FEATURE MAPPING, SCORE-, AND FEATURE-LEVEL FUSION FOR IMPROVED NORMAL AND WHISPERED SPEECH SPEAKER VERIFICATION

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

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

ABSTRACT

In this paper, automatic speaker verification using normal and whispered speech is explored. Typically, for speaker verification systems with varying vocal effort inputs, standard solutions such as feature mapping or addition of data during parameter estimation (training) and enrollment stages result in a trade-off between accuracy gains with whispered test data and accuracy losses (up to 70% in equal error rate, EER) with normal test data. To overcome this shortcoming, this paper proposes two innovations. First, we show the complementarity of features derived from AM-FM models over conventional mel-frequency cepstral coefficients, thus signalling the importance of instantaneous phase information for whispered speech speaker verification. Next, two fusion schemes are explored: score- and feature-level fusion. Overall, we show that gains as high as 30% and 84% in EER can be achieved for normal and whispered speech, respectively, using feature-level fusion.

Index Terms— Whispered speech, AM-FM model, i-vectors, speaker verification, system fusion, feature mapping.

1. INTRODUCTION

Biometrics based applications are helping to prevent fraud by combining mathematics and digital signal processing techniques. Such technologies are burgeoning for identity management as they eliminate the need for personal identification numbers, passwords, and security questions [1]. Notwithstanding, several challenges and unresolved problems are still present hampering widespread usage. For example, biometrics recognition is compromised by external factors that may alter the patterns that are being analyzed (e.g., cuts and burns to the finger in fingerprint-based systems; ambient noise in speech-based solutions), as well as by natural human physiological factors, such as aging and disease (e.g., in facial and speech-based systems) [1, 2].

According to recent statistics, speech-based biometrics have ranked highly in consumer preference, outranking fingerprint and iris scanning solutions [3, 4]. Despite speech-based biometrics gaining grounds, two factors have posed serious threats to its performance: ambient noise and varying vocal efforts. Ambient noise has detrimental effects on speaker recognition performance, particularly those trained with mel-frequency cepstral coefficients (MFCC). As an example, accuracies as low as 7% have been reported in very noisy environments [5]. As such, over the years several speech enhancement algorithms have been proposed for environment-robust speaker recognition applications [6, 5]. Varying vocal efforts, however, have received significantly less exposure, despite their severe detrimental effects on speaker verification performance. For example, accuracies as low as 20% have been reported for whispered-speech speaker identification [7] in clean conditions. In fact, it is highly likely that customers utilizing a mobile banking application on their smartphones will whisper sensitive information.

Over the last few years, a handful of strategies have been reported to improve the performance of whispered speech speaker recognition, particularly within training/test mismatch [8, 7, 9]. Improvements, however, have been minimal. Another strategy, which has not been widely explored, is to use feature mapping. A recent study showed that such an approach can be helpful in speaker identification scenarios when the input presented is shouted speech [10]. For feature mapping, neural networks and Gaussian mixture models have been widely used in the voice conversion and voiced speech reconstruction (from whispered speech) literature [11, 12, 13]. It is not clear, however, if such mappings alter speaker identity information relevant for automated speaker recognition. This paper explores the advantages of feature mapping alongside other mismatch compensation strategies.

Typically, two main strategies are used to handle the mismatch problem, namely, (1) multiple model recognizer, where dedicated speaker models are obtained for different vocal efforts (e.g., [14]) and (2) multi-style models, where each model is obtained from a combination of normal speech and small amounts of speech of varying vocal efforts [15, 14]. Notwithstanding, the two different methods were shown to have their advantages and disadvantages. For example, while both improve the performance of whispered speech [8, 14], multiple model training requires significant amounts of whispered speech data to obtain the speaker models, which can be hard to obtain in practice. Multi-style based systems, in turn, de-

This work was funded by the Natural Sciences and Engineering Research Council of Canada and the Administrative Department of Science, Technology and Innovation of Colombia (COLCIENCIAS).

spite requiring lower amounts of whispered speech to train the models, trade gains in whispered speech to losses in normal speech accuracy, often by the same amount [14].

The goal of this paper is to propose a practical strategy to design a speaker verification system capable of handling two different speaking styles by exploring the use of feature mapping, alternate feature representations, as well as two fusion strategies, namely score- and feature-level fusion.

