WHISPERED SPEAKER VERIFICATION AND GENDER DETECTION USING WEIGHTED INSTANTANEOUS FREQUENCIES

Milton Sarria-Paja, Tiago H. Falk and Douglas O'Shaughnessy

Institut National de la Recherche Scientifique (INRS-EMT), University of Quebec
Montréal, Quebec, Canada.

ABSTRACT

In this paper, automatic speaker verification and gender detection using whispered speech is explored. Whispered speech, despite its reduced perceptibility, has been shown to convey relevant speaker identity and gender information. This study compares the performance of a GMM-UBM speaker verification system trained with normal and whispered speech under different matched and mismatched conditions, and describes the benefits of adaptation in a speaking-style independent model to handle both vocal efforts. It is shown that performance improvements can be achieved by using speaking-style and gender dependent models, as well as by adding features based on the AM-FM signal representation. Moreover, the AM-FM based features showed to be more discriminative than classical MFCCs for whispered speech gender detection. Experimental results suggest that whispered speech carries sufficient information for reliable automatic speaker identification.

Index Terms— Whispered speech, gender detection, speaker verification, GMM, instantaneous frequency.

1. INTRODUCTION

Recent advances in speaker recognition technology have focused attention on robustness to different noise or recording conditions and channel or microphone effects. However there are still different research problems that have received little attention, and require more effort to make advances towards the understanding of speech communication. That is the case when there are changes in the vocal effort, which have proven to affect significantly the performance of automatic speech recognition and speaker recognition systems [1, 2, 3]. Particularly, whispered speech exhibits significant differences with normal phonated speech, being the main physical difference the complete lack of vocal folds vibration. Furthermore there are also changes in the vocal tract configuration, resulting in formant shifts toward higher frequencies, especially for the lower formants [4]. It is also well known that whispered speech has gained great attention in applications where users want to protect the content of their spoken words, e.g., in mobile telephone banking. However, despite the many applications and documented properties of whispered speech [5, 6, 7, 8], study of this particular vocal effort level is still limited and more research is needed to explore its properties and use in existing speech-enabled applications.

For automatic speaker recognition, Gaussian mixture model (GMM) based systems have become the dominant approach for text-independent speaker recognition using Maximum a Posteriori (MAP) adaptation and mel-frequency cepstral coefficients (MFCC) as feature vectors [9, 10]. Literature also shows that gender dependent models have better performance than gender-independent ones, especially with gender-unbalanced data. Notwithstanding, changes in vocal effort during the testing phase can result in significant reduction in system performance. Despite efforts to address the mismatch problem, recent studies have shown that the best solution is to include small portions of whispered speech to adapt the models [11, 2]. This strategy can improve significantly the performance of recognition systems, thus allowing for normal and whispered speech to be handled. Nevertheless, different authors suggest that for optimal applications, it is better to have dedicated models for each vocal effort and select the most likely model according to the detected vocal effort [5, 3].

In this study, besides the classical MFCC features, we explore two different approaches to characterize whispered speech for speaker verification (SV) and gender detection. Comparison of the characterization schemes is performed using a standard SV system based on GMM and MAP adaptation in two scenarios: *i*) Speaking-style independent model, using a fixed length of normal speech and variable length of whispered speech for training and *ii*) Dedicated models for whispered speech, using both gender independent and dependent models. Our experimental results show that accurate whispered-speech gender detection can be achieved and systems based on AM-FM features outperform those based on MFCCs. The developed speaker verification system using only whispered speech and gender specific models was shown to provide reliable accuracy, but additional efforts are still needed in order to achieve performance figures inline with those obtained with normally-phonated speech.

The remainder of this paper is organized as follows. Section 2 describes the SV system and feature representations. Sections 3 and 4 present the experimental setup and results, respectively. Lastly, conclusions are presented in Section 5.

