Strategies to Enhance Whispered Speech Speaker Verification: A Comparative Analysis

Milton Sarria-Paja and Tiago H. Falk *
Institut National de la Recherche Scientifique, Centre EMT, University of Quebec, Montreal, QC

Abstract

Today, automated speech–enabled tools are increasingly being used in everyday environments. This mobility has created new challenges for developers, who are now faced with input speech of varying styles (e.g. whispered) and corrupted by different noise sources. In this paper, special emphasis is placed on whispered speech, an underexplored yet burgeoning area due to the rapid proliferation of smartphones around the world. More specifically, this paper explores the performance boundaries achievable with whispered speech for a speaker verification task, both in matched and mismatched *train/test* conditions. Several strategies are investigated to improve the performance in the mismatched scenario, as well as in situations involving ambient noise. Our results agree with previously reported studies in adjacent areas, that significant gains could be obtained by training speaker models with both naturally voiced and whispered speech data. Moreover, additional gains could be achieved with speaking style and gender dependent systems. Overall, speaker verification performance inline with that obtained with naturally-voiced speech could be attained for whispered speech once specific strategies were put in place. Particularly, feature fusion showed to be an important strategy for practical applications in both clean and noisy conditions.

Keywords: Whispered speech, gender detection, speaker verification, instantaneous frequency, vocal effort classification, modulation spectrum.

Résumé

De nos jours, les outils tirant profit de l'analyse automatique de la parole sont de plus en plus utilisés au quotidien. Cette mobilité engendre de nouveaux défis pour les développeurs, qui doivent composer avec différents types de parole (par exemple, des chuchotements) et de sources de bruit. Dans cet article, une attention spéciale est accordée à la parole chuchotée, qui malgré son importance particulière dans le contexte d'une augmentation fulgurante de l'utilisation de téléphones intelligents dans le monde, demeure un champ inexploré. Plus spécifiquement, cet article explore les niveaux de performance atteignables lorsque la parole chuchotée est utilisée pour la vérification de locuteurs, à la fois dans des conditions correspondant et non-correspondant d'entraînement et de test. Plusieurs stratégies sont explorées afin d'améliorer la performance dans le cas non-correspondant, de même que dans des situations impliquant un bruit ambiant. Nos résultats confirment ceux obtenus dans des domaines connexes : des gains de performance significatifs peuvent être obtenus en développant des modèles de locuteurs basés sur la parole voisée et chuchotée. De plus, des gains additionnels peuvent être obtenus en considérant des modèles spécifiques à un style de parole et au sexe. Globalement, un niveau de performance semblable à celui obtenu avec la parole voisée a été atteint lors d'une tâche de vérification de locuteurs basée sur la parole chuchotée. En particulier, la fusion au niveau des traits caractéristiques (≪ feature fusion≫) s'est avérée une stratégie importante pour le succès d'applications pratiques dans des conditions de parole propre et bruitée.

Mots clefs: Parole chuchotée, détection de genre, vérification du locuteur, fréquence instantanée, classement de l'effort vocal, spectre de modulation

1 Introduction

Human speech is a natural and flexible mode of communication that not only conveys a message, but also traits such as identity, age, gender, social and region of origin, emotional, and health states, to name a few [1]. Under controlled conditions, speech processing systems have become useful across a number of domains. As examples, a number of applications have emerged that allow people to use their voices to interact with their devices (e.g., Apple's Siri), login to secure services (e.g., Bell Canada's Voice Identification Service), or even unlock their mobile devices (e.g., Baidu-I^{2}R Research Centre's Speaker Verification Service). Many such applications have thrived due to the recent proliferation of mobile devices. Notwithstanding, while the ubiquity of smartphones has opened a pathway for new speech applications, user mobility has created several challenges that still need to be addressed, such as the robustness to ambient noise or varying vocal efforts (e.g., whispering). While robustness to noise has been addressed numerous times in the past (e.g. [2–4]), little attention has been given to varying vocal efforts.

Here, special emphasis is given to whispered speech as, with the burgeoning of mobile speech applications, users have become more cautious about protecting the content of their spoken words, (e.g., during mobile telephone banking) specially when providing their credit card number, bank account number, or other personal information. One limiting factor in the widespread development of whispered speech applications lie on the lack of large amounts of training data [5–7], as is the case with normally-voiced speech. Notwithstanding, the increasing interest in this speaking style has led to the development of a few publicly-available databases, such as the CHAINS corpus [8]. Such initiatives open doors for speaking-style dependent models to be used and accurate whispered speech applications to emerge.

Existing automatic speech and speaker recognition systems do not perform well under whispered speech conditions, particularly if normal speech was used during training (i.e., training/testing mismatch conditions) [6, 9–11]. Despite this drop in performance of automated systems, subjective studies have suggested that whispered speech still conveys a significant amount of speaker identity information and degree of understanding [12, 13]. As such, recent studies looking at speaker identification have shown that the best solution is to include small portions of whispered speech during training to adapt the speaker models [6, 14]. Alternately, other studies have explored the benefits of developing automated sys-

*. falk@emt.inrs.ca

tems with dedicated speaker models for different vocal efforts (e.g., [9, 14, 15]), thus taking into account the particular characteristics of each vocal effort.

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 [10, 16]. Therefore, classical methods designed for normal speech characterization tend to fail for whispered speech, as commonly used features (e.g. Mel frequency cepstral coefficients - MFCC) are sensitive to such changes [6, 11]. The aim of this paper is to explore the performance envelope achievable with whispered speech, particularly within the scope of a small scale speaker verification (SV) task. To this end, we explore the benefits of different existing preprocessing methods, frequency warping strategies, feature representations, and SV strategies. The main goal of this paper is to comprehensively investigate which system configurations result in the best performance for whispered and normally-voiced speech, both in clean and noisy conditions. Ultimately, it is hoped that the insights reported herein will help the development of large scale applications in more realistic scenarios, and for future development of practical systems that can be used in everyday settings.

The remainder of this paper is organized as follows. Section 2 provides the background on whispered speech, emphasizing the main differences with normal speech. Section 3 describes the speaker verification problem, the corpus employed for speaker verification, the feature extraction approaches, as well as the baseline settings and results. Section 4 discusses different approaches and strategies to reduce the error rate in whispered speech speaker verification. Section 5 discusses the robustness of the best feature representations and system design to different levels of babble noise. Section 6 presents further discussion and analysis of the main results and describes future research directions. Lastly, Section 7 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. For example, topics such as pitch perception and the correlation between perceived pitch and formant location have been studied, as well as the measurement of the formant shifts towards higher frequencies [17, 18]. Moreover, perceptual studies have suggested that whispered speech still conveys a significant amount of speaker identity and gender information [12, 13, 19, 20].

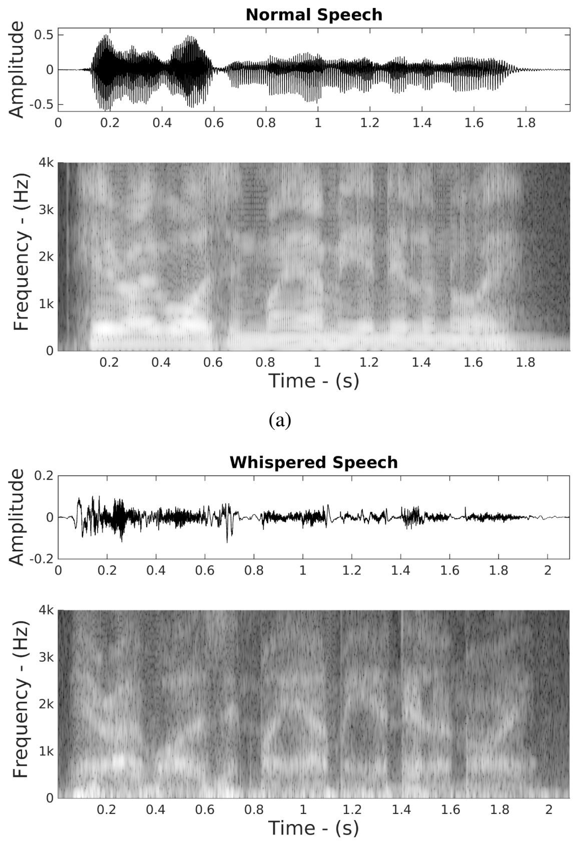
Using signal processing tools, acoustical studies have found that whispered speech has a lower and flatter power spectral density [10]. In [16], it was found that the duration of consonants in whispered speech is prolonged by about 10% relative to normally-voiced speech. In addition to the duration increase, the intensity of the whispered consonants is lower by about 12 dB. These significant changes have been documented only in voiced consonants. A recent study has also corroborated the perceptual findings regarding the formant shifts in whispered mode [21]. The above-mentioned insights have been used by the research community to tackle different challenges, such as reconstruction of normal speech from whispers [22–24], speech recognition [9, 10], and speaker identification [6–8, 14] with whispered speech.