2. AUTOMATED SPEAKER VERIFICATION

In speaker recognition the two most popular tasks are speaker identification (SI) and speaker verification (SV). Commonly, SV exhibits greater practical applications related to SI, especially in access control and identity management applications. Whispered speech, in the past, has been explored mainly for the SI problem within a reduced amount of speakers or using only female speakers [8, 7, 9]. This paper aims to fill this gap by combining three different databases with both male and female speakers.

2.1. Feature extraction

2.1.1. Mel-frequency cepstral coefficients (MFCC)

Standard MFCC features were extracted following conventional steps, including pre-emphasis, 27 triangular mel-spaced bandpass filters, 25ms windows, and 40% overlap. Thirteen coefficients were extracted, including the 0th coefficient (log-energy), and were appended by delta and double-delta coefficients, each found using a 9-point anti-symmetric filter to avoid phase distortion. Lastly, cepstral mean and variance normalization was utilized during active speech periods to remove unwanted linear channel effects.

2.1.2. AM-FM derived features

The AM-FM model decomposes the speech signal into bandpass channels and characterizes each channel in terms of its envelope and phase (instantaneous frequency) [7]. 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. Originally, in [7], it was proposed to use a 80-channel Gabor filterbank, however in our experiments a 27-channel gammatone filterbank was used instead, with filter center frequencies (fc_k) ranging from 100 Hz to 7000 Hz and bandwidths characterized by the mel scale. It was concluded from a pilot experiment that this is an optimal setting for our purposes. After filtering, each analytic subband signal $s_k(n)$ is uniquely related to a real–valued bandpass signal $y_k(n)$ by $s_k(n) = y_k(n) + j \cdot \hat{y}_k(n)$, where $\hat{y}_k(n)$ represents the Hilbert transform of $y_k(n)$. In this work, the Hilbert envelope approach is used to decompose each analytic signal in terms of its envelope and phase. For the sake of notation, let $a_k(n)$ denote the low–frequency modulator and $f_k(n)$ the instantaneous frequency for each bandpass signal.

Here, the set of features based on the AM-FM model is the so called Weighted Instantaneous Frequencies (WIF). These features are computed by combining the values of $a_k(n)$ and $f_k(n)$ using a short-time approach, more specifically:

$$F_k = \frac{\sum_{i=n_0}^{n_0+\tau} f_k(i) \cdot a_k^2(i)}{\sum_{i=n_0}^{n_0+\tau} a_k^2(i)}, \quad k = 1, \ldots, 27, \tag{1}$$

where n_0 is the starting sample point and τ is the length of the time frame. To maintain the analogy with the MFCCs, here the WIF features were also computed on a per-window basis using a 25 ms window with 40% overlap. Pre-emphasis and feature normalization were not required and WIF features are expressed in kHz, as suggested by [7].

2.2. Feature mapping

Two feature mapping techniques were evaluated in our experiments. The first is the classical Gaussian mixture model (GMM) regression [16]. Such method models both the source and the target feature vectors using a joint density GMM of aligned normal and whispered speech features. Model parameters are estimated using the standard expectation-maximization (EM) algorithm. With the estimated parameters a mapping function is formulated to compute the minimum mean square error estimate of the target feature vectors. This mapping can be used to transform whispered to normal speech features or vice-versa, by properly defining source and target feature vectors. Details can be found in [16]. Herein, full covariance matrices were used and the number of Gaussians was varied on the training stage. The best results are reported using a model with 128 components.

The second technique is based on neural networks, which have been shown useful in the voice conversion literature [11]. Here, we explore the use of emerging deep neural networks (DNN), which have achieved state-of-the-art results across several research domains. We explore their flexibility in learning the direct mappings between the whispered and normal speech features. Two stacked pre-trained autoencoders [17] with 512 hidden units each were used in our experiments. This technique is also used to transform whispered to normal speech features or vice-versa, depending on the specific setting.

2.3. i-vectors/PLDA approach

Current state-of-the-art speaker verification systems are based on i-vector extraction features and Probabilistic Linear Discriminant Analysis (PLDA) based scoring. A brief description of these two tools is given below for the sake of completeness. In the experiments herein, the open-source *Bob* signal processing toolbox was used [18]. For i-vector extraction, speaker and session-dependent supervectors of concatenated GMM means are modeled as $M = m + T\phi$, where m is the speaker- and channel-independent supervector, $T \in \mathbb{R}^{CF \times D}$ is a rectangular matrix of low rank covering the important variability (total variability matrix) in the supervector space.