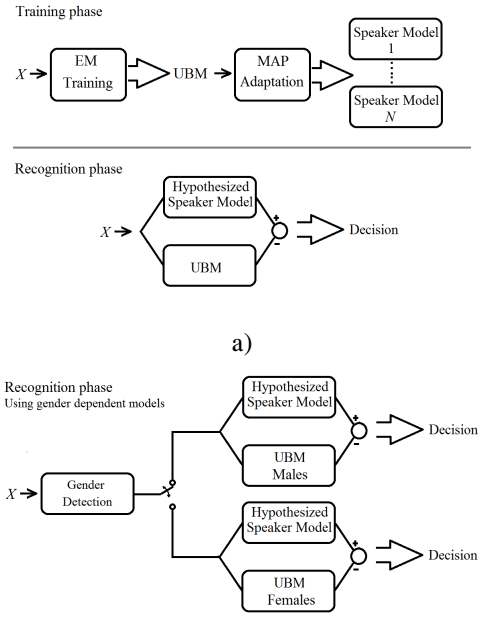

Fig. 1. a) General and b) gender-dependent SV systems.

2. SPEAKER VERIFICATION

This section describes the used SV system, as well as the feature representations explored here.

2.1. Pre-processing and SV system

In our experiments, pre-processing consisted of downsampling the speech data to 8 kHz, normalizing to -26 dBov (dB overload) using the ITU-T P.56 speech voltmeter [12], and pre-emphasizing using a first order FIR filter with constant $a = 0.97$. For the recognition system, the classical M-Gaussian universal background model (UBM) is constructed using the Expectation Maximization (EM) algorithm and the data available for training from all the speakers. Then a GMM per speaker is obtained by using MAP adaptation [10], as depicted by Fig. 1 (a). The detection error tradeoff (DET) curve is used for performance comparisons.

2.2. Feature extraction

For each speech recording, a sequence of N-dimensional feature vectors $X = \{\boldsymbol{x}_1, \cdots \boldsymbol{x}_n, \cdots \boldsymbol{x}_T\}$ were extracted, where T represents the total number of frames in the recording, and $\boldsymbol{x}_n$ the N-dimensional feature vector indexed at discrete time n. Here, two feature sets were explored:

2.2.1. Mel-frequency cepstral coefficients (MFCC)

The speech signal $s(n)$ is analyzed in short overlapping frames, each frame is multiplied by a Hamming window before the FFT computation. Then a set of 24 triangular bandpass filters spaced according to the Mel scale is used. Finally, the cosine transform is applied to the log energies obtained from the filter bank to obtain a set of cepstral coefficients. Only the first 19 cepstral coefficients are employed for training and testing. Dynamic or transitional features (Δ MFCC) derived from the difference of cepstral coefficients are not included in this study as they were shown to not provide advantage over static features in mismatch conditions [2].

2.2.2. Weighted Instantaneous Frequencies (WIFs)

Recently the AM-FM model has shown to be powerful in speaker recognition applications. The AM-FM model decomposes the speech signal to bandpass channels and characterizes each channel in terms of its envelope and phase (instantaneous frequency) [13, 14]. To decompose the speech signal in bandpass channels, the speech signal $s(n)$ is filtered by a bank of 23 critical-band gammatone filters [15, 16]. Filter center frequencies range from 50 Hz to 3528 Hz and their bandwidths are characterized by the equivalent rectangular bandwidth (ERB). Next, each analytic subband signal $s_k(n)$ is uniquely related to a real-valued bandpass signal $y_k(n)$, which is the output of the k-th gammatone filter, by the relation:

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

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. 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 subband phase in the carrier, coherent demodulation makes the important distinction between carrier and modulator phase.

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, whose values can be calculated by using either the Hilbert envelope approach or coherent demodulation. Next the values of $m_k(n)$ and $f_k(n)$ are combined to obtain the so called Weighted Instantaneous Frequencies (WIFs) using a short time approach:

$$F_k = \frac{\sum_{i=n_0}^{n_0+\tau} f_k(i) \cdot m^2(i)}{\sum_{i=n_0}^{n_0+\tau} m^2(i)}, \tag{2}$$

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 $\tau/2$. Since there are no published studies comparing each demodulation approach for speaker recognition tasks, here both approaches were investigated to test which better captures speaker and gender specific information from whispered speech.

