redin Revista Facultad de Ingeniería, Universidad de Antioquia, No. 79, pp. 50-62, 2016

An insight to the automatic categorization of speakers according to sex and its application to the detection of voice pathologies: A comparative study

Una mirada a la categorización automática de hablantes de acuerdo al sexo y su aplicación a la detección de patologías de voz: Un estudio comparativo

Jorge Andrés Gómez-García[1*], Laureano Moro-Velázquez[1], Juan Ignacio Godino-Llorente[1], César Germán Castellanos-Domínguez[2]

[1]Centro de Tecnología Biomédica (CTB), Universidad Politécnica de Madrid. Parque Científico y Tecnológico de la UPM, Crta. M40 km, 38. C. P. 28223. Pozuelo de Alarcón, España.
[2]Departamento de Ingeniería Electrónica, Eléctrica y Computación. Universidad Nacional de Colombia. Campus la Nubia, km. 7 vía al Magdalena. A. A. 127. Manizales, Colombia

ARTICLE INFO
Received May 31, 2015
Accepted September 23, 2015

KEYWORDS
Voice pathology detection, inverse filtering, GMM, UBM

Detección de la patología de voz, filtrado inverso, GMM, UBM

ABSTRACT: An automatic categorization of the speakers according to their sex improves the performance of an automatic detector of voice pathologies. This is grounded on findings demonstrating perceptual, acoustical and anatomical differences in males' and females' voices. In particular, this paper follows two objectives: 1) to design a system which automatically discriminates the sex of a speaker when using normophonic and pathological speech, 2) to study the influence that this sex detector has on the accuracy of a further voice pathology detector. The parameterization of the automatic sex detector relies on MFCC applied to speech; and MFCC applied to glottal waveforms plus parameters modeling the vocal tract. The glottal waveforms are extracted from speech via iterative lattice inverse filters. Regarding the pathology detector, a MFCC parameterization is applied to speech signals. Classification, in both sex and pathology detectors, is carried out using state of the art techniques based on universal background models. Experiments are performed in the Saarbrücken database, employing the sustained phonation of vowel /a/. Results indicate that the sex of the speaker may be discriminated automatically using normophonic and pathological speech, obtaining accuracy up to 95%. Moreover, including the a-priori information about the sex of the speaker produces an absolute performance improvement in EER of about 2% on pathology detection tasks.

RESUMEN: Una categorización automática de los hablantes de acuerdo con su sexo mejora el rendimiento de un detector automático de patologías de voz. Esto se fundamenta en hallazgos que demuestran diferencias perceptuales, acústicas y anatómicas en voces masculinas y femeninas. En particular, este trabajo persigue dos objetivos: 1) diseñar un sistema que discrimine automáticamente el sexo de hablantes utilizando habla normofónica y patológica, 2) estudiar la influencia que este detector de sexo tiene sobre el acierto de un posterior detector de patologías de voz. La parametrización del detector automático de sexo se basa en MFCC aplicados sobre señales de voz; y MFCC aplicados a formas de onda glotal junto a parámetros que modelan el tracto vocal. Las formas de onda glotal se extraen de la voz a través de un filtrado inverso iterativo en celosía. En cuanto al detector de patologías, una parametrización MFCC se aplica a señales de voz. La clasificación, tanto en los detectores de sexo como de patología, se lleva a cabo con técnicas del estado del arte basadas en modelos de base universal. Experimentos son realizados sobre la base de datos Saarbrücken empleando la fonación sostenida de la vocal /a/. Los resultados indican que el sexo del hablante puede ser discriminado automáticamente utilizando habla normofónica y

* Corresponding author: Jorge Andrés Gómez García
e-mail: jorge.gomez.garcia@upm.es
ISSN 0120-6230
e-ISSN 2422-2844

DOI: 10.17533/udea.redin.n79a06

patológica, obteniendo una precisión de hasta un 95%. Por otra parte, al incluir información a priori sobre el sexo del hablante se produce una mejora de alrededor del 2% de rendimiento absoluto en EER, en tareas de detección de patología.

1. Introduction

Systems that automatically detect pathologies by means of voice present potential advantages respect to traditional detection and evaluation procedures. In particular, they provide an objective assessment of the clinical state of patients, reduce the evaluation time and the cost of diagnosis and treatment [1]. Furthermore, they avoid invasive procedures by employing signals which are easily recorded by inexpensive means. Nonetheless, the usage of speech signals poses a difficulty due to the intrinsic variability of the voice which compromises the potential performance of automatic detection systems. In this regard, it is known that speech not only conveys linguistic but also a large amount of information about the speaker, such as sex (For discussions on the preferences of using this term over gender, please refer to [2]), age, regional origin, health, etc. [3]. Therefore, the design of automatic systems for the detection of voice pathologies should be carried out paying special attention to the influence of these acoustic and paralinguistic traits.

One major finding that facilitates the design process of automatic systems to classify speakers is described in [4]. In this work, accent and gender are identified as the most important sources of variability between speakers in speech recognition systems. Hence, differentiating speakers according to their sex and removing the influence of the accent should improve the results of any automatic system designed to categorize the speech. To this extent, a naive yet useful manner to counteract the influence of the accent in automatic pathology detection systems is to use sustained vowels instead of continuous speech. This also reduces the variability related to the prosody of the speakers since sustained vowels are relatively devoid of individual speech characteristics such as speaking rate, speaker's dialect, intonation, and idiosyncratic articulatory behavior. Also, variations due to the phonetic context and stress are reduced [5]. It might be argued that by using sustained vowels the phonetic richness of the speech is restrained. Nonetheless, several automatic voice pathology detectors have performed successfully when using this acoustic material [6-8].

On the other hand, the variability introduced by the sex of the speaker remains as a major concern in the design of speech recognition systems. Certainly, the literature reports that by exploiting a priori information about the sex of the speaker, the performance of speech recognition, identification or verification systems improves [9]. For instance, authors in [10] enhanced the accuracy of an automatic emotion recognizer by incorporating information about the sex of the speaker. In [11], a speaker recognition system obtained a 2% accuracy improvement at equal error rate (EER) when using sex-specific models. Likewise, in [12], a sex classification stage improved accuracy and decreased computational load of a speaker diarization system.

The nature of the variability introduced by the sex of the speaker stands on physiological, acoustic, and psychophysical factors [13]. Regarding *perceptual differences* parameters such as effort, pitch, stress, nasality, melodic patterns of intonation and co-articulation are used for characterizing female voices, while male voices are judged on the basis of effort, pitch and hoarseness [14]. It is also argued that female voices possess a "breathier" quality than male voices [15]. Regarding *physiological differences*, the human laryngeal anatomy differs between sexes at a variety of levels. Particularly, males tend to have a more acute thyroid angle; thicker vocal folds; a longer vocal tract; a larger pharyngeal-oral apparatus, thyroid lamina and skull compared to that of females [16, 17]. Studies of excised human larynges have shown that anteroposterior dimensions of the glottis are 1.5 times larger in men than in women [18]. Besides that, the female pharynx has been found to be shorter than of males during the production of the three cardinal vowels. This may be a key factor in distinguishing between male and female voice qualities during speech production [17]. In addition, the observation of the glottis during phonation has suggested the presence of a posterior glottal opening that persists throughout a vibratory cycle and which is common for female speakers, but occurs much less frequently among male speakers [19]. Indeed, about 80% of females and 20% of males have a visible posterior glottal aperture during the closed portion of a vocal period [15]. Regarding *acoustical differences,* the pitch is the most known trait differentiating sexes [14], with females' pitch higher than of males' by as much as an octave [20]. There are also several important acoustic consequences of the posterior glottal opening during the closed phase of phonation, and which is more frequent in females. A first consequence is a breathier voice quality which is the result of a larger amount of air passing through the vocal tract [16] and that affects the relative amplitude of the first harmonic of the speech spectra [18, 21]. A second consequence is the widening of the first-formant bandwidths. This is because the glottal aperture produces energy losses particularly at low frequencies, resulting in a bigger bandwidth of the first formant [19, 21]. A third acoustic consequence is the generation of turbulence noise in the vicinity of the glottis [21] perceived as a high level of aspiration noise in the spectral regions corresponding to the third formant, contributing to a breathier voice [20]. A final consequence is a lower spectral tilt due to the presence of aspiration noise [20], which turns out to be a significant parameter for differentiating between male and female speech samples [19].