To illustrate some of the significant differences between normal and whispered speech, their waveforms and spectrograms are depicted by Figure 1(a) and 1(b) respectively, for the utterance "*Here I was in Miami and Illinois*". From Figure 1(b), it can be observed that whispered speech is mostly turbulent noise modulated by the vocal tract with no clear structure. With normal speech (Figure1(a)), on the other hand, the glottal excitation is clear. Moreover, the time waveform for whispered speech is significantly lower in amplitude ; in this particular case about 15 dB lower. Figure 2(a) in turn, illustrates the average power spectrum for the same utterance, using 32 ms windows and a 12 order linear predictive model to estimate the spectral envelope. From Figure 2(a), it is evident that the differences lie mostly in the low frequencies. For normal speech, most of the energy is concentrated below 1 kHz, whereas for whispered speech it is concentrated below 500 Hz, with frequency shifts in the spectral peaks and valleys. Between 1 kHz and 4 kHz the two spectral envelopes follow a similar trend, where spectral peaks and valleys are located in approximately the same frequency values, however the differences in magnitude are not constant. Regarding frame energy distribution, the histogram in Figure 2(b) was computed using male and female speech and utterances of about 55 s from 36 speakers and shows that the concentration of high-energy frames is higher for normal speech, with 60% of the frames having energy between -10 dB and 10 dB. For whispered speech, on the other hand, 70% of the frames have energy between -35 dB and -10 dB. Combined, these findings show that significant differences exist between whispered and normal-voiced speech in terms of temporal, spectral and energy dynamics. As such, it is expected that any speech-based technology trained on normal speech will fail when tested on whispered speech. Clearly, strategies need to be devised to improve system performance. As mentioned previously, the focus of the present paper is to explore such strategies for a speaker verification task.

3 Baseline SV system characterization

3.1 Automatic speaker verification system

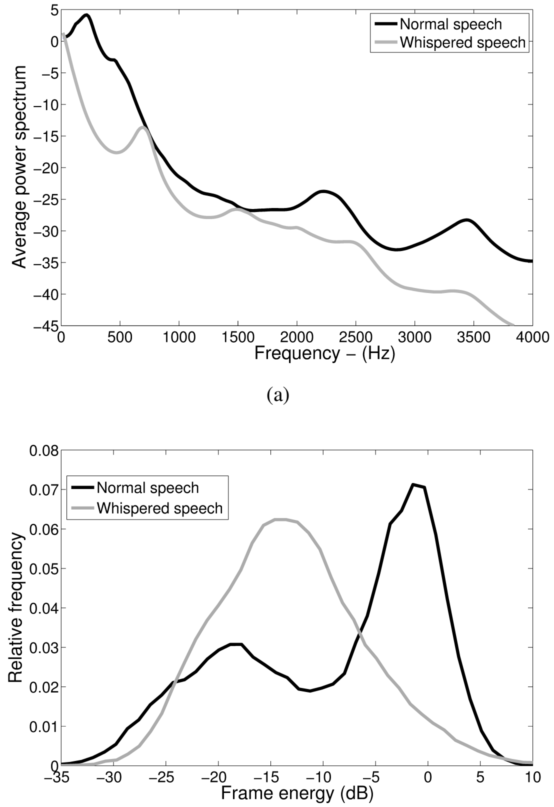
In automatic speaker recognition (SR) there are two classical tasks that can be performed : speaker identification (SI) and speaker verification (SV). Identification is the task of deciding, given a speech sample, who among a set of speakers said it. This is an N–Class problem (given N speakers), and the performance measure is usually the classification rate or accuracy. Verification, in turn, is the task of deciding, given a speech sample, whether the specified speaker really said it or not. The SV problem is a two class problem of deciding if it is the same speaker or an impostor requesting verification. Commonly, SV exhibits greater practical applications



(a)

(b)

Figure 1: Comparison of waveform and spectrogram of the speech signal "*Here I was in Miami and Illinois*" from the same speaker in (a) normal and (b) whispered speech mode.



(a)

(b)

Figure 2: Plots of average power spectrum and frame energy distribution. (a) average power spectrum comparison of the utterance "*Here I was in Miami and Illinois*" spoken by same speaker and (b) frame energy distribution for normal and whispered speech using combined male and female data across 36 speakers.

related to SI, specially in access control and identity management applications. In the past, whispered speech has only been explored within the SI problem [5–8, 14, 25], where the use of the accuracy metric does not give a clear picture of the actual impact of mismatch conditions between training and testing [26]. In addition, it is not clear whether the strategies proposed for SI systems can also be useful for SV systems.

Currently, state-of-the-art SV systems based on normal speech use highly elaborate techniques, such as the so-called i-vectors [27]. However, to properly train such systems, large amounts of training data are required [2, 28]. Unfortunately these amounts of data are hard to collect for whispered mode, which can affect the training and limit the advantages of these techniques over other strategies. Furthermore, these methods are heavily dependent of the data, i.e., the nature of the testing data should be the same with the one the i-vector extractor was trained on [29]. According to our experiments, a classification system based on Gaussian mixture models (GMM) and maximum a posteriori (MAP) adaptation, as depicted by Figure 3, was more suitable for dealing with mismatched scenarios. For the described system, the widely-used mel-frequency cepstral coefficients (MFCC) are used to implement a text–independent SV system [2, 30]. First an M-Component GMM is trained as an universal background model (UBM) using the Expectation – Maximization (EM) algorithm and the training data available from all speakers. Then, a GMM for each speaker is obtained using MAP adaptation, as depicted by top half diagram in Figure 3. During the recognition phase (bottom half of Figure 3), the hypothesized speaker model is scored against the UBM and a decision is made based on thresholding. More details can be found in [30].

3.2 Speech stimuli

In our experiments, the CHAINS (Characterizing Individual Speakers) speech corpus was used [8]. The corpus contains the recordings of 36 speakers obtained in two different sessions with a time separation of about two months, there are three different accents : 28 speakers from Ireland (16 male), 5 speakers from the USA (2 male) and 3 speakers from the United Kingdom (2 male). Additional details about the database can be found in [8]. Speech stimuli was generated under six speaking conditions, namely solo (natural rate reading), retelling without time constraints, two-person synchronous reading, repetitive synchronous imitation, accelerated-rate reading, and whispered.

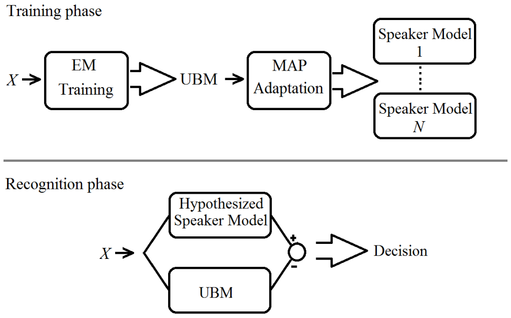


Figure 3: Block diagram of a general SV system. Top and bottom diagrams represent the training and testing stages, respectively, for a GMM-UBM SV based system

For our experiments, two speaking styles were used - solo and whispered - where the same text was read in both conditions. We used the speech stimuli generated from reading the paragraph of the *Cinderella story* (average duration : 55 seconds, minimum duration : 48 seconds) for training, and kept the stimuli generated from reading the *Rainbow Text* (average duration : 30 seconds ; minimum duration : 23 seconds) segmented in short sentences of 3 seconds, plus 32 individual sentences (nine selected from the CSLU Speaker Identification corpus and 23 from the TIMIT corpus) for testing. Data was originally recorded at 44.1 kHz sample rate but downsampled to 8 kHz, as motivated by [31].

3.3 Baseline performance in matched and mismatched conditions

Prior to feature extraction, in our experiments we normalized the speech data to -26 dBov (dB overload) using the ITU-T P.56 speech voltmeter [32], and pre-emphasized using a first order FIR filter with constant $a = 0.97$. Then 19 MFCC were computed on a per-window basis excluding the 0–th order cepstral coefficient, using a 32 ms window with 50% overlap and 24 triangular bandpass filters. Delta coefficients were also included to convey temporal dynamics information. Delta coefficients were computed by means of an anti-symmetric Finite Impulse Response (FIR) filter of length nine to avoid phase distortion of the temporal sequence. For all experiments herein, the training data was fixed to 35 seconds per speaker, and the number of Gaussian components per model was fixed to $M = 32$, showing a tradeoff between performance and computational burden.

Before presenting the results, we want to illustrate the effects of pre-emphasizing and normalizing the speech recording. Figure 4(a) and 4(b) depict the average spectrum and frame energy distribution, respectively, of amplitude-normalized and pre-emphasized recordings using male and female speech. As can be seen, the gap between the two speaking styles seen in Figure 2 has been greatly diminished, although most of the differences remain below 1.2 kHz.

Table 1 reports the Equal Error Rate (EER) obtained with the baseline system under different *train/test* conditions. In the table, 'c' stands for cepstral coefficients and 'Δ' for delta coefficients. As can be seen, for normal speech in the *normal/normal* (train/test) matched condition inclusion of delta coefficients did not provide any advantage over using only MFCCs. In fact, in the *normal/whisper* and *whisper/whisper* scenarios, inclusion of delta parameters had a negative impact on system performance, as previously reported by [6]. Only in the mismatch *whisper/normal* condition, was an improvement in EER with the inclusion of Δ parameters observed ; the gains, however, were modest and we can not considerate this as a significant advantage. In the Table, the values in bold represent the baseline performances with which improvements will be gauged against.

