

ORIGINAL ARTICLE

## On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices

JULIÁN DAVID ARIAS-LONDOÑO<sup>1</sup>, JUAN I. GODINO-LLORENTE<sup>1</sup>, MARIA MARKAKI<sup>2</sup> & YANNIS STYLIANOU<sup>2</sup>

<sup>1</sup>Universidad Politécnica de Madrid, Circuits & Systems Engineering, EUIT de Telecomunicación, Universidad Politécnica de Madrid, Ctra. Valencia, km 7, Madrid, 28031 Spain, and <sup>2</sup>University of Crete, Computer Science Department, Heraklion, Crete, 71409 Greece

### Abstract

This work presents a novel approach for the automatic detection of pathological voices based on fusing the information extracted by means of mel-frequency cepstral coefficients (MFCC) and features derived from the modulation spectra (MS). The system proposed uses a two-stepped classification scheme. First, the MFCC and MS features were used to feed two different and independent classifiers; and then the outputs of each classifier were used in a second classification stage. In order to establish the best configuration which provides the highest accuracy in the detection, the fusion of information was carried out employing different classifier combination strategies. The experiments were carried out using two different databases: the one developed by The Massachusetts Eye and Ear Infirmary Voice Laboratory, and a database recorded by the Universidad Politécnica de Madrid. The results show that the combination of MFCC and MS features employing the proposed approach yields an improvement in the detection accuracy, demonstrating that both methods of parameterization are complementary.

**Key words:** *Automatic assessment of voice, combining pattern classifiers, Gaussian mixture models, mel-frequency cepstral coefficients, modulation spectra, pathological voices, support vector machines*

### Introduction

The automatic assessment of voice quality based on acoustic analysis is an efficient tool for the objective support of the diagnosis and the screening of vocal and voice diseases. Most of the works carried out in this area have been oriented to the study of acoustic parameters and pitch perturbation measurements and noise (1,2). Nevertheless, the approaches based on the characterization of the spectral components to identify the abnormal glottal activity have also been shown to be reliable in the detection of pathological voices (3,4). In this sense, the state of the art reports two parameterization approaches previously used in this context that provided good results: mel-frequency cepstral coefficients (MFCC) (5), and features extracted from the modulation spectra (MS) (6).

MFCC have been formerly used for the detection of pathological voices with good results (3). The main advantage of the MFCC parameters is that, in contrast to other parameters found in the state of the art (2,7–9), its calculation does not require a previous pitch estimation, a difficult task in the presence of pathologies. Besides, the alterations related to the mucosal waveform due to an increase of mass are reflected in the low bands used to calculate the MFCC, whereas the higher bands are able to model the noisy components due to a lack of closure (4).

On the other hand, MS may be seen as a non-parametric way to represent the modulation present in the speech introduced by the presence of pathologies, since dysphonic voices are characterized by frequency-band-dependent time-varying amplitude fluctuations (11). Moreover, it offers a compact way

Correspondence: Julián David Arias-Londoño, Universidad Politécnica de Madrid, Circuits & Systems Engineering, EUIT de Telecomunicación, Universidad Politécnica de Madrid, Ctra. Valencia, km 7, Madrid, 28031 Spain. E-mail: jdarias@ics.upm.es

(Received 17 May 2010; accepted 28 September 2010)

to fuse the various phenomena observed during speech production (i.e. noise, frequency and amplitude perturbations, tremor, etc.), providing important and complementary dynamic information useful for the detection of pathological voices (4), since MS are able to capture the amplitude envelope fluctuations evident during sustained vowel phonations (11).

In our previous work (12), we fused the information from both families of parameters (i.e. MFCC and MS), and the improvements obtained in the detection accuracy led us to conclude that both parameterization approaches are complementary and provide important information for the characterization of pathological voices. Nevertheless, there are important differences related to the parameterization procedure of both families of features, complicating the necessary merging procedure that has to be followed to feed a single classifier. Due to this fact, this work investigates the use of several strategies to combine classifiers in order to fuse the information from MS and MFCC. The advantage of combining outputs of different classifiers instead of merging features is that the structure of the feature space used to feed each classifier is much simpler. Furthermore, although one of the classifiers would yield a better performance, the set of speech registers misclassified by each classifier would not necessarily overlap; therefore the combination of their outputs could improve the overall performance of the system (13).

This work employed a strategy based on combining classifiers to fuse information from MFCC and features extracted from the MS. Two classifiers based on Gaussian mixture models (GMM) and support vector machines (SVM) were used, since their modeling and discrimination capabilities have demonstrated their reliability for this task (3). The experiments were carried out using two different pathological speech databases, and the results are presented in terms of confusion matrices and figures of merit.

The paper is organized as follows: The Methods section describes the parameterization approaches used in this work, gives a brief overview of the pattern classification methods, and describes the strategies used to combine the classifiers. The Experiments and decision-making section describes the experiments carried out, and is followed by the Results section. And finally, Conclusions is dedicated to a brief discussion and the main conclusions.

## Methods

Figure 1 depicts a schematic diagram of the system developed for the automatic detection of pathological voices.

The speech signal is divided into frames in order to be parameterized by means of short-time analysis.

The optimum frame size required for MFCC and MS is very different, hence the proposed scheme uses two parallel feature extraction and classification procedures.

The final classification was carried out in two steps. First, two different and independent classifiers were trained from MFCC and MS features. Then, in order to establish the best configuration to merge the information, and with the aim of comparing the results, different strategies for combining the outputs of such classifiers were tested.

