Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 45 (2017) 437–456

www.elsevier.com/locate/csl

Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speech speaker verification[☆]

Milton Sarria-Paja[a,b], Tiago H. Falk[*,a]

[a] Institut National de la Recherche Scientifique (INRS-EMT), University of Quebec, Montreal, Quebec, Canada
[b] Universidad Santiago de Cali, Cali, Colombia

Received 15 October 2016; received in revised form 3 February 2017; accepted 25 April 2017

Abstract

Whispered speech is a natural speaking style that despite its reduced perceptibility, still contains relevant information regarding the intended message (i.e., intelligibility), as well as the speaker identity and gender. Given the acoustic differences between whispered and normally-phonated speech, however, speech applications trained on the latter but tested with the former exhibit unacceptable performance levels. Within an automated speaker verification task, previous research has shown that i) conventional features (e.g., mel-frequency cepstral coefficients, MFCCs) do not convey sufficient speaker discrimination cues across the two vocal efforts, and ii) multi-condition training, while improving the performance for whispered speech, tends to deteriorate the performance for normal speech. In this paper, we aim to tackle both shortcomings by proposing three innovative features, which when fused at the score level, are shown to result in reliable results for both normal and whispered speech. Overall, relative improvements of 66% and 63% are obtained for whispered and normal speech, respectively, over a baseline system based on MFCCs and multi-condition training.

Keywords: Whispered speech; Speaker verification; Modulation spectrum; Mutual information; System fusion

1. Introduction

According to recent statistics, speech-based biometrics have ranked highly in costumer preference, outranking fingerprint and iris scanning solutions (O’Neil King, 2014; Markets, 2015). Due to widespread usage of smartphones worldwide, speech-based biometrics are quickly gaining popularity, particularly in financial institutions (O’Neil King, 2014). Within such applications, customers can gain access to their secure banking and insurance services by simply speaking into their phones. For financial institutions, this ease-of-use enhances customer satisfaction, whilst reducing customer care costs through increased automation rates. Costumers, on the other hand, given the flexibility of speech based communication can pose big challenges to such applications by changing, for

[☆] This paper has been recommended for acceptance by Roger Moore.
[*] Corresponding author.
E-mail address: falk@emt.inrs.ca (T.H. Falk).
URL: http://www.musaelab.com (T.H. Falk).

http://dx.doi.org/10.1016/j.csl.2017.04.004
0885-2308/ 2017 Elsevier Ltd. All rights reserved.

example, their vocal effort based on the environment or the context they are in. This, together with ambient noise, has posed serious threats to speech enabled applications performance in general. Ambient noise has detrimental effects on speech based biometrics systems, particularly those trained with mel-frequency cepstral coefficients (MFCC). As an example, speaker identification accuracy as low as 7% has been reported in very noisy environments (Ming et al., 2007). As such, over the years, several speech enhancement algorithms have been proposed for environment-robust speaker recognition applications (Rao and Sarkar, 2014). Varying vocal efforts, however, have received significantly less exposure, despite its severe detrimental effects on speaker verification performance. For example, whispered-speech speaker identification accuracy as low as 20% has been reported (Grimaldi and Cummins, 2008) in clean conditions. In fact, it is highly likely that customers utilizing a mobile banking application on their smartphones will use a low vocal effort when providing sensitive information, and for the purposes of this research we are interested in whispered speech.

Here, special emphasis is placed on whispered speech because this speaking-style has gained great attention for security applications lately. With reduced perceptibility, whispered speech is a natural mode of speech production that conveys relevant and useful information for many applications. Just as normal-voiced speech, whispered speech not only conveys a message, but also traits such as identity, gender, emotional, and health states, to name a few (Lass et al., 1976; Tartter, 1991; Ito et al., 2005; Chenghui et al., 2009; Tsunoda et al., 2012). As previously mentioned, whispered speech is commonly used in public situations where private or discrete information needs to be exchanged, for example, when providing a credit card number, bank account number, or other personal information. Despite the amount of information present in whispered speech, there are certain characteristics that make this speaking style challenging when presented as a possible input to speech enabled applications. As an example, the most salient characteristic of whispered speech is the lack of vocal fold vibration. Furthermore, when a person whispers, several changes occur in the vocal tract configuration, thus altering not only the excitation source, but also the syllabic rate and the general temporal dynamics characteristics of the generated speech signal (Jovicic and Saric, 2008; Ito et al., 2005). Hence, it is expected that classical methods designed for normal-voiced speech characterization will fail when tested in atypical scenarios including whispered speech (Grimaldi and Cummins, 2008; Ito et al., 2005; Fan and Hansen, 2011; Zelinka et al., 2012).

Despite the limited research done in this field, different approaches attempting to overcome some of these disadvantages have been reported, particularly within training/test mismatch conditions where speaker models were trained with normal speech and tested with whispered speech (Fan and Hansen, 2011; Grimaldi and Cummins, 2008; Fan and Hansen, 2013). In previous work, we found that among different feature sets and strategies evaluated, including frequency warping and alternate feature representation such as MHEC (mean Hilbert envelope coefficients) or WIF (weighted instantaneous frequencies), invariant information between normal-voiced and whispered speech is not sufficient to achieve reliable performance in speaker verification tasks for both speaking styles (Sarria-Paja and Falk, 2015). In addition to this, it was also observed that the strategies with better performance for normal-voiced speech did not exhibit the same benefits for whispered speech. Finally, it was shown the need to include data from both speaking styles in order to allow for a speaker verification system to handle both normal and whispered speech for practical applications (Sarria-Paja and Falk, 2015). It was concluded that efforts should be directed towards new feature representations aimed at reducing the impact of the addition of whispered speech during training or enrollment and to extract more efficiently speaker specific information from whispered recordings.

This paper proposes just that. Here, we propose the computation of features aiming at extracting invariant information embedded within both speaking styles. This is achieved by computing modulation spectrum based features, which in the past have been shown to accurately separate speech from environment-based components (e.g., noise and reverberation) (Falk and Chan, 2010), thus adding robustness to speaker recognition systems. We also use mutual information (MI) as an analysis measure to identify invariant information between normal-voiced and whispered speech feature pairs. MI helps to analyze both linear and non-linear statistical dependencies between the two feature sets, and has shown to be an effective way to measure relevance and redundancy among features for feature selection purposes or even characterization (Peng et al., 2005; Estevez et al., 2009; Clerico et al., 2015). This, combined with system fusion, will help to not only reduce error rates when there are no whispered speech recordings from target speakers for enrollment, but also to reduce the observed negative impact of adding whispered speech during parameter estimation and enrollment. As an example, in speech recognition, gains in whispered speech accuracy were countered by losses in normal speech accuracy, often by the same amount (Zelinka et al., 2012). Such tradeoffs were attributed to excessive generality of the speech model (caused by large variations in the training set) and consequently reduced capability of discriminating among speech units (Zelinka et al., 2012). The new proposed system overcomes this limitation for a speaker verification task.

The remainder of this paper is organized as follows. Section 2 provides a brief background on whispered speech, emphasizing the main differences with normal speech and some approaches found in the literature related to the problem at hand. Section 3 describes the speaker verification problem, the corpus employed for speaker verification, and the baseline system characterization. Section 4 discusses different approaches and strategies to reduce the error rate in whispered speech speaker verification, presents and discusses the experimental results and the performance achieved by the proposed schemes. Lastly, Section 5 presents the conclusions.

2. Whispered speech

In the past, perceptual studies have been conducted to characterize major acoustic differences between whispered and normal-voiced speech; the lack of fundamental frequency being the most relevant difference. Nonetheless, it is not the only major change, and formant shifts towards higher frequencies (Thomas, 1969; Higashikawa et al., 1996), especially lower formants (Sharifzadeh et al., 2012) have also been reported. Whispered speech also has a lower and flatter power spectral density (Ito et al., 2005). In Tran et al. (2013), a statistical analysis was carried out to compare several acoustic and visual features between the two speaking styles. In total, 64 acoustic features referred as *low level descriptors* (LLDs) grouped in three categories, i.e., spectral LLDs, prosody LLDs and voice quality LLDs were compared and 56 showed to be statistically different. But not everything regarding normal-voiced and whispered speech is different; for example while it has been documented that characteristics of vowels and voiced consonants are significantly different, unvoiced consonants are relatively similar (Jovicic and Saric, 2008).

Findings of the above mentioned studies have lead researchers to explore different applications of whispered speech processing such as reconstruction of normal voiced speech from whispers (Sharifzadeh et al., 2010), whispered speech recognition (Ito et al., 2005) and also speaker recognition (Grimaldi and Cummins, 2008; Fan and Hansen, 2013) (SR) which is the focus of this work. Whispered speech, in the past, has been explored mainly for the speaker identification problem within a reduced amount of speakers or using only female speakers (Grimaldi and Cummins, 2008; Fan and Hansen, 2009; Jin et al., 2007; Fan and Hansen, 2008; 2011; 2013). In these works, low accuracy in mismatched training/testing conditions has been reported. Moreover, several strategies to address the intrinsic limitations of current speech enabled systems to process whispered speech have been proposed, such as robust features, modified linear cepstral coefficients (LFCC) and feature mapping (Fan and Hansen, 2009), feature warping over MFCC and score combination at the frame level (Jin et al., 2007), model adaptation schemes, such as maximum linear regression (MLLR), and feature transformation from normal to whispered speech to be used during map adaptation (Fan and Hansen, 2013). Despite the many different efforts, relative improvements range from 8% to 46% and for all cases, the achieved classification results are still not useful for practical applications (e.g. accuracy $< 89\%$ for speaker identification). In addition, in Jin et al. (2007) and Fan and Hansen (2011) it has been documented that the less expensive and most effective strategy is to add small amounts of whispered speech from target speakers during enrollment. This approach, known as multi-style modeling, is one of the most popular approaches proposed to handle multiple vocal effort inputs in a speech enabled system (Ito et al., 2005; Zelinka et al., 2012).