In addition to the acoustic differences reported from the study of the raw speech, there are some differences in the glottal components among sexes. On one hand and regarding the female glottal waveform, this presents a shorter period, lower peaks and lower peak-to-peak flow amplitude than of males [22]. Likewise, the derivative of the glottal waveform does not present an abrupt discontinuity

during the closing time due to the incomplete closure of the vocal folds [13]. In general, it is stated that female glottal components are symmetric, with opening and closing portions of the waveform tending toward equal duration [23]. Conversely, and regarding the glottal waveform of male speakers, it is found that the open quotient is smaller and the maximum flow declination rate is greater than of females [19]. Moreover, the closing portion of the waveform generally occupies 20%–40% of the total period and it might not exist an easily identifiable closed period [14]. In general, it is stated that male glottal waveform are asymmetrical and present a hump in the opening phase.

The abovementioned differences evidence that the design of an automatic sex recognition system is feasible, either from the speech or from the glottal waveform. Indeed, automatic sex recognition in normophonic voice has been discussed before in the literature. In [24] the authors employ cepstral features and support vector machines (SVM) for sex recognition, obtaining 100% classification accuracy when using English allophones as acoustic material. In [25] the authors develop a methodology based on relative spectral perceptual linear predictive coefficients and Gaussian mixture models (GMM). Experiments are performed in noisy and clean utterances of different languages, providing classification accuracy up to 98% for clean speech and 95% for noisy speech. In [26], Mel frequency cepstral coefficients (MFCC) are used in connected speech. The MFCC are employed to characterize speech signals, glottal waveforms and deglottized voices. At the end, a performance up to 99% is obtained by using principal component analysis and quadratic linear discrimination analysis. In spite of the performance of abovementioned studies, and even when they take into consideration different languages or noise levels, the acoustic material is restrained to normophonic speakers. Hence, the application of these systems for pathological voices is not demonstrated, being this a challenging problem due to the presence of perturbations inherent to pathological states. Respecting pathological voice detection, it is evidenced in [27] that a manual segmentation of the speakers' database according to their sex improves accuracy in an automatic pathology detection system. Furthermore, authors in [28] determined a significant sex-specific separation of control and pathological classes in the Saarbrücken database, by using statistical analysis and a series of acoustical and spectral measurements. Above studies suggest that profiting from the a priori information about the sex of speakers might be helpful in normophonic vs. pathological discrimination tasks.

The present paper proposes in a first instance to automatically differentiate the sex of the speaker to latter categorize normophonic and pathological speakers. The objective of this cascading procedure is to improve classification accuracy by simplifying the statistical models used in the identification of the presence/absence of voice pathologies. For this end, a state of the art pattern recognition framework based on GMM for classification and MFCC features for characterization is considered. For comparison purposes', the parameterization is applied to the raw speech and to glottal waveforms extracted by following the inverse filtering approach in [29, 30].

The expected contributions by following the aforementioned scheme are:

- To test out if it is feasible to automatically extract information about the sex of speakers from normophonic and pathological voice. To our knowledge this work is one of the first attempts to extract this type of information automatically from such acoustic material.
- To test out if separating glottal source and vocal tract components produce an increment in automatic sex identification compared to just using the voice signal. This is supported by the mentioned differences between vocal and glottal signals among sexes.
- To test out if including automatically extracted information about the sex of the speaker might improve the performance of automatic pathology detectors.

The structure of the paper is as follows: Section 2 includes the theoretical review of some of the tools employed in this study. Section 3 presents the methodological setup employed in the different experiments of the paper. Section 4 includes the results. Finally, section 5 presents the discussions, conclusions and future work.

2. Theoretical background

The scheme proposed in this paper is divided in two main stages running in cascade: the automatic detection of the speaker's sex, followed by the automatic detection of pathologies. For the sake of comparison, the sex detection stage employs raw speech and glottal waveform extracted using inverse filtering techniques, whereas the pathology detection stage employs only raw speech signals. Firstly, a short view of the inverse filtering procedure for glottal waveforms extraction is presented. Latter, the parameterization approach is outlined. Finally, the pattern recognition approach used for classification is described.

2.1. Inverse filtering

Voice is formed by a glottal excitation that is filtered by the vocal tract to yield the air-flow at the mouth, which in turn is converted to a pressure waveform at the lips and propagated as sound waves. Since the glottal flow and the vocal tract can be assumed to be linearly separable [31], several methods have been proposed to extract the glottal excitation waveform from the speech, most of them based on inverse filtering. One succeeding inverse filtering algorithm is the iterative adaptive inverse filtering technique [32], which employs an iterative refinement of the vocal tract model and the glottal signal to produce a better estimate of the glottal waveform. This method depends on a correct modeling of the vocal tract by means of Linear Predictive Coefficients (LPC). However, due to deficiencies of LPC

with high pitched voices, variations have been proposed employing lattice filters [29]. This technique allows a precise reconstruction of the pressure and flow variables along the tube and of glottal signals. Such approach has been successfully used in pathological voice applications [29, 30], and therefore is considered in the present paper. The iterative inverse filter process based on lattice filters is illustrated in Figure 1 and includes the following stages [29]:

- *Elimination of the lip-radiation effects:* The input voice $s(n)$ is filtered using an inverse radiation model filter $H_r(z)$ to compensate the radiation effects at the lips to produce a trace of radiation compensated voice $s_r(n)$.
- *Elimination of the glottal source spectral fingerprint of the input voice*: a simple glottal pulse inverse model filter $H_g(z)$ is used to cancel the behavior of the glottal source in the radiation compensated voice, producing a trace of deglottalized voice $s_v(n)$.
- *Estimation of the vocal tract transfer function by inverse linear predictive filtering using adaptive paired lattices:* The previous signal is inverse-filtered using lattice filters to extract the model of the vocal tract given by the transfer function $F_v(z)$.
- *Elimination of the vocal tract transfer function on input voice:* The inverse of the function in previous step is applied to the radiation compensated voice $s_r(n)$, producing a residual trace containing only information on the glottal source derivative $v_g(n)$.
- *Estimation of the glottal source transfer function to be applied in 2.*

The process is iterated 2-3 times to obtain a refined glottal source derivative $v_g(n)$, which is then integrated to obtain the glottal source waveform. Besides that waveform, the vocal tract model $F_v(z)$ obtained in stage 3 is considered a set of parameters by itself.

2.2. Parameterization

The MFCC [33] are calculated following a method based on the human auditory perception 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. Such mapping converts real into perceived frequency. MFCC features are examined in this study due to their ability to model acoustic signals and their widespread usage in speech technology applications. MFCC are applied on one hand to the raw speech, and on the other to the glottal waveforms extracted via inverse filtering. With respect to the latter, the MFCC are concatenated with the vocal tract model $(F_v(z))$ to form a new set of parameters.