### Parameterization

The speech signal is parameterized following a frame basis approach. For the case of MFCC, the frames were segmented with 40 ms Hamming windows extracted with a 50% frame overlap. This length ensures that every frame contains, at least, two consecutive pitch periods (3). On the other hand, the MS was estimated using longer windows (250 ms) shifted 50 ms. According to a criterion based on singular value decomposition and mutual information, the most

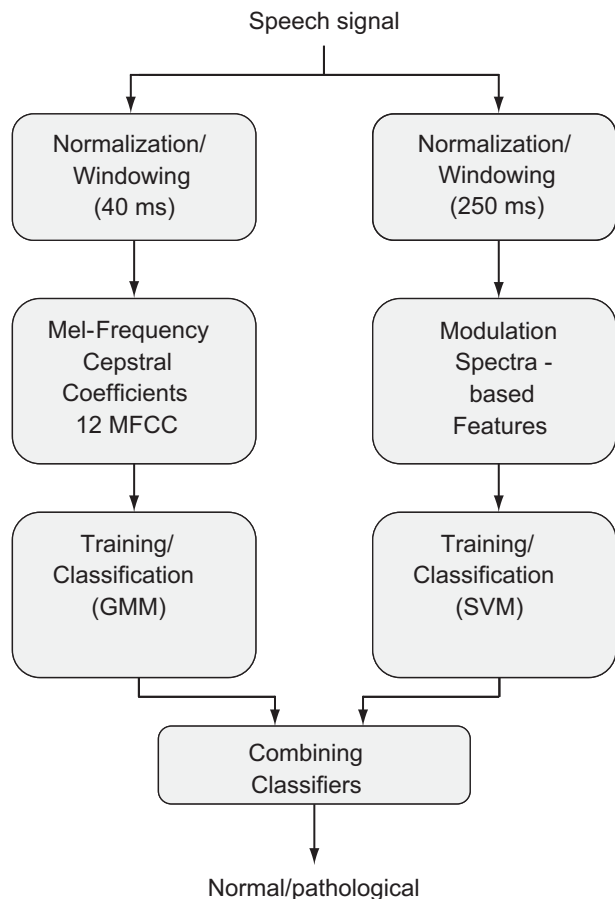


Figure 1. General scheme of the system developed for the automatic detection of pathological voices.

relevant measures derived from MS were selected as features to feed the classifier.

The following sections give an overview of the parameterization procedures used in this work.

*Modulation spectra.* The most common modulation frequency analysis framework (6) for a discrete signal,  $x(n)$ , employs a short-time Fourier transform (STFT) that for each frame under analysis,  $m$ , is given by  $X_k(m)$ :

$$X_k(m) = \sum_{n=-\infty}^{\infty} h(mM-n)x(n)e^{-j\frac{2\pi}{K}kn} \quad (1)$$

$$k = 0, \dots, K-1$$

where  $h(\cdot)$  is the acoustic frequency analysis window with a hop size of  $M$  samples and length of  $K$  samples, with  $m$  denoting time. At this stage, the resolution in the frequency domain is also of  $K$  samples, but a mel scale filtering can be employed to reduce the resulting number of frequency bins,  $K$ . The distribution of envelope amplitudes of voiced speech has a strong exponential component. Hence we use a log transformation of the amplitude values  $|X_k(m)|$  and subtract their mean log amplitude:

$$\hat{X}_k(m) = \log|X_k(m)| - \overline{\log|X_k(m)|} \quad (2)$$

where  $\{\bar{\cdot}\}$  denotes the average operator over  $m$ .

Next, a second STFT is used to detect the frequency content of  $|\hat{X}_k(m)|$ :

$$X_l(k, s) = \sum_{m=-\infty}^{\infty} g(lL-m)|\hat{X}_k(m)|e^{-j\frac{2\pi}{S}sm} \quad (3)$$

$$s = 0, \dots, S-1$$

where  $g(\cdot)$  is the modulation frequency analysis window, and  $L$  is the hop size (in samples);  $k$  and  $s$  are referred to as the ‘acoustic’ and ‘modulation’ frequencies, respectively. The tapered windows,  $h(\cdot)$  and  $g(\cdot)$ , are used to reduce the side lobes of both frequency estimates.

Then, the representation of a MS displays the modulation spectral energy  $|X_l(k, s)|$  (magnitude of the sub-band envelope spectra) in the joint acoustic/modulation frequency plane. In order to enable cross-database portability of the classification system, a feature sub-band normalization has been employed according to Markaki et al. (14). Every acoustic frequency sub-band was normalized with the marginal of the modulation frequency representation:

$$X_{l, norm}(k, s) = \frac{X_l(k, s)}{\sum_s X_l(k, s)} \quad (4)$$

The modulation spectra were computed on a frame-by-frame basis using 250 ms windows shifted by 50 ms. A mel scale filtering was applied with 53 bands, while the size of the Fourier transform for the time domain transformation was set to 257. Therefore, each modulation spectrum consisted of  $K = 53$  acoustic frequencies and  $S = 257$  modulation frequencies, resulting therefore in a  $53 \times 257$  image per frame. The normalized modulation spectra computed for each frame were stacked to obtain a third-order real tensor  $D \in R^{K \times S \times F}$ , where  $F$  is the number of frames.

Figure 2 depicts two MS obtained respectively for a normal and a pathological speech signal. The figure shows that the energy at the modulations corresponding to fundamental frequency and its harmonics is localized in the lower acoustic frequencies for pathological speech signals. Such behaviour is not observed for normal voices.

Since MS contains a large amount of data and is estimated using a frame-by-frame strategy, it is necessary to reduce the dimensionality to obtain the most relevant features. This work employed a methodology based on a third-order generalization of singular value decomposition (HOSVD) (15) which projects the features over the singular vectors of the acoustic and modulation frequency subspaces with the highest energy. This approach has been successfully used for discriminative tasks in speech processing (16). The projection of the MS features over the principal axes with the highest energy in each subspace results in a compact set of features with minimum redundancy.

Further, a criterion based on mutual information (MI) was used to select the more relevant features for the classification task. The relevance is defined as the mutual information  $I(x_j; c)$  between feature  $x_j$  and class  $c$ . The maximum relevance (*MaxRel*) feature selection criterion simply selects the most relevant features to the target class  $c$  (17). Through a sequential search, which does not require estimation of multivariate densities, the most important  $q$  features of  $I(x_j; c)$  in descending order are selected.