3. Automatic speaker verification (SV) system characterization

Traditionally, speaker recognition systems have been based on identity vectors (*i-vectors*) extraction (Dehak et al., 2011) and matching between a test utterance and a target speaker is done using either a fast scoring method based on cosine distance between i-vectors or probabilistic linear discriminant analysis (PLDA) (Sizov et al., 2014) based scoring. More recently, deep neural network (DNN) approaches have shown to be useful either during feature extraction or computing statistics by replacing the role of the classical Gaussian mixture models (Matejka et al., 2016). However, to properly train a DNN based system, large amounts of training data are required (Matejka et al., 2016). To the best of our knowledge, no large scale corpus with annotated whispered speech is available to train a DNN based system, which poses a challenge in this high variability scenario and limits the advantages of these techniques over other strategies for the task at hand. For this reason we use classical approaches as briefly described below for the sake of completeness. For the experiments herein, the open-source *Bob* signal processing toolbox was used (Anjos et al., 2012).

3.1. Mel Frequency Cepstral Coefficients − MFCC

The Mel Frequency Cepstral Coefficients (MFCCs) are the most widely used features in today's speech enabled systems. Here, prior to MFCC computation, each speech recording was down-sampled to 16 kHz. Next the recordings were pre-emphasized using a first order finite impulse response filter with constant $a = 0.97$. In our experiments, 27 triangular bandpass filters spaced according to the mel scale were used in the computation of 13 MFCC features including the 0−th order cepstral coefficient (log-energy). The MFCC were computed on a per-window basis using a 25 ms window with 40% overlap. Dynamic or transitional features (Δ and $\Delta\Delta$ MFCC) were computed by means of an anti-symmetric Finite Impulse Response (FIR) filter of length nine to avoid phase distortion of the temporal sequence. MFCC are widely used in speech applications, thus a detailed description of the signal processing steps involved is beyond the scope of this paper. After dropping frames where no vocal activity was detected, cepstral mean and variance normalization was applied per recording to remove linear channel effects. These features are then used to compute the so-called i-vectors, as described next.

3.2. i-vectors/PLDA approach

The i-vector extraction technique was proposed to map large-dimensional input data to a small-dimensional feature vector while retaining most relevant speaker information. First, a C-Component Gaussian mixture model (GMM) is trained as an universal background model (UBM) using the Expectation − Maximization (EM) algorithm and the data available from all speakers from the train set or background data. Speaker and session-dependent supervectors of concatenated GMM means are modeled as $M = m + T\phi$, where m is the speaker- and channel-independent supervector, $T \in \mathbb{R}^{CF \times D}$ is a rectangular matrix of low rank covering the important variability (total variability matrix) in the supervector space. C, F and D represent, respectively, the number of Gaussians in the UBM, the dimension of the acoustic feature vector and the dimension of the total variability space. Finally $\phi \in \mathbb{R}^{D \times 1}$ is a random vector with density $\mathcal{N}(0, I)$ and referred to as the identity vector or *i-vector* (Dehak et al., 2011). This procedure is complemented with some post-processing techniques such as linear discriminant analysis (LDA), whitening, and length normalization. These techniques can be used to remove nuisance effects in the total variability space. The interested reader is referred to Dehak et al. (2011) for more complete details. For the experiments herein, an i-vector is computed per enrollment utterance and then they were averaged to obtain a single i-vector per target speaker.

The PLDA model (Prince and Elder, 2007; Sizov et al., 2014), on the other hand, splits the total data variability into within-individual and between-individual variabilities, both residing in small-dimensional subspaces. Originally introduced for face recognition, PLDA has become a standard in speaker recognition. PLDA was formulated in Prince and Elder (2007) as $\phi_{ij} = \mu + Vy_i + Ux_{ij} + \varepsilon_{ij}$, where ϕ_{ij} is the i-th feature vector associated to the j-th speaker, the matrices $V \in \mathbb{R}^{D \times P}$ and $U \in \mathbb{R}^{D \times M}$ span the between- and within- individual spaces, μ is a global mean, $y_i \sim \mathcal{N}(0, I)$ and $x_{ij} \sim \mathcal{N}(0, I)$ are hidden variables in the spaces spaned by V and U, respectively, and the residual $\varepsilon_{ij} \sim \mathcal{N}(0, \Sigma)$ is defined to be Gaussian with zero mean and diagonal covariance Σ. In a verification scenario, there are two possible hypotheses: 1) ϕ_{test} and ϕ_{enrol} share the same class, and 2) ϕ_{test} and ϕ_{enrol} are from different classes. Lastly, the corresponding score can be obtained by computing the log-likelihood between the two hypotheses, which is given by $s = \ln(P(\phi_{test}, \phi_{enrol})) - \ln(P(\phi_{test})P(\phi_{enrol}))$; details can be found in Prince and Elder (2007) and Sizov et al. (2014). For the experiments herein, the dimensionality of matrices V and U were set to $P = dim_{LDA}$ and $M = 0$, where dim_{LDA} represents the dimensionality of the LDA model, which is tuned accordingly per feature set.

Fig. 1, shows the protocol typically followed during training, enrollment and testing stages. As can be seen, three different sets of speech recordings are needed. First, large amounts of speech data are needed to train e.g., the so-called GMM universal background model (UBM) and estimate other parameters needed for i-vector extraction (e.g., the T matrix estimation). During enrollment, a separate set is needed from each speaker to allow for e.g., maximum a posteriori adaptation in GMM-based systems or for i-vector extraction to match with testing samples. Lastly, a third unseen data set is needed for system accuracy calculation. In the case of multi-style training, whispered speech data can be available in one or multiple datasets.

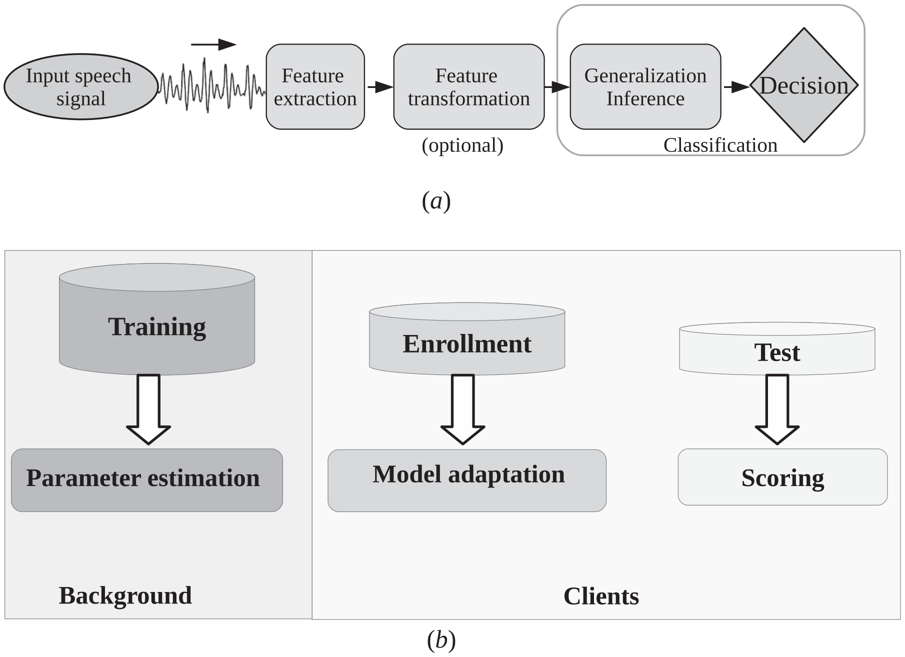


Fig. 1. (a) Building blocks for a general purpose pattern recognition system that can be applied to speaker verification and (b) data recording sets involved during training, enrollment and testing of a speaker verification system.

3.3. Corpus description

In our experiments, three different databases were used, the CHAINS (Characterizing Individual Speakers) speech corpus (Grimaldi and Cummins, 2008), wTIMIT (whispered TIMIT) (Lim, 2011) and TIMIT databases (Garofolo and Consortium, 1993). The CHAINS and wTIMIT databases contain normal and whispered speech. Table 1 presents details about the number of speakers and recordings per speaker available in the datasets.

Speakers from the three databases were divided in two disjoint sets for training, to be used as background data, and a second set for enrollment and testing. Recordings from 462 speakers from TIMIT database and 14 speakers from wTIMIT, 476 in total, are included in the training set. Recordings from 100 speakers from TIMIT database, 24 speakers from wTIMIT and 36 speakers from CHAINS are included in the enrollment and testing set. Average duration for each speech recording is 4.5s. To characterize the baseline system we included only normal speech recordings for training and enrollment, for testing we used two recordings per speaker, and if there are whispered speech recordings available then two additional sentences were included per speaker. Details can be found in Table 2.

Table 1
Details about the three databases used in our experiments.

Database	Num. of speakers		Recordings/speaker	
	Female	Male	Norm.	Whsp.
TIMIT	192	438	10	–
wTIMIT	24	24	450	450
CHAINS	16	20	37	37

Table 2
Number of speakers and total number of recordings per database for training, enrollment and testing sets.

	Num. of speakers/database			Total record.	
	TIMIT	wTIMIT	CHAINS	Norm.	Whsp.
UBM estimation	462	0	0	3696	0
T matrix estimation	462	14	0	9996	6300
LDA and PLDA training	462	14	0	9996	6300
Enrollment	100	24	36	1280	480
Testing	100	24	36	320	120
Fusion system	68	10	0	780	230

3.4. Baseline SV system characterization

To characterize the performance of a valid baseline system to compare performances to, we follow the steps suggested in Khoury et al. (2014). In Khoury et al. (2014), the authors used a GMM-UBM + MAP (maximum a posteriori) adaptation based system and provided the lists for background, enrollment and test sets using the TIMIT dataset. By using the same lists we report equal error rate (EER) results in Table 3 using two scoring strategies, i.e., cosine kernel and PLDA based scoring, and as feature vectors we used MFCC. According to the results reported in Table 3, MFCC seems to be a better choice than LFCC, which was a feature set proposed in Khoury et al. (2014). These results corroborate those from previous studies in the context of i-vectors, where the PLDA based system is preferable to the cosine distance based one.