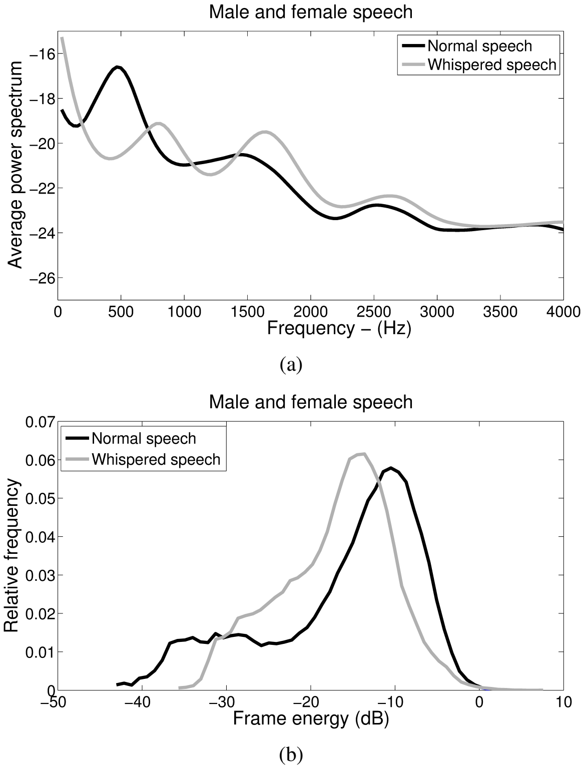


Figure 4: Plots of (a) average power spectrum and (b) frame energy distribution after preprocessing for normal and whispered speech (averaged over 36 speakers).

Table 1: EER(%) comparison for different *training/testing* conditions after power normalization and pre-emphasis. Results in bold represent the baseline systems with which the tested improvements will be gauged against.

		EER(%)	
Training	**Testing**	c	$c + \Delta$
Normal	Normal	**2.13**	2.33
Normal	Whisper	**35.75**	38.62
Whisper	Normal	29.81	28.18
Whisper	Whisper	2.90	3.12

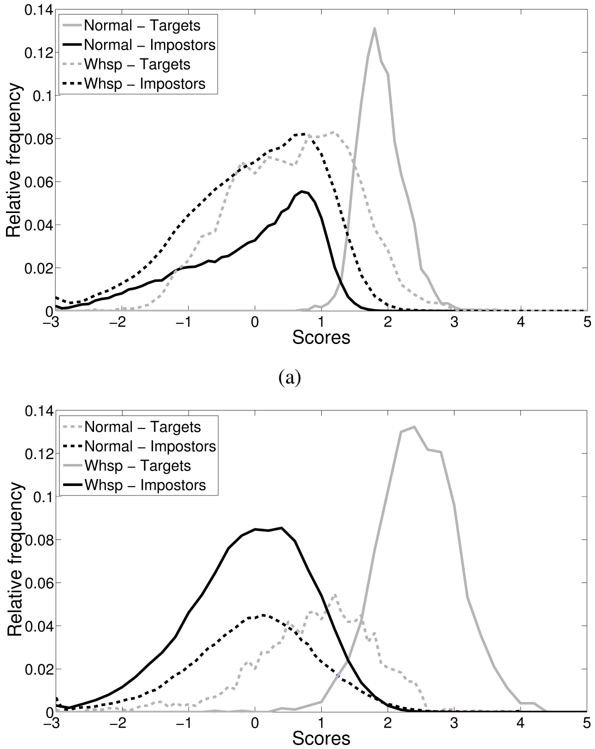
Overall, it can be seen that significant performance degradation occurs in the mismatch conditions. When testing with whispered speech, the obtained EER for the mismatch condition was more than 10 times greater than in the matched condition. Moreover, a gap of approximately 6 – 9% can be seen in mismatched cases, depending on what speaking style is used for training. As can be seen, lower EER is achieved when training with whispered speech and testing with normal. This was expected, as in our dataset, approximately 70/30% of the normal-speech training data was comprised of voiced/unvoiced speech segments. When training with normal speech, it is likely the GMMs became biased towards voiced characteristics which are not present in whispered speech. On the other hand, when training with whispered speech, the GMMs could more accurately represent unvoiced normal-speech segments, as only small differences have been observed between unvoiced consonants in whispered and normal speech modes [16]. To better illustrate this point, Figure 5 shows the plots of the scores distribution for target speakers and impostors under the two training conditions. Continuous lines represent the speaking style used for training (i.e., normal speech in subplot (a) and whispered speech in subplot (b)).

Figure 5(a) shows that by using normal speech for training the scores of normal speech are less scattered than those for whispered speech, which, in turn, show a high degree of overlap. Figure 5(b), on the other hand, shows the scores obtained when training only with whispered speech. As can be seen, scores from whispered speech testing recordings are still more scattered than those for normal speech, but the overlap has been reduced. Overall, as expected the matched *normal/normal* scenario resulted in the lowest EER. Together these findings suggest that alternate strategies are needed to improve the performance of SV systems based on whispered speech, particularly in mismatched cases. This is the focus of the sections to follow.

4 Strategies to improve system performance in mismatched train/test scenarios

4.1 Frequency and feature warping

Different frequency warping strategies have been proposed and can be used in lieu of the classical mel scale. These frequency warpings allow greater resolution to be placed at certain frequency ranges. Commonly used scales include : linear, exponential and the whisper sensitive scale (WSS) [33], in addition to the widely used mel scale. Previous studies using the exponential and linear scales showed that relative improvements of around 20% could be achieved ; however, for further improvements some knowledge about the speaking style was needed for testing [5, 25]. Furthermore, the improvements were shown only for the whispered speech speaker identification task, thus there is no evidence about the effects of this front-end in the speaker verification task. Table 2 shows the mappings between the original (f) and warped ($\hat{f}$) frequencies used in our experiments. The linear scale is omitted from the Table, as $\hat{f} = f$.



(a)

(b)

Figure 5: Plots of score distributions for target and impostor speakers using normal and whispered speech files. The scores were computed using two different systems, the system in (a) was trained only with normal speech and the system in (b) was trained only with whispered speech. Continuous lines are representative of the speaking style used for training.

Table 2: List of frequency warping strategies used in the experiments. Cepstral coefficients derived are MFCC (mel), EFCC (exponential - Exp. in the table) and WSSCC (WSS).

Scale	Frequency warping
Mel	$\hat{f} = 2595 \times \log_{10}(1 + \frac{f}{700})$
Exp.	$\hat{f} = 10610 \times (10^{f/50000} - 1)$
WSS	$\hat{f} = \begin{cases} \frac{2475 f^4}{1220^4 + f^4}, & 0 < f < 2000 \\ 4100 - \frac{2000}{1+e^{(f-300)/310}}, & 2000 \leq f < 4000 \end{cases}$

Using the same settings as before, 19 cepstral coefficients were computed using the above described frequency warping strategies, along with the delta coefficients. Cepstral coefficients derived are MFCC (mel), EFCC (exponential), WSSCC (WSS), and LFCC (linear). This experiment allows us to determine which frequency warping strategy can better reduce the negative impact of train/test mismatch. Additionally, to mitigate the effects of linear channel mismatch, a widely accepted method is called *feature warping*, which maps the distribution of the cepstral features to a normal distribution ($\mathcal{N}(0, 1)$) by using a 3-second sliding window, also known as short-time Gaussianization (STG) [34]. For the sake of comparison, the different feature sets are evaluated in the two possible scenarios : with and without STG.

Results are shown in Table 3 where two *training/testing* conditions are evaluated, namely *normal/normal* and *normal/whisper* (represented in the table as N/N and N/W, respectively). Whilst the negative impact of mismatch is still evident, all frequency warping strategies have improved the MFCC performance. As an example, by using the whisper sensitive scale and appending delta coefficients it is possible to reduce the EER by approximately 13% relative to the baseline in mismatch condition without using feature warping. Furthermore, STG can result in additional improvements in the mismatch condition, leading to improvements up to 31% relative to the baseline. Notwithstanding, one disadvantage of frequency and feature warping is the drop in performance obtained in the matched N/N condition. For example, with MFCCs the EER doubles after STG. The other frequency warping strategies, on the other hand, resulted in more stable results after STG. As before, no significant advantages were observed by appending the delta coefficients.

