interspeech 2008 incorporating SST 08

Spectro-Temporal Features for Robust Far-Field Speaker Identification

Tiago H. Falk and Wai-Yip Chan

Department of Electrical and Computer Engineering
Queen's University, Kingston, Ontario, Canada
E-mail: {falkt, chan}@ee.queensu.ca

ABSTRACT

Features derived from an auditory spectro-temporal representation of speech are proposed for robust far-field speaker identification. The auditory representation is obtained by first filtering the speech signal with a gammatone filterbank. A modulation filterbank is then applied to the temporal envelope of each gammatone filter output. Compared to commonly used mel-frequency cepstral coefficients (MFCC), the proposed features are shown to be more robust to mismatched conditions between enrollment and test data and are less sensitive to increasing reverberation time (RT). Experiments with simulated and recorded far-field speech show that a Gaussian mixture model based identification system, trained on the proposed features, attains an average improvement in identification accuracy of 15% relative to a system trained on MFCC. Improvements of up to 85% are attained for larger RT.

Index Terms: Speaker identification, reverberation, modulation spectrum, Gaussian mixture model, reverberation time.

1. INTRODUCTION

In today's fast-paced society, mobility and the ability to multitask have been the driving forces behind the advances in hands-free speech communications. Applications include voice activated controls in automobiles, personal computers and cellphones, as well as teleconferencing. Far-field hands-free applications, however, introduce a series of performance degrading factors for automatic speaker recognition systems. The two most prominent factors include channel mismatch (mismatch between train and test data) and reverberation.

To compensate for channel mismatch conditions, techniques such as cepstral mean substraction (CMS), mean substraction and variance normalization (CMSVN), and relative spectral (RASTA) filtering, have been proposed [1]. In reverberant environments, mismatch conditions can also occur due to different room transfer functions and varying acoustic reverberation levels[1]. In [2], room transfer functions are assumed to be time-invariant and improved speaker verification performance is attained by designing speaker models for several different room transfer functions. Online, a "room transfer function classifier" is used to determine, from the speech signal, which speaker model to use. The major limitation of such an approach is that, in practice, room transfer functions are time-varying and can change considerably with acoustic source positioning or placement of room furnishings.

Alternately, speech enhancement (dereverberation) techniques can be used to reduce the detrimental effects of reverberation. In [3], microphone arrays combined with CMS are used to improve speaker recognition performance. More recently, the use of reverberation compensation, feature warping, CMS, and multiple microphone combination was proposed [4]. Dereverberation, however, is a difficult and often ill-conditioned problem, in particular if only a single far-field microphone is available. In fact, a recent study has shown that only modest improvement in *speech* recognition performance is attained for dereverberated speech [5].

In this paper, an alternate approach is taken to improve automatic speaker identification (ASI) performance in far-field reverberant environments. In particular, a novel feature set, shown to be insensitive to increasing reverberation time (RT) *and* robust to mismatch reverberation conditions between enrollment and test data, is presented. The features are derived from an auditory spectro-temporal representation of speech. Experiments carried out with single-channel simulated reverberant speech and with multi-channel recorded reverberant speech illustrate the gains obtained by using the proposed features. Comparisons with a baseline system show that improvements in identification accuracy of up to 85% for large RT can be attained with the proposed system.

2. REVERBERANT SPEECH DATABASES

In this section, a description is given of the two reverberant speech databases used in our experiments.

2.1. Single-Channel Simulated Reverberant Speech

The SIREAC (SImulation of REal ACoustics) tool [6] is used to artificially generate reverberant speech with different RTs. The room impulse responses represent typical office environments and RT values between 0.2-0.5 s (0.1 s increments)

[1] Reverberation levels are often quantified by means of the so-called reverberation time (RT) which is the interval required for the sound energy to decay by 60dB after the sound source is turned off.