The lower part of Table 3 also reports the results from similar experiments by using the three databases and using only normal speech for parameter estimation and enrollment, thus helping to quantify the effects of whispered speech on a standard SV system during the testing stage. As can be seen, when testing with normal speech, the PLDA scoring based system outperforms the cosine kernel based scoring. When testing with whispered speech, from the tabulated EER values, it can be observed that significant performance degradation occurs in mismatched training/test conditions. There is a gap in performance between normal and whispered speech higher than 20% for the two cases. Finally, by comparing the tabulated results only for TIMIT database (top of Table 3) and the results when including CHAINS and wTIMIT databases (bottom), the addition of new speakers during enrollment whose recordings have been obtained in different conditions can affect the performance of the system. Henceforth, the MFCC + i-vector/PLDA based system will be the baseline system and will be referred to as **S0**. Results reported in the Table were obtained using the following parameters for the UBM, T matrix, PLDA and LDA: $C = 256$, $D = 400$, and $P = 288$. When combining the three databases the total number of target and non-target trials are 320 and 50,880, respectively, for normal speech, and for whispered speech 120 target and 19,080 non-target trials. All comparisons will be relative to this benchmark.

Table 3
EER (%) comparison with the baseline system using only TIMIT database (Top), and including CHAINS and wTIMIT (Bottom). For these results $C = 256$, $D = 400$ and $P = 288$.

Scoring	EER (%)	
approach	Norm.	Whsp.
Only TIMIT database		
Baseline system (LFCC) (Khoury et al., 2014)	2.68	–
i-vector/cosine kernel (MFCC)	2.91	–
i-vector/PLDA (MFCC)	1.79	–
Including CHAINS and wTIMIT databases		
i-vector/cosine kernel	5.00	27.16
i-vector/PLDA (S0)	2.81	27.31

4. Proposed strategies to improve system performance when testing with whispered speech

Results presented in the previous section have shown that standard SV systems have serious deficiencies when facing atypical scenarios. As such, for the task at hand, it is necessary to devise strategies aiming to compensate for the negative effects when whispered speech is considered into the possible testing scenarios. The approach taken in this work is to improve the feature representation used for i-vector extraction. From the baseline experiments it is clear that the total variability matrix can map to a highly discriminative space as long as there is sufficient statistics to learn from. This is the case for normal-voiced speech, but not for whispered speech. Thus, it is necessary to calculate new features aiming to maximize the invariant information present in both speaking styles. As was previously discussed in Sarria-Paja and Falk (2015), often the strategies that performed better in the mismatched scenario did not offer good results in the matched case and vice-versa. Hence, the compromise of keeping high performance levels for normal-voiced speech while reducing error rates for whispered speech is hard to maintain.

The lack of sufficient whispered speech recordings for parameter estimation is one of the problems that has been discussed before. Even if a large number of recordings were collected, it would not suffice as long as the number of speakers is small. While this is the case for the experiments herein, this scenario can allow us to evaluate how efficiently a system uses the data available in different stages not only during training but also during enrollment. First of all, it can be assumed that there are no whispered speech recordings from target speakers but there is data from a reduced set of background speakers in the training set. These recordings can be used in two different ways: 1) Include whispered speech recordings in the training set such that those recordings can be used during parameter estimation, or 2) These recordings can also be used to perform different analysis helpful to find invariant information present in both normal-voiced and whispered speech related to speaker identity. The former case, relies on the assumption that there is enough inter-speaker variability in the training set for both speaking styles, which for the task at hand, is true only for normal speech. The latter case, in turn, relies on finding invariant information between normal and whispered speech, which is a challenging task due to the acoustic differences between the two vocal efforts. In this work we make use of both approaches aiming to compensate for the lack of whispered speech recordings during parameter estimation or enrollment with improved features aiming to maintain invariant information between normal-voiced and whispered speech, as well as improve performance for whispered speech whilst not affecting (or even improve) performance in normal-voiced speech.

4.1. Proposed auditory-inspired amplitude modulation features − AAMF

For the analysis in this section we assume that an observed time-domain signal is the result of multiplying a low-frequency modulator (temporal envelope) by a high-frequency carrier. Hence, the modulation spectrum characterizes the rate of change of long-term speech temporal envelopes (Falk et al., 2012), and the analysis is carried out by using acoustic subbands. The modulation frequency (modulation domain) represents the frequency content of the subband amplitude envelopes and it potentially conveys information about speaking rate and other speaker specific attributes (Kinnunen and Li, 2010). Auditory-inspired amplitude modulation features have been effectively used in the past to improve automatic speaker identification in the presence of room reverberation (Falk and Chan, 2010). It was shown that modulation spectral features accurately separate speech from environment-based components. In that case, the technique relied on identifying the modulation frequency channels that remained unaffected by environmental noise based on energy levels, and disregarding those that presented significant changes when affected by noise. A similar idea can be applied for the task at hand, however, identification of channels or variables containing invariant information cannot be based on energy levels given their inherent differences between normal-voiced and whispered speech. For this reason, mutual information (MI) is chosen to compare pairs of variables coming from the two speaking styles and determine whether a specific channel contains shared information that can be useful for speaker recognition purposes.

The approach to compute the auditory-inspired amplitude modulation features is based on the approach presented in Kinnunen et al. (2008). However, the algorithm has been adapted to fit our needs. Thus for the sake of completeness, we present the details of our implementation. More specifically, the speech signal $s(n)$ is first processed by an N-point short-time discrete Fourier transform (STDFT) to generate $S(nL, f_a)$

given by:

$$S(nL_a, f_a) = \sum_{m=-\infty}^{\infty} s(m) w_a(nL_a - m) e^{-i\frac{2\pi k}{N}m}, \tag{1}$$

where $w_a(n)$ is an acoustic frequency analysis window and L_a denotes the frame shifts, the subscript a stands for acoustic domain. Acoustic frequency components (termed f_a) are aligned in time to form the conventional time-frequency representation. In order to emulate human cochlear processing, the squared magnitudes of the obtained acoustic frequency components are grouped into 27 subbands ($|S_j(\cdot)|, j = 1, \ldots, 27$), spaced according to the perceptual mel scale as depicted in Fig. 2 (top plot). A second transform is then performed across time for each of the 27 subband magnitude signals to yield:

$$S_j(mL_m, f_m) = \sum_{n=-\infty}^{\infty} |S_j(n)| w_m(mL_m - n) e^{-j\frac{2\pi k}{N}n}, \tag{2}$$

where $w_m(m)$ is a modulation frequency analysis window, L_m the frame shift, the subscript m stands for modulation domain, j indexes the acoustic frequency bands, and f_m represents modulation frequency bins. Following recent physiological evidence of an auditory filterbank structure in the modulation domain (Xiang et al., 2013), we further group squared modulation frequency bins into eight subbands using logarithmically-spaced triangular bandpass filters distributed between 0.01−80 Hz modulation frequency as depicted in Fig. 2 (bottom plot). The speech modulation spectrum results in a high-dimensional feature representation (e.g., 27 acoustic bands × 8 modulation bands = 216 dimensions), finally log_{10} compression is applied. Fig. 3 summarizes the above described process. As can be seen, each recording is represented as a 3-dimensional array with dimensions being acoustic frequency, modulation frequency and time. For a given time context, which for this work is 100 ms, a two dimensional array represents the energy distribution across the different channels in both frequency domains. The evolution through time of a particular point with acoustic frequency j and modulation frequency i represents the variable $\xi_{(i,\, j)}$.

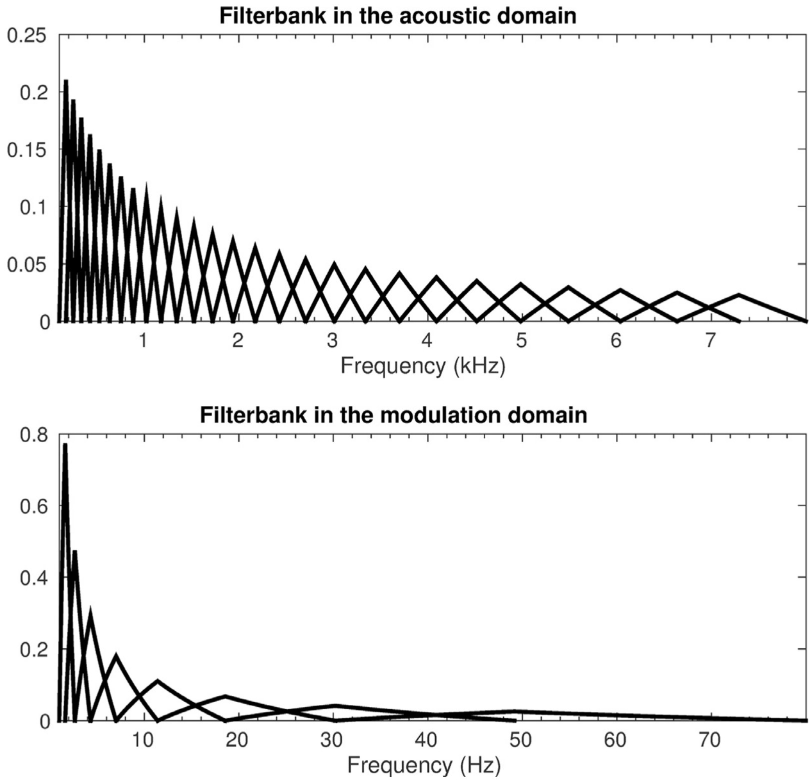


Fig. 2. Plots of frequency response of the 27- (top) and 8-channel (bottom) filterbanks used in the experiments herein.