C, F and D represent, respectively, the number of Gaussians in the universal background models (UBM), the dimension of the acoustic feature vector and the dimension of the total variability space. Finally $\phi \in \mathbb{R}^{D\times 1}$ is a random vector with density $\mathcal{N}(0, I)$ and referred to as the identity vector or *i-vector* [19]. This procedure is complemented with some post-processing techniques such as linear discriminant analysis (LDA), whitening, and length normalization to remove nuisance effects in the total variability space. The interested reader is referred to [19] for more complete details.

Probabilistic linear discriminant analysis (PLDA) based scoring is used to compare a test utterance and a target speaker. This approach was formulated in [20] as $\phi_{ij} = \mu + Vy_i + Ux_{ij} + \varepsilon_{ij}$, where ϕ_{ij} is the i-th feature vector associated to the j-th speaker, the matrices $V \in \mathbb{R}^{D\times P}$ and $U \in \mathbb{R}^{D\times M}$ span the between- and within- individual spaces, μ is a global mean, $y_i \sim \mathcal{N}(0, I)$ and $x_{ij} \sim \mathcal{N}(0, I)$ are hidden variables in the spaces spanned by V and U, respectively, and the residual $\varepsilon_{ij} \sim \mathcal{N}(0, \Sigma)$ is defined to be Gaussian with zero mean and diagonal covariance Σ. In a verification scenario, there are two possible hypotheses: 1) ϕ_{test} and ϕ_{enrol} share the same class, and 2) ϕ_{test} and ϕ_{enrol} are from different classes. Lastly, the corresponding score can be obtained by computing the log-likelihood between the two hypotheses, which is given by $s = \ln(P(\phi_{test}, \phi_{enrol})) - \ln(P(\phi_{test})P(\phi_{enrol}))$; details can be found in [20, 21].

2.4. Fusion strategies

Two fusion schemes were also investigated in this paper: *i) score-level fusion* and *ii) feature fusion*. In the former, separate data (different from background and target speakers) is needed and two systems trained on MFCC and WIF feature sets, respectively, are evaluated using an unseen evaluation set (see Section 3.1). A logistic regression function is then found to map the evaluation scores into a final decision using the Brosaris toolkit [22]. With feature fusion, in turn, MFCC and WIF features are concatenated into a final 66-dimensional feature vector (39 MFCC + 27 WIF). Principal component analysis is then performed to remove redundant features and only the top-30 components are kept as features.

3. EXPERIMENTAL RESULTS

3.1. Corpus description

In our experiments, three different databases were used, the CHAINS (Characterizing Individual Speakers) speech corpus [7], wTIMIT (whispered TIMIT) [23] and TIMIT databases [24]. The CHAINS and wTIMIT databases contain normal and whispered speech. Table 1 presents details about the number of speakers and recordings per speaker.

Speakers from the three databases were divided in two disjoint sets, one for development (parameter estimation of GMM, T-matrix and PLDA) and the other for enrollment and testing (target speakers). Recordings from 462 speakers from TIMIT database and 14 speakers from wTIMIT were included in the development set. Recordings from 100 speakers from the TIMIT database, 24 speakers from wTIMIT and 36 speakers from CHAINS, in turn, were included in the target speakers set. Average duration for all speech recordings is 4.5 seconds. To characterize the baseline system, we included only normal speech recordings from both the development and target speakers sets. During enrollment eight recordings per speaker were used; for testing, however, we used two recordings per speaker, and if there are whispered speech recordings available, then two additional utterances were included.

Database	**No. of speakers**		**Recordings/speaker**	
	Female	Male	Normal	Whisper
TIMIT	192	438	10	–
wTIMIT	24	24	450	450
CHAINS	16	20	37	37

Table 1. Details of the three databases used in this work

To train the score-level fusion system we selected an independent set of speakers, 68 from the TIMIT database and 10 from the wTIMIT database, to create a new evaluation list. For enrollment, a configuration similar to the one used for the original evaluation list was used, including eight additional recordings of whispered speech for the 10 speakers of wTIMIT. For the new evaluation list, in order to have approximately the same amount of target and impostor scores from each speaking style, two recordings of normal speech and 15 recordings of whispered speech per speaker were used.