4.2 Frequency sub-band analysis

Results presented in Tables 1 and 3 suggest that whispered speech conveys information highly related to each speaker, but significant differences are still present between the two speaking styles. Motivated by the results in Figure 4(a), we also explore the use of only a sub-band of the speech signal in which their difference is minimized. According to Figure 4(a), this sub-band ranges from approximately 1.2 kHz to 4 kHz. As such, the frequency-warpings are calculated between 1.2 and 4 kHz. This frequency band comprises mostly information from the second and third formants (F2 and F3). EER performance results are shown in Table 4. As observed, further gains are obtained in the mismatch condition, but at the cost of reduced performance in the matched scenario. Notwithstanding, these findings corroborate previously-reported cues showing a significant amount of speaker-specific information in the second and third formants [35, 36]. An additional advantage of focusing within this sub-band is that for whispered speech, shifts in F2 of 2 - 24% and in F3 of 1 - 10% have been observed relative to normal-voiced speech [21]. This is a rather low variation when compared with the shift for F1 that can be 50% or higher [21]. The most relevant improvement in mismatch condition is achieved using MFCC ; when comparing with the results in Table 3, a relative reduction in the error rate of approximately 38% is achieved using STG and without appending delta coefficients. It is important to emphasize that in the matched condition the error rate is three times higher than that reported in Table 3. Together, these results show the high relevance of speaker identity information contained below 1.2 kHz, particularly for normal speech.

4.3 Alternate feature representations

Some authors have proposed to use features completely different in nature to cepstral coefficients. As an example, features derived from the AM-FM signal representation have proven to be more robust in noisy conditions and perform at the same level as cepstral coefficients [8, 37]. The main difference is that cepstral coefficients are based on power spectrum estimation (i.e., frequency domain) whilst features derived from the AM-FM signal representation are computed in the time domain. More specifically, the AM-FM model decomposes the speech signal into bandpass channels and characterizes each channel in terms of its envelope and phase (instantaneous frequency) [8, 38]. The speech signal $s(n)$ is filtered through a bank of N_K filters, resulting in the bandpass signal $y_k(n) = s(n) * h_k(n)$, where $h_k(n)$ corresponds to the impulse response of the k-th filter. There are different approaches for filter design that have been used in speech applications. In this study, two approaches were tested : a gammatone filterbank [39], and the Gabor filterbank [8], each with 23 channels. Filter center frequencies range from 50 Hz to 3528 Hz and their bandwidths are characterized by the mel frequency scale. After filtering, each analytic sub-band signal $s_k(n)$ is uniquely related to a real–valued bandpass signal $y_k(n)$ by the relation :

$$s_k(n) = y_k(n) + j \cdot \hat{y}_k(n) \qquad (1)$$

where $\hat{y}_k(n)$ stands for Hilbert transform of $y_k(n)$. There are two approaches to decompose each analytic signal in terms of its envelope and phase : *i*) the Hilbert envelope approach (non–coherent demodulation) and *ii*) coherent demodulation [38]. The main difference between these two approaches is in the allocation of phase between the envelope and carrier. Whereas the Hilbert envelope places all of the sub-band phase in the carrier, coherent demodulation makes the important distinction between carrier and modulator phase. In our previous work, it was found that the Hilbert envelope approach resulted in improved performance relative to the coherent demodulation approach [40], hence in this work only the Hilbert envelope approach is used. For the sake of notation, let $m_k(n)$ denote the low–frequency modulator and $f_k(n)$ the instantaneous frequency for each bandpass signal. Figure 6 depicts the general process to decompose the speech signal into bandpass channels and their respective modulator and instantaneous frequencies.

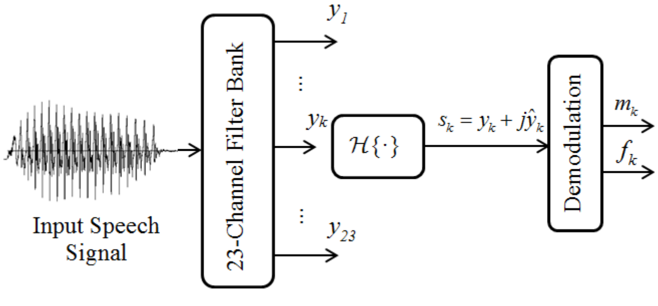


Figure 6: AM-FM signal representation. Block diagram to decompose the speech signal in bandpass channels and compute the low frequency modulator and the instantaneous frequency per channel.

Table 3: EER(%) comparison for matched and mismatched *training/testing* condition, using different frequency warping strategies and comparing the effects of using STG as feature warping. N/N and N/W correspond to training with normal speech and testing with normal or whispered speech, respectively. All feature representations where computed from the full 0 to 4 kHz band. EER values in bold highlight the best performance achieved in matched and mismatched conditions.

	without STG				**with STG**			
Cepstral	c		$c+\Delta$		c		$c+\Delta$	
Coefficients	**N/N**	**N/W**	**N/N**	**N/W**	**N/N**	**N/W**	**N/N**	**N/W**
MFCC	**2.13**	35.75	2.33	38.62	5.08	32.23	4.78	35.23
LFCC	4.88	31.04	4.60	30.20	4.17	**24.33**	5.20	25.82
EFCC	5.09	31.36	5.21	30.10	4.18	24.57	5.26	25.64
WSSCC	6.01	31.02	6.21	29.08	6.17	25.70	7.50	27.26

Table 4: EER(%) comparison for matched and mismatched *training/testing* condition using the sub-band from 1.2 kHz to 4 kHz to compute the different feature sets with different frequency warping strategies and comparing the effects of using STG as feature warping. N/N and N/W correspond to training with normal speech and testing with normal or whispered speech, respectively. EER values in bold highlight the best performance achieved in matched and mismatched conditions.

	without STG				**with STG**			
Cepstral	c		$c+\Delta$		c		$c+\Delta$	
Coefficients	**N/N**	**N/W**	**N/N**	**N/W**	**N/N**	**N/W**	**N/N**	**N/W**
MFCC	8.64	26.50	9.02	26.82	**7.14**	**21.81**	9.20	24.51
LFCC	9.58	27.54	9.53	25.96	7.44	21.81	9.62	22.89
EFCC	9.39	27.18	9.45	26.24	7.74	22.47	9.38	23.43
WSSCC	8.36	27.75	8.85	26.93	8.89	24.87	11.62	25.58

Here, two features are explored based on the AM-FM signal decomposition. The first is the so called Weighted Instantaneous Frequencies (WIF). These features are computed by combining the values of $m_k(n)$ and $f_k(n)$ using a short-time approach :

$$F_k = \frac{\sum_{i=n_0}^{n_0+\tau} f_k(i) \cdot m_k^2(i)}{\sum_{i=n_0}^{n_0+\tau} m_k^2(i)}, \quad k = 1, \ldots, 23, \qquad (2)$$

where τ is the length of the time frame. F_k is calculated over the full length of each $m_k(n)$ with increments of $\tau/2$.

The second feature set is the mean Hilbert envelope coefficients (MHEC) proposed in [37] and shown to perform better than traditional MFCC features under noisy conditions for normal speech for speaker verification. In this case, the envelope $m_k(n)$ is blocked into frames and the mean Hilbert envelope for a specific frame in channel k is calculated as :

$$E_k = \frac{\log\left(\frac{1}{\tau}\sum_{i=n_0}^{n_0+\tau} w(i-n_0+1) \cdot m_k(i)\right)}{\bar{E}_k}, \quad k = 1, \ldots, 23 \qquad (3)$$

where $w(n)$ is a Hamming window of length τ, and the term $\bar{E}_k$ represents the long-term average in each channel which normalizes the values of E_k. Finally, for a specific frame and using all 23 E_k values, a discrete cosine transform (DCT) is applied to produce the MHEC features [37].

Table 5 reports the EER obtained with the different filterbank characterizations, considering both the full band and the limited sub-band (1.2–4 kHz) components. In the matched condition, MHEC and WIF perform better than cepstral coefficients without STG and at the same level using STG. However, in mismatched condition both WIF and MHEC achieve error rates similar to the ones achieved with cepstral coefficients. These results suggest that the information present in the slowly varying envelope of the bandpass signals is highly discriminative, but extremely sensitive to changes in the vocal effort. By limiting the analysis frequency band to 1.2–4 kHz, a significant reduction of approximately 36% could be achieved relative to the baseline system in mismatched condition (see Table 1). This, however came at a severe penalty for the matched scenario, as was similarly observed with the cepstral coefficients (see Table 4).

4.4 Feature combination

Since cepstral coefficients, WIF, and MHEC extract complementary information, we explored feature combination as an alternate strategy to improve SV performance in mismatched scenarios. For this experiment, and based on the results presented in Table 4, the mel and linear scales were selected to

Table 5: EER(%) comparison for matched and mismatched *training/testing* conditions, using features derived from the AM-FM signal representation. Limited band corresponds to 1.2–4 kHz. Norm/Norm and Norm/Whsp correspond to training with normal speech and testing with normal or whispered speech, respectively. For each feature representation (WIF and MHEC) EER values in bold highlight the best performance per train/test condition.

	Filter Bank	EER–Full band		EER–limited band	
		N/N	N/W	N/N	N/W
WIF	Gammatone	**1.63**	33.73	5.87	24.63
	Gammatone + STG	4.48	29.48	7.86	23.19
	Gabor	2.18	35.65	6.53	24.27
	Gabor + STG	4.17	30.92	7.99	**22.77**
MHEC	Gammatone	2.06	42.24	9.80	26.72
	Gammatone + STG	5.51	41.34	10.71	28.78
	Gabor	**1.57**	36.73	9.13	**26.24**
	Gabor + STG	4.23	34.09	11.62	26.78

compute the cepstral coefficients in the 1.2–4kHz sub-band with STG. Moreover, motivated by results in Table 5, the WIF features using the Gammatone filter bank and the MHEC features using the Gabor filter bank were selected as they showed to be more effective in the matched condition without STG.

