HILBERT ENVELOPE BASED FEATURES FOR ROBUST SPEAKER IDENTIFICATION UNDER REVERBERANT MISMATCHED CONDITIONS

Seyed Omid Sadjadi and John H.L. Hansen[⋆]

Center for Robust Speech Systems (CRSS),
The University of Texas at Dallas, Richardson, TX 75080–3021, USA
{sadjadi, john.hansen}@utdallas.edu

ABSTRACT

It is well known that MFCC based speaker identification (SID) systems easily break down under mismatched training and test conditions. One such mismatch occurs when a SID system is trained on anechoic speech data, while test is carried out using reverberant data collected via a distant microphone. In this study, a new set of feature parameters based on the Hilbert envelope of Gammatone filterbank outputs is proposed to improve SID performance in the presence of room reverberation. Considering two distinct perceptual effects of reverberation on speech signals, i.e., coloration and long-term reverberation, two different compensation strategies are integrated within the feature extraction framework to effectively suppress the effects of reverberation. Experimental evaluation is performed using speech material from the TIMIT, four different measured room impulse responses (RIR) from Aachen impulse response (AIR) database, and a GMM-based SID system. Obtained results indicate significant improvement over the baseline system with MFCCs plus cepstral mean subtraction (CMS), confirming the effectiveness of the proposed feature parameters for SID under reverberant mismatched conditions.

Index Terms— Gammatone filterbank, Hilbert envelope, mismatched conditions, reverberation suppression, speaker identification

1. INTRODUCTION

Recent advances in DSP manufacturing technology have enabled automatic speech systems to be integrated to virtually every electronic/mobile component of an individual's daily life. Nevertheless, providing robustness to these systems still remains a challenge because of the variety of acoustic mismatch scenarios that may occur between training and test conditions due to background noise, reverberation, accent, language, emotions, vocal effort, etc.

Specifically, performance of automatic speaker identification (SID) engines has been shown to severely degrade under reverberant mismatched conditions [1], [2]. Reverberation has various destructive effects on spectro-temporal characteristics of speech signals, most notably including temporal smearing, filling dips and gaps in the temporal envelope, increasing the prominence of low-frequency energy, and flattening the formant transitions [3]. These effects in turn can mask higher frequencies in the speech spectrum and blur the spectral details, both of which are useful acoustic cues for speaker identification.

Several compensation techniques to alleviating the adverse impact of room reverberation on performance of SID systems have been reported in the literature, most of which were first developed for automatic speech recognition or speech enhancement. The techniques have been applied at different stages of SID systems, i.e., signal [4], feature [5], model [6], and scoring stages [5], [6]. At the signal level, multichannel (e.g., microphone arrays) speech processing techniques have been employed to provide robustness to SID systems in reverberant and/or noisy conditions [4], although this imposes additional hardware requirements and more complexity on SID systems, and is not applicable in cases where only a single-channel signal (e.g., telephone) or prerecorded mono speech data is available. At the feature level, despite its simplicity, cepstral mean subtraction (CMS) has been shown to be helpful, but only for small reverberation times (a.k.a. T_{60}) where the length of analysis windows is comparable to that of the room impulse response (RIR) [2]. In [5], it was assumed that the effect of reverberation on the speech signal can be modeled as an additive noise, and a spectral subtraction method was adopted to suppress the reverberation before applying CMS. A feature warping method was also applied and a significant SID accuracy improvement was obtained over the baseline system. At the model level, assuming that there is access to RIRs and that a rough estimate of T_{60} can be calculated, reverberation classification and acoustic model matching based on reverberant background model (RBM) have been successfully employed [6], [7]. At the scoring level, similar to methods used for channel mismatched conditions, in [6] a combination of different normalization strategies were used to remove possible biases in the calculated likelihoods.