Applying the HOSVD algorithm, the near-optimal projections or principal axes (PC) of features were detected among those contributing more than 0.1% to the ‘energy’ of  $D$ . For MEEI database (Massachusetts Eye and Ear Infirmary Voice Laboratory), 44 PC were detected in the acoustic frequency and 29 PC in the modulation frequency subspace, resulting in a reduced space of  $44 \times 29 = 1,276$

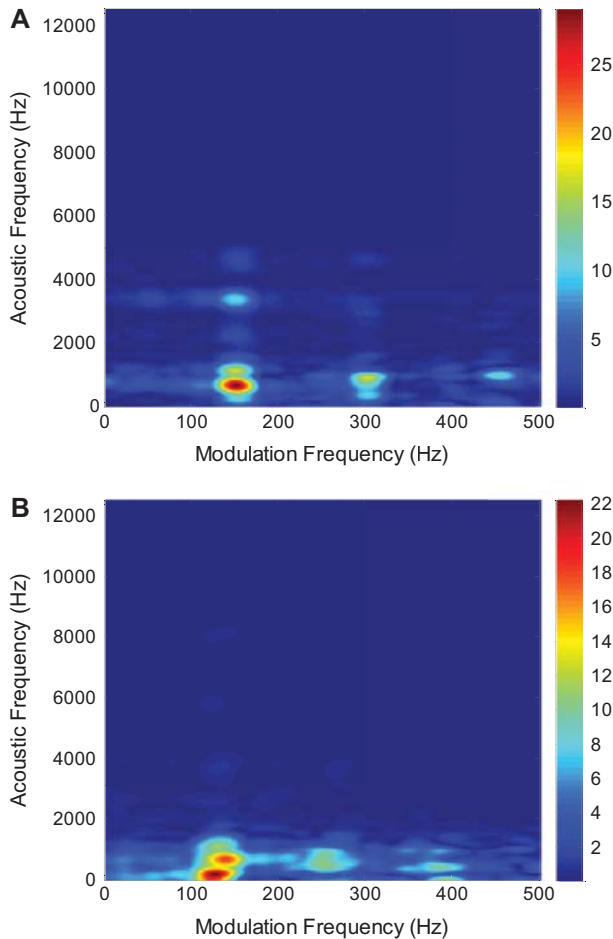


Figure 2. MS plot. A: normal voice (file BJB1NAL.NSP taken from the MEEI database). B: pathological voice (file DAP17AN.NSP taken from the MEEI database).

features. For UPM (Universidad Politécnica de Madrid), the reduced space had dimensions of  $53 \times 36 = 1,908$ . Next, the features which were more correlated to the voice pathology detection task were selected for each database, using the maximal relevance criterion (*MaxRel*). For details about the application of the *MaxRel* criterion for this task, please refer to Markaki et al. (14).

Henceforth the features obtained by employing this methodology will be referred as MS maximum relevance (MSMR) features.

*Mel-frequency cepstral coefficients.* MFCC (5) are a family of parameters that can be either estimated using a parametric approach derived from linear predictive coefficients or using a non-parametric (Fast fourier transform (FFT)-based) approach. The definition has been included in the text. approach. The non-parametric approach allows modelling of the effects induced by the presence of pathology over the excitation (vocal folds) and the system (vocal tract) (3). Another justification for using the MFCC parameters

is that this measure follows a transformation in the frequency domain to a perceptual scale. This transformation shares processing principles with the human auditory system response (18) and matches well with the fact that an experienced speech therapist can very often detect the presence of a disorder just by listening to it.

MFCC parameters are obtained, for each frame under analysis of the input signal,  $m$ , by calculating the discrete cosine transform (DCT) over the logarithm of the energy in several frequency bands as shown in:

$$c_p(m) = \sum_{b=1}^{M_B} \log(X_b(m)) \cos \left[ p \left( k - 0.5 \right) \frac{\pi}{N} \right] \quad (5)$$

where  $b = (1, 2, \dots, M_B)$ ,  $M_B$  being the number of the mel bands in the mel scale;  $p = (1, 2, \dots, P)$ ,  $P$  being the number of MFCC coefficients extracted; and  $X_b(m)$  given by:

$$X_b(m) = \sum_{k=1}^{NFFT} h_{mb}(k) X_k(m) \quad (6)$$

where  $h_{mb}(k)$  is a triangular weighting function associated with the mel band in the mel scale, and  $X_k(m)$  is the STFT of the input signal  $x(n)$ —defined in Equation 1.

Each band in the frequency domain is bandwidth-dependent on the centre frequency of the filter. The higher the frequency, the wider the bandwidth is. This method is based on the human perception system, establishing a logarithmic relationship between the real frequency scale (Hz) and the perceptual frequency scale (mels). The suggested formula that models this relationship is as follows (19):

$$F_{\text{mel}} = 2595 \log_{10} \left( 1 + \frac{F_{\text{Hz}}}{700} \right) \quad (7)$$

Following the aforementioned procedure, and in line with other works found in the state of the art (3,12), 12 MFCC were extracted for each frame.

Figure 3 shows the MFCC plot obtained for a normal and a pathological speech signal respectively. The pathological voice shows light differences in the upper coefficients, and clear differences in the second and third coefficients.

