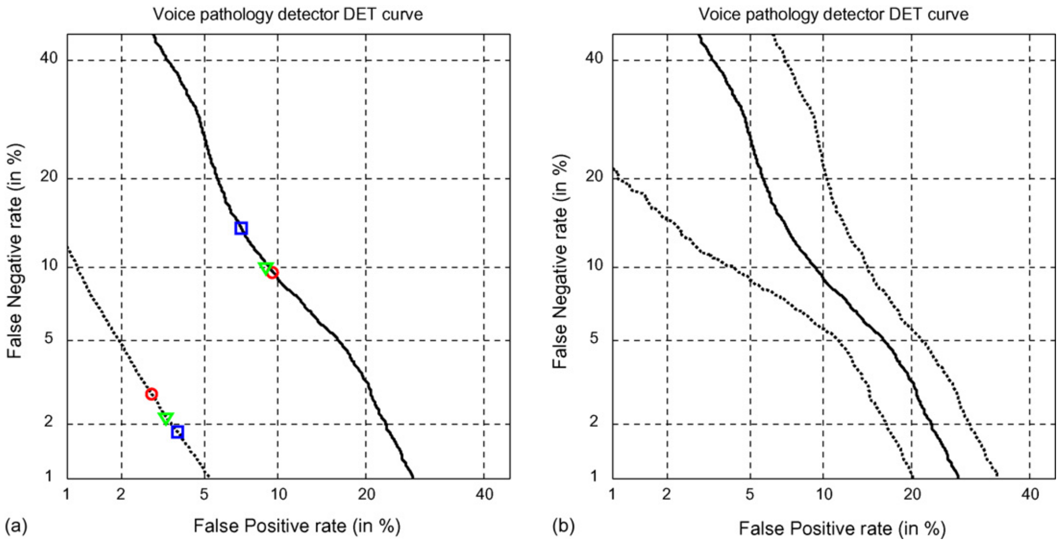

Fig. 2. (a) Averaged DET plots for the designed detector. The dotted line corresponds to the training set and the solid line to the test set. (b) Averaged DET plot for the test set of the designed detector along with confidence intervals based on standard deviation.

represent the system's performance. Fig. 2a shows two *DET* plots reflecting the overall performance of the detector. The dotted line has been calculated with the averaged training data (i.e., the scores obtained with the 10 training sets), while the solid line corresponds to the averaged testing data. The distance between these two lines is related to the generalization capabilities of the system. There are three different points marked in the figure. The point marked with a round circle is the *equal error rate* point (*EER*) [47], that is, the point for which the false positives rate equals the false negatives rate. The point marked with a square is the operating point of the system, that is, the point that reflects the performance at the chosen threshold. In this case, the operating point corresponds to threshold 0. An ordinary decision is to use the same threshold for which the *EER* point was obtained on the training set. Another point of interest is the *detection cost function* (*DCF*) [42], which is the point on the curve that minimizes the classification error, having in mind the global costs of false positives and false negatives and the a priori probabilities of both classes. This point is marked in Fig. 2a with a triangle, considering that these costs and probabilities are the same for normal and pathological voices. The *DCF* should be the ideal point of operation. The distance between the actual operating point and the *DCF* is also an indication of the degradation of the system performance due to the unknown data in the test set.

Fig. 2b shows the same *DET* plot, along with the confidence intervals obtained with 1 standard deviation of the 10 experiments. These bands give an idea of the confidence of the decisions made by the detector.

Fig. 3 shows the *ROC* curve corresponding to the averaged data from the 10 test sets. When the performance of the system is high, the curve is close to the upper-left corner, so it is difficult to compare visually several curves. For this matter, the *area under the curve* is a useful statistic. In this case, the *AUC* is 0.9578.