Another way of dealing with reverberant mismatched conditions in SID is to design acoustic features that are less susceptible to the destructive effects of reverberation. Although not optimal, MFCCs have been the most widely used acoustic features for SID. However, it is well known that speech systems that use MFCCs are vulnerable to training and test mismatch [2], and this has motivated extensive research efforts to find more robust acoustic features capable of capturing speaker identity conveyed in the speech signal. In particular, feature parameters obtained from subband Hilbert envelopes have shown promise for automatic speech and speaker recognition tasks under reverberant mismatched conditions [8], [9].

In this study, inspired by the human perception mechanism in reverberation known as the *precedence effect* [10], a new set of acoustic feature parameters based on the Hilbert envelope of Gammatone filterbank outputs is proposed. Given that the reverberation possesses two distinct perceptual effects on speech signals known as coloration and log-term reverberation, two different compensation strategies are integrated within the feature extraction framework to effectively suppress the two effects. Performance of a GMM based SID system [11] with the proposed feature parameters is benchmarked against that obtained with MFCCs along with CMS under four different reverberant mismatched conditions simulated using

[⋆] This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen. Approved for public release; distribution unlimited.

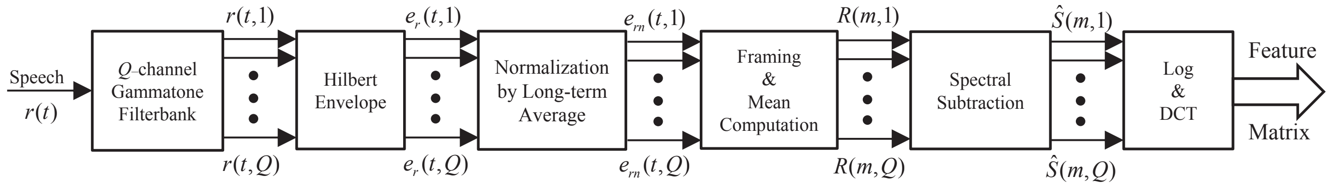

Fig. 1. Block diagram of the proposed feature extraction and reverberation compensation scheme

measured room impulse responses extracted from the Aachen impulse response (AIR) database [12].

2. HILBERT ENVELOPE BASED FEATURE EXTRACTION

2.1. Mathematical model of reverberation

In a reverberant environment, the signal received at the microphone is a delayed sum of a direct sound and its reflections from walls and objects in the acoustic enclosure, and hence can be modeled as the convolution of the RIR with the speech signal,

$$r(t) = s(t) * h(t), \tag{1}$$

where $r(t)$ and $s(t)$ are the reverberant and anechoic signals, respectively, and $h(t)$ is the RIR. The RIR $h(t)$ can be partitioned into two parts $h_e(t)$ and $h_l(t)$ as,

$$h(t) = \begin{cases} 0, & t < 0 \\ h_e(t), & 0 \le t < t_e \\ h_l(t), & t_e \le t < L \end{cases} \tag{2}$$

where L is the length of $h(t)$, and t_e is a time window threshold chosen such that $h_e(t)$ consists of the direct path signal and a few early reflections, while $h_l(t)$ consists of all late reflections. Late reflections that smear the speech spectra and reduce signal quality, are characterized by T_{60}. These have a long-term effect on speech signals and therefore cannot be compensated for using CMS within the short-term speech analysis framework. On the other hand, early reflections that cause coloration distortion and increase the prominence of low-frequency energy, are characterized by the direct-to-reverberant ratio (DRR) which is dependent on the distance between the sound source and microphone.

In [13], it was shown that temporal envelopes of subband reverberant signals obtained from speech decomposition through a bank of bandpass filters can be written as,

$$e_r(t,j) \approx \frac{1}{2} e_s(t,j) * e_h(t,j), \tag{3}$$

where $e_r(t,j)$ is the temporal envelope of the reverberant signal at the j^{th} subband. Eq. (3) suggests that the temporal envelope of subband reverberant signals can be approximated as the convolution of the temporal envelope of the direct sound with that of the RIR in that subband, assuming that the analysis window length is longer than the RIR length L. It is worth mentioning that the signal envelope is also a good measure for detecting the direct sound (first wave-front), as suggested by a computational model for the *precedence effect* [10].