### Classification

A first decision about the presence or absence of pathology for each set of features was taken by using

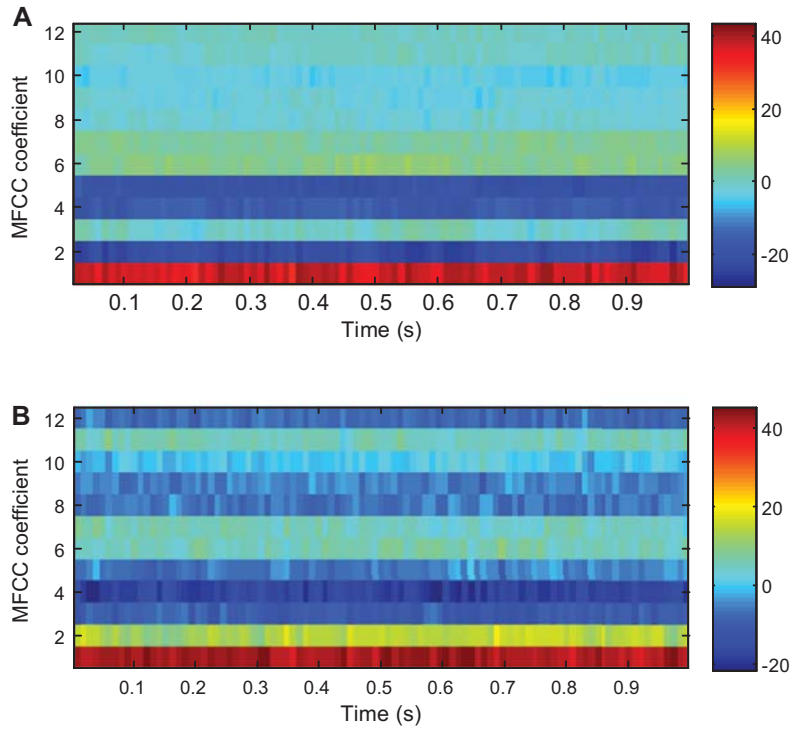


Figure 3. MFCC plot. A: normal voice (file BJB1NAL.NSP taken from MEEI database). B: pathological voice (file DAP17AN.NSP taken from MEEI database).

different classifiers. For the case of MFCC, a classifier based on GMM was employed. This classifier has previously been used for the same task with good results (3). On the other hand, the set of MSMR features was classified by using a SVM which is a class of discriminative classifiers with high generalization capabilities. The SVM has also been successfully used for voice pathology detection in previous works. The criterion to use a different classifier for each family of parameters was based on previous experiments in which it was empirically demonstrated that the MFCC improve the representation capabilities when they are modelled with GMM, and MSMR performs better using a SVM in the classification stage. Moreover, GMM present problems with the MSMR features due to the fact that *MaxRel* selects more than 100 relevant features, and the number of parameters to be estimated for the GMM exponentially increases in relation to the dimensionality of the feature space. This phenomenon is known as ‘curse of dimensionality’ and affects the training stage.

*Gaussian mixture models.* The central idea of the GMM is to estimate the probability density function of a data set  $\mathbf{x} \in \mathbb{R}^d$ , by means of a set of Gaussian weighted functions. The model can be expressed as (17):

$$p(\mathbf{x} | c) = \sum_{i=1}^Q \alpha_i \Theta_i(\mathbf{x}) \quad (8)$$

where  $c = \{\zeta_n, \zeta_p\}$  indicates the class to model,  $\Theta_i(\mathbf{x})$ ,  $i = 1, 2, \dots, Q$  are the component densities, and  $\alpha_i$  are the component weights. Each component density is a  $d$ -variate Gaussian function, so the different components act together to model the overall *pdf*. The most common estimation method of the parameters of the model is the expectation maximization (EM) algorithm (20).

For each class to be recognized, the parameters of a different GMM are estimated. Thus, the evaluation is made calculating for each GMM the a-posteriori probability of an observation. The score given to each sequence is obtained by calculating the logarithm of the ratio between the likelihoods given by both models (called log-likelihood ratio) (21).

*Support vector machines.* A SVM is a two-class classifier. The problem in approaching a SVM (22) is analogous to solving the problem of finding a linear function that satisfies:

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b \text{ with } \mathbf{w} \in \chi, b \in \mathbb{R} \quad (9)$$

where  $\chi$  corresponds to the space of the input patterns  $\mathbf{x} \in \mathbb{R}^d$ . The function  $f(\cdot)$ , is calculated following an optimization problem from sums of a kernel function (22). The support vector algorithm looks for the hyperplane that separates the two classes with the largest margin of separation.

Considering a radial basis function kernel, the training implies adjusting the aperture of the kernel,  $\gamma$ , and a penalty parameter,  $C$  (22).

The output given by the SVM for each speech sample can be interpreted as the likelihood that the sample belongs to a specific class. Henceforth, the log-likelihood (likelihood in the log domain) will be called ‘score’.

### Combining classifiers

There are two important aspects that have to be kept in mind to fuse information from different classifiers: the score normalization technique used and the combination rule. The first one is needed to eliminate the bias due to the different dynamic ranges of the scores given by the different classifiers, and also for the use of combination rules which make the assumption that the scores are taken from a probability density function in the interval  $[0,1]$ . On the other hand, the problem of selecting the combination rule is related to choosing the function which provides the highest accuracy by merging the scores given by the different classifiers.

*Score normalization.* The simplest normalization technique is the *Min-Max* normalization (23). This normalization is best suited for the case where the bounds (maximum and minimum values) of the scores given by a classifier are known. However, even if the scores are not bounded, it is possible to estimate the minimum and maximum values for a set of scores and then applying the normalization given by:

$$s_k^* = \frac{s_k - \min}{\max - \min} \quad (10)$$

where  $s_k$  are the raw scores and  $s_k^*$  are the normalized scores. The drawback of this method is that it is highly sensitive to outliers in data used for estimation.

Another widely used normalization method is a logistic transformation which allows reducing the effects due to the dispersion in the scores. The transformation function is given by:

$$s_k^* = \frac{1}{1 + \exp\left(-\left(\omega_0 + \omega s_k\right)\right)} \in [0,1] \quad (11)$$

where:

$$\omega = \frac{\mu_{c_1}^2 - \mu_{c_2}^2}{2\sigma^2}, \quad \omega_0 = \frac{\mu_{c_2} - \mu_{c_1}}{\sigma^2} \quad (12)$$

$\mu_{c_1}$  and  $\mu_{c_2}$  being the mean of the class 1 and class 2, respectively. This transformation assumes that the distribution of both scores is Gaussian with a common variance  $\sigma$ . However, the scatter does not have to be equal, so the value of  $\sigma = 0.5(\sigma_{c_1} + \sigma_{c_2})$  is a good trade-off,  $\sigma_{c_1}$  and  $\sigma_{c_2}$  being the estimate of the standard deviation for each class. This normalization provides a linear transformation of the scores in the overlapping region, while the scores outside this region are non-linearly transformed, and therefore this normalization does not retain the shape of the original score distribution (23).