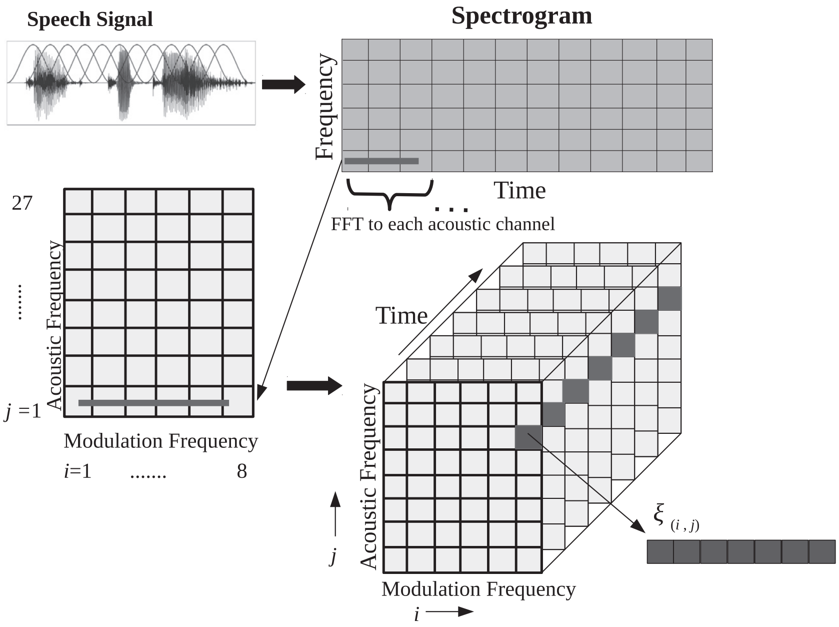


Fig. 3. Decomposition of a speech recording in terms of acoustic and modulation frequency components in a short time basis.

Each time context (a matrix with 216 elements) can be collapsed into a vector and used as standard features. However, given the high dimensionality of the resulting space and correlation among different dimensions, each feature vector is projected to a lower dimensional space using principal component analysis (PCA) with 40 components retaining 98.7% of cumulative variance, which according to our experiments showed to be an optimum value. These 40 components are then used for i-vector calculation. In order to validate the discriminative capabilities of this feature representation, we carried out an experiment by comparing with the standard MFCC feature vectors using different configurations of the SV system. Results are presented in Table 4.

According to these results, AAMF not only performs better in the matched condition for all cases but also helps to reduce error rates in the mismatched condition. These results consider that there is no information about whispered speech in any stage, showing that AAMF features have more discriminative power and can generalize better than standard MFCC. Next, we propose to use the modulation spectrum representation to perform an additional analysis aiming at a better representation for the two vocal efforts taking into account information from whispered speech recordings in the training set. The analysis is described below.

4.1.1. Feature selection using mutual information (MI)

Having the modulation spectrum representation, we applied the following analysis using MI. First, recordings from 14 speakers, seven female and seven male, from the wTIMIT database were used. Each speaker uttered

Table 4
EER (%) comparison between MFCC and AAMF using different values for the number of Gaussian components in the UBM and T matrix dimension.

Feature set	UBM	Normal			Whispered		
		T matrix dimension					
		200	300	400	200	300	400
MFCC	**128**	3.23	3.44	3.38	30.00	29.17	28.56
	256	3.05	2.92	2.81	29.43	28.52	27.31
AAMF	**128**	1.25	1.07	1.25	22.85	25.71	23.33
	256	0.94	0.94	1.04	24.17	26.85	25.00

approximately 450 different sentences, each of them in normal-voiced and then in whispered mode, in total 6298 pairs of utterances were included into the analysis. Since phonemes uttered by the same speaker have different duration in time for whispered and normal-voiced speech, we need to ensure that training utterances are phonetically aligned such that during the MI analysis all sentences have the same duration and the analysis can be performed between two equivalent frames. To guarantee this, we used a similar approach as the one in Hanilci et al. (2013), where alignment is achieved using dynamic time warping (DTW).

To compute the MI information, it is assumed that given two random variables X and Y with probability mass functions $p(x)$ and $p(y)$ respectively, and joint distribution $p(x, y)$, then the mutual information between X and Y is defined as:

$$I(X,Y) = \sum_{x,y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} \tag{3}$$

To derive $p(x)$, $p(y)$ and $p(x, y)$ the variables X and Y were partitioned in N uniform intervals and then the observed values for x and y were discretized. Having the number of data pairs, the number of intervals and the discretized values, then the probability mass functions are represented by a histogram (Moddemeijer, 1989).

The alignment algorithm, on the other hand, needs to be adapted in order to be useful for whispered speech. Initially, we compared three feature representations to compute the distance matrix between the two recordings to be aligned, then we selected the two best alignments and compare the final system performance. For the sake of completeness, here we describe the three approaches and how the best approach was selected. The first feature representation is the standard MFCC previously described in Section 3.1. The second approach is based on results presented in previous studies (Fan and Hansen, 2009; Sarria-Paja and Falk, 2015), where it has been shown that on average comparing the spectral envelope of both speaking styles, the frequency band where there are less differences is approximately between 1.2−4 kHz. Then for our purposes, by using a linear spaced filterbank within this acoustic band, 12 linear frequency cepstral coefficients were used to compute the distance matrix. A linear spaced filterbank is preferred for this purpose in order to not emphasize any particular frequency band. Finally, taking into account that the first two approaches can modify the dynamics of the speech recording and it can be reflected in the modulation domain, we decided also to do the alignment using the previously described AAMF. Fig. 4 compares spectrograms before and after warping for the first two approaches and illustrate how the choice in feature representation affects considerably the final result. Fig. 5 on the other hand, compares the three alignment paths; as can be seen, the second and third approaches (Fig. 5(b) and (c), respectively) can be considered as optimal because the lowest-cost paths are always close to the diagonal, which allows that the replicated frames are evenly distributed along the whole recording and not in a single area, as is the case with the MFCC.

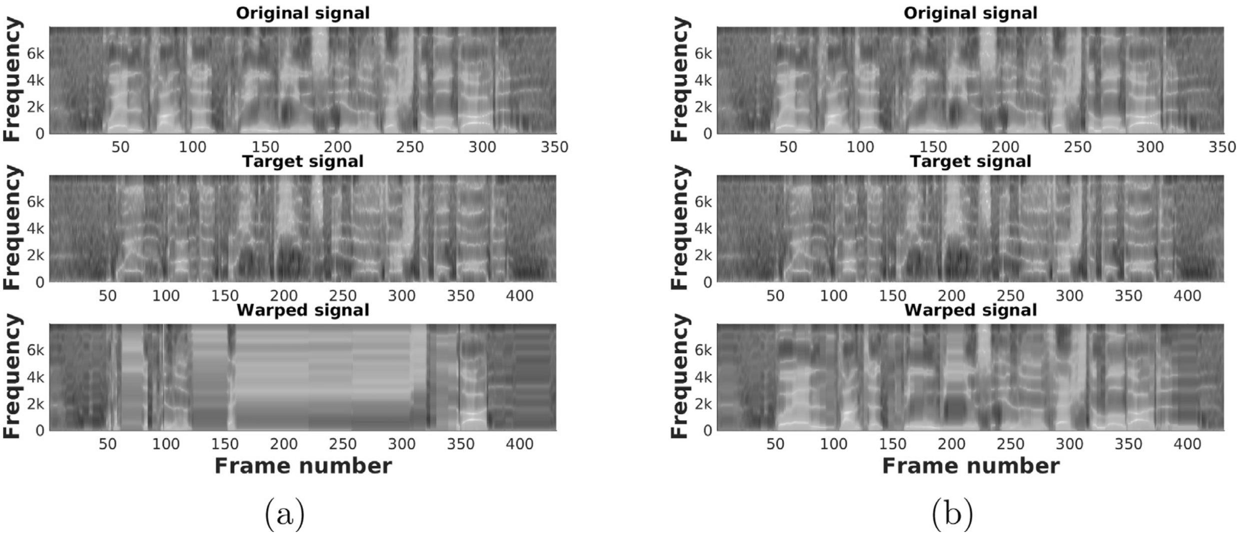
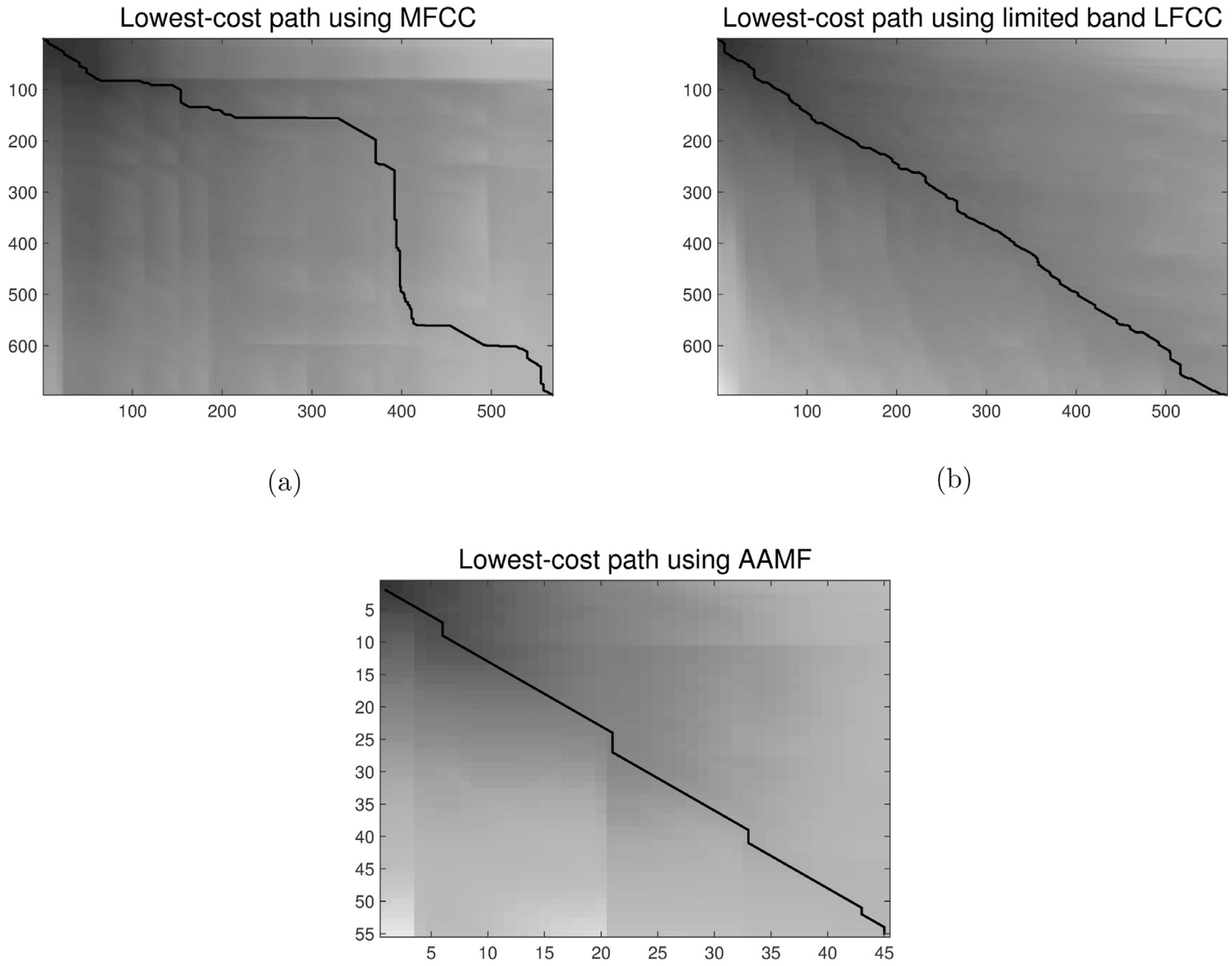