2.3. Classification

The parameterization module provides for each speaker a set of feature vectors $\boldsymbol{x}$ of dimension d, one for each time window under analysis. Each of these feature vectors is analyzed and categorized using a GMM-based detector.

A GMM models the probability density of a random variable. For a particular class i, Eq. (1) defines a finite mixture of G multivariate Gaussian components, where λ_r^i are scalar mixture weights, $\mathcal{N}(\cdot)$ are Gaussian density functions

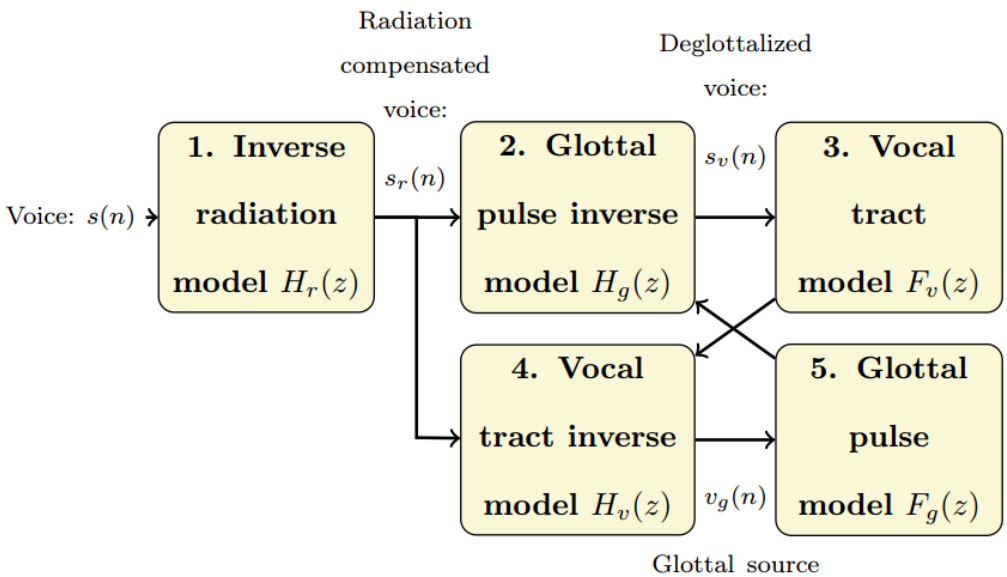


Figure 1 Outline of the iterative inverse filtering approach based on lattice filters employed in this paper

with mean $\boldsymbol{\mu}_r^i$ of dimension d and covariance matrix $\boldsymbol{\psi}_r^i$ of dimension $d \times d$. $\Theta^i = \{\lambda_r^i, \boldsymbol{\mu}_r^i, \boldsymbol{\psi}_r^i\}|_{r=1}^G$ is formed by comprising the above mentioned set of parameters and can be estimated by using the *expectation-maximization* algorithm in a *maximum likelihood* maximization scheme.

$$p(\boldsymbol{x}|\Theta^i) = \sum_{r=1}^G \lambda_r^i \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_r^i, \boldsymbol{\psi}_r^i) \quad (1)$$

Since the design of a pathology or a sex detector is a two-class problem, two models are required, one representing the target class $p(\boldsymbol{x}|\Theta^c)$, and the non-target class $p(\boldsymbol{x}|\Theta^{\bar{c}})$.

An enhancement of the GMM that has improved the results in speaker identification systems is the GMM-UBM [34]. The idea is to employ a generic GMM model, referred as *Universal Background Model* (UBM), which is trained using some background data typically belonging to a different database. The UBM serves as a well-trained initialization model, for which is possible to adapt specific models using the provided, and scarcer, training data. The adaptation procedure is typically applied to the mean of the UBM rather than to the whole set Θ^i. Literature on speech processing reports different algorithms to adapt the UBM. However, in this paper the adaptation is carried out using the *maximum a posteriori* (MAP) algorithm. Given a collection of training data $\mathcal{X}^i = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_N\}$, the MAP technique adapts the mean $\hat{\boldsymbol{\mu}}_r^i$ for the r-th Gaussian as in Eq. (2); where $\boldsymbol{\mu}_r^{ubm}$ is the UBM mean and $\rho_r = \eta_r/(\eta_r + \beta)$ is a data-dependent adaptation coefficient that controls the balance between old and new estimates (β is usually set to $E_r(\mathcal{X}^i)$ and η_r are sufficient statistics defined in Eq. (3), and $p(r|\boldsymbol{x}_t)$ are responsibilities representing the probability of the component r in explaining the value $\boldsymbol{x}_t$ as defined in Eq. (4).

$$\hat{\boldsymbol{\mu}}_r^i = \rho_r E_r(\mathcal{X}^i) + (1 + \rho_r)\boldsymbol{\mu}_r^{ubm} \quad (2)$$

$$E_r(\mathcal{X}^i) = \frac{1}{\eta_r}\sum_{t=1}^N p(r|\boldsymbol{x}_t)\boldsymbol{x}_t$$
$$\eta_r = \sum_{t=1}^N p(r|\boldsymbol{x}_t) \quad (3)$$

$$p(r|\boldsymbol{x}_t) = \frac{\lambda_r^i \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_r^i, \boldsymbol{\psi}_r^i)}{\sum_{j=1}^G \lambda_j^i \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_j^i, \boldsymbol{\psi}_j^i)} \quad (4)$$

The result of this adaptation procedure is a GMM-UBM model as in Eq. (5).

$$p(\mathcal{X}^i|\Theta^i) = \prod_{t=1}^N \sum_{r=1}^G \lambda_r^i \mathcal{N}(\boldsymbol{x}_t; \hat{\boldsymbol{\mu}}_r^i, \boldsymbol{\psi}_r^i) \quad (5)$$

In general it is possible to derive a log-likelihood decision function $\Lambda(\cdot)$ for discriminating if a test data $\boldsymbol{y}$ belongs to the target class or to the complementary. This function is presented in Eq. (6).

$$\Lambda(\boldsymbol{y}) = \log[p(\boldsymbol{y}|\Theta^c)] - \log[p(\boldsymbol{y}|\Theta^{\bar{c}})] \quad (6)$$

A further improvement of the GMM-UBM is the GMM-UBM-iVector [35]. The method relies on mapping from an utterance to a $G \times d$-dimensional vector $\boldsymbol{m}$ called *supervector*, which is usually obtained by stacking the mean vectors of UBM models: $\boldsymbol{m}^{ubm} = [\boldsymbol{\mu}_1^{ubm}, \cdots \boldsymbol{\mu}_r^{ubm}, \cdots \boldsymbol{\mu}_G^{ubm}]$ [36]. The objective is to model the nuisances introduced in the database due to speakers, recording conditions, etc., into a space of *total variability*, $\boldsymbol{T}$. This variability is later compensated by using, typically, linear discriminant analysis and probabilistic linear discriminant analysis. The variability-dependent supervector for a speaker t of a class i is modeled as in Eq. (7), where $\boldsymbol{T}$ is a rectangular matrix of low rank and $\boldsymbol{w}$ is a random vector called *iVector* having a standard normal distribution $\mathcal{N}(0,1)$ [35].

$$\boldsymbol{m}_t^i = \boldsymbol{m}^{ubm} + \boldsymbol{T}\boldsymbol{w} \quad (7)$$

The GMM-SVM is another GMM improvement which combines the discriminating power of a SVM with the GMM modeling capabilities. A SVM is a discriminative binary classifier constructed from sums of a kernel function $\mathcal{K}(\cdot,\cdot)$ which is of the form of Eq. (8), where s_l are ideal outputs taking values -1 or 1, τ_l are weights such that $\sum_{l=1}^L \tau_l s_l = 0, \tau_l > 0$; δ is a learned constant; and $\boldsymbol{z}_l$ are the L support vectors obtained from a training set by an optimization process.

$$f(\boldsymbol{x}) = \sum_{l=1}^L \tau_l s_l \mathcal{K}(\boldsymbol{x}, \boldsymbol{z}_l) + \delta \quad (8)$$

This Kernel function must fulfill the Mercer condition presented in Eq. (9), where $\boldsymbol{b}(\cdot)$ is a mapping from the input space to a possibly infinite dimensional expansion space.