A more robust non-linear transformation is the Hampel score normalization given by (23):

$$s_k^* = \frac{1}{2} \left\{ \tanh \left( 0.01 \left( \frac{s_k - \mu_{c_1}}{\sigma_{c_1}} \right) \right) + 1 \right\} \quad (13)$$

where  $\mu_{c_1}$  and  $\sigma_{c_1}$  are the mean and standard deviation of the objective class (in our case the pathological class). This method reduces the influence of the tail points in the score distribution and is not sensitive to outliers.

*Score combination rules.* The basic rules to consolidate the evidence obtained from multiple classifiers along with their theoretical framework were proposed in Kittler et al. (13). Here only the rules that can be applied for merging at the score level are considered (see (13,24) and cites therein).

Consider the problem of classifying an input pattern  $z$  into one of  $c$  possible classes based on the evidence provided by  $R$  different classifiers. Let  $\mathbf{x}_i$  be the feature vector presented to the  $i$ -th classifier. Let the outputs of the individual classifiers be  $P(\zeta_j | \mathbf{x}_i)$ , the following rules can be used to determine  $c$  (23):

*Product rule:* This rule is based on the assumption of statistical independence of the representations used to feed each classifier. The input pattern is assigned to class  $c$  such that:

$$c = \arg \max_j \prod_{i=1}^R P(\zeta_j | \mathbf{x}_i) \quad (14)$$

*Sum rule:* Apart from the assumption of statistical independence of the multiple representations used in

the product rule, the sum rule also assumes that the a-posteriori probabilities computed by individual classifiers do not deviate much from the a-priori probabilities. The sum rule assigns the input pattern to class  $c$  such that:

$$c = \arg \max_j \sum_{i=1}^R P(\zeta_j | \mathbf{x}_i) \quad (15)$$

This rule is applicable in presence of a high level of noise leading to ambiguity in the classification problem. A modification of the sum rule is the weighted sum in which a different weight is assigned to each classifier and, in this way, the classifier with the best individual accuracy will have more influence in the final decision. This method requires learning the specific weights of the classifier but, due to the fact that, in the particular task addressed in this work, there are only two classes, the weights ( $w_1, w_2$ ) can be found by a simple search taking into account that they are varied over the interval [0,1] and the sum  $w_1 + w_2$  must be equal to 1.

*Max rule:* The max rule approximates the mean of the a-posteriori probabilities by the maximum value. In this case, it is assigned the input pattern to class  $c$  according to:

$$c = \arg \max_j \max_i P(\zeta_j | \mathbf{x}_i) \quad (16)$$

*Min rule:* The min rule is derived by bounding the product of the a-posteriori probabilities. Here, the input pattern is assigned to class  $c$  such that:

$$c = \arg \max_j \min_i P(\zeta_j | \mathbf{x}_i) \quad (17)$$

Finally, another simple scheme that can be used for combining scores is to use the outputs of the classifiers as features to feed a second layer of classification. This approach combines the generalization capabilities of several classification techniques in order to find an optimum final decision boundary. In this work, this second layer has been implemented by means of a  $k$ - $nn$  classifier, a multilayer perceptron (MLP), and a SVM.

## Experiments and decision-making

### Databases

The first database used was developed by The Massachusetts Eye and Ear Infirmary Voice Laboratory

(MEEI) (25). Due to the different sampling rates of the recordings stored in this database, a down-sampling with a previous half-band filtering was carried out, when needed, in order to adjust every utterance to a 25 kHz sampling rate; 16 bits of resolution were used for all the recordings. The registers contain the sustained phonation of the /ah/ vowel from patients with a variety of voice pathologies: organic, neurological, and traumatic disorders. According to the criteria explained in Parsa et al. (26), a subset of 173 pathological and 53 normal speakers were used for the experimentation. The subset selected ensures a balance according to gender, age, and pathologies.

The second database was recorded by the Universidad Politécnica de Madrid (UPM) (27). This database stores 200 pathological voices (74 men and 126 women, aged 11–76) with a wide variety of organic pathologies (nodules, polyps, oedemas, carcinomas, etc.), and 199 normal voices (87 men and 112 women, aged 16–70). The data set contains the sustained phonation of the /aa/ Spanish vowel with a sampling rate of 50 kHz and 16 bits of resolution. All the speakers stored in the database were used for the experimentation.

### Experimental set-up

Different normalization techniques and combination rules were employed in order to establish the best configuration to fuse the information obtained from MFCC + GMM and MSMR + SVM. During the experiments the most appropriate values of the SVM and GMM parameters ( $\gamma$  and  $C$  for the SVM, and  $M$  for the GMM) were determined to achieve the best possible accuracies for each corpus. With respect to the SVM, the optimum working point was searched inside the grid  $C = [10^0, 10^3]$  and  $\gamma = [10^{-4}, 10^{-2}]$ . Regarding the GMM, the parameter  $Q$  has been evaluated into  $Q = [2, 6]$ .

The methodology proposed in Sáenz-Lechón et al. (28) was used for the evaluation of the system. The generalization abilities of the system have been tested following a stratified cross-validation scheme with four different sets for training and validation, repeated four times using 75% of the utterances for training and 25% for testing. The results are presented giving the following rates: true positive rate ( $tp$ ) (or *sensitivity*, is the ratio between pathological files correctly classified and the total number of pathological voices) and true negative rate ( $tn$ ) (or *specificity*, is the ratio between normal files correctly 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.