In this study, inspired by the human perception mechanism, the Gammatone filterbank is employed for signal decomposition into subbands, and the squared magnitude of the Hilbert envelope in each subband is calculated as the temporal envelope. A normalization strategy that functions as a form of automatic gain control (AGC) is used in each subband to suppress any spectral coloration effect of the reverberation in that subband [14]. In addition, since the long-term reverberation corresponding to late reflections can be treated as an additive Gaussian uncorrelated random process, a spectral subtraction technique is adopted to suppress the long-term effect of reverberation on the speech signal.

2.2. Mean Hilbert envelope coefficients

In this section, the procedure for extracting a new set of acoustic feature parameters, based on the Hilbert envelope of Gammatone filterbank outputs, for robust SID under reverberant mismatched conditions is described.

The block diagram of the proposed feature extraction scheme is depicted in Fig. 1. First, the preemphasized reverberant speech signal $r(t)$ is filtered using a 32-channel Gammatone filterbank to simulate the effect of auditory filtering which takes place in the cochlea [15]. The filterbank center frequencies are uniformly spaced on equivalent rectangular bandwidth (ERB) scale between 50 and 8000 Hz (assuming a sampling rate of $F_s = 16$ kHz). Next, the temporal envelope of the j^{th} channel output $r(t,j)$ is computed as the squared magnitude of analytical signal obtained using the Hilbert transform. More specifically, let

$$r_a(t,j) = r(t,j) + i\hat{r}(t,j), \tag{4}$$

denote the analytical signal, where $\hat{r}(t,j)$ is the Hilbert transform of $r(t,j)$, and i is the imaginary unit. The temporal envelope $e(t,j)$ is thus calculated as,

$$e_r(t,j) = r^2(t,j) + \hat{r}^2(t,j)\,. \tag{5}$$

$e_r(t,j)$ is also called the Hilbert envelope of the signal $r(t,j)$. As a particular requirement of the envelope convolution model in Eq. (3) remarked in [13], the Hilbert envelope $e_r(t,j)$ is smoothed using a low-pass filter with a cut-off frequency of 20 Hz. Next, in each channel, the smoothed Hilbert envelope is normalized by the long-term average computed over the entire utterance as,

$$e_{rn}(t,j) = \frac{e_r(t,j)}{\frac{1}{N}\sum_{t=0}^{N-1} e_r(t,j)}\,, \tag{6}$$

with N being the signal length in samples. As stated earlier, this functions as an AGC and is used to suppress any spectral coloration effect of the reverberation in different frequency channels.

In the next stage, the low-pass filtered normalized Hilbert envelope $e_{rn}(t,j)$ is blocked into frames of 25 ms duration with a skip rate of 10 ms. A Hamming window is applied to each frame to minimize discontinuities at the edges. To estimate the temporal envelope amplitude in frame m, the sample means are computed as,

$$R(m,j) = \frac{1}{M}\sum_{t=0}^{M-1} v(t) e_{rn}(t,j)\,, \tag{7}$$

where $v(t)$ denotes the Hamming window and M is the frame size in samples. Note that $R(m,j)$ is a measure of the spectral energy at the center frequency of the j^{th} channel, and therefore provides a short-term spectral representation of the speech signal $r(t)$.