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{b}(\boldsymbol{x})\boldsymbol{b}(\boldsymbol{z})^T \quad (9)$$

By defining $\boldsymbol{m}^c$ and $\boldsymbol{m}^{\bar{c}}$ as supervectors of the target and non-target classes respectively, a linear sequence kernel might be employed. This follows the form of Eq. (10) [37].

$$\mathcal{K}(\cdot,\cdot) = \sum_{r=1}^G \left(\sqrt{\lambda_r}\boldsymbol{\psi}_r^{-1/2}\boldsymbol{m}_r^c\right)^T \left(\sqrt{\lambda_r}\boldsymbol{\psi}_r^{-1/2}\boldsymbol{m}_r^{\bar{c}}\right) \quad (10)$$

Using this scheme for a test utterance $\boldsymbol{y}$, the classification is carried out as the inner product between the target model and the GMM supervector as in Eq. (11).

$$f(\boldsymbol{y}) = [\sum_{l=1}^L \tau_l s_l \boldsymbol{b}(\boldsymbol{z}_l)]^T \boldsymbol{b}(\boldsymbol{y}) + \delta \quad (11)$$

The aforementioned classification schemes (i.e. GMM, GMM-UBM, GMM-UBM-iVector and GMM-SVM) have been applied to the automatic detection of sex and voice pathologies, aiming to identify the best classification framework.

3. Experimental Set–Up

3.1. Saarbrücken Database

The Saarbrücken voice disorders database [28, 38] holds a collection of speech registers from more than 2000 normal and pathological German speakers. It contains the recordings of the sustained phonation of vowels /*i*/,/*a*/ and /*u*/ produced at normal, high and low pitch, as well as with rising-falling pitch. Voice is recorded at a sampling frequency 50 kHz and 16-bits of resolution. For the purpose of this work, only the /*a*/ vowel at normal pitch is considered. Additionally, a subset of the database is segmented by a speech therapist, removing recordings with a low dynamic range or interferences and selecting registers according to an age balance. Table 1 comprises the final distribution of the data according to sex and condition after the speech therapist assessment. Similarly, the database employed for training the UBM models is presented in [27].

Table 1 Distribution of patients according to sex and pathology in the Saarbrücken database

Condition	Female	Male	Total
Normophonic	366	202	568
Pathological	547	431	978
Total	913	633	1546

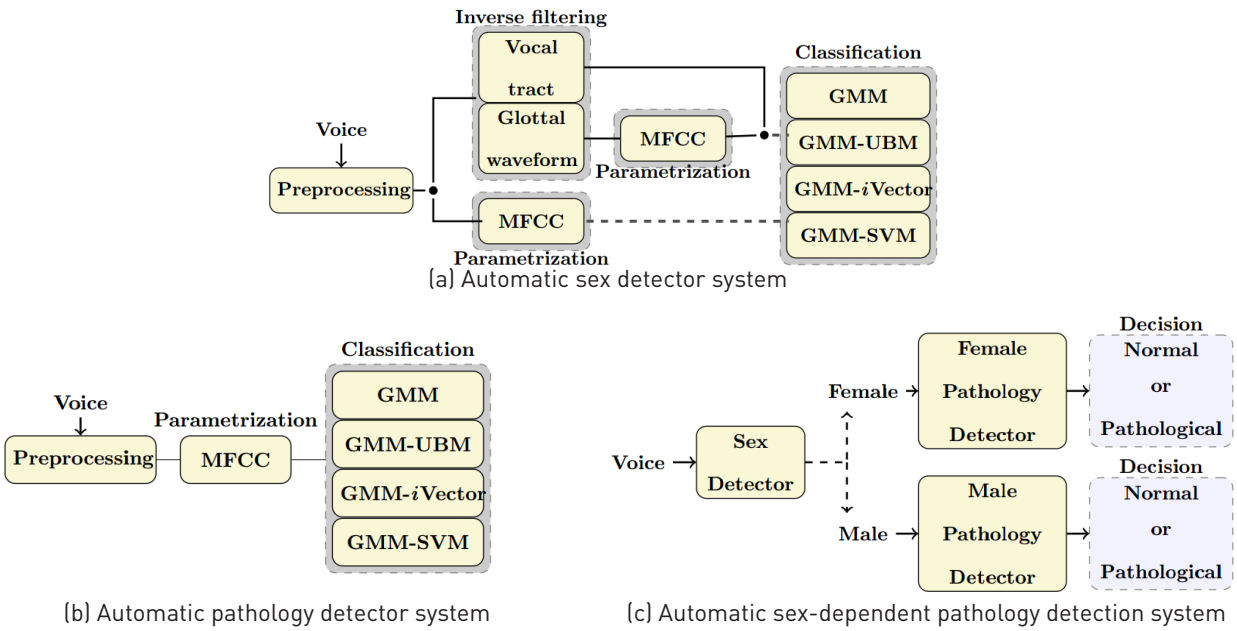
3.2. Methodology

The proposed methodology comprises two main stages: an automatic sex detector and an automatic detector of voice pathologies. The combination in cascade of these subsystems forms a sex-dependent pathology detector. In this manner, the sex detector identifies if the speaker is male/female, and then feeds the sex-dependent pathology detector to check for the speakers' condition.

Sex detector

Figure 2(a) represents the scheme of the automatic sex detector developed in this study. A detailed description of each one of the stages is presented next.

- In the *preprocessing* stage, all speech signals are down-sampled to 25 kHz. In addition, a [–1, 1] normalization is carried out to homogenize the amplitude of processed recordings.
- In the *inverse filtering stage*, the glottal waveform and the vocal tract model $F_v(z)$ are further extracted from the speech via inverse filtering. For modeling the vocal tract $(F_s/1000)+2$ coefficients are used, whereas 4 coefficients are used for modeling the glottal waveform, being F_s the sampling frequency of the input voice recording. $F_v(z)$ is obtained by inverse filtering the deglottalized voice using adaptive paired lattices, as explained in the stage 3 of the section 2.1.
- In the *parameterization* stage two experiments are considered. Firstly, MFCC features are applied to



(a) Automatic sex detector system

(b) Automatic pathology detector system

(c) Automatic sex-dependent pathology detection system

Figure 2 Outline of the proposed systems employed in the paper: (a) presents an automatic sex detector, whereas (b) presents a pathology detector. Both schemes are combined in a cascading framework, forming the proposed automatic sex-dependent pathology detection system in (c)

speech signals. Secondly, MFCC features are applied to glottal waveform and concatenated with the vocal tract model $F_v(z)$, resulting from the inverse filtering procedure. In all cases, the number of MFCC coefficients is varied as: [4:2:22].

- In the *classification stage*, a GMM, a GMM-UBM, a GMM-UBM-iVector and a GMM-SVM are employed. The number of Gaussians is varied in the following set: {4, 8, 16, 24, 32, 64, 128, 200, 256}. Training of the UBM model is carried out using a private database belonging to *"Universidad Politécnica de Madrid"* [27]. The performance of the classifiers is assessed using a 10-fold cross-validation strategy, calculating the classifier accuracy α within a given confidence interval (q). Assuming a 95% value, the q range is estimated as $q = \pm 1.96\sqrt{\alpha(1+\alpha)/M}$, where M is the total number of patterns classified. Moreover, *detection error trade-off curves* (DET) are employed, as well as specificity (s_p) and sensitivity (s_e) at the EER point.