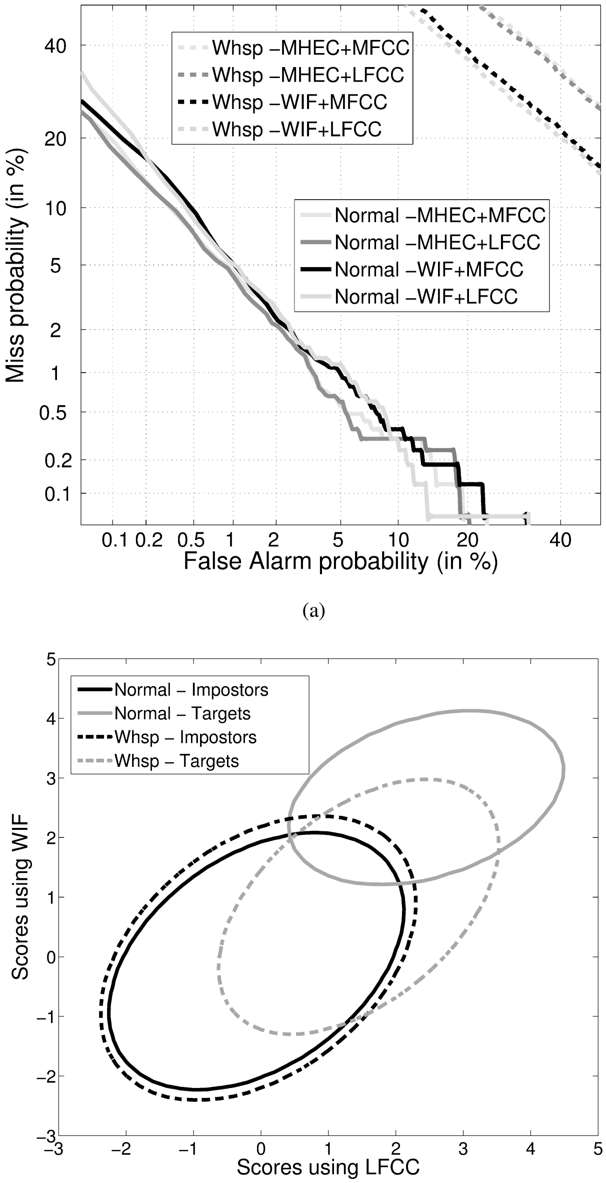
Results for feature combination are shown in Figure 7(a) and Table 6. In the table, the features labelled in the columns are combined with the features labelled in the rows to produce a new feature space and the EER corresponding to each testing condition is presented in the respective intersection. According to these results, feature combination does not help to obtain further reductions of the EER in mismatch condition (N/W). Notwithstanding, combining WIF and LFCC and comparing the results with the baseline system, this combination can help to maintain the performance inline with the baseline system for the match condition, whilst achieving relative reduction of the EER in the mismatch condition by approximately 21% . To extend the analysis, the scores of target speakers and impostors were calculated separately using WIF and LFCC. These scores were used to estimate the parameters of a 2 dimensional full covariance Normal distribution. The contours of the distributions are depicted in Figure 7(b) with continuous lines for normal speech and dashed lines for whispered speech. As can be seen, the overlap between target speakers and impostors for normal speech is minimum, however for whispered speech the scores are more scattered and higher overlap exists. As such, any decision boundary minimizing the error rate for normal speech will not necessarily be optimal for whispered speech. Such findings suggest the need for speaking-style dependent models, as will be described in Section 4.7.

4.5 Gender dependency analysis

Male and female voices are different from each other in terms of physical characteristics (pitch and vocal track length), linguistics and style. As such, some authors recommend to train separate systems per gender [41, 42]. To test if this trend also

Table 6: EER(%) comparison with different feature combination, where the best features from Tables 4 and 5 were selected. EER values in bold represent the best performance per train/test condition.

Cepstral	WIF		MHEC	
Coefficients	N/N	N/W	N/N	N/W
MFCC	2.17	29.35	2.29	36.96
LFCC	2.29	**28.16**	**2.05**	36.60



(a)

(b)

Figure 7: Plots of (a) DET curves for feature combination and (b) contours of an estimated Gaussian distribution for the scores of testing utterances. These Plots were obtained by using only normal speech for training and normal and whispered speech for testing.

occurs with whispered speech, we tested a gender-dependent system as well. This is possible, as systems have been shown to accurately discriminate genders from whispered speech in clean conditions [12,40]. Results are presented in Table 7 for individual features and Table 8 for feature combination. In the latter case, the combined feature sets used are the same as in Section 4.4. As can be seen, in the matched condition there are some differences relative to Table 6. First, for normal speech the feature representations that performed best for male speech did not perform at the same level for female speech, thus corroborating previous findings [41, 42]. Next, feature combination (Table 8) in gender dependent models does not help to reduce the impact in the mismatch condition relative to the results shown Table 6. This is also corroborated when comparing the results from Table 6 and the overall error rates presented in Table 8, thus suggesting that feature combination is more effective in the mismatched train/test condition for gender independent systems. It is possible that the models can learn some specific structures about whispered speech when both genders are involved into the parameter estimation. Together, these findings suggest that speaking style and gender dependencies are present in the whispered speech SV task. Such scenario will be further explored in Section 4.8.

4.6 Training with combined *normal/whisper* data

Results presented so far have shown that reliable performance can be achieved in matched conditions, but significant drop in performance occurs in mismatched conditions. As an alternate solution, here we explore the use of both normal and whispered speech during training and model adaptation as has been done in previous studies for speaker ID [6, 14]. This allows speaker-specific information represented in whispered speech features to be properly modeled. Since whispered speech training data can be sparse, it is not clear how much whispered speech material is necessary to achieve acceptable performance levels for practical applications. In order to be able to perform a comparison with the baseline system, we investigate the effects of adding small amounts of whispered speech to the training set, using a MFCC–GMM system (without delta coefficients). Experiments were conducted using a fixed duration length of normal speech (35 seconds per speaker) and different duration lengths of whispered speech for training.

Results of these experiments are illustrated in Figure 8 and Table 9. As can be seen, there is significant improvement by adding as little as 5 seconds of whispered speech per speaker relative to the mismatch performance reported in Table 1. By gradually increasing the duration length of whispered speech, the performance of the system also gradually improves, thus corroborating previous speaker identification findings [6, 14]. Nevertheless, using the same amount of data (35 s) for both vocal efforts shows that improved performance is still achieved with normal speech with respect to whispered speech (11% lower EER). In addition, it is necessary to pay attention to the slight losses induced by the addition of whispered speech, which slightly increases the EER for normal speech. For example, using only normal speech for training, an EER of 2.13 % was reported in Table 1. Here, in the case of using the same amount of data for both vocal efforts, an EER of 3.05 % (i.e., 43% higher) was found. According to these results, for a practical SV verification task improved performance can be achieved for whispered test speech, but at the cost of lower performance for normal test speech.

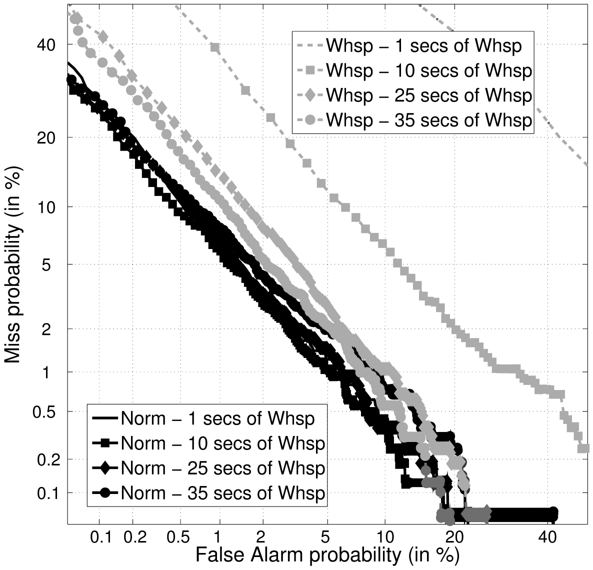


Figure 8: DET curves exploring the effects of adding different amounts of whispered speech to the 35 s of normal speech during the training phase.

4.7 Speaking–style dependent SV systems

Up to now speaking-style *independent* SV systems have been described to handle both vocal efforts. Recent literature on SI and speech recognition, on the other hand, has recommended the use of speaking-style dependent models [9, 10, 14], as depicted by Figure 9. The method builds on the previously described MFCC-GMM algorithm and takes into account the different subclasses that can be modelled in order to build a complete speaker verification system. In this section, two classes are investigated : normal and whispered modes. In order to develop a speaking-style dependent SV system, a classification stage is needed in order to detect specific speaking styles. For example, a recently proposed *normal/whispered* speech classifier can be used, as it was shown to perform accurately even in noisy conditions [43].