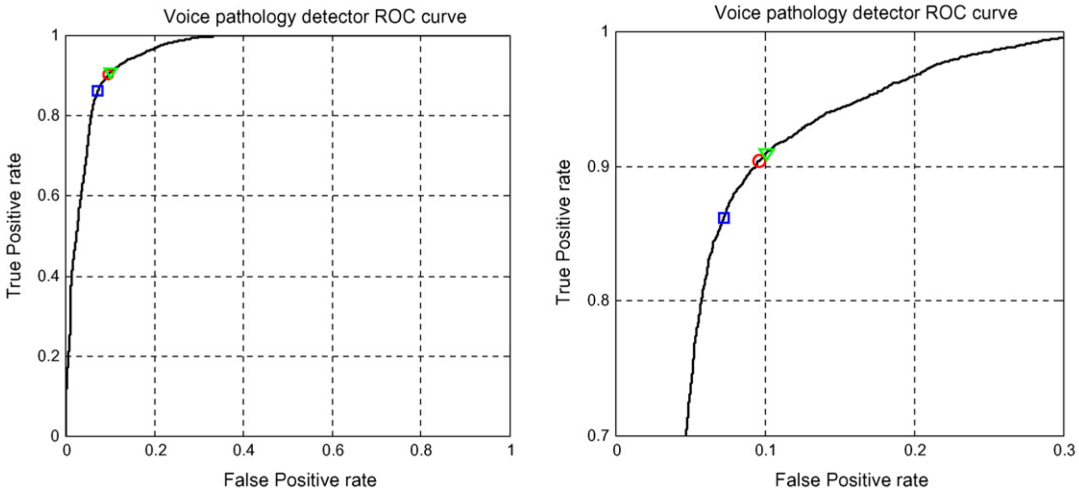

Fig. 3. ROC plot of the designed detector. The right plot is a zoom of the region of interest. The points marked in the plots correspond to those marked in the DET plots.

Table 6
Results of the classification (in %) given in a frame basis (mean ± standard deviation)

Detector's decision	Actual diagnosis	
	Pathological	Normal
Pathological	92.73 ± 5.24	13.66 ± 6.49
Normal	7.27 ± 5.24	86.34 ± 6.49

Once we have chosen a point of operation, we can compute the final performance measures. Table 6 shows the confusion matrix of the system with the mean and standard deviation values obtained averaging the results for each individual experiment.

The total accuracy of the system is 89.6 ± 2.49%. The accuracy on a file basis (the percentage of correctly classified recordings) is 90.42 ± 0.04%. This measure is computed for each experiment, and then averaged, by setting up a threshold to the number of classified frames. If more than 50% of the frames of a file are assigned to a certain class, then the whole file is assumed to belong to that class.

6. Conclusions

The only way to improve and to profit from other works is to have an objective means to measure the efficiency of different approaches. In this paper, we have described a set of requirements that an automatic detector of voice pathologies should meet to allow comparisons with other systems.

The database of pathological voices is an essential point in any research, and so we have suggested the use of a well described subset of the only commercially available one. We have adopted a cross-validation strategy based on several partitions of the whole dataset, in order to obtain averaged classification ratios along with confidence intervals for every measure. We prefer to present the results by means of a *DET* curve, because when averaging different folds of tests, the curves tend to be linear and this allows us to compare several systems at a glance more easily than with *ROC* curves. Measurements of the area under the *ROC* curve are also valuable for objective comparisons.

As far as we know, there were no previous works in existing literature addressing these issues. We intend to continue the research in pathological voice detection and classification using the presented methodology. In any case, it seems evident that new publicly available and well designed databases are needed. In all probability it will be necessary to carry out a public evaluation of pathological voice assessment systems, such as the NIST's Speaker Recognition ones.

Acknowledgements

This work was supported by the grants: TIC-2003-08956-C02-00 and TIC-2002-2273 from the *Ministry of Science and Technology* and AP2001-1278 from the *Ministry of Education* of Spain. The authors wish to thank the attendants at Maveba'05 Workshop for their discussion on these topics.

References

[1] L. Gavidia-Ceballos, J.H.L. Hansen, Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection, IEEE Trans. Biomed. Eng. 43 (4) (April 1996) 373–383.

[2] H. Kasuya, K. Masubuchi, S. Ebihara, H. Yoshida, Preliminary experiments on voice screening, J. Phonetics 14 (1986) 463–468.

[3] G. Banci, S. Monini, A. Falaschi, N. de Sario, Vocal fold disorder evaluation by digital speech analysis, J. Phonetics 14 (1986) 495–499.

[4] F. Plante, J. Borel, C. Berger-Vachon, I. Kauffmann, Acoustic detection of laryngeal diseases in children, in: Proceedings of Eurospeech '93, Berlin, Germany, (1993), pp. 1965–1968.