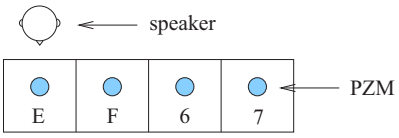

Fig. 1. PZM microphone setup at ICSI meeting room.

and 1 s are simulated; the latter simulates a larger meeting room. In our experiments, reverberant speech is generated by corrupting a subset of the TIMIT database with the SIREAC tool. Speech files are downsampled to 8 kHz and utterances from 340 of the 630 speakers are used. Of the ten available utterances per speaker, eight are used to train the speaker model and two are kept for testing. After reverberation is introduced, this amounts to 3400 test speech signals.

2.2. Multi-Channel Recorded Reverberant Speech

The ICSI Meeting Corpus [7] is used to test the proposed algorithm on real multi-channel reverberant audio recordings. The four table microphones are omnidirectional pressure zone microphones (PZM) and are arranged according to Fig. 1. Meetings involved anywhere from three to 10 participants (averaging six) with levels of English language fluency ranging from fluent to "hard-to-transcribe." The full corpus contains speech files for 53 speakers; however, data for only 29 speakers has been made freely available by ICSI (the full corpus is licensed by LDC). Of these 29 speakers, speech files from 23 speakers are used in our experiments. The remaining six speakers did not provide sufficient material to accurately train speaker models. Of the data available, 80% is used for training and 20% is left for testing.

3. SYSTEM DESCRIPTION

In this section, the proposed auditory spectro-temporal features are described as well as the proposed ASI system.

3.1. Spectro-Temporal Features

The proposed features are extracted from an auditory spectro-temporal representation of speech commonly termed *modulation spectrum*. To obtain such a representation, the speech signal is first filtered by a bank of critical-band filters. A critical-band gammatone filterbank, with 23 filters, is used to emulate the processing performed by the cochlea. The filter center frequencies range from 125 Hz to 3.5 kHz and filter bandwidths are characterized by the equivalent rectangular bandwidth. As examples, the first and last filters have bandwidths of 38 Hz and 410 Hz, respectively. The Hilbert temporal envelope is then obtained for each of the 23 cochlear filter outputs.

Temporal envelopes are multiplied by a 256 ms window with 32 ms shifts and analyzed with an eight-filter modulation filterbank. The center frequencies and bandwidths of the modulation filters are described in Table 1. Here, 256 ms frames are used to obtain appropriate modulation frequency resolution. The auditory representation, for a given frame j, is denoted as $\mathbf{X}_j = \{\mathbf{x}_{1,j}, \ldots \mathbf{x}_{8,j}\}$, where $\mathbf{x}_{m,j}, m = 1, \ldots, 8$ is a 23 dimensional vector containing one energy value (normalized by the maximum energy obtained over all modulation frequency bands) for each cochlea frequency band. Henceforth, the notation $\bar{\mathbf{X}}$ will be used to denote the auditory representation $\mathbf{X}_j$ averaged over active speech frames.

Table 1. Modulation filter center frequencies (f_c) and bandwidths (BW) expressed in Hz.

	Modulation Frequency Band Index							
	1	2	3	4	5	6	7	8
f_c	4.0	6.5	10.7	17.6	28.9	47.5	78.1	128.0
BW	2.4	3.9	6.5	11.0	18.2	29.0	47.6	78.8

The motivation for using a spectro-temporal representation of speech originates from the known fact that the diffuse reverberant tail can be modeled as an exponentially damped Gaussian white noise process [8]. As RT increases, the signal attains more Gaussian white-noise like properties. In addition, it is known that the Hilbert envelope can contain frequencies up to the bandwidth of its originating signal [9]. For clean (unreverberated) speech, Hilbert envelopes contain frequencies ranging from 2 Hz - 20 Hz [10, 11] with peaks at approximately 4 Hz, corresponding to the syllabic rate of spoken speech [12]. With reverberant speech, however, higher Hilbert envelope frequencies (henceforth refereed to as modulation frequencies) are also expected due to the "whitening" effects of the diffuse tail. In fact, since the cochlear filter centered at the lowest acoustic frequency (125Hz) has a bandwidth of 38Hz, it is expected that reverberation effects be more pronounced beyond such frequencies (i.e., modulation frequency bands 5-8, c.f. Table 1).