Fig. 4. Plots comparing two alignment strategies (a) Using full band MFCC and (b) using limited band LFCC.



(a)

(b)

(c)

Fig. 5. Plots comparing the lowest cost path computed with three feature representations. (a) Using standard MFCC, (b) using limited band LFCC (1.2−4 kHz), and (c) using AAMF.

Finally, having the time series aligned, MI values were computed per variable per speaker; this results in 216 MI values. Each value is normalized using the sum of entropies as $\widehat{mi} = 2 \cdot mi/(H_1 + H_2)$, where $\widehat{mi}$ is the normalized MI value, H_1 and H_2 are the entropy values of the two variables being compared. Next, having all MI values for a given speaker, they are re-scaled to the range [0−1] by using the transformation $\widehat{x} = \left(\frac{x - min_x}{max_x - min_x}\right)$, where x is the original value of a given variable to be re-scaled and $\widehat{x}$ the scaled value. Finally, all MI values were averaged over the 14 speakers. Results from this analysis allow to identify the variables with high degree of shared information between normal-voiced and whispered speech by thresholding, i.e., variables with MI values higher than such a threshold were kept while the others disregarded. For the experiments herein, the threshold was set to 0.4, which resulted in the selection of 141 and 157 variables depending on which alignment approach is used, i.e., LFCC or AAMF, respectively.

Fig. 6(a) and (b), summarizes the process to compute the resulting modulation spectrum based features having the MI analysis results. Fig. 7, on the other hand, depicts the selected acoustic and modulation bands after applying thresholding. As can be seen both approaches eliminate the lower acoustic band, the approach based on LFCC, on the other hand, also eliminates more information in the band 2−5 kHz. Finally, principal components analysis (PCA) was used to reduce the high-dimensional feature set to 40 dimensions, accounting for 99.3% and 99.1% of

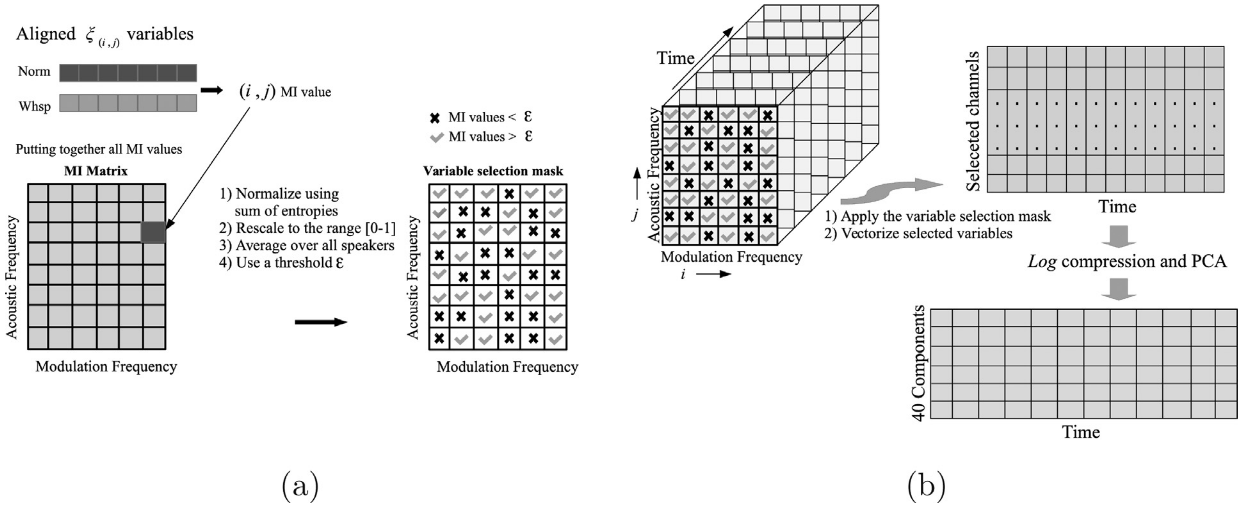


Fig. 6. Process to compute modulation spectrum based features (a) Identification of relevant variables (acoustic and modulation channels) using MI and (b) feature selection and decorrelation using PCA.

accumulated variance when using LFCC and AAMF for alignment, respectively. Even though there were no big differences in terms of error rate when testing with normal speech, the alignment based on AAMF showed to have better performance when testing with whispered speech. As such, results reported henceforth will be based on such an approach. After modifying the AAMF as the result of disregarding the channels with low mutual information, we call the resulting feature set as AAMF(FS), which are then used for i-vector calculation.

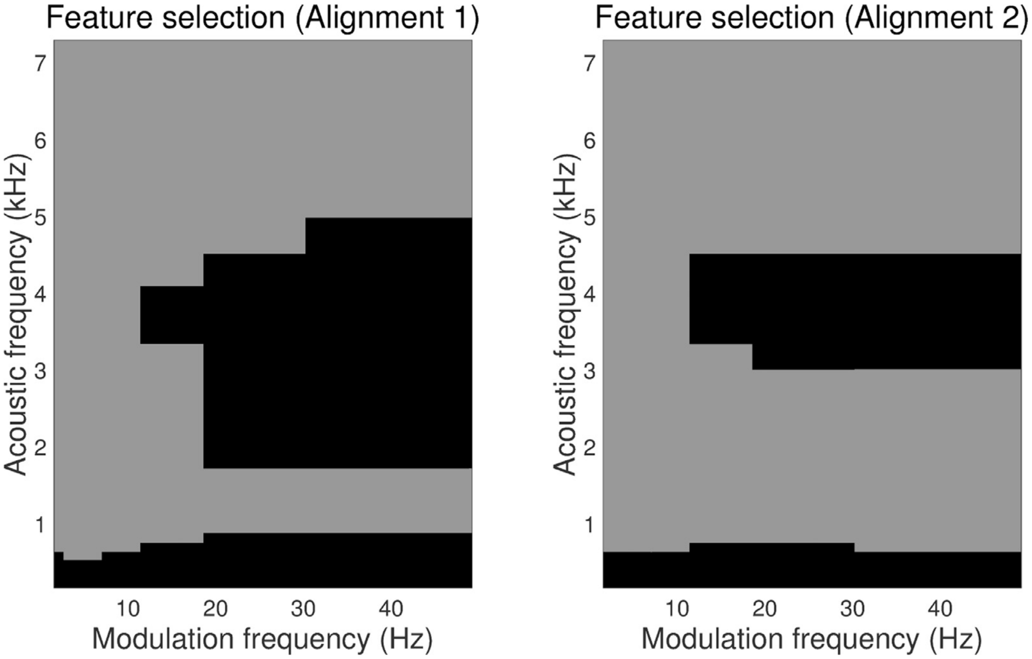


Fig. 7. Acoustic and modulation bands selected, these bands contain high degree of information that is common for both, normal-voiced and whispered speech. Alignment 1: LFCC. Alignment 2: AAMF. Grey areas correspond to selected channels, while the black ones to the disregarded channels.

4.2. Residual features and limited band MFCC

Based on the source-filter model for the process of human speech generation, it is possible to split the speech signal in two components: an excitation signal and a transfer function which models the vocal tract configuration (O'Shaughnessy, 2000). The excitation can be visualized as the combination of two different signal generators; one for voiced-speech and another for voiceless (noise-like) speech. The excitation signal is also known as residual. In the past, features extracted from the residual have been shown to contain important speaker-dependent information useful for speaker recognition tasks (Drugman and Dutoit, 2010; Chetouani et al., 2009). This is relevant for whispered speech because by removing the influence of the vocal tract, then differences related to the spectral envelope are no longer a nuisance factor affecting SV performance. The process to compute the residual or excitation signal is as follows: Given the speech signal, using linear predictive analysis, it is possible to rebuild the vocal tract transfer function by estimating the parameters of a low-order all-pole filter. By definition, linear prediction analysis uses the redundancy in the speech signal to predict the current sample as a linear combination of past p samples. The residual is the prediction error obtained as the difference between the predicted speech sample and the actual sample (O'Shaughnessy, 2000). Since the excitation signal is spectrally flat, uncorrelated white noise, the transfer function of this all-pole model represents the spectral envelope of the speech signal. Having the transfer function, inverse filtering is used to recover the excitation or residual signal (O'Shaughnessy, 2000).