3. EXPERIMENTAL SETUP

In our experiments the CHAINS Speech Corpus developed in [13] was used. This corpus contains the recordings of 36 speakers with three different accents: 28 speakers from Ireland (12 females, 16 males), 5 speakers from USA (3 females, 2 males) and 3 speakers from the United Kingdom (2 females, 1 male). The recordings were collected in two different sessions, the first recording session was carried out in a professional recording studio, whereas the second recording session was carried out in a quiet office environment. Additional details about the recording equipment can be found in [13]. In this particular study two speaking styles were used: *normal speech*, in which speakers read a prepared text alone at a comfortable rate, and *whispered speech* where speakers read the same prepared text but whispering. Material selected from normal and whispered speech is used as follows: the first paragraph of the *Rainbow Text* for training (average duration of 30 seconds, minimum length approximately 23 seconds), and a version of the *Cinderella story* is used for testing (average 55 seconds, minimum length approximately 48 seconds). Normal speech was recorded during the first session whereas whispered speech was recorded during the second session.

The three abovementioned feature representations are computed on a per–window basis, 32 ms windows and 50 % overlap. The estimate of short–time WIFs is expressed in kHz in order to overcome the problem associated with the nodal variances of the GMM [1]. Mean and variance normalization were used, with the assumption that channel effects are constant over the entire utterance. Prior to system evaluation and for all the experiments, the testing data was divided into overlapping segments of fixed lengths (5 seconds, 4.5 seconds overlap), and each segment is treated as a separate test utterance. This procedure is illustrated in Fig. 2 and is commonly used in speaker verification tasks [9, 1].

$$X = \{\underbrace{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \cdots \boldsymbol{x}_n}_{segment_1}, \cdots \boldsymbol{x}_T\}$$
$$\underbrace{\qquad\qquad\qquad}_{segment_2}$$

Fig. 2. Utterances for testing

4. RESULTS AND DISCUSSION

In this section, we present and discuss our obtained speaker verification experimental results.

4.1. Effects of adding whispered speech to the training set

In order to investigate the effects of adding small amounts of whispered speech to the training set, an experiment was performed with the baseline MFCC-GMM system. Experiments were conducted using fixed length of normal speech data (23 seconds) combined with variable amounts (lengths) of whispered speech per speaker for training. The number of Gaussians was adjusted to 35, which in our pilot experiments showed to be an optimal value for both speaking styles. Results of these experiments are illustrated in Fig. 3 where solid line curves show DET curves that correspond to training with normal and whispered speech and tested with normal speech. Dashed curves, in turn, correspond to training with normal and whispered speech, but tested with whispered speech (represented as "Whsp" in the figure). As can be seen, there is a significant reduction in system performance in mismatch conditions, i.e, when the system is evaluated with whispered speech and there is no data available for training of this particular speaking style. By gradually increasing the amount of whispered speech available for training, the performance of the system gradually increases without affecting the performance of the system when testing with normal speech; this corroborates previous findings [11, 2]. Nevertheless, using approximately the same amount of data of both vocal efforts we can see that better performance is achieved with normally-phonated speech than with whispered speech.

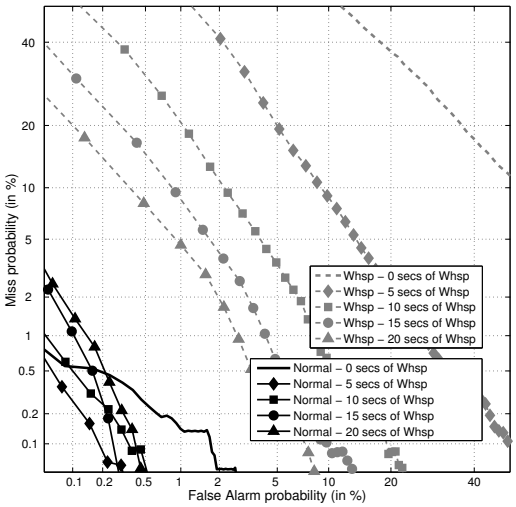

Fig. 3. DET curve to investigate the effects of different lengths of whispered speech for speaker model training