As a figure of merit two curves may be plotted using the scores given by each classifier to show the performance of the proposed architecture: the detector

Table I. Results obtained with MFCC + GMM and MSMR + SVM using the MEEI database.

| Features   | Sensitivity | Specificity | Accuracy |
|------------|-------------|-------------|----------|
| MFCC + GMM | 95.20%      | 91.04%      | 94.22%   |
| MSMR + SVM | 97.38%      | 79.72%      | 93.22%   |

error trade off (DET), and the receiver operating characteristic (ROC). The ROC is a popular tool in medical decision-making (29). It reveals diagnostic accuracy expressed in terms of sensitivity and 1-specificity or false positive rate (fp). The DET plot (30) has been used widely for the assessment of detection performance in speaker identification tasks. The DET curve plots error rates (false positive rate and false negative rate) on both axes, giving uniform treatment to both types of error.

## Results

Tables I and II show the results using MFCC and MSMR features independently for MEEI and UPM

Table II. Results obtained with MFCC and MSMR using the UPM database.

| Features   | Sensitivity | Specificity | Accuracy |
|------------|-------------|-------------|----------|
| MFCC + GMM | 77.00%      | 83.04%      | 80.01%   |
| MSMR + SVM | 80.50%      | 82.91%      | 81.70%   |

databases, respectively. The Tables show the best results obtained after tuning the parameters of the classifiers. Both approaches (MFCC + GMM and MSMR + SVM) yielded a similar accuracy. However, the results showed different correct classification rates for MEEI and UPM databases. This behaviour has already been shown in a previous work (12) and is attributable to the bigger number of speakers of the second database and to the differences of the recording conditions that introduce a larger variability. In this sense, and in view of the linearity of the DET plots, the results using UPM appear to be more consistent.

Table III shows the results obtained using the different combination strategies studied. The best results reported are highlighted in bold. A score

Table III. Results combining classifiers with different strategies.

| Normalization | Combination rule   | Sensitivity |       | Specificity |       | Accuracy     |              |
|---------------|--|-------------|-------|-------------|-------|--------------|--------------|
|               |  | MEEI        | UPM   | MEEI        | UPM   | MEEI         | UPM          |
| Min-Max       | Sum  | 97.97       | 83    | 79.25       | 84.42 | 93.56        | 83.71        |
| Min-Max       | Product  | 96.80       | 84.38 | 83.96       | 81.03 | 93.78        | 82.71        |
| Min-Max       | Max  | 97.38       | 79.75 | 79.72       | 84.80 | 93.22        | 82.27        |
| Min-Max       | Min  | 95.78       | 86.38 | 90.57       | 74.87 | 94.56        | 80.64        |
| Min-Max       | Weighted sum<br>$w_1$ (MFCC) = 0.93 <sup>a</sup>   0.53 <sup>b</sup><br>$w_2$ (MSMR) = 0.07 <sup>a</sup>   0.47 <sup>b</sup> | 96.08       | 83.50 | 91.98       | 84.80 | 95.11        | <b>84.15</b> |
| Min-Max       | k-nn (k = 1)   | 96.8        | 86.13 | 75.0        | 78.14 | 91.67        | 82.14        |
| Min-Max       | MLP (3 hidden nodes)   | 96.51       | 81.13 | 79.72       | 83.54 | 92.56        | 82.33        |
| Min-Max       | SVM  | 97.38       | 85    | 80.66       | 81.28 | 93.44        | 83.15        |
| Logistic      | Sum  | 98.11       | 82.28 | 80.19       | 83.79 | 93.89        | 83.33        |
| Logistic      | Product  | 97.53       | 86.38 | 80.19       | 77.39 | 93.44        | 81.89        |
| Logistic      | Max  | 95.78       | 76.75 | 90.09       | 88.32 | 94.44        | 82.52        |
| Logistic      | Min  | 97.38       | 87    | 79.72       | 72.61 | 93.22        | 79.82        |
| Logistic      | Weighted sum<br>$w_1$ (MFCC) = 0.93 <sup>a</sup>   0.41 <sup>b</sup><br>$w_2$ (MSMR) = 0.07 <sup>a</sup>   0.59 <sup>b</sup> | 96.66       | 83.25 | 93.40       | 84.55 | <b>95.89</b> | 83.90        |
| Logistic      | k-nn (k = 1)   | 97.67       | 84.25 | 81.13       | 80.28 | 93.78        | 82.27        |
| Logistic      | MLP (3 hidden nodes)   | 97.27       | 82.75 | 90.09       | 83.29 | 95.56        | 83.02        |
| Logistic      | SVM  | 97.24       | 83.25 | 86.32       | 83.79 | 94.67        | 83.52        |
| Hampel        | Sum  | 96.51       | 83    | 89.15       | 82.04 | 94.78        | 82.52        |
| Hampel        | Product  | 96.51       | 83    | 89.15       | 82.04 | 94.78        | 82.52        |
| Hampel        | Max  | 97.82       | 75.88 | 83.02       | 89.07 | 94.33        | 82.46        |
| Hampel        | Min  | 94.62       | 88    | 94.34       | 72.36 | 94.56        | 80.20        |
| Hampel        | Weighted sum<br>$w_1$ (MFCC) = 0.45 <sup>a</sup>   0.31 <sup>b</sup><br>$w_2$ (MSMR) = 0.55 <sup>a</sup>   0.69 <sup>b</sup> | 96.66       | 83.75 | 90.09       | 84.42 | 95.11        | 84.09        |
| Hampel        | k-nn (k = 1)   | 97.97       | 85.75 | 81.13       | 78.27 | 93.33        | 82.02        |
| Hampel        | MLP (3 hidden nodes)   | 96.08       | 80.87 | 89.62       | 82.79 | 94.56        | 81.83        |
| Hampel        | SVM  | 95.49       | 85    | 91.51       | 78.64 | 94.56        | 81.23        |

<sup>a</sup>Weights obtained with MEEI database.<sup>b</sup>Weights obtained with UPM database.

