Mean Hilbert Envelope Coefficients (MHEC) for Robust Speaker Recognition

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

Center for Robust Speech Systsems (CRSS)
Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas, Richardson, TX 75080-3021, USA
{sadjadi, taufiq.hasan, john.hansen}@utdallas.edu

Abstract

The recently introduced mean Hilbert envelope coefficients (MHEC) have been shown to be an effective alternative to MFCCs for robust speaker identification under noisy and reverberant conditions in relatively small tasks. In this study, we investigate the effectiveness of these acoustic features in the context of a state-of-the-art speaker recognition system. The i-vectors are used to represent the acoustic space of speakers, while modeling is performed via probabilistic linear discriminant analysis (PLDA). We report speaker verification performance on the NIST SRE-2010 extended telephone and microphone trials for both female and male genders. Experimental results confirm consistent superiority of MHECs to traditional MFCCs within i-vector speaker verification, particularly under microphone and telephone training-test mismatch conditions. In addition, fusion of subsystems trained with the individual front-ends proves that the two acoustic features (i.e., MHEC and MFCC) provide complimentary information for recognizing speakers.

Index Terms: Mean Hilbert Envelope Coefficients (MHEC), mismatch conditions, NIST SRE, speaker recognition

1. Introduction

Current state-of-the-art speaker recognition systems are primarily focused on channel and session mismatch compensation techniques in the back-end. The research trend in this domain has gradually migrated from joint factor analysis (JFA) based methods, which attempt to model the speaker and channel subspaces separately [1], towards the i-vector approach that models both speaker and channel into a single space termed the total variability space [2]. Various classifiers, models, and scoring methods are conveniently applied to i-vectors. These include support vector machines (SVM), probabilistic linear discriminant Analysis (PLDA) [3], [4], and the simple yet effective cosine distance (CD) based scoring which is typically combined with LDA followed by within-class covariance normalization (WCCN) [5].

In spite of porgress seen in back-end advancements, several research efforts have been made recently that target at developing acoustic features (or front-ends) which are not only capable of capturing speaker identity conveyed in the speech signal, but also robust to environmental distortion (e.g., see [6], [7], [8]). Although originally designed to represent acoustic spaces of different phonemes for ASR, MFCCs have been the most widely used features for speaker recognition tasks, probably because they provide acceptable performance in NIST SRE related applications. However, it is well-known that MFCC based systems are susceptible to training and test mismatch in environmental noise and reverberation. There are several factors contributing to this susceptibility, among which the following are most dominant: 1) the spectrum estimation in standard MFCC extraction is not robust to noise and channel distortions [7], and 2) the auditory model used in MFCC is neither accurate, nor optimal for speaker recognition. This has been our motivation in the design of robust acoustic features that are less affected by background noise [9] or room reverberation [10]. In this study, we evaluate the effectiveness of our recently introduced Mean Hilbert Envelope Coefficients (MHEC) [10], in the context of a state-of-the-art i-vector speaker verification system with PLDA modeling.

We have previously demonstrated that, when compared to MFCCs, employing MHECs as acoustic features results in substantial gains in performance of a GMM based speaker identification system under noisy and/or reverberant mismatched conditions in relatively small tasks. This study represents our first attempt to evaluate MHECs versus MFCCs on a speaker verification task at the scale of the NIST SRE. Evaluations are performed using NIST SRE-2010 extended telephone and microphone trials (core conditions 1 through 5) for bothe female and male genders. In addition to evaluating speaker recognition performance individually for each front-end, we investigate fusion of the subsystems trained with each feature. Results of the fusion experiment shall reveal whether or not the two acoustic representations provide complimentary information for speaker recognition applications.

2. Mean Hilbert Envelope Coefficients

In this section, the procedure for extracting the acoustic feature parameters based on the Hilbert envelope of Gammatone filterbank outputs, is described. A block diagram illustrating the proposed feature extraction scheme is depicted in Figure 1.