4.2. Performance of different feature representations

Next, we explore the gains obtained by using different features in the matched testing condition (i.e., train/test on whispered speech). Fig. 4 depicts the DET curves for the three features. It is observed that WIFs using coherent demodulation exhibit the worst performance. On the other hand, WIFs using the Hilbert envelope approach achieved performance inline with those obtained with the classical MFCC features. This suggests that there is speaker specific information in the phase of the acoustic signal and that an approach based on Hilbert envelopes can be used to characterize such information, at least as reliably as MFCCs. Moreover, since

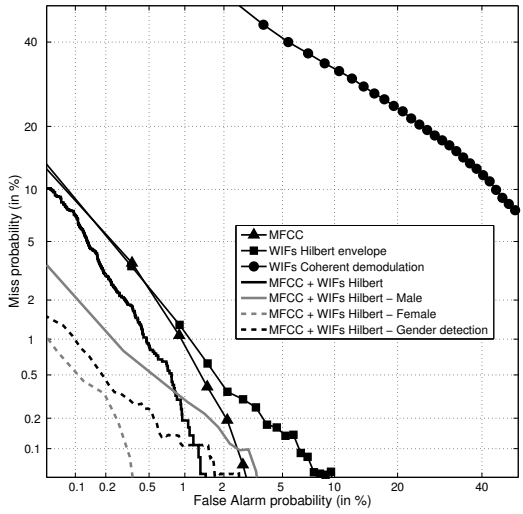

Fig. 4. DET curve to investigate the effects of different feature representations and fusion strategies in a matched train/test whispered-speech condition

MFCCs and WIFs extract complementary information from the speech signal, we explored their fusion to see if it resulted in improved performances. The results shown in Fig. 4 (continuous black line) show that this is indeed the case.

4.3. Effects of gender detection for gender-dependent SV

Lastly, to investigate the benefits of using gender dependent models, the recordings were separated by gender in the training phase. Then two UBMs are obtained and their respective speaker specific models after MAP adaptation, which allows having two independent recognition systems. The gender-dependent recognition systems are evaluated separately for each gender and results showed that there are additional benefits for female speech and a small gain for male speech (see solid gray and dashed DET curves in Fig. 4). According to these findings, it is suggested that dedicated models be used not only per speaking-style but also per gender. Hence, in order to have a completely automated system using gender-dependent models, an additional gender detection stage is needed prior to the speaker verification stage, as is illustrated in Fig. 1 (b). For this purpose, an M-component GMM per gender was trained; gender detection results for different feature representations and number of Gaussian components M are summarized in Table (1).

As can be seen, for gender detection WIF features (obtained via both Hilbert envelope and coherent demodulation approaches) outperform MFCC features. Moreover, WIF features obtained via the Hilbert envelope approach achieve close to perfect accuracy even with only two Gaussian components. 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, and even without the glottal excitation gender discrimination is a feasible task using whispered speech [17, 18].

Feature Set	**Number of Gaussians (M)**		
	2	**5**	**10**
MFCC	94.61	95.14	96.73
WIFs – Hilbert envelope	99.61	99.72	99.99
WIFs – Coherent Demodulation	85.66	95.67	97.57
MFCC + WIFs – Hilbert envelope	97.73	99.05	99.69

Table 1. Gender detection accuracy for different feature representations and number of Gaussian components

5. CONCLUSIONS

This paper has addressed the issue of speaker verification based on whispered speech. A standard GMM-UBM model was used and trained using three different feature representations, namely 1) mel-frequency cepstral coefficients, and weighted instantaneous frequencies (WIFs) obtained via 2) a Hilbert envelope approach and 3) via a coherent demodulation approach. Experimental results using a speaking-style independent approach showed that incorporation of whispered speech during training was beneficial to the task at hand. Reliable performances could be achieved once equal amounts of normally-phonated and whispered speech data were used for training of the speaker models. Notwithstanding, the obtained performance on normal speech was higher than that obtained with whispered speech.