It is widely known that the residual signal of normal speech contains quasi-period pulses corresponding to glottal closure/opening instances during vocal fold vibration. Such pulses are not present in whispered speech, thus exemplify a major difference between the two vocal efforts. As shown in the experiments herein, however, despite their lower SV performance for both normal and whispered speech, the information they convey is complementary to existing features. More specifically, unvoiced speech segments, which are also present in normally-phonated speech, results in a noise-like residual signal (O'Shaughnessy, 2000). Unvoiced sounds have been shown to remain unaffected during whispering (Jovicic and Saric, 2008) and to also convey important speaker-dependent information for whispered speech speaker recognition (Fan and Hansen, 2011; Xu and Zhao, 2012). As such, it is expected that the residual signal will convey useful complementary information for the task at hand.For our experiments, the standard MFCC were computed over the residual signal in a similar setting as the one previously described in Section 3.1. These features are termed RMFCC and are used for i-vector computation.

In addition to this, in a previous work it was shown that by using the sub-band from 1.2 kHz to 4 kHz to compute the different feature sets (Sarria-Paja and Falk, 2015) it was possible to improve performance in the mismatch condition, but at the cost of reduced performance in the matched scenario. In a different study, it was shown that for text-dependent speaker identification, higher frequency channels were more relevant for speaker recognition than those located at lower frequencies (Besacier and Bonastre, 1997). It was reported that the lowest identification rates were associated to channels containing information of first and second formant, and that there was a high negative impact in performance when removing channels containing information from the frequency band between 5 kHz and 8 kHz (Besacier and Bonastre, 1997). By using these insights, we propose to compute the standard MFCC features using the sub-band from 1.2 kHz to 8 kHz. By doing this, the sub-band that comprises mostly information from the first formant (F1) is removed, which for whispered speech the shift in F1 can be more than 50% relative to normal speech. Hence, most of the speaker specific information relevant for speaker recognition tasks is preserved and the performance in normal speech should not be affected. These features are termed LMFCC and are used for i-vector computation.

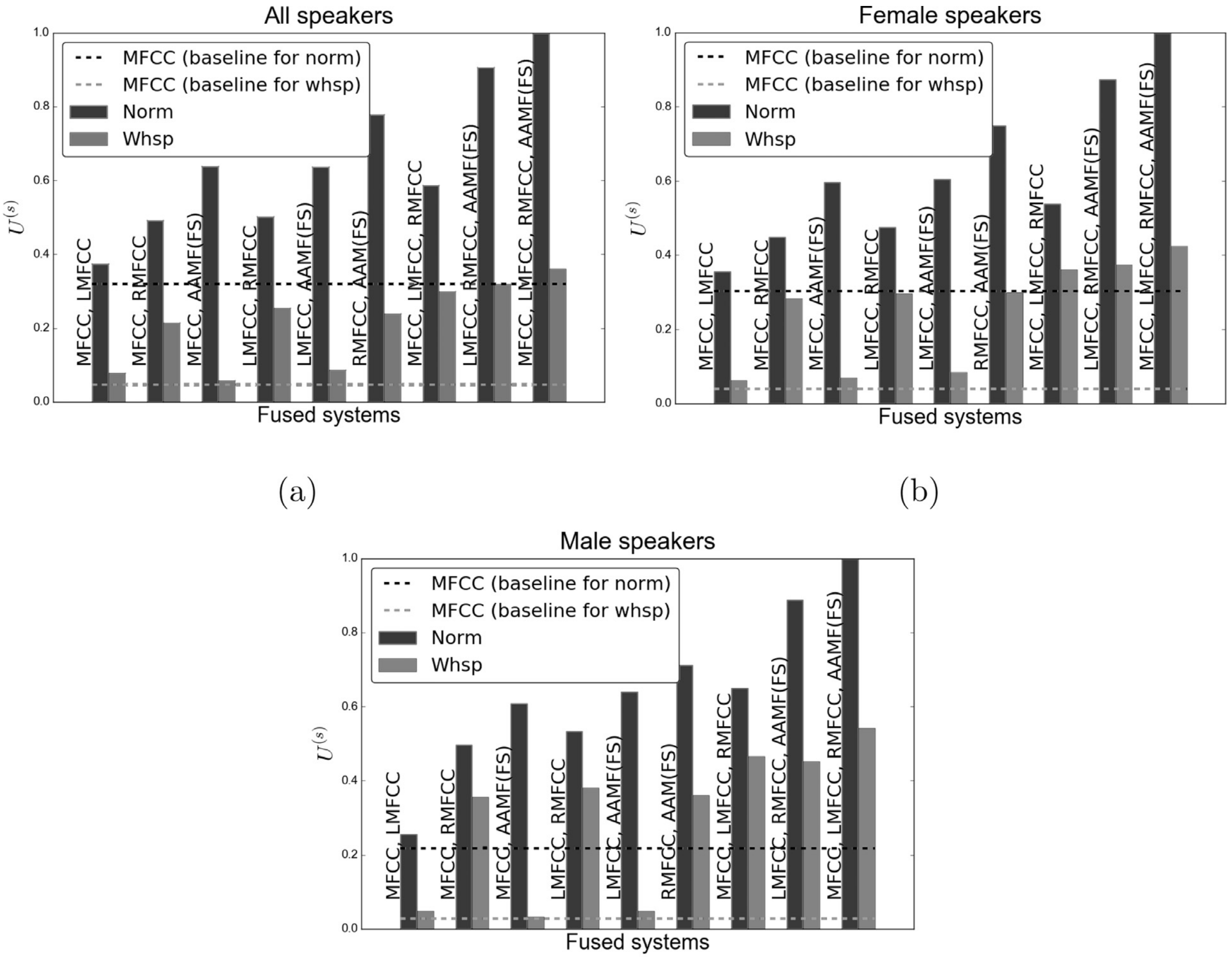
4.3. Score-domain analysis

In order to perform an analysis and explore what are the contributions of each feature set, we perform an analysis using as features the output scores of the systems trained on the feature sets described in the sections above. In previous work, separability using statistical analysis has been studied in the feature space to identify relevance of acoustic subbands for speaker recognition tasks (Lei and Lopez-Gonzalo, 2009; Gallardo et al., 2014). However, A comparison in the score domain is more feasible and easier to interpret than a comparison in the feature space, mostly because each feature set extracts information from recordings in different ways and it is hard to decide when any particular measure is actually describing the contribution of a given feature representation and can be considered better against another feature set. For the analysis we use the Lawley-Hotelling statistic (Rencher, 2003), a commonly used measure in MANOVA (multivariate analysis of variance) analysis when we wish to compare the mean vectors of k

groups of samples for significant differences. In this case, we want to test whether or not the mean of impostors scores equals the mean of target speakers. The hypotheses are, therefore: $\mathcal{H}_0 : \mu_i = \mu_t$, vs. $\mathcal{H}_1 : \mu_i \neq \mu_t$, where μ_i and μ_t stand for impostors and target speakers mean, respectively. The Lawley-Hotelling statistic is defined as (Rencher, 2003):

$$U^{(s)} = \mathrm{tr}(E^{-1}H) = \sum_{i=1}^{s} \lambda_i, \tag{4}$$

where E and H are the “between” and “within” matrices, respectively, λ_i are the eigenvalues of $E^{-1}H$, and $s = \min\{p, k\}$, being p the number of variables and k the number of groups or classes. The main advantage of this test is that the multivariate information in E and H about separation of mean vectors is summarized into a single scale, on which we can determine if the separation of mean vectors is significant. We reject $\mathcal{H}_0$ for large values of $U^{(s)}$. This test is carried out by combining different systems in an incremental way, and separating the scores from normal and whispered speech. This allows us to better understand the effect that the addition of a particular system has in the separability of impostors and target speakers scores for each speaking style. The analysis is also carried out per gender and the results are summarized in Fig. 8.



(a)

(b)

(c)

Fig. 8. Lawley-Hotelling statistic analysis using combination of different systems to explore contributions of individual feature sets. (a) Gender independent, (b) female speakers and (c) male speakers.

In Fig. 8(a), bars represent the $U^{(s)}$ measure for the combined systems, normalized by the max value because we are interested on the relative improvements from the baseline and not in the absolute value. In the plots, normal scores are in dark grey, whispered speech scores, in turn, are in clear grey. Dashed lines represent the same measure for the baseline system, black for normal and grey for whispered speech. As can be seen, for all cases the addition of a new system should increase the separability for normal speech scores, but the same effect is not observed for whispered speech. When combining with the baseline system, the feature set that seems to be more beneficial for whispered speech is RMFCC; AAMF(FS) on the other hand, seems to add more separability to normal speech scores. When we combine only the proposed feature sets, i.e., LMFCC, RMFCC and AAMF(FS) benefits occur for both normal and whispered speech, thus suggesting that limited band and residual cepstral coefficients mainly increase separation for whispered speech recordings and AAMF(FS) helps mostly in the separation of normal speech scores. Finally, combining all systems, i.e., baseline and the systems trained with the proposed feature sets, we expect to have the maximal separation not only for normal speech scores but also for whispered speech. These results however, give an idea of how each system contributes to the separation of impostors and target speakers scores, but it is necessary to corroborate these predictions by training a fusion system to properly evaluate the overall performance of the proposed strategies. Lastly, in order to explore possible gender dependencies, i.e., that some systems may offer different advantages when testing only with speech recordings from male or female speakers. Fig. 8(b) and (c) shows the results for female and male speakers, respectively. In this case, we can observe that for both genders the tendency of separability values follows a similar behavior as the one observed by pooling speakers together in a gender independent analysis (Fig. 8(a)). It can be observed that gains over the baseline for whispered speech in male speakers are higher. This is a positive observation as it has been reported that male speech is highly affected in whispered mode (Sharifzadeh et al., 2012), thus most recent whispered speech speaker verification studies are reported based on female speakers (Fan and Hansen, 2010; 2011; 2013). These observations show that proposed features improve whispered speech separability for target and impostor speakers for male speech recordings. Given limitations in our available whispered speech datasets (e.g., gender imbalance with roughly twice as many male data points as female), however, gender-specific models are not explored herein and are left for a future study. As such, results in the following section refer only to gender independent systems.