3.2. Results and discussion

Table 2 reports the equal error rate (EER) results obtained with the standard i-vector/PLDA based SV paradigm using the conventional MFCC features. Three cases are reported: *Baseline* illustrates the scenario where only normal speech is available for training and enrollment, no feature mapping is applied and no whispered speech features were used for parameter estimation. *Case a)*: illustrates the case where normal speech features from the enrollment set were mapped to whispered ones using GMM or DNN mapping functions. *Case b)*, in turn, exemplifies the scenario where whispered speech features in the test set were mapped to normal speech ones using the GMM/DNN mapping functions. Latter case assumes an oracle normal/whisper classification system, thus the results for normal speech are unaffected. For clean conditions, this is not an unrealistic assumption [25]. In both *Case a)* and *b)*, whispered speech features from the development set were also included during parameter estimation because by using only the mapped features slight improvements were observed (in the order of 2%). The three columns represent no feature mapping (none) and GMM or DNN based mapping.

As can be seen from the Table 2, both feature mappings add some gains when testing with whispered speech, with relative improvements of up to 37%. Table 3 in turn, compares the feature mappings in terms of mean cepstral distance and

Scenario	Normal			Whispered		
	Feature Mapping					
	none	GMM	DNN	none	GMM	DNN
Baseline	**2.93**	–	–	**28.00**	–	–
Case a	4.06	8.75	6.25	19.15	24.17	20.00
Case b	4.06	4.06	4.06	19.15	17.50	21.07

Table 2. EER comparison with the baseline system and the two feature mappings in different scenarios. For these results $C = 256$, and $D = 200$.

Evaluation Measures	Norm to Whsp		Whsp to Norm	
	GMM	DNN	GMM	DNN
MCD	13.84	12.78	13.96	12.75
ε_{rms}	0.644	0.596	0.649	0.595

Table 3. Evaluation measures comparison between the two feature mapping techniques. MCD - Mean Cepstral Distance and ε_{rms} - root mean square error

root mean square error. In terms of these measures the DNN performs better than the GMM; however this is not reflected in the EER results. As such, the GMM mapping seems to be optimal to compensate when whispered speech is present during testing. These results also show that the addition of whispered speech during parameter estimation does not suffice to boost performance when testing with this speaking style because, in this case, whispered speech data does not contain enough inter-speaker variability. As has been shown before, addition of recordings from target speakers, even in small amounts, is the solution that seems to effectively close the gap in performance between normal and whispered speech [8, 26]. This, however, comes with an increase in the error rate for normal speech, the goal in score and feature fusion is to overcome this limitation and develop a system that performs well with both whispered and normal speech.

Table 4, in turn, shows the EER values for the two proposed fusion schemes under two cases and compares it with the standard MFCC based system. *Case 1* exemplifies the scenario where whispered speech is available during both training and testing, but not during enrollment. *Case 2*, on the other hand, exemplifies the setting where whispered speech is available in all three stages. Relative to results for MFCC features, the gains attained with score-level fusion were approximately 61% and 20% for normal and whispered speech, respectively. By adding whispered speech during training and enrollment, (i.e., Case 2) score-level fusion resulted in slight increases in EER for normal speech, but in significant drops in EER for whispered speech. Overall, the final performance attained with score-level fusion was 53% and 38% lower than using only MFCC, for normal and whispered speech respectively. As can be seen, score level fusion and adding small amounts of whispered speech to training and enrollment stages has shown to be a useful strategy to achieve reliable results for both normal and whispered speech.

Scenario	Normal			Whisper		
	Fusion level					
	MFCC	SCF	FF	MFCC	SCF	FF
Case 1	4.06	1.56	0.74	19.15	15.49	17.69
Case 2	5.56	2.57	2.03	8.90	5.45	4.35

Table 4. EER comparison between the two fusion schemes: score and feature fusion with the standard MFCC based system. For these results $C = 256$, and $D = 200$. SCF stands for score fusion and FF stands for feature fusion.

For feature-level fusion, also significant gains were observed for normal speech under Case 1, 80% lower EER relative to MFCC alone, and 7% for whispered speech. Case 2, once whispered speech was incorporated during training and enrollment, the obtained EER was 63% lower for normal speech, relative to using only MFCC features and 51% lower for whispered speech. Results from standalone WIF features were not included as they were slightly better than MFCC but the benefits of their use was observed in the fusion schemes.