Up to this stage, only the coloration distortion due to early reflections has been suppressed. The long-term effect of reverberation (due to late reflections) can be modeled as an uncorrelated additive noise [16], and hence can be compensated via spectral subtraction. As suggested in [17], it is assumed that the power spectrum of late components of the RIR is a smoothed shifted version of that of the reverberant speech, and thus the power spectrum of the clean speech signal is estimated as,

$$|\hat{S}(m,j)|^2 = |R(m,j)|^2 \cdot \max\left[\frac{|R(m,j)|^2 - \gamma w(m-\rho) * |R(m,j)|^2}{|R(m,j)|^2}, \varepsilon\right] \quad (8)$$

where $R(m,j)$ is the short-term spectral energy obtained from the previous stage, the symbol $*$ denotes the convolution in the time domain, and $w(m)$ is a smoothing function which is chosen as the Rayleigh distribution. The parameters ρ and ε are the relative delay of the late RIR components, and the maximum attenuation floor, respectively. The relative delay has been shown to be independent of reverberation characteristics and is commonly set to 50 ms for speech which corresponds to 5 frames in our case. The flooring parameter ε is fixed to 0.01 which is equivalent to a maximum attenuation of -20 dB.

To compress the dynamic range of the estimated spectral parameters $\hat{S}(m,j)$, the natural logarithm is applied. Finally, the discrete cosine transform (DCT) is applied to: 1) convert the spectrum to cepstrum, and 2) decorrelate the various feature dimensions. The latter is important because GMMs with *diagonal* covariance matrices can then be used to model the acoustic space of each speaker (as opposed to *full* covariance matrices). The output is a matrix of 32-dimensional cepstral features, entitled the mean Hilbert envelope coefficients (MHEC).

3. EXPERIMENTS

Performance of the proposed feature extraction and reverberation compensation scheme is evaluated in the context of a GMM based closed-set SID system. SID accuracies are used as a measure to compare performance of the proposed features with that of MFCCs plus CMS under reverberant mismatched conditions. Training and test speech material is obtained from the TIMIT database that contains signals from 630 speakers including 192 female and 438 male speakers. There are 10 sentences per speaker recorded under clean laboratory conditions at a sampling rate of 16 kHz. A total of 8 sentences ($\sim$ 24 s) are used to train the speaker models, while the remaining 2 sentences ($\sim$ 6 s) test the models. To simulate different reverberant conditions, RIR samples extracted from the AIR database are convolved with test material. Four RIRs with distinct source to microphone distances (d_{SM}) are used including studio booth, meeting, office, and lecture rooms. More information about the RIRs is summarized in Table 1.

Table 1. Properties of the four RIRs extracted from the AIR database for experiments.

Room Type	Dimension (m^3)	d_{SM} (m)	T_{60} (s)	DRR (dB)
Studio booth	$3.0 \times 1.8 \times 2.2$	1.0	0.11	8.78
Meeting	$8.0 \times 5.0 \times 3.1$	2.80	0.25	2.89
Office	$5.0 \times 6.4 \times 2.9$	3.0	0.48	-0.89
Lecture	$10.8 \times 10.9 \times 3.15$	10.2	0.83	-5.62

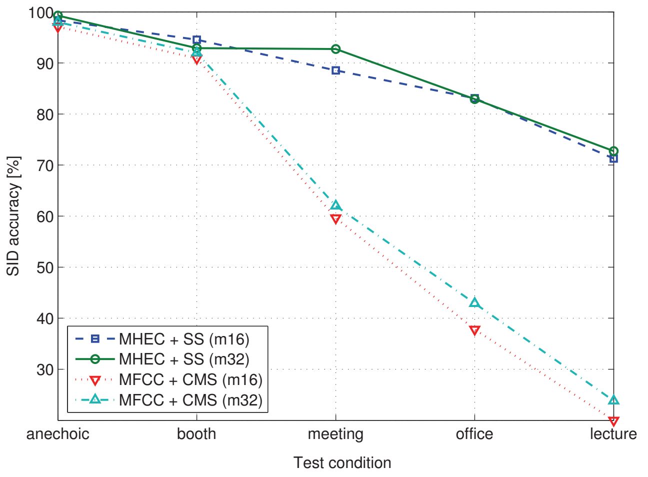

Fig. 2. SID accuracy for the proposed acoustic features with spectral subtraction and for MFCCs with CMS, under anechoic and four different reverberant test conditions.