4.4. System fusion

As shown above, the proposed features can extract important invariant information and are expected to help reduce error rates in mismatch condition. Additionally, each feature set is expected to extract information that is complementary to the other sets according to the Lawley-Hotelling statistic analysis performed in the previous section. With this in mind, it is necessary to implement a strategy to take advantage of this complementarity. For this purpose, system fusion schemes have shown to be effective at improving performance in many applications by combining the strengths of different systems (Khoury et al., 2014; Doddington et al., 2000). To train the fusion strategy it is necessary to have an independent development set, in which the enrollment and testing recordings are from a separate set of speakers, which are not included in the original training or testing sets. Herein, we selected 10 speakers from the wTIMIT database and 68 from the TIMIT database in order to create the development evaluation list. For the new evaluation list, in order to have approximately the same amount of target and impostors scores from each speaking style, two recordings per speaker were used of normal speech and 14 recordings per speaker of whispered speech (see Table 2). Next, each SV system is evaluated to obtain the respective scores for each trial in the new development evaluation list. Finally, these scores are labeled as target/non-target and are used as features to train a linear binary classifier using the scheme depicted in Fig. 9. The Bosaris toolkit was used for this purpose; details can be found in Brummer and de Villiers (2011).

Results are presented in Table 5 where each individual system has been labeled "**S*i***" for brevity. First, we want to analyse the effect of adding whispered speech during the training stage (**S1**) to the baseline system and compare it with the individual systems using the proposed feature sets. When comparing the system **S1** with results reported in Table 3 (**S0**), it can be seen that there is a slight increment of error rate for normal speech from 2.81% to 3.13%, whilst for whispered speech we observe an error rate reduction from 27.31% to 20.83%. This represents a 23% relative improvement just by adding whispered speech recordings to the training set, which is the main difference between **S0** and **S1**. Now, by comparing system **S1** with the three proposed systems, we can see for example that **S2** maintains the performance in normal speech, but in addition to this, a relative reduction of 38% is achieved for

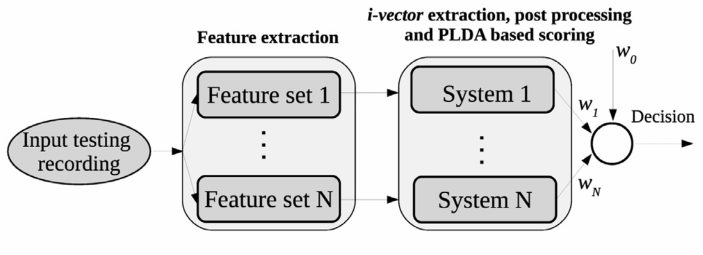


Fig. 9. Schematic representation of the fusion system.

Table 5
EER (%) comparison for different feature sets and the fusion systems under two testing conditions. LMFCC limited band mel cepstral coefficients, RMFCC − residual mel cepstral coefficients and AAMF(FS) − Auditory-inspired amplitude modulation features (using feature selection). For these results $C = 256$ and $T = 400$.

SV system	Normal	Whispered
Individual systems		
MFCC (S0)	2.81	27.31
MFCC (S1)	3.13	20.83
LMFCC (S2)	3.13	16.67
RMFCC (S3)	6.70	22.95
AAMF(FS) (S4)	0.94	17.80
Fused systems		
S5: S2 + S3	2.50	11.67
S6: S3 + S4	1.44	12.64
S7: S2 + S3 + S4	0.94	10.04
S8: S1 + S2 + S3 + S4	1.25	12.43

whispered speech compared to **S0**. The system based on the RMFCC (**S3**) feature set has a more discrete performance, mostly because in this case all information related to spectral envelope has been removed, but according to the achieved performance, these results show that there is useful information related to speaker identity in the residuals. In the mismatch condition, the improvement relative to **S0** is of 15%. Finally, the system based on AAMF(FS) (**S4**) also present substantial error reductions relative to **S0**, not only in mismatch condition but also in the matched condition; improvements are 66% and 34% for normal and whispered speech are observed, respectively.

Table 5 also reports error rates comparing the fusion of different systems, which resulted in improved performance relative to the baseline. In the Table the notation ***Si* + *Sj*** means that the fusion is done by using the outputs of systems ***Si*** and ***Sj***. According to the Lawley-Hotelling statistic shown in Fig. 8, fusion of systems based on LMFCC and RMFCC (**S5**) should benefit both, match and mismatch condition. This is corroborated by results in Table 5; when comparing with **S1** there is a relative improvement of 20% for normal speech and 43% for whispered speech. If we compare these results with the baseline system **S0** (Table 3), the relative improvement in the mismatch condition is of 57%. Moreover, from Fig. 4, the fusion of RMFCC and AAMF(FS) (**S6**) based systems should also benefit the task at hand. Indeed, by combining **S3** and **S4**, the relative improvements are 53% and 40% for normal and whispered speech, respectively, when comparing with **S1**, and 48% and 54% relative to the baseline system **S0**.

Lastly, fusion of all proposed feature sets, with (**S8**) and without (**S7**) the baseline MFCC; achieved the lowest EER. Unlike the theoretical separation analysis reported in Fig. 4, the results in Table 5 actually suggest that adding MFCCs reduce overall performance. This is likely due to the fact that the limited amount of data available to train the linear function for score fusion did not model the boundary found with the Lawley-Hotelling analysis. Notwithstading, the fusion analysis results from Table 5 follow the general tendencies observed in Fig. 8 and show the proposed features extracting complementary information from speech recordings, thus helping not only to reduce error rates when testing with whispered speech, but to also improve system performance for normal speech. Overall, fusion of the systems using the proposed feature sets (**S2 + S3 + S4**) achieves relative improvements of 66% and 63% for normal and whispered speech, respectively, when comparing to **S0**, and 69% and 51% when comparing to **S1**, respectively. Lastly, is important to emphasize that despite these substantial gains, there is still a gap in performance between normal and whispered speech which is not desirable for practical applications. Until now, no information from whispered speech recordings of target speakers has been included. As reported in Sarria-Paja and Falk (2015), such an approach is needed.

As such, in the following experiment, whispered speech data from target speakers was gradually added during the enrollment stage. Results are presented in Table 6. The column labeled *Wshp. speech utt. in enrollment* represents the number of utterances, each in average 4.5 s, from target speakers that were added during enrollment. As can be seen, in the *Fusion* **S7** system, only by adding one utterance the performance for whispered speech is already inline with the performance of the baseline system when using eight utterances. In addition to this, for each case there is a relative difference of about 50% between the baseline system and the proposed system, hence less data is needed to improve performance when testing with whispered speech, which clearly represents an advantage. The other aspect to highlight is the degradation in performance for normal speech as more utterances of whispered speech are present

Table 6
EER (%) comparison for different feature sets and the fusion systems under two *Training/Testing* conditions with varying amounts of whispered speech during enrollment. For these results $C = 256$ and $T = 400$.

Number of Whsp. utt. in enrollment	Baseline S1		Fusion S7		Fusion S8	
	Norm.	Whsp.	Norm.	Whsp.	Norm.	Whsp.
0	3.13	20.83	0.94	10.04	1.25	12.43
1	3.03	17.14	0.97	8.33	1.25	10.40
2	3.44	14.17	1.02	7.50	1.25	9.07
3	3.75	13.07	1.16	6.51	1.56	7.50
4	4.23	11.56	1.14	5.72	1.64	6.59
5	4.69	10.80	1.25	4.61	1.56	5.45
6	4.87	9.58	1.56	4.47	1.56	4.86
7	5.31	8.92	1.56	3.63	1.80	3.71
8	5.31	8.25	1.61	3.33	1.56	3.33

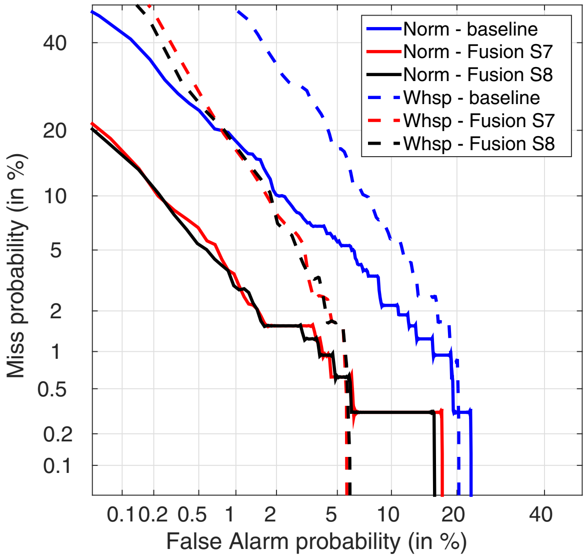


Fig. 10. DET curves comparing the baseline system and the fusion system 2 when whispered speech is added during training and enrollment, these DET curves correspond to the last row of Table 6.

during enrollment. This is a problem that affects both the baseline and the proposed system, but is more noticeable in the baseline system. Overall, the proposed system keeps the error rate below 2% for normal speech, which is actually better than the performance achieved by **S0**. An additional important aspect is that the final error rate achieved for *whispered* speech by *Fusion* **S7** is inline with the attained performance for *normal* speech by the **S1** system with no whispered enrollment utterances.