With speaking style dependent systems, the concept of "mismatch" shifts from one of *train/test* mismatch to one of errors in speaking style classification. In order to analyse the benefits of having dedicated speaker models for each speaking style, this first set of experiments will assume an "oracle" system in which perfect *normal/whisper* classification is achieved. For clean conditions, this is not an unrealistic assumption [43]. Within this scenario, we are particularly in-

Table 7: EER(%) comparison with different feature representation using gender dependent models and the best features from Tables 4 and 5. Best results are highlighted per gender and per training/testing condition.

	WIF		MHEC		MFCC		LFCC	
Gender	**N/N**	**N/W**	**N/N**	**N/W**	**N/N**	**N/W**	**N/N**	**N/W**
Female	**2.15**	38.52	2.72	40.79	7.60	28.18	8.46	**25.07**
Male	2.19	40.91	**1.04**	38.94	7.42	**25.85**	6.90	26.16
Overall	2.17	39.84	1.78	39.76	7.50	26.88	7.59	25.67

Table 8: EER(%) comparison with different feature combination using gender dependent models and combining the best features from Tables 4 and 5. Best results are highlighted per gender and per train/test condition.

	Female				Male				Overall			
Cepstral	**WIF**		**MHEC**		**WIF**		**MHEC**		**WIF**		**MHEC**	
Coefficients	**N/N**	**N/W**	**N/N**	**N/W**	**N/N**	**N/W**	**N/N**	**N/W**	**N/N**	**N/W**	**N/N**	**N/W**
MFCC	3.15	32.71	3.44	41.64	2.40	36.44	**1.25**	39.66	2.73	34.78	2.22	40.54
LFCC	**2.86**	**32.43**	3.29	41.35	2.40	**35.51**	1.35	39.77	2.60	34.14	2.21	40.47

Table 9: Effects of adding different amounts of whispered speech to the normal speech training set.

Amount of whispered	**EER(%)**	
training data (s)	**Normal**	**Whispered**
1	2.54	30.97
5	2.53	13.25
10	2.49	7.91
15	2.60	5.47
20	2.62	4.24
25	2.66	3.94
30	2.63	3.52
35	3.05	3.45

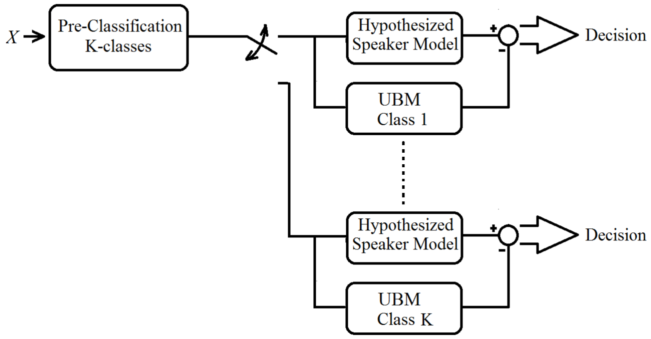


Figure 9: Multimodel framework for SV. Block diagram for a K-class speaking style dependent SV system

terested in the performance obtained with the whispered test speech files. Tables 10 and 11 show the EER comparison for different frequency warpings and AM-FM feature representations, respectively. As can be seen from Table 10, inclusion of delta coefficients degrades performance of the system. Overall, the Linear-Frequency Cepstral Coefficients (LFCC) and MFCC showed to be the two sets of feature vectors that can achieve the lowest error rates, outperforming the WSS scale, which was developed specifically for whispered speech [33]. From Table 11, in turn, it can be seen that the AM-FM based features provide a modest improvement over the cepstral-based features. When using the gammatone filterbank, WIF features outperformed the MHEC ones. The opposite behaviour was observed with the Gabor filter bank. In both cases (cepstral and AM-FM based features), the EER results obtained with whispered test speech files are slightly higher than those obtained with the normal-voiced files in Table 3, where an EER of 2.13% was reported with MFCCs.

Table 10: EER(%) comparison in W/W condition using speaking style dependent models. Results are for whispered test files and using different warping strategies to compute cepstral coefficients.

	EER(%)	
Cepstral coefficients	c	$c + \Delta$
MFCC	2.90	3.12
LFCC	2.90	3.08
EFCC	3.12	4.15
WSSCC	4.22	6.02

Subsequently, feature combination was explored. Motivated by the results presented in Tables 10 and 11, the mel and linear scales were chosen to compute the MFCC and LFCC features, respectively. The gammatone filterbank was used to compute the WIF features and the Gabor filterbank to compute the MHEC features. Since the inclusion of delta coefficients did not present any advantage for the considered feature sets, they were not included in this feature combina-

Table 11: EER(%) comparison in W/W condition using speaking style dependent models. Results are for whispered test files and using AM-FM based features. Highlighted results are the best EER values per feature representation.

	AM-FM features	
Filter Bank	**WIF**	**MHEC**
Gammatone	**2.55**	3.10
Gabor	2.62	**2.60**

tion analysis. Results are shown in the Table 12. According to these results, significant improvements can be achieved by combining features, thus corroborating their complementarity. A relative reduction of the EER of approximately 33% can be seen when comparing the best results from Tables 10 and 11, and outperforming those for normal speech reported in Table 1.

Table 12: EER(%) comparison in W/W condition with different feature combination, where the best features from Tables 10 and 11 were selected.

Cepstral	**AM-FM features**	
Coefficients	**WIF**	**MHEC**
MFCC	1.79	2.03
LFCC	1.91	1.85

4.8 Gender and speaking–style dependent SV systems

For these experiments, recordings were separated by gender in the training phase. Following the scheme presented in Figure 9, besides the speaking style detection, it would be necessary to detect two additional classes, i.e., male and female speech. For these experiments, two UBMs were obtained (one for each gender) as well as their respective speaker-specific models from MAP adaptation. Mel and linear scales were chosen to compute the cepstral coefficients, and from the AM-FM features the WIF using the Gammatone filter bank and the MHEC using the Gabor filter bank were selected. EER results are reported in Table 13. As can be seen, gender dependent systems provide advantages only for female speakers. Feature combination, on the other hand, did not provide further advantages as can be observed by Table 14. Interestingly, in the matched N/N scenario shown in Tables 7 and 8, male speech was shown to result in improved performance related to female speech. With the matched W/W scenario shown in Tables 13 and 14, the inverse is seen and female speech results in better performance, with AM-FM based features resulting in optimal performance.

As illustrated by Figure 9, when using class-specific models, gender classification needs to be performed prior to the verification stage. For this purpose, an M-component GMM was trained. Initially, different amounts of training data and different values of M were evaluated to analyse how these values affect gender detection error rates. Figure 10 shows that while the amount of training data does not have a significant effect on EER, the number of Gaussian components does. From Figure 10 it can be seen that there is a settling point using M=30 and 40 seconds of training data per speaker. Table 15 summarizes the gender detection error rates for different feature sets and Gaussian components M. As can be seen, for gender detection WIF and MHEC features, both using the Gabor filter bank, outperform MFCC features. Moreover, MHEC and WIF features achieve close to perfect accuracy even with only 10 Gaussian components. With MFCC, on the other hand, this performance is only achieved using $M = 30$. This suggests that there is gender-specific information in the phase of the acoustic signal and that an approach based on Hilbert envelopes can be used to characterize such information. This corroborates previously-reported subjective findings that whispers not only carry information about speaker identity but also about the gender [12, 19]. Hence, even without glottal excitation, gender discrimination has been shown to be a feasible task using whispered speech. Note that cepstral coefficients using other frequency scales or feature combination did not show any advantage for gender detection, hence they were not included in Table 15.

Table 13: EER(%) comparison in W/W condition for gender and speaking style dependent models. Results are for whispered speech test files, and best EER values are highlighted per gender.

Gender	**WIF**	**MHEC**	**MFCC**	**LFCC**
Female	1.27	**0.99**	1.41	1.55
Male	**1.86**	2.18	3.73	3.21
Overall	1.59	1.65	2.69	2.47

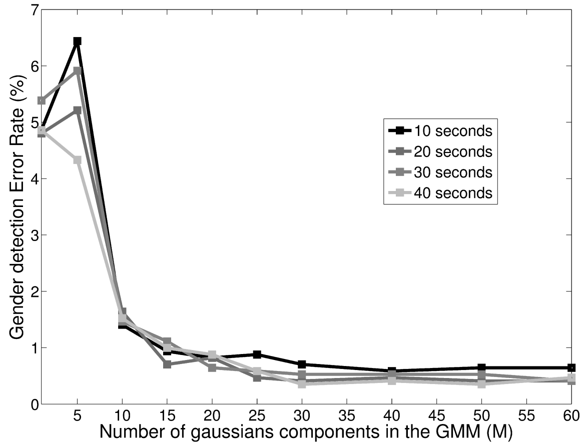


Figure 10: Gender detection error rate using MFCC as a function of of Gaussian components (M) and amount of training data per speaker.