The plots in Fig. 2 assist in illustrating the effects of increasing RT on different modulation frequency bands. Subplots (a)-(d) depict $\bar{\mathbf{X}}$ for a female speaker in clean and reverberant conditions with $RT = 0.4, 0.7$ and 1s, respectively. Similar behavior is observed for utterances spoken by male speakers. Note the increase in energy at higher modulation frequency bands as RT increases. As conjectured above, more pronounced reverberation effects are witnessed for modulation frequency bands 5-8. Hence, in order to devise an ASI system that is robust to increasing RT, we propose to use features extracted from the first four modulation frequency channels. The proposed ASI system is described next.

3.2. Proposed ASI System

Gaussian mixture (GM) densities are used to model $\mathbf{x}_{m,j}$ for the lowest four modulation frequency bands ($m = 1, \ldots, 4$). A GM density consists of a weighted sum of M component densities $p(\mathbf{x}|\mathbf{\Lambda}) = \sum_{i=1}^{M} \alpha_i b_i(\mathbf{x})$, where $\alpha_i, i = 1, ..., M$ are the mixture weights, with $\alpha_i \geq 0$ and $\sum_{i=1}^{M} \alpha_i = 1$,

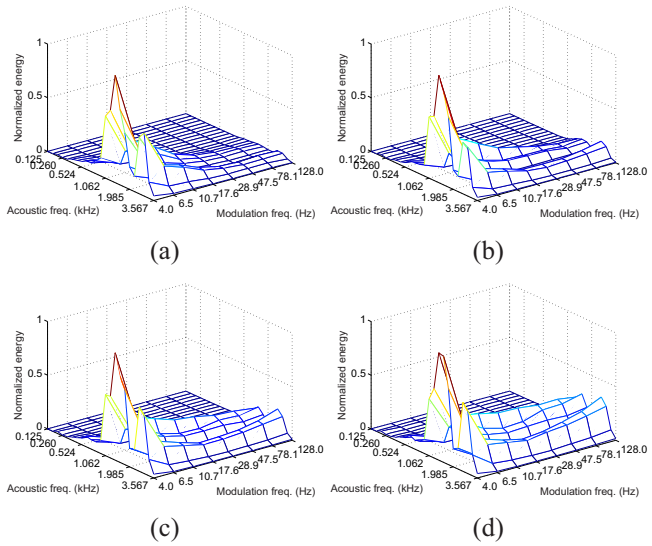

Fig. 2. $\bar{\mathbf{X}}$ for (a) clean and (b) reverberant speech with $RT = 0.4$ s, (c) 0.7 s, and (d) 1 s, for a female speaker.

and $b_i(\mathbf{x})$ are Gaussian densities with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. The parameter list, $\boldsymbol{\Lambda}$=$\{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_M\}$, defines a particular GM model, where $\boldsymbol{\lambda}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i\}$.

For each speaker, one GM model is trained for each modulation frequency band, for a total of four models over the four lowest bands; each model comprises 32 diagonal components. Clean speech is used for training in order to demonstrate the robustness of the system to mismatched reverberation conditions. Identification is based on the log-likelihood ($LL_{m,k}$) measure computed for N active speech frames

$$LL_{m,k} = \sum_{j=1}^{N} \log(p_m(\mathbf{x}_{m,j}|\boldsymbol{\Lambda}_{m,k})),$$

where m indexes the modulation frequency band, p_m and $\boldsymbol{\Lambda}_{m,k}$ represent, for such band, the GM model and GM parameters for speaker k, respectively. Given a group of N_S speakers, the identified speaker $\hat{S}$ is obtained using the following log-likelihoood test

$$\hat{S} = \arg \max_{1 \le k \le N_S} \{\max(LL_{1,k}, LL_{2,k}, LL_{3,k}, LL_{4,k})\}.$$

4. EXPERIMENT SETUP

In this section, the baseline system and experiments on simulated and recorded reverberant speech are described.