Sex-dependent voice pathology detector

Figure 2(c) presents an outline of the sex-dependent voice pathology detector used in this work. This system is composed by two modules connected in cascade, so the above-described sex detector is used to categorize the speakers according to their sex to further discriminate between normophonic and pathological speakers with two different detectors, one for each sex using the scheme in 2(b).

In addition to the proposed sex-dependent pathology detector, two extra experiments are considered. These constitute the baseline systems used to contrast the performance of the proposed scheme. On one hand, a sex-independent pathology detection system is employed to compare whether or not sex influences the performance of the automatic detector of pathology. The sex-independent system follows the scheme presented in 2(b), leaving aside the sex of the speaker. On the other, the sex-dependent pathology detector of 2(c) is used, but fed with speakers manually categorized according to their sex. The advantage of this scheme is the avoidance of errors committed by the automatic sex detector.

The main stages of the experiments followed are described next, remarking that all three systems follow the same methodology but differ only on the manner on which the a priori information about the sex of the speaker is exploited.

- In the *preprocessing*, the methods are the same as in the automatic sex detector.
- Regarding *parameterization*, the speech is characterized using MFCC coefficients, varying the number in the interval [4:2:22].
- Finally, the *classification and validation* is carried out following the same approach as in the sex detector.

4. Results

4.1. Sex detector

The sex detector is fed with parameters extracted from the raw speech signal and from the glottal waveform extracted by inverse filtering. In this respect Figure 3 presents some samples of the speech and extracted glottal signals for normophonic and pathological speakers.

The MFCC coefficients are used to characterize the raw speech signal, whereas the glottal waveforms are parameterized using MFCC coefficients concatenated with the parameters modelling the vocal tract obtained during the inverse filtering process. Figure 4 shows the DET curve that represents the performance of the sex detector for these two approaches. Additionally, Table 2 summarizes the performance in terms of α, s_e and s_p.

4.2. Sex-dependent voice pathology detector

Firstly, a sex independent pathology detector is used as baseline. It represents a system with no a priori sex information taken into account. Thus, the classifier has two statistical models, one for normophonic speakers and the other for pathological. The DET curve evaluating the performance is shown in Figure 5, while Table 3 summarizes the results in terms of α, s_e and s_p.

Finally, a baseline sex-dependent pathology detection system developed segmenting manually the database according to the sex of the speaker is considered and contrasted with the proposed sex-dependent pathology detector employing automatic sex detection. In both cases, the classifier has four statistical models, two for normophonic speakers (one for males and another for females) and two for pathological. Regarding the proposed sex-dependent voice pathology detector, the operation point which produced the best performance in the sex detector is employed for this further stage. This operation point is obtained with 22 MFCC and using the GMM-SVM classifier with 4 Gaussians.

The DET curves for the baseline and the proposed system are shown in Figure 6, whereas the performance summarized in Table 4.

5. Discussions and conclusions

The present work studies the influence of the speaker's sex in the performance of voice pathology detection systems. Two parameterization schemes have been employed for the design of the sex detector. In particular, a comparison is carried out using MFCC coefficients extracted from the raw speech, and MFCC coefficients extracted from the glottal waveform fused with parameters of the vocal tract model obtained via inverse filtering. Performance has been assessed by means of GMM, GMM-UBM, GMM-UBM-iVector and GMM-SVM classifiers. These classifiers represent

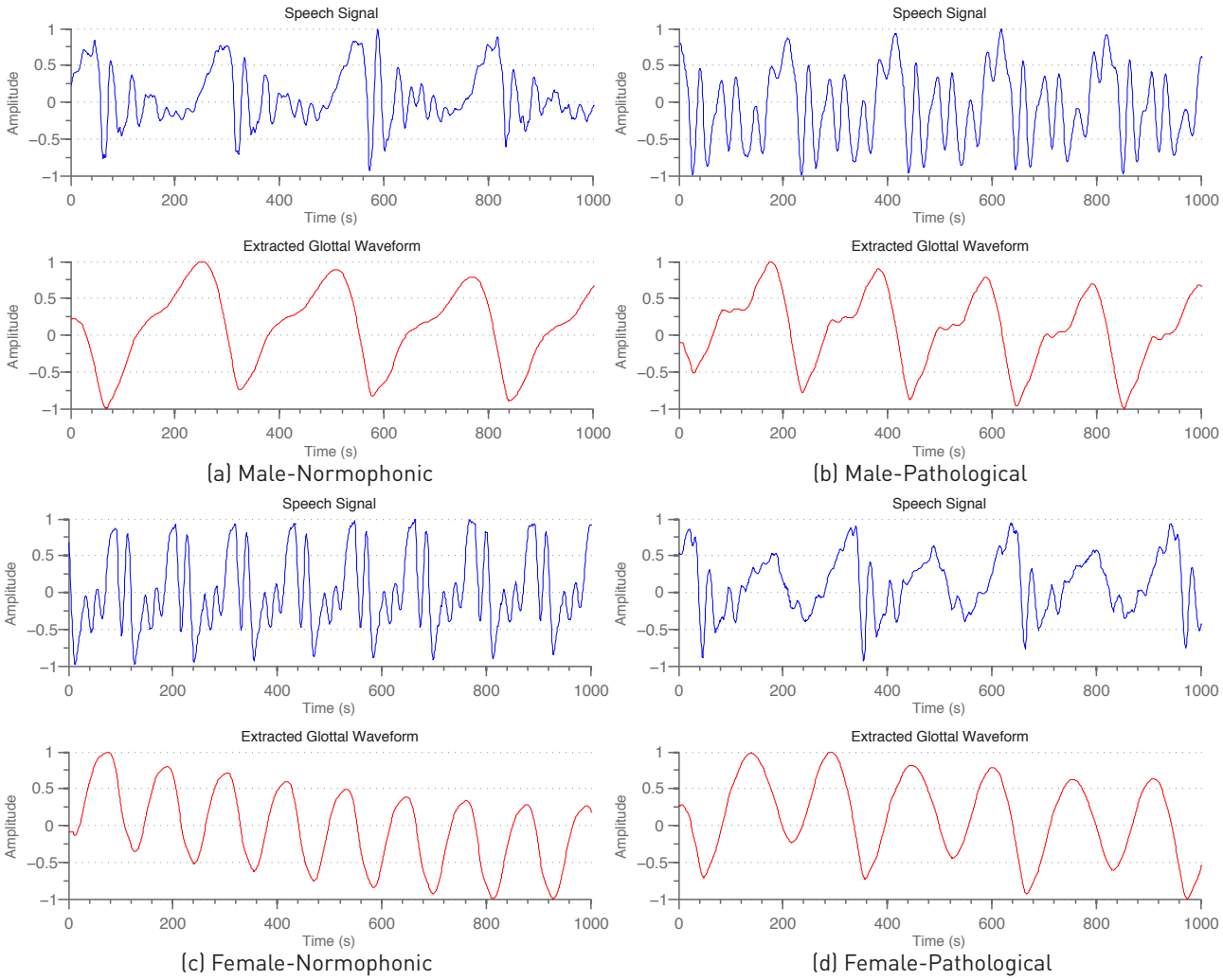
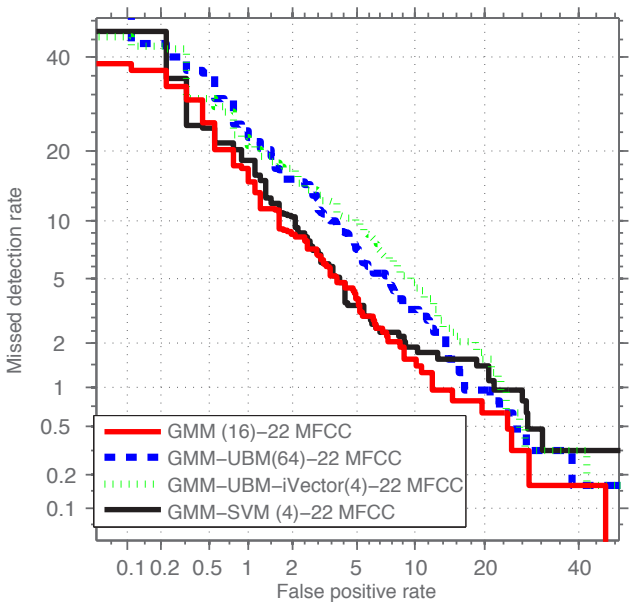