Table 14: EER(%) comparison in W/W condition for gender and speaking style dependent models and feature combination. Results are for whispered speech test files, and best EER values are highlighted by gender.

Cepstral	Female		Male		Overall	
Coefficients	WIF	MHEC	WIF	MHEC	WIF	MHEC
MFCC	1.41	**1.13**	2.59	2.90	2.06	2.11
LFCC	1.27	**1.13**	**2.28**	2.69	1.83	1.99

5 Robustness to noise

As mentioned previously, whispered speech based SV is burgeoning due to the popularity of smartphones. But user mobility has also created several challenges that need to be addressed, one of them is robustness to ambient noise. Hence, it is important to analyse the robustness of the investigated features to noise. For these experiments, speaker models were trained with clean whispered speech and testing data was contaminated with three different signal-to-noise ratios (SNR) of babble noise : 5, 10 and 15 dB. Babble noise was chosen due to its challenging speech-like nature, as well as its likely presence in places where whispered speech SV is bound to be used. Using the speaking-style dependent system proposed in Section 4.7, experimental results are shown for both normal and whispered speech in Table 16. As can be seen for whispered speech, EER in noisy conditions increased for all feature representations as the SNR decreased, thus suggesting the sensitivity of the features to ambient noise. The benchmark feature MFCC is the feature set with worse performance at all SNR levels. LFCC and exponential frequency cepstral coefficients (EFCC), on the other hand, have better performance when tested alone, thus suggesting that a proper selection of frequency warping can improve robustness against noise. Interestingly, while in clean conditions the cepstral coefficients extracted from the WSS-warped spectra (i.e., WSSCC) did not result in optimal results, they outperformed all other cepstral-based features with noisy speech. A similar dependency on noise levels was observed with the MHEC features, which were outperformed by the WIF features. Regarding these latter features, the use of the gammatone filter bank showed improved robustness against noise relative to Gabor filter bank. Overall, our results suggest that WSSCC combined with WIF are the most appropriate setup for whispered speech SV under noisy environments.

Table 15: Gender detection error rates (%) for different feature sets and number of Gaussian components (M). Best results are highlighted by number of Gaussians.

Feature	M Gaussians						
Set	1	5	10	20	30	40	50
MFCC	4.85	4.33	1.52	0.99	0.58	0.40	0.38
WIF	3.45	1.22	**0.52**	**0.46**	0.46	0.39	0.35
MHEC	**3.04**	**1.11**	0.70	0.60	**0.38**	**0.35**	**0.30**

Similar noise sensitivity of the various features was also observed for normal speech. As seen previously, the MFCC features were most affected by noise. Interestingly, the WSSCC features also showed to be optimal in the very noisy scenario (SNR=5dB) for normal speech, thus showing the importance of frequency warping strategies for improved robustness against babble noise. Overall, AM-FM based features, as well as their combination with different cepstral features, were not as beneficial for normal-speech speaker verification in noisy settings as they were with whispered speech.

Note that results presented in Table 16 were obtained by assuming perfect *normal/whispered* speech classification in the speaking-style dependent system (i.e., an oracle system). However, different levels of noise can also affect this stage prior to speaker verification. In order to be able to quantify the total effect on system performance by the inclusion of noise, a second experiment was performed. Here, the speaking-style classifier described in [43] was used. EER comparison is shown in Table 17. The last row in the table shows the speaking style classifier error rates for different noise level scenarios. As can be seen, *normal/whisper* classification errors result in 20%, 16% and 10% relative increases in EER for SNR of 15, 10 and 5 dB, respectively. Despite this drop in performance, the speaking-style dependent system exhibits reliable performance even in noisy conditions. It is important to emphasize that results are not reported for the gender and speaking style dependent systems from Section 4.8 as the gender detection classifier was shown to be very sensitive to babble noise.

6 Discussion

There is evidence based on subjective studies suggesting that invariant speaker identity across different vocal efforts exists [13], i.e., a listener can recognize a speaker without training, using only the experience with normally voiced speech of the same speaker. Despite different strategies, such as frequency warping, preprocessing, and alternate feature representations, our results suggest that the invariant information between normal and whispered speech is not sufficient to achieve reliable performance in an SV task for *both* speaking styles. A compromise must be kept in order to guarantee system performance in normal and whispered speech. Notwithstanding, for most of the cases evaluated herein, improvements in the mismatched condition were accompanied with reduced performance in the matched scenario. Moreover, the strategies that performed better for normal speech did not exhibit the same

Table 16: EER(%) comparison for different feature representations under different ambient noise levels. Best EER values are highlighted in bold per feature group for the tested SNR levels and the two train/test conditions.

	W/W			N/N		
	SNR level			SNR level		
Feature set	**15 dB**	**10 dB**	**5 dB**	**15 dB**	**10 dB**	**5 dB**
MFCC	13.42	22.53	31.82	12.34	19.18	26.98
LFCC	7.13	13.42	21.27	7.20	9.13	13.67
EFCC	7.25	13.30	21.21	**6.96**	**9.07**	13.43
WSSCC	**6.35**	**9.59**	**15.78**	7.20	9.38	**12.95**
WIF (Gamma.)	**5.33**	**8.87**	**14.80**	16.33	22.14	28.80
WIF (Gabor)	7.43	12.22	20.61	**11.07**	**16.27**	**23.65**
MHEC (Gamma.)	16.48	27.44	36.49	18.81	27.47	35.33
MHEC (Gabor)	13.24	23.37	32.59	12.76	18.63	26.74
LFCC+WIF (Gamma.)	5.51	10.19	18.45	**7.86**	**10.95**	16.70
EFCC+WIF (Gamma.)	5.21	10.07	18.15	8.17	11.07	16.27
WSSCC+WIF (Gamma.)	**5.03**	**8.09**	**13.78**	8.23	11.07	**15.91**

benefits for whispered speech. This makes it difficult to find a speaker feature representation that stores speaker identity information invariant across both vocal efforts. More research is needed to find vocal effort invariant features.

Table 17: EER(%) comparison in W/W condition using the two feature representations more robust to noise (see Table 16) and *normal/whispered* speech detector in [43]. Last row reports detection error rate for the *normal/whispered* speech detector

	SNR level		
Feature set	**15 dB**	**10 dB**	**5 dB**
WIF (Gammatone)	6.90	10.12	16.40
WSSCC+WIF (Gammatone)	6.09	9.43	15.23
N/W detection error (%)	1.03	2.01	5.54

Frequency warping strategies, in the matched condition for whispered speech showed interesting results. Simple approaches such as mel and linear scales showed to outperform the WSS scale, which was designed specifically for whispered speech. This WSS scale divides the frequencies into several critical bands from 0 Hz to 4 kHz giving more emphasis to the frequencies where the resonance peaks of F1 and F3 are located. We found that the only advantage given by this strategy is an error rate reduction in the mismatched condition. While the mel scale places emphasis on lower frequencies around F1 and F2, WSS can better handle the mismatch condition due to the lower variation of the third formant between normal and whispered speech relative to F1 and F2 [21]. Notwithstanding, the WSS scale showed useful in scenarios involving babble noise for both whispered and normally-voiced speech.

In addition, we found that whispered speech speaker verification performance was higher for female speakers. This suggests that female whispered speech carries more speaker-specific information that is captured by the investigated features. In fact, most of the recent published research in the field has been done only with females [6, 7], thus making the improvements seem more noticeable. This gender-dependency may be due to the fact that formant shifts are more noticeable in male speech than in female. As seen in Figure 4, and as previously reported in the literature [21], F1 shifts can be up to 71% for men and 52% for women ; F2 shifts can be up to 24% for men and 20% for women ; and F3 shifts can be of 10% and 4.8%, respectively [21]. Further investigation into this gender dependency is still needed.

Regarding robustness to noise, we can observe that LFCC and EFCC outperform MFCC features. One explanation can be that babble noise highly affects low frequencies, mostly between 0 Hz and 1 kHz. As a consequence, frequency warping strategies placing more emphasis in such band (such as the mel scale) will suffer higher degradation. The linear scale, in turn, gives equal relevance to all frequencies. Moreover, linear and exponential scales are not very different in the range between 0 and 4 kHz, as shown in [6]. The fact that WSSCC does not emphasize lower frequencies but place more emphasis in certain bands where there is highly discriminative information, helps to explain why WSSCC achieved the best performance amongst the tested cepstral based features in a noisy environment. Additionally, WIF features also showed high performance in noisy environments thus suggesting that phase information assists with noise robustness for whispered speech.

7 Conclusions

In this paper, the speaker verification (SV) task based on whispered speech recordings was addressed. More specifically, the performance bounds of a standard GMM–UBM SV system were obtained using several strategies, such as frequency warping, sub-band analysis, alternate feature representations, feature combination, as well as class-dependent modeling (i.e., speaking-style and gender-specific). Our experimental evaluation shows that mismatch *train/test* conditions can highly affect the performance of a SV system, independent of the feature representation used. As in previous studies in adjacent areas, it was shown that in order for a SV system to handle both normal and whispered speech for practical applications, speaker model training had to involve data of both vocal efforts. Such approach, however, resulted in poorer verification performance for normal speech. To overcome this limitation, speaking–style dependent models and gender-specific models where used. In the latter scenario, female speakers were seen to benefit the most. Overall, feature representations evaluated here have been mainly proposed for normal-voiced speech applications, thus suggesting that alternate feature representations, tuned for whispered speech speaker verification, are still needed.