First, the preemphasized speech signal $s(t)$ is decomposed into 24 bands through a 24-channel Gammatone filterbank [11]. The filterbank center frequencies are uniformly spaced on equivalent rectangular bandwidth (ERB) scale between 300 and 3400 Hz (assuming a telephone bandwidth at a sampling rate of $F_s = 8$ kHz). Next, since we are mostly interested in slowly varying amplitude modulations rather than the fine structure,the temporal envelope of the j^{th} channel output $s(t, j)$ is computed as the squared magnitude of the analytical signal obtained using

[⋆] This project was funded by AFRL under contract FA8750-12-1-0188 (approved for public release, distribution unlimited: 88ABW-2012-1820), and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

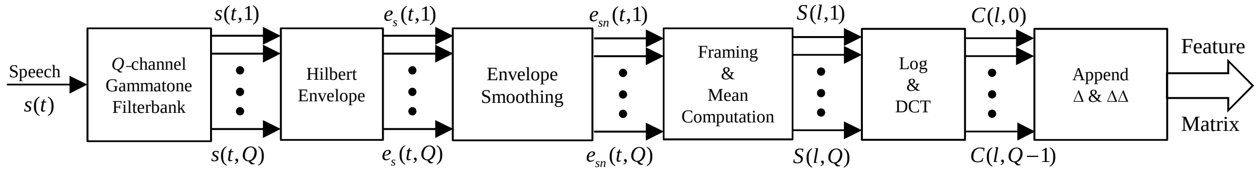

Figure 1: *Block diagram of the proposed feature extraction scheme. The symbols represent the output signals at each stage.*

the Hilbert transform. More specifically, let

$$s_a(t,j) = s(t,j) + i\hat{s}(t,j), \tag{1}$$

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

$$e_s(t,j) = s^2(t,j) + \hat{s}^2(t,j)\ . \tag{2}$$

Here, $e_s(t,j)$ is also called the Hilbert envelope of the signal $s(t,j)$. At the next stage, in order to further suppress the remaining redundant high frequency components, the Hilbert envelope $e_s(t,j)$ is smoothed using a low-pass filter with a cut-off frequency of $f_c = 20$ Hz as,

$$e_{sn}(t,j) = (1-\alpha)\, e_s(t,j) + \alpha\, e_{sn}(t-1,j), \tag{3}$$

where α is a smoothing factor (inversely) exponentially proportional to the cut-off frequency as,

$$\alpha = \exp\left(\frac{-2\pi f_c}{F_s}\right). \tag{4}$$

Next, the smoothed Hilbert envelope $e_{sn}(t,j)$ 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 l, the sample means are computed as,

$$S(l,j) = \frac{1}{N}\sum_{t=0}^{N-1} w(t) e_{sn}(t,j)\ , \tag{5}$$

where $w(t)$ denotes the Hamming window and N is the frame size in samples. Note that $S(l,j)$ is also 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 $s(t)$. To compress the dynamic range of the estimated spectral parameters $S(l,j)$, the natural logarithm is applied. In addition, the discrete cosine transform (DCT) is applied to: 1) convert the spectrum to the 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 therefore a matrix of 24-dimensional cepstral features $C(l,j)$, entitled the mean Hilbert envelope coefficients (MHEC). For our speaker verification experiments, only the first 12 coefficients are retained after DCT (excluding c_0).

Finally, because the cepstral representation of the speech spectrum is only a measure of the local pattern of the signal at a given frame, the first and second temporal cepstral derivatives are computed and appended to the static features to capture the dynamic pattern of speech over time. This results in 36-dimensional feature vectors.

Before concluding this section, it seems worthwhile to make a few remarks on the advantages of using the Gammatone filter-bank. First, the Gammatone filterbank simulates more accurately the effect of auditory filtering, which takes place along the basilar membrane in the cochlea, and is closely coupled with perception in humans [11]. Second, it provides a direct way to compute the so called "subjective spectrum" from the speech signal. In other words, it obviates any need for FFT calculation and, when compared to MFCCs, mel-band integration. More specifically, it helps reduce the computational load of taking a typical 256-point FFT (25 ms frames at $F_s = 8$ kHz), as well as integration of the estimated magnitude/power spectrum into 24 mel-bands for each frame.

3. Experimental Setup

In this section, the configuration for our i-vector speaker recognition system is briefly described.

3.1. Feature Extraction

MHECs and MFCCs are extracted as acoustic features from speech material for speaker verification experiments. In order to have a fair and meaningful comparison, the same configuration parameters are used to extract the two features. We use the popular HMM toolkit (HTK) implementation of MFCCs: 36-dimensional feature vectors (12 MFCC $+\Delta + \Delta\Delta$) are extracted using 25 ms frames with 10 ms shift. A total of 24 filters are used in the mel-filterbank which cover the frequency range 300–3400 Hz (i.e, the telephone bandwidth). Both features are normalized toward a Gaussian distribution through feature warping over a 3-second sliding window [12]. To remove silence and low energy speech segments, a two stage voice activity detection (VAD) is performed. In the first stage, which is used before feature extraction, a soft VAD based on several voicing measures is utilized to remove the non-speech segments. This strategy saves large amount of computation, since in this manner features are only extracted from speech segments. In the second stage, which is applied after the feature extraction, an energy based method is employed to drop the low-energy speech frames as well as the residual non-speech frames from the soft VAD in the first stage. These low energy frames are easily affected by noise and channel variabilities, and do not carry much speaker-dependent information .