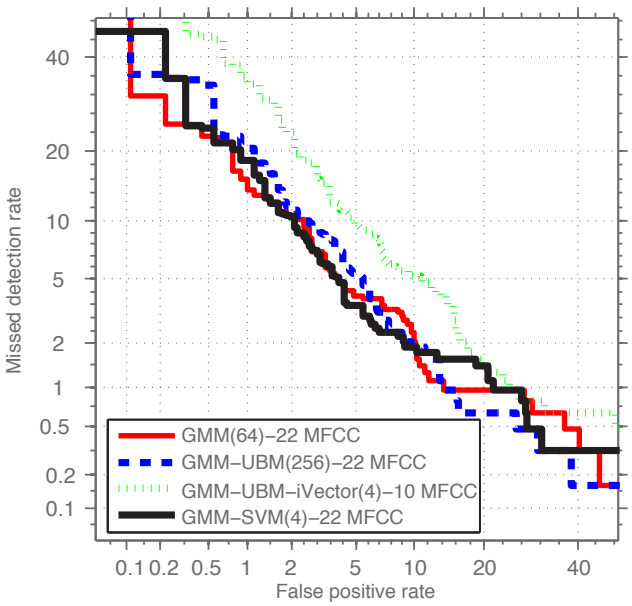
Figure 3 Speech and glottal waveform extracted via inverse filtering. For a male speaker: (a) normophonic and (b) pathological. And for a female speaker: (c) normophonic and (d) pathological

Table 2 Performance of the sex detector using the two proposed sets of parameterizations: MFCC applied to speech and MFCC applied to glottal waveforms along with the vocal tract model

Parameterization	System	$a \pm q$	s_e	s_p
MFCC Speech	GMM	95.66 ± 1.02	0.96	0.96
	GMM-UBM	93.85 ± 1.20	0.93	0.95
	GMM-UBM-iVector	92.95 ± 1.28	0.93	0.93
	GMM-SVM	95.79 ± 1.00	0.96	0.96
MFCC Glottal + Vocal tract	GMM	95.66 ± 1.02	0.95	0.96
	GMM-UBM	94.88 ± 1.10	0.94	0.95
	GMM-UBM-iVector	93.14 ± 1.26	0.93	0.93
	GMM-SVM	95.66 ± 1.02	0.95	0.96



(a) Sex detector using MFCC extracted from speech



(b) Sex detector using MFCC extracted from glottal waveform and including the vocal tract model

Figure 4 DET curve for the sex detector. The features used in (a) are MFCC extracted from voice signals, and in (b) MFCC extracted of glottal waveform along with vocal tract coefficients. The legend shows in parentheses the number of Gaussians used for each classifier and the number of MFCC coefficients that reported the best results

Table 3 Performance of the sex-independent pathology detector using the Saarbrücken database

System	$a \pm q$	s_e	s_p
GMM	71.65 ± 2.25	0.74	0.68
GMM-UBM	70.68 ± 2.27	0.72	0.69
GMM-UBM-iVector	69.05 ± 2.30	0.69	0.69
GMM-SVM	71.06 ± 2.26	0.71	0.71

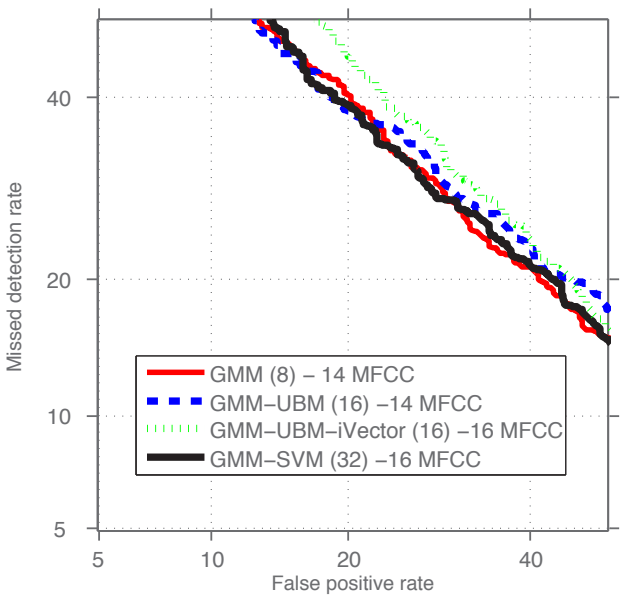
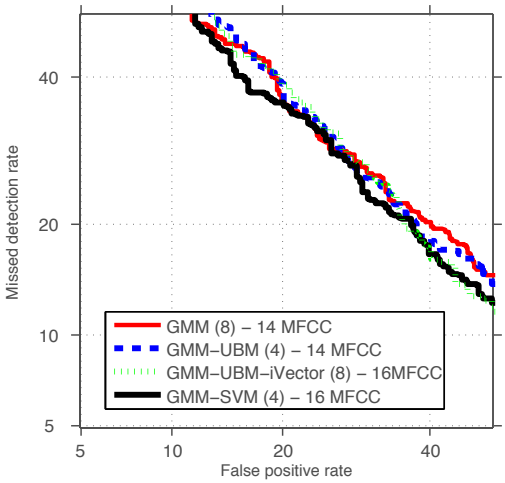
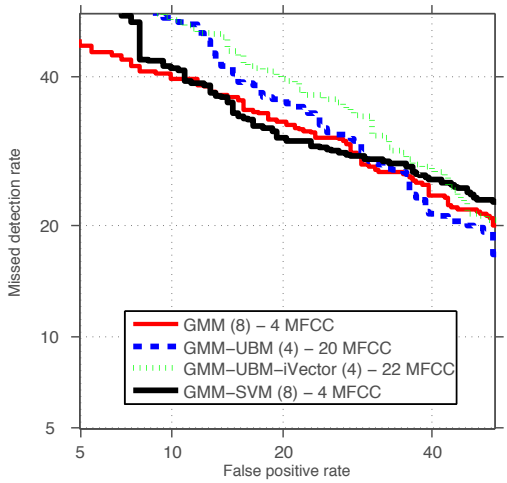


Figure 5 Baseline sex-independent pathology detector. The legend shows in parentheses the number of Gaussians used for each classifier and the number of MFCC coefficients that reported the best results



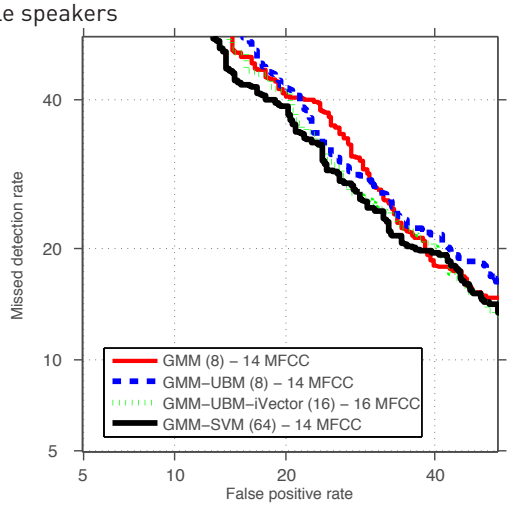
(a) Pathology detector using manual segmentation.