4.1. Baseline System

The widely used GM model based speaker identification system is used as the baseline [13]. Feature vectors consist of 12^{th} order mel-frequency cepstral coefficients (MFCC) appended with 12^{th} order delta MFCC; in pilot experiments, it was observed that the inclusion of double-delta coefficients

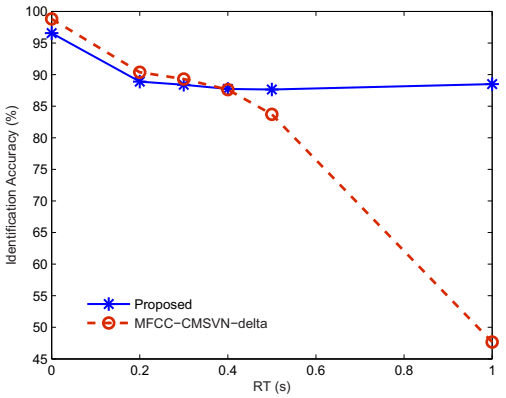

Fig. 3. Identification accuracy versus RT.

reduced identification accuracy at higher RT. MFCC are derived from a 26-channel mel-scale filterbank and the zeroth order coefficient (log-energy) is kept to form a 25 dimensional feature vector. Coefficients are computed from 25 ms frames with 10 ms shifts; only informative active speech frames are kept. Additionally, three channel compensation schemes are investigated: CMS, CMSVN, and RASTA. GM models with 64 diagonal components are used per speaker.

4.2. Experiment 1 - Simulated Data

The proposed algorithm is first tested on the simulated reverberant speech signals described in Section 2.1. The plots in Fig. 3 depict identification accuracy versus RT for the proposed scheme as well as for the baseline algorithm with CMSVN channel compensation, as it resulted in superior performance. As observed, baseline performance degrades almost linearly for $RT \geq 0.4$ s. The performance of the proposed system, on the other hand, is shown to be insensitive to increasing RT. Interestingly, baseline performance is slightly superior to that of the proposed system for lower RT values (≤ 0.3 s). It is conjectured that this gap in performance is due to the difference in complexity of the GM speaker models; recall the baseline system uses $M = 64$, whereas the proposed system $M = 32$. Since the proposed method uses longer windows (and window shifts) relative to the baseline system, roughly three times less training vectors are available. Hence, more complex speaker models are harder to obtain with short duration training data, as is the case with TIMIT. Strategies to overcome this limitation, such as maximum *a posteriori* (MAP) adaptation for GM training, are currently being investigated. Nonetheless, an average improvement of approximately 15% in identification accuracy is attained with the proposed scheme; 85% improvement is attained for $RT = 1$ s.

4.3. Experiment 2 - Real Data

The ICSI database is a multichannel database recorded in a noisy and reverberant meeting room with $RT \sim 0.3$ s. Noise

Table 2. Matched and mismatched ASI performance.

	Matched (%)		Mismatched (%)	
Channel	Baseline	Proposed (gain %)	Baseline	Proposed
E	82.1	95.7 (16.6)	78.2	86.9 (11.1)
F	81.2	92.8 (14.3)	75.1	89.9 (19.7)
6	79.6	95.7 (20.2)	76.7	85.5 (11.5)
7	78.7	91.3 (16.0)	75.8	84.1 (10.9)

sources include low-level hum of meeting room lights and fans (in particular for microphones no. 6 and 7), as well as noise from nearby elevators, hallway conversations, and laughter from other meeting participants. It is observed that the proposed feature set is robust to (quasi-)stationary noises, such as those produced by the lights and fan. Quasi-stationary noises show up in low modulation frequency channels ($<$ 1Hz), thus are not captured by the proposed feature set. Speech-like noise and competing speakers, on the other hand, may affect the proposed features. Hence, as is common with other ASI systems, a noise suppression algorithm is applied to reduce non-stationary noise. Noise-suppressed speech signals are used to test both the proposed and the baseline system.