Second, we explored the scenario of speaking-style dependent speaker verification and showed that improved results can be achieved relative to the speaking-style independent case. It was shown that the approach adopted to estimate the phase and envelope information from the speech signal has a significant impact on the discriminative capabilities of the WIF features. More specifically, using the Hilbert envelope approach resulted in improved performance relative to the coherent demodulation approach. Lastly, we explored the benefits of gender-dependent speaker verification. This step required the development of an automated whispered-speech gender identification system. Towards this end, we found that WIF features extracted via the Hilbert envelope approach achieved almost perfect gender classification. Future work will explore the robustness of the system and investigated features under noisy environmental conditions.

6. ACKNOWLEDGEMENTS

The authors acknowledge funding from the Natural Sciences and Engineering Research Council of Canada.

7. REFERENCES

[1] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 6, pp. 1097 –1111, aug. 2008.

[2] Xing Fan and J.H.L. Hansen, "Speaker identification within whispered speech audio streams," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1408 –1421, july 2011.

[3] Petr Zelinka, Milan Sigmund, and Jiri Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 732 – 742, 2012.

[4] Masahiro Matsuda and Hideki Kasuya, "Acoustic nature of the whisper," in *EUROSPEECH'99*, 1999, pp. –1–1.

[5] Taisuke Ito, Kazuya Takeda, and Fumitada Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139 – 152, 2005.

[6] M. Higashikawa, K. Nakai, A. Sakakura, and H. Takahashi, "Perceived pitch of whispered vowels-relationship with formant frequencies: A preliminary study," *Journal of Voice*, vol. 10, no. 2, pp. 155 – 158, 1996.

[7] Slobodan T. Jovicic and Zoran Saric, "Acoustic analysis of consonants in whispered speech," *Journal of Voice*, vol. 22, no. 3, pp. 263 – 274, 2008.

[8] Hamid Reza Sharifzadeh, Ian V. McLoughlin, and Martin J. Russell, "A comprehensive vowel space for whispered speech," *Journal of Voice*, vol. 26, no. 2, pp. 49 – 56, 2012.

[9] Douglas A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91 – 108, 1995.

[10] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19 – 41, 2000.

[11] Qin Jin, Szu-Chen Stan Jou, and T. Schultz, "Whispering speaker identification," in *Multimedia and Expo, 2007 IEEE International Conference on*, july 2007, pp. 1027 –1030.

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

[13] Fred Cummins, Marco Grimaldi, Thomas Leonard, and Juraj Simko, "The chains corpus: Characterizing individual speakers," in *International Conference on Speech and Computer SPECOM-2006*, 2006, pp. 431–435.

[14] P. Clark and L.E. Atlas, "Time-frequency coherent modulation filtering of nonstationary signals," *Signal Processing, IEEE Transactions on*, vol. 57, no. 11, pp. 4323 –4332, nov. 2009.

[15] Malcolm Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," Tech. Rep., Apple Computer – Perception Group, 1993.

[16] R.F. Lyon, A.G. Katsiamis, and E.M. Drakakis, "History and future of auditory filter models," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, 30 2010-june 2 2010, pp. 3809 –3812.

[17] Martin F. Schwartz and Helen E. Rine, "Identification of speaker sex from isolated, whispered vowels," *Journal of the Acoustical Society of America*, vol. 44, no. 6, pp. 1736–1737, 1968.

[18] Norman J. Lass, Karen R. Hughes, Melanie D. Bowyer, Lucille T. Waters, and Victoria T. Bourne, "Speaker sex identification from voiced, whispered, and filtered isolated vowels," *Journal of the Acoustical Society of America*, vol. 59, no. 3, pp. 975–678, 1976.