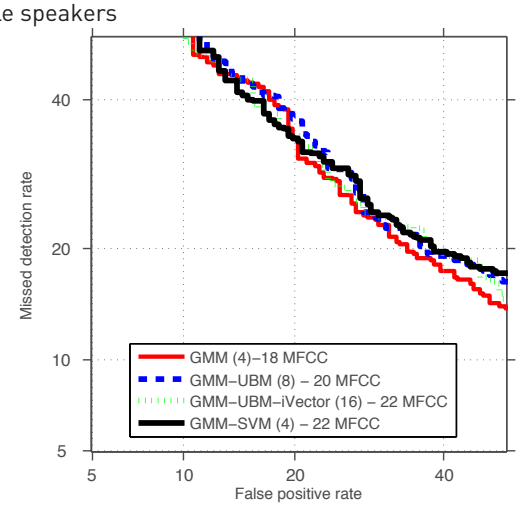
(b) Pathology detector using manual segmentation.

Female speakers

Male speakers



(a) Pathology detector using automatic sex detection.



(b) Pathology detector using automatic sex detection.

Female speakers

Male speakers

Figure 6 DET curve for the sex-dependent pathology detector using manual segmentation for (a) females and (b) males. Similarly, using automatic sex detection the results for (c) females and (d) males are shown.

the state of the art in other application domains such as speaker recognition. In a second stage and with the results obtained in the first phase of the study, a sex-dependent pathology detector is considered. MFCC features and GMM, GMM-UBM, GMM-UBM-iVector and GMM-SVM classifiers are employed. Additionally, and for the sake of completeness, a sex-independent pathology detector, and a manually segmented sex-dependent pathology detector are also tested.

In relation to the glottal waveforms obtained by inverse filtering, the glottal components obtained for male speakers are characterized by an asymmetrical shape, presenting a hump in the open phase of the glottal cycle for both normal and pathological speakers. This behavior match well with the observations reported in the literature. Similarly, female glottal waveforms exhibit a symmetrical appearance, which is also expected, for both normal and pathological speech. This suggests the usefulness of the inverse filtering algorithm even for pathological speech, thus ensuring a more trustworthy parameterization process. However, results are not significantly better than those obtained with a parameterization of the raw speech. This, combined with the higher computational cost of having to inverse filter voice signals, suggest that studying the glottal waveform via inverse filtering might not worthwhile for sex detection tasks in normal and pathological voices.

With respect to the voice pathology detector, results indicate that a manual segmentation of the database produces a slight improvement in performance compared to not using a priori information about the sex of the speaker. However this improvement is not significant. In particular, the accuracy in the mixed scenario for the

Table 4 Performance of the sex-dependent pathology detector using the Saarbrücken database and manual and automatic segmentation of the sex of the speaker. Results are given for male, female and mixed sexes, where the latter refers to combining male's and female's performance measures

Sex	System	Manual segmentation			Automatic sex detection		
		$a \pm q$	S_e	S_p	$a \pm q$	S_e	S_p
Male	GMM	71.87 ± 3.50	0.74	0.68	70.96 ± 2.97	0.72	0.68
	GMM-UBM	71.09 ± 3.53	0.72	0.69	70.52 ± 2.98	0.71	0.68
	GMM-UBM-*i*Vector	68.72 ± 3.61	0.69	0.69	72.97 ± 2.90	0.73	0.72
	GMM-SVM	70.00 ± 3.57	0.71	0.67	72.41 ± 2.92	0.72	0.72
Female	GMM	72.73 ± 2.89	0.76	0.68	76.16 ± 3.29	0.80	0.68
	GMM-UBM	72.07 ± 2.91	0.72	0.72	74.61 ± 3.36	0.78	0.68
	GMM-UBM-*i*Vector	72.00 ± 2.91	0.71	0.73	73.68 ± 3.40	0.75	0.70
	GMM-SVM	72.29 ± 2.90	0.72	0.73	73.68 ± 3.41	0.75	0.70
Mixed	GMM	72.38 ± 2.22	0.75	0.68	73.14 ± 2.21	0.76	0.68
	GMM-UBM	71.67 ± 2.24	0.72	0.71	72.23 ± 2.23	0.74	0.68
	GMM-UBM-*i*Vector	70.63 ± 2.27	0.70	0.71	73.27 ± 2.20	0.74	0.71
	GMM-SVM	71.35 ± 2.25	0.72	0.71	72.94 ± 2.21	0.74	0.71

manually segmented sex-dependent system is 72.38%, mildly higher than the baseline sex-independent pathology detector which performed 71.65%. Results are in line with those in [27] where the classification accuracy of an automatic detector of pathology is lightly improved by using a manual segmentation of the database according to the sex of the speakers. Concerning the proposed sex-dependent pathology detector, and in comparison with the two previous baselines, a light performance improvement is also observed. Specifically, an accuracy rate of 73.27% is achieved, implying an absolute improvement of 1% compared to using manual segmentation, and 2% when no a priori information about the sex of the speaker is included.

On the other hand, the results suggest that the classification approaches tested based on UBM have not provided clear improvements. This might be attributable to the number of speakers of the secondary database used to train the universal models. In particular, and in view of the results, the material used to train the UBM models and to train the total variability space matrix might have been insufficient.

Regarding the corpus of speakers used in this study, it is important to remark that the Saarbrücken database represents an interesting challenge to study the effect of pathologies in the speech. Although it remains almost unexplored in the literature, this database is of great interest due to the size, wide range of pathologies, and variety of speakers. It also illustrates the difficulty that the voice pathology detection problem encompasses. Indeed, the Saarbrücken database has been used for highlighting the necessity for a differentiated classification of normal and pathological phonation into additional subgroups since a strong overlapping between normal and pathological phonation is evidenced [28].

Regarding automatic detection of pathologies on the Saarbrücken database, other study reveals performances that are in line with those obtained in the present work. In particular, the accuracy rates in [39] are about 70%, when employing MFCC coefficients, noise related features, GMM classifiers and the vowel /a/ at normal pitch. Fusing vowels at different conditions (normal, high, low and rising-falling pitch), accuracy raised to 72%. In addition, the literature reports that fusing information from other vowels at different pitch conditions, and in over-optimistic scenarios, accuracies could reach 90% [39, 40].

It is also worth mentioning that the performance of the pathology detection system turns out to be better in male speakers than in female ones. This follows the results in [27] where similar findings are presented: the authors hypothesized that using a cepstral analysis, female voices presented wider distributions making the detection scenario more troublesome.

To sum up, results evidence that sex might be effectively distinguished from normal and pathological speech using sustained vowels with the proposed methodology. Also, there are not significant differences when analyzing speech and glottal and vocal tract components for sex detection, suggesting that the sole speech signal is as informative as its decomposition when employing inverse filtering. Results also provide evidences that despite the limitations attributable to the secondary database employed to train the UBM, to the used acoustic material, and to the simple parameterization methods utilized in this study, an a priori automatic categorization of the speakers according to their sex improves the performance of the automatic detectors of pathology. With that in mind, it would be interesting to study other stratification strategies which in turn might affect speech (e.g. age) to design new hierarchical pathology detection systems. The influence of other approaches to stratify the speakers remains as future work.