[5] J.B. Alonso, J. de León, I. Alonso, M.A. Ferrer, Automatic detection of pathologies in the voice by HOS based parameters, EURASIP J. Appl. Signal Process. 2001 (4) (2001) 275–284.

[6] E.J. Wallen, J.H.L. Hansen, A screening test for speech pathology assessment using objective quality measures, in: Proceedings of ICSLP '96, vol. 2, Philadelphia, PA, USA, (October 1996), pp. 776–779.

[7] R.T. Ritchings, G.V. Conroy, M.A. McGillion, C.J. Moore, N. Slevin, S. Winstanley, H. Woods, A neural network based approach to objective voice quality assessment, in: Proceedings of the 18th International Conference on Expert Systems ES '98, Cambridge, UK, (1998), pp. 198–209.

[8] R. Tadeusiewicz, W. Wszolek, M. Modrzejewski, The evaluation of speech deformation treated for larynx cancer using neural network and pattern recognition methods, in: Proceedings of EANN '98, 1998, pp. 613–617.

[9] D.G. Childers, K. Sung-Bae, Detection of laryngeal function using speech and electroglottographic data, IEEE Trans. Biomed. Eng. 39 (1) (January 1992) 19–25.

[10] B. Boyanov, S. Hadjitodorov, Acoustic analysis of pathological voices. A voice analysis system for the screening of laryngeal diseases, IEEE Eng. Med. Biol. Mag. 16 (4) (July/August 1997) 74–82.

[11] J.B. Alonso, F. Díaz, C.M. Travieso, M.A. Ferrer, Using nonlinear features for voice disorder detection, in: Proceedings of NOLISP '05, Barcelona, Spain, (April 2005), pp. 94–106.

[12] J.P. Campbell, D.A. Reynolds, Corpora for the evaluation of speaker recognition systems, in: Proceedings of ICASSP '99, vol. 2, Phoenix, AZ, USA, (March 1999), pp. 829–832.

[13] Massachusetts Eye and Ear Infirmary, Voice Disorders Database, Version.1.03 [CD-ROM], Kay Elemetrics Corp., Lincoln Park, NJ, 1994.

[14] J.P. Campbell, Speaker recognition: a tutorial, Proc. IEEE 85 (9) (September 1999) 1437–1462.

[15] J. Cheol-Woo, K. Dae-Hyun, Classification of pathological voice into normal/benign/malignant state, in: Proceedings of Eurospeech '99, vol. 1, Budapest, Hungary, (1999), pp. 571–574.

[16] V. Parsa, D.G. Jamieson, Identification of pathological voices using glottal noise measures, J. Speech Language Hearing Res. 43 (2) (April 2000) 469–485.

[17] G. de Krom, Consistency and reliability of voice quality ratings for different types of speech fragments, J. Speech Hearing Res. 37 (5) (October 1994) 985–1000.

[18] Y. Horii, Jitter and shimmer in sustained vocal fry phonation, Folia Phoniatrica 37 (1985) 81–86.

[19] J.L. Fitch, Consistency of fundamental frequency and perturbation in repeated phonations of sustained vowels, reading, and connected speech, J. Speech Hearing Disorders 55 (May 1990) 360–363.

[20] R.B. Reilly, R. Moran, P.D. Lacy, Voice pathology assessment based on a dialogue system and speech analysis, in: Proceedings of the American Association of Artificial Intelligence Fall Symposium on Dialogue Systems for Health Communication, Washington, DC, USA, 2004.

[21] M. Hirano, Clinical Examination of Voice, Springer Verlag, New York, 1981.

[22] P. Carding, E. Carlson, R. Epstein, L. Mathieson, C. Shewell, Formal perceptual evaluation of voice quality in the United Kingdom, Logopedics Phoniatrics Vocol. 25 (3) (2000) 133–138.

[23] V. Wolfe, J. Fitch, D. Martin, Acoustic measures of dysphonic severity across and within voice types, Folia Phoniatrica et Logopaedica 49 (6) (1997) 292–299.

[24] J. Kreiman, B.R. Gerratt, Validity of rating scale measures of voice quality, J. Acoustical Soc. Am. 104 (Pt 1 (3)) (September 1998) 1598–1608.