Lastly, regarding noise robustness, alternative frequency warping techniques to extract cepstral coefficients and AM-FM based features showed to offer more advantages in noisy environments than conventional MFCC features.

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

[1] D. O'Shaughnessy. *Speech communications - human and machine (2. ed.)*. IEEE Press Editorial Board, Piscataway, NJ, 2000.

[2] T. Kinnunen and H. Li. An overview of text-independent speaker recognition : From features to supervectors. *Speech Communication*, 52(1) :12–40, January 2010.

[3] K. K. Paliwal and K. Yao. Robust speech recognition under noisy ambient conditions. In H. Aghajan, R. Lopez-Cozar, and J.C. Augusto, editors, *Human-Centric Interfaces for Ambient Intelligence*, volume 1, pages 135–162. Elsevier, Academic Press, San Diego, California, 1st edition, 2010.

[4] S. Squartini, E. Principi, R. Rotili, and F. Piazza. Environmental robust speech and speaker recognition through multichannel histogram equalization. *Neurocomputing*, 78(1) :111–120, February 2012.

[5] X. Fan and J. H. L. Hansen. Speaker identification with whispered speech based on modified LFCC parameters and feature mapping. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4553–4556, April 2009.

[6] X. Fan and J. H. L. Hansen. Speaker identification within whispered speech audio streams. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5) :1408–1421, July 2011.

[7] X. Fan and J.H.L. Hansen. Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams. *Speech Communication*, 55(1) :119–134, January 2013.

[8] M. Grimaldi and F. Cummins. Speaker identification using instantaneous frequencies. *IEEE Transactions on Audio, Speech, and Language Processing,*, 16(6) :1097–1111, August 2008.

[9] P. Zelinka, M. Sigmund, and J. Schimmel. Impact of vocal effort variability on automatic speech recognition. *Speech Communication*, 54(6) :732–742, July 2012.

[10] T. Ito, K. Takeda, and F. Itakura. Analysis and recognition of whispered speech. *Speech Communication*, 45(2) :139–152, February 2005.

[11] C. Hanilci, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, and F. Ertas. Speaker identification from shouted speech : Analysis and compensation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8027–8031, May 2013.

[12] N.J. Lass, L.T. Waters, and V.L. Tyson. Speaker sex identification from voiced, whispered, and filtered isolated vowels. *Journal of the Acoustical Society of America*, 59(3) :975–678, 1976.

[13] V.C. Tartter. Identifiability of vowels and speakers from whispered syllables. *Perception & Psychophysics*, 49(4) :365–372, April 1991.

[14] Q. Jin, S.-C. Jou, and T. Schultz. Whispering speaker identification. In *IEEE International Conference on Multimedia and Expo*, pages 1027–1030, July 2007.

[15] A. Avila, M. Sarria-Paja, F. Fraga, and T. Falk. The effect of speech rate on automatic speaker verification : a comparative analysis of gmm-ubm and i-vector based methods. In *Proc. Audio Engineering Conference (AES-Brazil)*, pages –, May 2014.

[16] S.T. Jovicic and Z. Saric. Acoustic analysis of consonants in whispered speech. *Journal of Voice*, 22(3) :263–274, May 2008.

[17] I.B. Thomas. Perceived pitch of whispered vowels. *Journal of the Acoustical Society of America*, 46(2B) :468–470, 1969.

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

[19] M.F. Schwartz and H.E. Rine. Identification of speaker sex from isolated, whispered vowels. *Journal of the Acoustical Society of America*, 44(6) :1736–1737, 1968.

[20] V.C. Tartter. What's in a whisper ? *Journal of the Acoustical Society of America*, 86(5) :1678–1683, 1989.

[21] H.R. Sharifzadeh, I.V. McLoughlin, and M. Russell. A comprehensive vowel space for whispered speech. *Journal of Voice*, 26(2) :49–56, March 2012.

[22] R.W. Morris and M.A. Clements. Reconstruction of speech from whispers. *Medical Engineering & Physics*, 24(7-8) :515–520, October 2002.

[23] H.R. Sharifzadeh, I.V. McLoughlin, and F. Ahmadi. Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec. *IEEE Transactions on Biomedical Engineering*, 57(10) :2448–2458, October 2010.

[24] F. Ahmadi, I.V. McLoughlin, and H.R. Sharifzadeh. Analysis-by-synthesis method for whisper-speech reconstruction. In *Proc. IEEE Asia Pacific Conference on Circuits and Systems*, pages 1280–1283, December 2008.

[25] X. Fan and J.H.L. Hansen. Speaker identification for whispered speech based on frequency warping and score competition. In *Proc. INTERSPEECH*, pages 1313–1316, 2008.

[26] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds. The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective. *Speech Communication*, 31(2-3) :225–254, June 2000.

[27] 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*, 19(4) :788–798, May 2011.

[28] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5) :980–988, July 2008.

[29] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel. An i-vector extractor suitable for speaker recognition with both microphone and telephone speech. In *Proc. Odyssey Speaker and Language Recognition Workshop*, June 2010.

[30] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3) :19–41, January 2000.

[31] L. Besacier and J.-F. Bonastre. Subband approach for automatic speaker recognition : Optimal division of the frequency domain. In *Proc. First International Conference on Audio- and Video-based Biometric Person Authentication*, Lecture Notes in Computer Science, pages 195–202, March 1997.

[32] ITU-T P.56. *Objective measurement of active speech level.* International Telecommunication Union, 1993.

[33] Z. Tao, X.-J. Zhang, H.-M. Zhao, and W. Kulesza. Noise reduction in whisper speech based on the auditory masking model. In *Proc. International Conference on Information Networking and Automation*, volume 2, pages 272–277, October 2010.

[34] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *Proc. Odyssey Speaker and Language Recognition Workshop*, June 2001.

[35] K. McDougall. Speaker-specific formant dynamics : An experiment on Australian english /aI/. *Speech, Language and the Law.*, 11(1) :103–130, June 2004.

[36] K. McDougall and F. Nolan. Discrimination of speakers using the formant dynamics of /u :/ in British english. In *Proc. 16th International Congress of Phonetic Sciences*, pages 1825–1828, August 2007.

[37] S.O. Sadjadi and J.H.L. Hansen. Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions. In *Proc. INTERSPEECH*, pages 2138–2141, 2010.

[38] P. Clark and L.E. Atlas. Time-frequency coherent modulation filtering of nonstationary signals. *IEEE Transactions on Signal Processing*, 57(11) :4323–4332, November 2009.

[39] R.F. Lyon, A.G. Katsiamis, and E.M. Drakakis. History and future of auditory filter models. In *Proc. IEEE International Symposium on Circuits and Systems*, pages 3809–3812, June 2010.

[40] M. Sarria-Paja, T.H. Falk, and D. O'Shaughnessy. Whispered speaker verification and gender detection using weighted instantaneous frequencies. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7209–7213, May 2013.

[41] M. Senoussaoui, P. Kenny, N. Brummer, E. de Villiers, and P. Dumouchel. Mixture of plda models in I-Vector space for Gender-Independent speaker recognition. In *Proc. INTERSPEECH*, pages 25–28, August 2011.

[42] J. Alam, P. Kenny, and D. O'Shaughnessy. Low-variance multitaper mel-frequency cepstral coefficient features for speech and speaker recognition systems. *Cognitive Computation*, 5(4) :533–544, 2013.

[43] M. Sarria-Paja and T.H. Falk. Whispered speech detection in noise using auditory-inspired modulation spectrum features. *IEEE Signal Processing Letters*, 20(8) :783–786, August 2013.

[44] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. Automatic speech recognition and speech variability : A review. *Speech Communication*, 49(10-11) :763–786, October-November 2007.

[45] T.H. Falk and W.-Y. Chan. Modulation spectral features for robust far-field speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1) :90–100, January 2010.

[46] J.-W. Hung, W.-H. Tu, and C.-C. Lai. Improved modulation spectrum enhancement methods for robust speech recognition. *Signal Processing*, 92(11) :2791–2814, November 2012.

[47] M. Matsuda and H. Kasuya. Acoustic nature of the whisper. In *Proc. EUROSPEECH*, pages 133–136, 1999.

[48] H.R. Sharifzadeh, I.V. McLoughlin, and F. Ahmadi. Spectral enhancement of whispered speech based on probability mass function. In *Proc. Sixth Advanced International Conference on Telecommunications*, pages 207–211, 2010.

[49] M. Slaney. An efficient implementation of the patterson-holdsworth auditory filter bank. Technical report, Apple Computer – Perception Group, 1993.

Larson Davis
PCB
PCB
PCB
PCB
Total Customer Satisfaction
PCB PIEZOTRONICS