According to the results, feature-level fusion showed to play an important role for both normal and whispered speech speaker verification, this strategy helps to avoid negative effects in normal speech performance while adding whispered speech during training and enrollment. This is an important finding for practical systems relying on either type of vocal effort input. An additional advantage of the feature-level fusion scheme is that it relies only on training of the PCA dimensionality reduction step and of one speaker verification system. The score-level fusion scheme, on the other hand, relies on the training of two automated SV systems (one for each feature set), as well as one score fusion scheme. Finally, experiments suggest that whispered speech can carry as much speaker identity information as normal speech, but such information has to be properly extracted and WIFs appear to be a good set of features to complement the classical MFCC and extract additional information from speech recordings.

4. CONCLUSION

This paper has addressed the issue of speaker verification based on whispered speech. Train/test mismatch conditions have been shown to be a serious challenge for automated SV systems and previous studies have suggested the inclusion of whispered speech during training and enrollment stages. Here, we have shown that simple inclusion of small amounts of whispered data combined with feature mapping techniques does not suffice and two additional fusion schemes are proposed: score- and feature-level fusion. Over simple addition of whispered speech during training, gains by as much as 34% and 36% in EER could be achieved with score-level fusion and feature fusion, respectively. Moreover, we have shown the importance of features that rely on instantaneous phase information for the task at hand. When combined with conventional MFCC features, complementary speaker identity information was observed.

5. REFERENCES

[1] J. Unar, W. Chaw Seng, and A. Abbasi, "A review of biometric technology along with trends and prospects," *Pattern Recognition*, vol. 47, no. 8, pp. 2673 – 2688, 2014.

[2] K. Saeed, "A note on problems with biometrics methodologies," in *Proc. ICBAKE*, Sept 2011, pp. 20–22.

[3] R. O'Neil King, "Speech and voice recognition white paper," Biometrics Research Group, Inc., Tech. Rep., May 2014.

[4] "Global mobile biometrics market 2015-2019," CompaniesandMarkets.com, Tech. Rep., 2015.

[5] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Audio, Speech, Language Process.*, vol. 15, no. 5, pp. 1711–1723, July 2007.

[6] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun*, vol. 52, no. 1, pp. 12–40, January 2010.

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

[8] X. Fan and J. H. L. Hansen, "Speaker identification within whispered speech audio streams," *IEEE Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1408–1421, July 2011.

[9] ——, "Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams," *Speech Commun*, vol. 55, no. 1, pp. 119–134, January 2013.

[10] C. Hanilci, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, and F. Ertas, "Speaker identification from shouted speech: Analysis and compensation," in *Proc. ICASSP*, May 2013, pp. 8027–8031.

[11] S. Desai, A. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 954–964, July 2010.

[12] Z. Tao, J.-H. Gu, X.-D. Tan, Y.-S. Xu, T. Han, and H.-M. Zhao, "Reconstruction of normal speech from whispered speech based on rbf neural network," in *Proc. IITSI*, April 2010, pp. 374–377.

[13] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1859–1872, Dec. 2014.

[14] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Commun*, vol. 54, no. 6, pp. 732–742, July 2012.

[15] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Commun*, vol. 45, no. 2, pp. 139–152, February 2005.

[16] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, vol. 1, May 1998, pp. 285–288.

[17] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. ICML*, July 2008, pp. 1096–1103.

[18] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," in *20th ACM Conference on Multimedia Systems (ACMMM)*. ACM Press, Oct. 2012.

[19] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[20] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*, Oct 2007, pp. 1–8.

[21] A. Sizov, K. Lee, and T. Kinnunen, "Unifying probabilistic linear discriminant analysis variants in biometric authentication," in *Proc. S+SSPR*, 2014.

[22] N. Brummer and E. de Villiers, "The BOSARIS Toolkit User Guide: Theory, algorithms and code for binary classifier score processing," CAGNITIO Research, South Africa, Tech. Rep., 2011.

[23] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, University of Illinois, 2011.

[24] J. S. Garofolo, L. D. Consortium *et al.*, "Timit: acoustic-phonetic continuous speech corpus," 1993.

[25] M. Sarria-Paja and T. Falk, "Whispered speech detection in noise using auditory-inspired modulation spectrum features," *IEEE Signal Processing Letters*, vol. 20, no. 8, pp. 783–786, August 2013.

[26] M. Sarria-Paja, T. Falk, and D. O'Shaughnessy, "Whispered speaker verification and gender detection using weighted instantaneous frequencies," in *Proc. ICASSP*, May 2013, pp. 7209–7213.