MFCC features are extracted from frames of 25 ms duration with a skip rate of 10ms. Out of 27 filterbank log-energies, the first 12 cepstral coefficients are retained after applying the DCT (excluding c_0) , and delta features are appended to form a 24-dimensional feature vector for each frame. CMS is applied in an effort to help reduce the mismatch due to reverberation. MFCCs plus CMS serve as baseline acoustic features for subsequent comparisons. MHEC features are obtained using the procedure described in Section 2.2. A 31-dimensional feature vector is formed after excluding the energy term. Speech data from a total of 80 speakers including 37 females and 43 males are used as a development set to find the optimum parameters for the spectral subtraction stage, i.e., $\gamma = 0.1$ and $\varepsilon = 0.01$.

Two GMM based SID systems (one per feature type) are trained for evaluations, using only anechoic speech contained frames. An energy-based thresholding algorithm is adopted for silence frame removal. For each feature type, both 16 and 32-mixture GMMs are considered.

4. RESULTS

In this section, performance evaluation of the proposed acoustic feature parameters and reverberation compensation strategies is reported in terms of accuracy obtained from the GMM based SID system.

Fig. 2 represents identification rates obtained with the proposed acoustic feature parameters and reverberation compensation techniques, under anechoic and four different reverberant test conditions including studio booth, meeting , office, and lecture rooms, with T_{60} ranging from approximately 0 to 0.83 s. Speaker models with 16 and 32 Gaussian mixtures are employed to further investigate the robustness of the proposed features to over and underestimation effects imposed by the speaker models. It is observed that when MFCCs are used as acoustic features, even in the presence of small reverberation, the SID accuracy tends to drop significantly toward an unacceptable level. It is also evident that as the reverberation time increases and the distance between sound source and microphone becomes longer, CMS is less effective, giving rise to a significant performance degradation. On the other hand, the proposed feature parameters are much more robust to changes in room reverberation, even with 16-mixture speaker models. It is worth noting that because the compensation techniques, i.e., CMS and spectral subtraction, themselves introduce mismatch between training and test conditions, they should be applied to both training and test stages for the best performance to be achieved.

In order to better assess the impact of the two reverberation suppression stages in the proposed feature extraction algorithm, i.e., subband normalization and spectral subtraction, we consider the office room environment ($T_{60} \approx 0.5$ s), and perform SID with and without the compensation methods. Results are summarized in Table 2. It can be seen that even without the compensations, MHECs consistently outperform MFCCs under this environment. The increase in accuracy with only the subband normalization is higher than that with only the spectral subtraction. However, the two strategies have a synergistic effect on performance and further improvement occurs when applied together. Also, it is obvious that CMS helps reduce reverberation effects on MFCCs, although the performance is still quite low.

Table 2. Effectiveness evaluation of the proposed acoustic features with and without the compensation strategies. N refers to subband normalization, and SS denotes the spectral subtraction.

	Accuracy (%)	
Feature + Compensation	m16	m32
MHEC	62.91	60.91
MHEC + N	78.55	78.18
MHEC + SS	78.00	75.82
MHEC + N + SS	83.09	82.91
MFCC	26.18	28.55
MFCC + CMS	37.82	42.91

5. CONCLUSION

In this study the problem of SID under reverberant mismatched conditions has been studied. It was confirmed that the performance of GMM based SID systems with MFCCs plus CMS drops significantly in the presence of room reverberation, especially when the reverberation time increases. Inspired by the evidences observed from the human perception mechanism, a new set of feature parameters based on the Hilbert envelope of Gammatone filterbank outputs was proposed, and a spectral subtraction method for compensating the long-term effect of reverberation was adopted. The proposed feature was shown to be consistently superior to MFCCs in performance under mismatched training and test conditions due to reverberation, while providing the same SID accuracy under clean matched conditions. Further improvement in SID accuracy can be achieved with the feature parameters introduced in this study, by applying other compensation techniques such as feature warping and acoustic model adaptation (e.g., MAP adaptation).