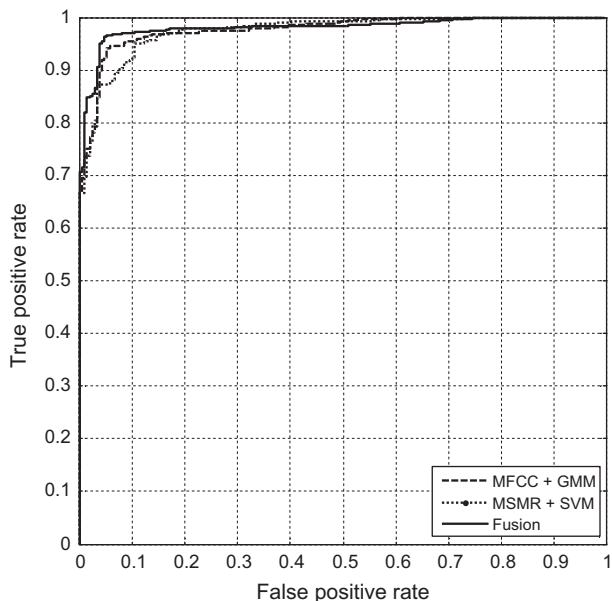


Figure 4. ROC curve for the best results using MFCC + GMM, MSMR + SVM, and fusing both classifiers (MEEI database).

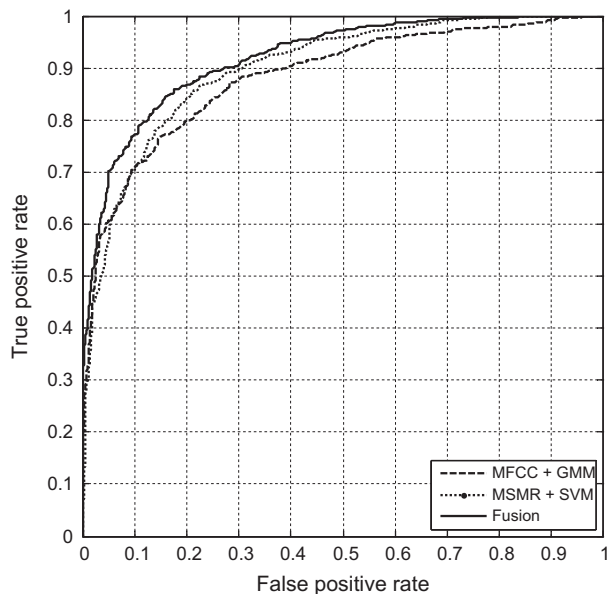


Figure 6. ROC curve for the best results using MFCC+GMM, MSMR+SVM, and fusing both classifiers (UPM database).

combination based on a weighted sum provided very good results with the different normalization criteria used (especially with *Min-Max* and logistic normalizations).

Figures 4–7 show ROC and DET plots of the best results obtained by using MFCC + GMM and MSMR + SVM independently and fusing both classifiers with the best approaches reported in Table III. For both databases there is an improvement in the accuracy fusing both classifiers. In the case of the UPM database, the DET plots show a more steady

behaviour showing similar false positive and false negative rates and an approximately Gaussian distribution for the scores.

### Conclusions

Pathological voice is characterized by an increase of the vocal folds mass, a subsequent lack of closure, or an elasticity change of the vocal folds and surrounding tissues. Previous experiments confirmed that MSMR provide complementary information to MFCC, since the low bands of the MFCC reflect alterations related

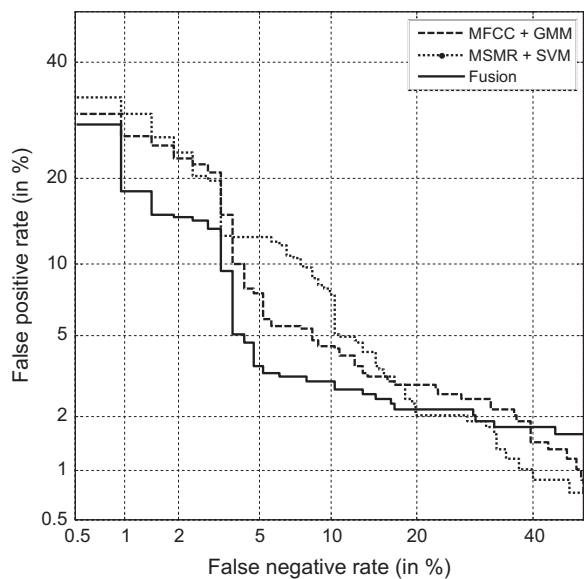


Figure 5. DET plot for the best results obtained using MFCC+GMM, MSMR+SVM, and fusing both classifiers (MEEI database).

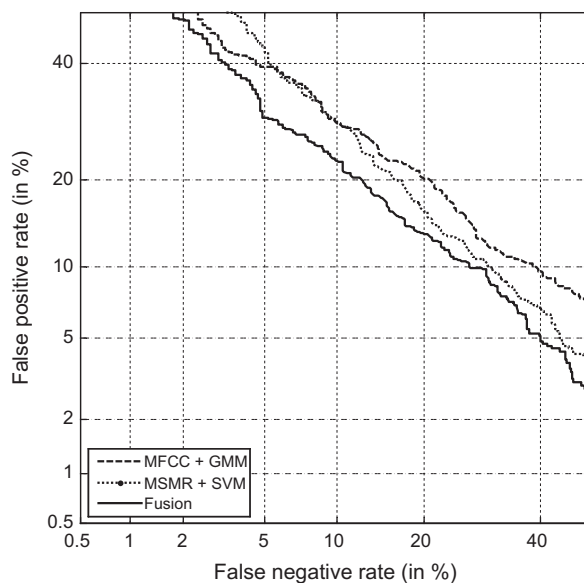


Figure 7. DET plot for the best results using MFCC+GMM, MSMR+SVM, and fusing both classifiers (UPM database).

to the mucosal waveform due to an increase of mass, the noisy components induced by lack of closure are modelled by the higher bands (3), and the MSMR features capture the amplitude envelope fluctuations present during sustained vowel phonation. Thus, as expected, fusing the information given by the MFCC and MSMR features yields an improvement of the accuracy, demonstrating the complementarity of both parameterization approaches for the detection of voice disorders. However, although the results improve the detection accuracy combining both classifiers, the improvement is lower than in our previous experiments fusing features (12). Anyway, these results are in concordance with the statements in other works found in the state of the art, which ensure that integrating information at an early stage of processing is more effective than integrating at a later stage, because the features contain more information about the problem to be solved than the outputs of the classifiers (23).