Table 2 reports identification accuracy for matched *and* mismatched conditions. Matched train-test conditions indicate that speaker models were trained and tested from signals captured by the same microphone. Mismatched performance is the average over the three remaining train-test combinations. Performance is compared to that of the baseline system with CMSVN channel compensation. As can be seen, the proposed scheme outperforms the baseline system by an average 17% for matched and 13.3% for mismatched conditions.

5. DISCUSSION AND CONCLUSION

The plot in Fig. 3 shows that a slight improvement in identification performance is attained with the proposed system for $RT = 1$ s relative to $RT \leq 0.5$ s; in contrast, the baseline system performance degrades monotonically with increasing RT. Although counter intuitive at first, this behavior can be explained by the insights described in [14]. Envelopes of $\mathbf{x}_{1,j}$ are shown to resemble spectral envelopes obtained from higher order ($\geq$20) linear prediction analysis of the speech signal. As such, for smaller RT, the reflections create irregular-period pitch pulses, thus distorting the excitation spectrum and hence the linear prediction (LP) envelope. With increasing RT, the excitation looks more Gaussian-noise like, thus having less impact on the LP envelope. Combined with information from multiple modulation frequency bands, slightly higher identification performance is attained for larger RT. In summary, the proposed ASI system is shown to be insensitive to increasing RT *and* robust to mismatched conditions between enrollment and test data, two desirable properties for reverberant far-field speaker recognition applications.

6. REFERENCES

[1] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Trans. Audio, Speech and Language Proc.*, vol. 15, no. 7, pp. 1979–1986, Sept. 2007.

[2] J. Gammal and R. Goubran, "Combating reverberation in speaker verification," in *Proc. IEEE Conf. Instrumentation and Measurement Tech.*, May 2005, pp. 687–690.

[3] J. Gonzalez-Rodriguez, J. Ortega-Garcia, C. Martin, and L. Hernandez, "Increasing robustness in GMM speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays," in *Proc. Intl. Conf. Spoken Language Proc.*, 1996.

[4] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 15, no. 7, pp. 2023–2032, Sept. 2007.

[5] K. Eneman and M. Moonen, "Multimicrophone speech dereverberation: experimental validation," *EURASIP Journal on Audio, Speech, and Music Proc.*, 2007.

[6] H. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems," in *Proc. Intl. Conf. Speech and Lang. Proc.*, 2005.

[7] A. Janin *et al*, "The ICSI meeting corpus," in *Proc. Intl. Conf. on Acoustics, Speech, Signal Processing*, April 2003, vol. I, pp. 364–367.

[8] R. Ratnam, D. Jones, B. Wheeler, W. O'Brien, C. Lansing, and A. Feng, "Blind estimation of reverberation time," *Journal Acoustical Soc. America*, vol. 114, no. 5, pp. 2877–2892, Nov. 2003.

[9] Z. Smith, B. Delgutte, and A. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Letters to Nature*, vol. 416, pp. 87–90, March 2002.

[10] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *Journal Acoustical Soc. America*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994.

[11] R. Drullman, J. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *Journal Acoustical Soc. America*, vol. 95, no. 5, pp. 2670–2680, May 1994.

[12] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Proc. Intl. Conf. Speech and Lang. Proc.*, Oct. 1996, pp. 2490–2493.

[13] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, Aug. 1995.

[14] T. H. Falk, H. Yuan, and W.-Y. Chan, "Spectro-temporal processing for blind estimation of reverberation time and single-ended quality measurement of reverberant speech," in *Proc. Intl. Conf. Speech and Lang. Proc.*, Sept. 2007, pp. 514–517.