Lastly, by comparing the two fusion schemes, we can see that there are no significant differences between the two. This can be corroborated by the DET curves shown in Fig. 10. Notwithstanding, *Fusion* **S8** performance is somewhat lower than that of *Fusion* **S7** for both speaking styles. Even though the gap is not considerable, this shows that it is not necessary to include MFCC features into the fusion scheme, and the proposed feature sets are capable of handling both speaking styles.

5. Conclusions

This paper has addressed the issue of speaker verification (SV) based on whispered speech. Two different approaches were proposed in order to reduce error rates for SV with whispered speech while maintaining performance with normal speech. First, three innovative features were proposed: AAMF, RMFCC and LMFCC, each taking into account complementary characteristics of whispered and normal-voiced speech signals. Second, a score fusion scheme based on systems trained on the three feature sets showed improvements in SV performance of 66% and 63% over a state-of-the-art baseline for normal and whispered speech, respectively.

Finally, if a SV system is aimed to handle two speaking styles it is evident the need to collect data from target speakers in both speaking styles. This is an easy task for normal-voiced speech, either for background or target speakers, as normal speech is available in daily broadcasts and standard phone conversations. For whispered speech, however, it would be a more difficult task and the system has to deal with limited data from both background and target speakers. With the proposed fusion scheme, it is shown that only 4.5 s (aprox.) of whispered enrollment data is needed to achieve the same performance as the baseline system, with 22.5 s (aprox.) of whispered enrollment data. Hence, the proposed features are well posed to handle vocal effort variations and low resource speaker verification tasks.

Acknowledgments

The authors acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Administrative Department of Science, Technology and Innovation of Colombia (COLCIENCIAS).

References

Anjos, A., Shafey, L.E., Wallace, R., Günther, M., McCool, C., Marcel, S., 2012. Bob: a free signal processing and machine learning toolbox for researchers. In: Proceedings of the 20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan. ACM Press, pp. 1449–1452.

Besacier, L., Bonastre, J.-F., 1997. Subband approach for automatic speaker recognition: optimal division of the frequency domain. In: Proceedings of the First International Conference on Audio- and Video-based Biometric Person Authentication, Lecture Notes in Computer Science, pp. 195–202.

Brummer, N., de Villiers, E., 2011. The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing. Technical report. CAGNITIO Research, South Africa.

Chenghui, G., Heming, Z., Wei, Z., Yanlei, W., Min, W., 2009. A preliminary study on emotions of chinese whispered speech. In: Proceedings of the International Forum on Computer Science-Technology and Applications, Vol. 2, pp. 429–433.

Chetouani, M., Faundez-Zanuy, M., Gas, B., Zarader, J., 2009. Investigation on lp-residual representations for speaker identification. Pattern Recognit. 42 (3), 487–494.

Clerico, A., Gupta, R., Falk, T., 2015. Mutual information between inter-hemispheric eeg spectro-temporal patterns: a new feature for automated affect recognition. In: Procceedings of IEEE/EMBS NER. IEEE/EMBS, pp. 2106–2109.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Audio Speech Lang. Process. 19 (4), 788–798.

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

Drugman, T., Dutoit, T., 2010. On the potential of glottal signatures for speaker recognition. In: Proceedings of INTERSPEECH, pp. 2106–2109.

Estevez, P., Tesmer, M., Perez, C., Zurada, J., 2009. Normalized mutual information feature selection. IEEE Trans. Neural Netw. 20 (2), 189–201.

Falk, T., Chan, W.-Y., 2010. Modulation spectral features for robust far-field speaker identification. IEEE Trans. Audio Speech Lang. Process. 18 (1), 90–100.

Falk, T., Chan, W.-Y., Shein, F., 2012. Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. Speech Commun. 54 (5), 622–631.

Fan, X., Hansen, J., 2008. Speaker identification for whispered speech based on frequency warping and score competition. In: Proceedings of INTERSPEECH, pp. 1313–1316.

Fan, X., Hansen, J., 2013. Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams. Speech Commun. 55 (1), 119–134.

Fan, X., Hansen, J.H.L., 2009. Speaker identification with whispered speech based on modified LFCC parameters and feature mapping. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 4553–4556.

Fan, X., Hansen, J.H.L., 2010. Acoustic analysis for speaker identification of whispered speech. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 5046–5049.

Fan, X., Hansen, J.H.L., 2011. Speaker identification within whispered speech audio streams. IEEE Trans. Audio Speech Lang. Process. 19 (5), 1408–1421.

Gallardo, L., Wagner, M., Möller, S., 2014. Advantages of wideband over narrowband channels for speaker verification employing mfccs and lfccs. In: Proceedings of INTERSPEECH.

Garofolo, J. S., Consortium, L. D., et al., 1993. Timit: Acoustic-phonetic Continuous Speech Corpus.

Grimaldi, M., Cummins, F., 2008. Speaker identification using instantaneous frequencies. IEEE Trans. Audio Speech Lang. Process. 16 (6), 1097–1111.

Hanilci, C., Kinnunen, T., Saeidi, R., Pohjalainen, J., Alku, P., Ertas, F., 2013. Speaker identification from shouted speech: analysis and compensation. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8027–8031.

Higashikawa, M., Nakai, K., Sakakura, A., Takahashi, H., 1996. Perceived pitch of whispered vowels-relationship with formant frequencies: a preliminary study. J. Voice 10 (2), 155–158.

Ito, T., Takeda, K., Itakura, F., 2005. Analysis and recognition of whispered speech. Speech Commun. 45 (2), 139–152.

Jin, Q., Jou, S.-C., Schultz, T., 2007. Whispering speaker identification. In: Proceedings of IEEE International Conference on Multimedia and Expo, pp. 1027–1030.

Jovicic, S., Saric, Z., 2008. Acoustic analysis of consonants in whispered speech. J. Voice 22 (3), 263–274.

Khoury, E., El Shafey, L., Marcel, S., 2014. Spear: an open source toolbox for speaker recognition based on Bob. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 1655–1659.

Kinnunen, T., Lee, K., Li, H., 2008. Dimension reduction of the modulation spectrogram for speaker verification. In: Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2008), pp. 30–34.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: from features to supervectors. Speech Commun. 52 (1), 12–40.

Lass, N., Waters, L., Tyson, V., 1976. Speaker sexidentification from voiced, whispered, and filtered isolated vowels. J. Acoust. Soci. Am. 59 (3), 975–978.

Lei, H., Lopez-Gonzalo, E., 2009. Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition. In: Proceedings of INTERSPEECH.

Lim, B.P., 2011. Computational Differences Between Whispered and Non-whispered Speech. Ph.d. thesis. University of Illinois.

Markets, R.a., 2015. Global Mobile Biometrics Market 2015−2019. Technical report. Research and Markets.

Matejka, P., Glembek, O., Novotny, O., Plchot, O., Grezl, F., Burget, L., Cernocky, J., 2016. Analysis of DNN approaches to speaker identification. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 5100–5104.

Ming, J., Hazen, T., Glass, J., Reynolds, D., 2007. Robust speaker recognition in noisy conditions. IEEE Audio Speech Lang. Process. 15 (5), 1711–1723.

Moddemeijer, R., 1989. On estimation of entropy and mutual information of continuous distributions. Signal Process. 16 (3), 233–248.

O'Neil King, R., 2014. Speech and Voice Recognition White Paper. Biometrics Research Group, Inc Technical report.

O'Shaughnessy, D., 2000. Speech Communications − Human and Machine. 2nd ed. IEEE.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27 (8), 1226–1238.

Prince, S., Elder, J., 2007. Probabilistic linear discriminant analysis for inferences about identity. In: Proceedings of IEEE 11th International Conference on Computer Vision, ICCV, pp. 1–8.

Rao, K., Sarkar, S., 2014. Robust speaker recognition in noisy environments. SpringerBriefs in Speech Technology. Springer International Publishing.

Rencher, A., 2003. Methods of Multivariate Analysis. John Wiley & Sons. vol. 492.

Sarria-Paja, M., Falk, T., 2015. Strategies to enhance whispered speech speaker verification: a comparative analysis. J. Can. Acoust. Assoc. 43 (4), 31–45.

Sharifzadeh, H., McLoughlin, I., Ahmadi, F., 2010. Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec. IEEE Trans. Biomed. Eng. 57 (10), 2448–2458.

Sharifzadeh, H., McLoughlin, I., Russell, M., 2012. A comprehensive vowel space for whispered speech. J. Voice 26 (2), 49–56.

Sizov, A., Lee, K., Kinnunen, T., 2014. Unifying probabilistic linear discriminant analysis variants in biometric authentication. In: Proceedings of Structural, Syntactic, and Statistical Pattern Recognition, S+SSPR, pp. 464–475.

Tartter, V., 1991. Identifiability of vowels and speakers from whispered syllables. Percept. Psychophys. 49 (4), 365–372.

Thomas, I., 1969. Perceived pitch of whispered vowels. J. Acoust. Soc. Am. 46 (2B), 468–470.

Tran, T., Mariooryad, S., Busso, C., 2013. Audiovisual corpus to analyze whisper speech. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8101–8105.

Tsunoda, K., Sekimoto, S., Baer, T., 2012. Brain activity in aphonia after a coughing episode: different brain activity in healthy whispering and pathological aphonic conditions. J. Voice 26 (5), 668.e11–668.e13.

Xiang, J., Poeppel, D., Simon, J., 2013. Physiological evidence for auditory modulation filterbanks: Cortical responses to concurrent modulations. J. Acoust. Soc. Am. 133 (1), EL7–El12.

Xu, J., Zhao, H., 2012. Speaker identification with whispered speech using unvoiced-consonant phonemes. In: Proceedings of International Conference on Image Analysis and Signal Processing, IASP, pp. 1–4.

Zelinka, P., Sigmund, M., Schimmel, J., 2012. Impact of vocal effort variability on automatic speech recognition. Speech Commun. 54 (6), 732–742.