### Acknowledgements

This work was supported by COST (European Cooperation in Science and Technology), action 2103; and TEC2009-14123-C04-02 from the Ministry of Science and Technology of Spain.

**Declaration of interest:** The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

### References

1. Hadjitodorov S, Mitev P. A computer system for acoustic analysis of pathological voices and laryngeal disease screening. *Med Eng Phys.* 2002;24(6):419–29.
2. Baken RJ, Orlikoff R. Clinical measurement of speech and voice. 2nd ed. Singular Publishing Group, San Diego, CA, USA; 2000.
3. Godino-Llorente JI, Gomez-Vilda P, Blanco-Velasco M. Dimensionality reduction of a pathological voice quality assessment system based on GMMs and short-term cepstral parameters. *IEEE Trans Biomed Eng.* 2006;53(10):1943–53.
4. Markaki M, Stylianou Y. Using modulation spectra for voice pathology detection and classification. In: Proc. of IEEE EMBS, Minneapolis, MN, USA; 2009. p. 2514–7.
5. Davis SB, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans Acoust.* 1980;28(4):357–66.
6. Atlas L, Shamma SA. Joint acoustic and modulation frequency. *EURASIP J Appl Signal Processing.* 2003;7:668–75.
7. Fejoo S, Hernández C. Short-term stability measures for the evaluation of vocal quality. *J Speech Hear Res.* 1990; 33:324–34.
8. Yumoto E, Sasaki Y, Okamura H. Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness. *J Speech Hear Res.* 1984;27(1):2–6.
9. de Krom G. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *J Speech Hear Res.* 1993;36(2):254–66.
10. Bou-Ghazale SE, Hansen JHL. A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Trans Speech Audio Proc.* 2000; 8(4):429–42.
11. Malyska N, Quatieri TF, Sturim D. Automatic dysphonia recognition using biologically inspired amplitude-modulation features. In: Proc. ICASSP, Philadelphia, PA, USA; 2005. p. 873–6.
12. Markaki M, Stylianou Y, Arias-Londoño JD, Godino-Llorente JI. Dysphonia detection based on modulation spectral features and cepstral coefficients. In: Proc. of IEEE ICASSP, Dallas, TX, USA; 2010.
13. Kittler J, Hatef M, Duin RPW, Matas J. On combining classifiers. *IEEE Trans Pattern Anal Mach Intell.* 1998; 20(3):226–39.
14. Markaki M, Stylianou Y. Normalized modulation spectral features for cross-database voice pathology detection. In: Proc. Interspeech, Brighton, UK; 2009. p. 935–8.
15. De Lathauwer L, De Moor B, Vandewalle J. A multilinear singular value decomposition. *SIAM J Matrix Anal Appl.* 2000;21:1253–78.
16. Markaki M, Stylianou Y. Dimensionality reduction of modulation frequency features for speech discrimination. In: Proc. of InterSpeech, Brisbane, Australia; 2008. p. 646–9.
17. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell.* 2005;27:1226–38.
18. Huang X, Acero A, Hon HW. Spoken language processing. New Jersey: Prentice Hall PTR; 2001.
19. Rabiner LR, Juang BH. Fundamentals of speech recognition. Englewood Cliffs, NJ: Prentice-Hall; 1993.
20. Moon TK. The expectation-maximization algorithm. *IEEE Signal Process Mag.* 1996;Nov:47–60.
21. Reynolds DA, Quatieri TF, Dunn RB. Speaker verification using adapted Gaussian mixture models. *Digit Signal Process.* 2000;10:19–41.
22. Vapnik V. An overview of statistical learning theory. *IEEE Trans Neural Netw.* 1999;10(5):988–1000.
23. Jain A, Nandakumar K, Ross A. Score normalization in multimodal biometric systems. *Pattern Recognit.* 2005;38: 2270–85.
24. Kuncheva L. Combining pattern classifiers. Pattern recognition. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2004.
25. Massachusetts Eye and Ear Infirmary. Voice disorders database. Version 1.03. [CD-ROM]. Lincoln Park, NJ: Kay Elemetrics Corp. 1994.
26. Parsa V, Jamieson D. Identification of pathological voices using glottal noise measures. *J Speech Language Hear Res.* 2000;43(2):469–85.
27. Godino-Llorente JI, Osma-Ruiz V, Sáenz-Lechón N, Cobeta-Marco I, González-Herranz R, Ramírez-Calvo C. Acoustic analysis of voice using WPCVox: a comparative study with multi dimensional voice program. *Eur Arch Otolaryngol.* 2008;265(4):465–76.
28. Sáenz-Lechón N, Godino-Llorente JI, Osma-Ruiz V, Gómez-Vilda P. Methodological issues in the development of automatic systems for voice pathology detection. *Biomed Signal Process Control.* 2006;1(2):120–8.
29. Hanley JA, McNeil BJ. “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” *Radiology.* 1982;143(1):29–36.
30. Martin A, Doddington GR, Kamm T, Ordowski M, Przybocki M. “The DET curve in assessment of detection task performance,” in Proc. of 5th European Conference on Speech Communication and Technology - Eurospeech 97, Rhodes, Crete. 1997;4:1895–98.