6. REFERENCES

[1] P. Castellano, S. Sridharan, and D. Cole, “Speaker recognition in reverberant enclosures,” in *Proc. IEEE ICASSP’96*, Atlanta, GA, May 1996, vol. 1, pp. 117–120.

[2] Y. Pan and A.Waibel, “The effects of room acoustics on MFCC speech parameter,” in *Proc. ICSLP’00*, Beijing, China, Oct. 2000, pp. 129–132.

[3] P.F. Assmann and A.Q. Summerfield, “The perception of speech under adverse conditions,” in *Speech Processing in the Auditory System*, S. Greenberg, W.A. Ainsworth, A.N. Popper, and R.R. Fay, Eds. New York: Springer-Verlag, 2004, Chap. 5, pp. 231–308.

[4] 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. ICSLP’96*, Philadelphia, PA, Oct. 1996, pp. 1333–1336.

[5] Q. Jin, T. Schultz, and A. Waibel, “Far-Field Speaker Recognition”, *IEEE TASLP*, vol. 15, no. 7, pp. 2023–2032, Sept. 2007

[6] I. Peer, B. Rafaely, and Y. Zeigel,“Reverberation matching for speaker recognition,” in *Proc. IEEE ICASSP’08*, Las Vegas, NV, Apr. 2008, pp. 4829–4832.

[7] J. Gammal and R. Goubran, “Combating reverberation in speaker verification,” in *Proc. IEEE Conf. Instrum. Meas. Technol., IMTC’05*, Ottawa, Canada, May 2005, pp. 687–690.

[8] T.H. Falk and W.-Y. Chan, “Modulation spectral features for robust far-field speaker identification,” *IEEE TASLP*, vol. 18, no. 1, pp. 90–100, Jan. 2010.

[9] S. Thomas, S. Ganapathy, H. Hermansky, “Hilbert envelope based features for far-field speech recognition,” in *Proc. 5^{th} intl. workshop Machine Learning for Multimodal Interaction, MLMI’08*, Utrecht, Netherlands, Sept. 2008, pp. 119–124.

[10] K. D. Martin, “Echo suppression in a computational model of the precedence effect,” in *Proc. IEEE WASPAA’97*, New Paltz, NY, Oct. 1997.

[11] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” *Speech Commun.*, vol. 17, pp. 91–108, Aug. 1995.

[12] M. Jeub, M. Schafer, and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in *Proc. IEEE DSP’09*, Greece, Jul. 2009, pp. 1–5.

[13] J. Mourjopoulos and J.K. Hammond, “Modelling and enhancement of reverberant speech using an envelope convolution method,” in *Proc. IEEE ICASSP’83*, Boston, MA, Apr. 1983, pp. 1144–1147.

[14] S. Greenberg and B.E.D. Kingsbury, “The modulation spectrogram: In pursuit of an invariant representation of speech,” in *Proc. IEEE ICASSP’97*, Munich, Germany, Apr. 1997, pp. 1647–1650.

[15] R.D. Patterson *et al.*, “Complex sounds and auditory images,” in *Auditory Physiology and Perception*, Y. Cazals, L. Demany, and K. Horner Eds. Oxford: Pergamon Press, 1992, pp. 429–446.

[16] K. Lebart, J. M. Boucher, and P. N. Denbigh, “A new method based on spectral subtraction for speech dereverberation,” *Acta Acustica*, vol. 87, pp. 359–366, 2001.

[17] M. Wu and D. Wang, “A two-stage algorithm for one-microphone reverberant speech enhancement,” *IEEE TASLP*, vol. 14, no. 3, pp. 774–784, May 2006.