[25] M. Fröhlich, D. Michaelis, E. Kruse, Image sequences as necessary supplement to a pathological voice database, in: Proceedings of Voicedata '98, Utretch, The Netherlands, (January 1998), pp. 64–69.

[26] R.T. Ritchings, M.A. McGillion, C.J. Moore, Pathological voice quality assessment using artificial neural networks, Med. Eng. Phys. 24 (8) (2002) 561–564.

[27] Y. Qi, R.E. Hillman, Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals, J. Acoustical Soc. Am. 102 (1) (1997) 537–543.

[28] J. Cheol-Woo, K. Dae-Hyun, Analisys of disordered speech signal using wavelet transform, in: Proceedings of ICSLP '98, Sydney, Australia, 1998.

[29] M. Wester, Automatic classification of voice quality: comparing regression models and hidden Markov models, in: Proceedings of Voicedata '98, Utretch, The Netherlands, (January 1998), pp. 92–97.

[30] S. Hadjitodorov, P. Mitev, A computer system for acoustic analysis of pathological voices and laryngeal disease screening, Med. Eng. Phys. 24 (6) (January 2002) 419–429.

[31] A.A. Dibazar, S. Narayanan, T.W. Berger, Feature analysis for automatic detection of pathological speech, in: Proceedings of the Second Joint EMBS/BMES Conference, vol. 1, Houston, TX, USA, (November 2002), pp. 182–183.

[32] C. Maguire, P. de Chazal, R.B. Reilly, P.D. Lacy, Identification of voice pathology using automated speech analysis, in: Proceedings of MAVEBA 2003, Florence, Italy, (December 2003), pp. 259–262.

[33] J.I. Godino-Llorente, P. Gómez-Vilda, Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors, IEEE Trans. Biomed. Eng. 51 (2) (February 2004) 380–384.

[34] M. Marinaki, C. Kotropoulos, I. Pitas, N. Maglaveras, Automatic detection of vocal fold paralysis and edema, in: Proceedings of ICSLP '04, Jeju Island, South Korea, November 2004.

[35] K. Umapathy, S. Krishnan, V. Parsa, D.G. Jamieson, Discrimination of pathological voices using a time–frequency approach, IEEE Trans. Biomed. Eng. 52 (3) (March 2005) 421–430.

[36] V. Parsa, D.G. Jamieson, Acoustic discrimination of pathological voice: sustained vowels versus continuous speech, J. Speech Language Hearing Res. 44 (2) (April 2001) 327–339.

[37] R. Moran, R.B. Reilly, P. de chazal, P.D. Lacy, Telephone based voice pathology assessment using automated speech analysis and Voice XML, in: Proceedings of the Irish Signals and Systems Conference, Belfast, Ireland, 2004.

[38] J.I. Godino-Llorente, R.T. Ritchings, C. Berry, The effects of inter and intra speaker variability on pathological voice quality assessment, in: Proceedings of MAVEBA 2003, Florence, Italy, (December 2003), pp. 157–160.

[39] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., Wiley Interscience, 2000.

[40] J. Ferreiros, J.M. Pardo, Improving continuous speech recognition in Spanish by phone-class semicontinuous HMMs with pausing and multiple pronunciations, Speech Commun. 29 (1) (September 1999) 65–76.

[41] G.R. Doddington, M.A. Przybocki, A.F. Martin, D.A. Reynolds, The NIST speaker recognition evaluation—overview, methodology, systems, results, perspective, Speech Commun. 31 (2–3) (June 2000) 225–254.

[42] A.F. Martin, G.R. Doddington, T. Kamm, M. Ordowski, M.A. Przybocki, The DET curve in assessment of detection task performance, in: Proceedings of Eurospeech '97, vol. IV, Rhodes, Crete, (1997), pp. 1895–1898.

[43] J.A. Swets, R.M. Dawes, J. Monahan, Better decisions through science, Scientific Am. (October 2000) 82–87.

[44] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (April 1982) 29–36.

[45] J.A. Hanley, B.J. McNeil, A method of comparing the areas under receiver operating characteristics curves derived from the same cases, Radiology 148 (3) (September 1983) 839–843.

[46] S. Haykin, Neural Networks, Macmillan, New York, 1994.

[47] D.A. Reynolds, Speaker identification using Gaussian mixture speaker models, Speech Commun. 17 (1995) 91–108.