6. Acknowledgement

This work has been funded by the Spanish Ministry of Economy and Competitivity under grant TEC2012-38630-C04-01 and *Ayudas para la realización del doctorado* (RR01/2011) from Universidad Politécnica de Madrid.

7. References

1. J. Godino, N. Sáenz, V. Osma, S. Aguilera and P. Gómez, "An integrated tool for the diagnosis of voice disorders", *Medical Engineering & Physics*, vol. 28, no. 3, pp. 276-289, 2006.
2. World Health Organization (WHO), *Gender mainstreaming for health managers: a practical approach.* Geneva, Switzerland: WHO; Department of Gender, Women and Health, 2011.
3. M. Benzeghiba *et al.,* "Automatic speech recognition and speech variability: A review", *Speech Communication*, vol. 49, no. 10-11, pp. 763-786, 2007.
4. C. Huang, T. Chen, S. Li, E. Chang and J. Zhou, "Analysis of speaker variability", in *2nd* INTERSPEECH, Aalborg, Denmark, 2001, pp. 1377-1380.
5. V. Parsa and D. Jamieson, "Acoustic discrimination of pathological voice: sustained vowels versus continuous speech", *Journal of Speech, Language, and Hearing Research,* vol. 44, no. 2, pp. 327-339, 2001.
6. N. Sáenz, J. Godino, V. Osma and P. Gómez, "Methodological issues in the development of automatic systems for voice pathology detection", *Biomedical Signal Processing and Control,* vol. 1, no. 2, pp. 120-128, 2006.
7. J. Godino, P. Gómez and M. Blanco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters", *IEEE Trans. Biomed. Eng.,* vol. 53, no. 10, pp. 1943-1953, 2006.
8. J. Arias, J. Godino, N. Sáenz, V. Osma and G. Castellanos, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients", *IEEE Trans. Biomed. Eng.,* vol. 58, no. 2, pp. 370-379, 2011.
9. D. Childers, K. Wu, K. Bae and D. Hicks, "Automatic recognition of gender by voice", in *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), New York, USA, 1988, pp. 603-606.
10. T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiation", in *Language Resources and Evaluation Conference* (LREC), Genoa, Italy, 2006, pp. 1123-1126
11. W. Andrews, M. Kohler, J. Campbell, J. Godfrey and J. Hernández, "Gender-dependent phonetic refraction for speaker recognition", in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), Orlando, USA, 2002, pp. 149-152.
12. S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557-1565, 2006.
13. D. Childers and K. Wu, "Gender recognition from speech. Part II: Fine analysis", *The Journal of the Acoustical Society of America,* vol. 90, pp. 1841-1856, 1991.
14. K. Wu and D. Childers, "Gender recognition from speech. Part I: Coarse analysis", *The Journal of the Acoustical Society of America,* vol. 90, pp. 1828-1840, 1991.
15. D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820-857, 1990.
16. T. Hixon, G. Weismer and J. Hoit, *Preclinical Speech Science: Anatomy, Physiology, Acoustics, Perception,* 1st ed. San Diego, USA: Plural Publishing, Inc., 2008.
17. A. Behrman, *Speech and Voice Science, 1st* ed. San Diego, USA: Plural Publishing, Inc., 2007.
18. M. Södersten, S. Hertegård and B. Hammarberg, "Glottal closure, transglottal airflow, and voice quality in healthy middle-aged women", *Journal of Voice,* vol. 9, no. 2, pp. 182-197, 1995.
19. H. Hanson and E. Chuang, "Glottal characteristics of male speakers: acoustic correlates and comparison with female data", *The Journal of the Acoustical Society of America,* vol. 106, no. 2, pp. 1064-1077, 1999.
20. E. Mendoza, N. Valencia, J. Muñoz and H. Trujillo, "Differences in voice quality between men and women: use of the long-term average spectrum (LTAS)", *Journal of Voice*, vol. 10, no. 1, pp. 59-66, 1996.
21. H. Hanson, "Glottal characteristics of female speakers: acoustic correlates", *The Journal of the Acoustical Society of America,* vol. 101, no. 1, pp. 466-481, 1997.
22. E. Holmberg, R. Hillman and J. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice", *The Journal of the Acoustical Society of America,* vol. 84, no. 2, pp. 511-529, 1988.
23. R. Monsen, and E. Engebretson, "Study of variations in the male and female glottal wave", *The Journal of the Acoustical Society of America,* vol. 62, no. 4, pp. 981-993, 1977.
24. L. Walawalkar, M. Yeasin, A. Narasimhamurthy and R. Sharma, "Support vector learning for gender classification using audio and visual cues: A comparison", in *1st International Workshop on Pattern Recognition with Support Vector Machines* (SVM), Niagara Falls, Canada, 2002, pp. 144-159.
25. Y. Zeng, Z. Wu, T. Falk and W. Chan, "Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech", in *International Conference on Machine Learning and Cybernetics,* Dalian, China, 2006, pp. 3376-3379.
26. C. Muñoz, R. Martínez, A. Álvarez, L. Mazaira and P. Gómez, "Discriminacion de genero basada en nuevos parámetros MFCC", in *1st WTM-IP: Workshop de Tecnologías Multibiométricas para la Identificación de personas,* Las Palmas de Gran Canaria, Spain, 2010, pp. 22-25.
27. R. Fraile, N. Sáenz, J. Godino, V. Osma and C. Fredouille, "Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex", *Folia Phoniatrica et Logopaedica,* vol.

61, no. 3, pp. 146-152, 2009.
28. M. Putzer and W. Barry, "Instrumental dimensioning of normal and pathological phonation using acoustic measurements", *Clinical Linguistics & Phonetics,* vol. 22, no. 6, pp. 407-420, 2008.
29. P. Gómez *et al.,* "Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters", *Journal of Voice,* vol. 21, no. 4, pp. 450-476, 2007.
30. P. Gómez *et al.,* "Evidence of vocal cord pathology from the mucosal wave cepstral contents", in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), Montreal, Canada, 2004, pp. 437-440.
31. M. Airas, "TKK Aparat: an environment for voice inverse filtering and parameterization", *Logopedics Phoniatrics Vocology,* vol. 33, no. 1, pp. 49-64, 2008.
32. P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering", *Speech Communication,* vol. 11, no. 2-3, pp. 109-118, 1992.
33. P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental", in *Joint Workshop on Pattern Recognition and Artificial Intelligence*, Hyannis, USA, 1976, pp. 91-103.
34. D. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.
35. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification", *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 4, pp. 788-798, 2011.
36. T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors", *Speech Communication,* vol. 52, no. 1, pp. 12-40, 2010.
37. W. Campbell, D. Sturim and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification", *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308-311, 2006.
38. M. Pützer and W. Barry, *Saarbrücken voice database*, Saarland University. [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de. Accessed on: Aug. 29, 2009.
39. D. Martínez, E. Lleida, A. Ortega, A. Miguel and J. Villalba, "Voice pathology detection on the Saarbrücken voice database with calibration and fusion of scores using multifocal toolkit", in *IberSPEECH: "VII Jornadas en Tecnología del Habla"* and *III Iberian SLTech Workshop,* Madrid, Spain, 2012, pp. 99-109.
40. D. Martínez, E. Lleida, A. Ortega and A. Miguel, "Score level versus audio level fusion for voice pathology detection on the Saarbrücken Voice Database", in *IberSPEECH: "VII Jornadas en Tecnología del Habla" and III Iberian SLTech Workshop*, Madrid, Spain, 2012, pp. 110-120.