3.2. UBM Training

Gender dependent 1024-mixture universal background models (UBM) are trained using only the English telephone data selected from the NIST SRE 2004, 2005, 2006, as well as the Switchboard 2 (Phase III) and Switchboard Cellular (Part 1 and 2). These corpora are available through LDC [13] or by participating in the SRE evaluations [14, 15]. There are a total of 9676 conversations from 951 male speakers, and 12490 conversations

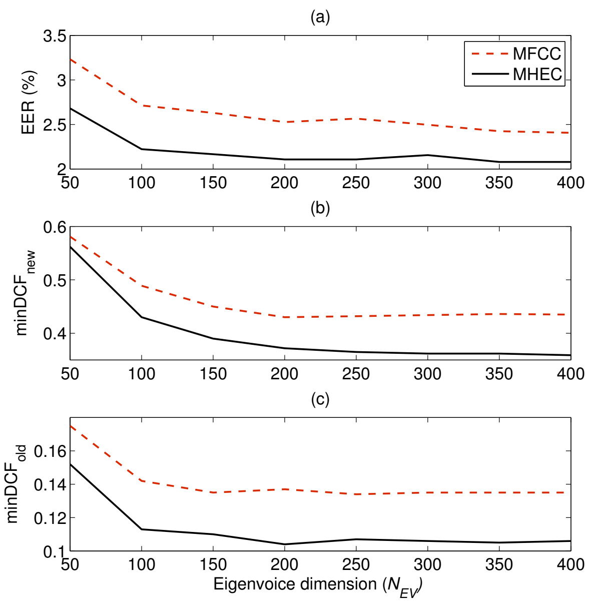

Figure 2: *MFCC vs MHEC performance with varying number of columns in the eigenvoice matrix Φ for extended telephone-telephone trials (core condition 5).*

from 1168 female speakers. The HTK is employed for UBM training with 15 expectation maximization (EM) iterations per binary split. The UBM is later used to extract the zeroth and first order Baum-Welch statistics for i-vector extraction.

3.3. I-vector Extraction

I-vectors are extracted for both MFCC and MHEC features using the front-end factor analysis scheme as described in [2]. This is accomplished in a similar manner to eigenvoice learning with the exception that instead of labeling speaker and channel information for subspace modeling, each utterance is assumed to be produced by a unique speaker. The total subspace matrices with 400 columns are estimated from the same data used for UBM construction.

3.4. PLDA

A Gaussian PLDA with a full-covariance noise model is used for both session variability compensation and scoring [4]. In this generative model, an arbitrary D dimensional i-vector η_r extracted from a speech utterance is expressed as,

$$\eta_r = m + \Phi\beta + \epsilon_r, \tag{6}$$

where m is the $D \times 1$ speaker independent mean vector, Φ is the $D \times N_{EV}$ rectangular matrix representing a basis for the speaker-specific subspace/Eigenvoices, β is an $N_{EV} \times 1$ latent vector having a standard normal distribution, and ϵ_r is the $D \times 1$ random vector representing the full covariance residual noise. Here, the only free parameter is the number of eigenvoices N_{EV}, which denotes the number of columns in the matrix Φ, and is set to 200 in our experiments. To train the PLDA model, the i-vectors extracted from the UBM dataset as well as microphone data from NIST SRE 2005 and 2006 are utilized.

4. Results and Discussion

To evaluate the effectiveness of MHECs in large-scale speaker recognition applications, as well as to compare their performance against MFCCs, speaker verification experiments are conducted using the NIST SRE-2010 extended microphone and telephone trials (conditions 1–5) for both female and male genders.

In our first experiment, we investigate the effect of number of columns in the eigenvoice matrix Φ on the performance by varying the dimension from 50 to 400 with an increment of 50. Here, only the extended telephone-telephone male trials (core condition 5) are considered. Results are illustrated in Fig. 2 in terms of equal error rate (EER) as well as the new and old minimum detection cost functions (minDCF). It is evident from the figure (Fig. 2 (a)-(c)) that the subsystem trained with MHECs consistently outperforms the one trained using MFCCs, for all eigenvoice dimensions in the matrix Φ. For eigenvoice dimensions greater than 100, the MHEC based subsystem exhibits a

Table 1: *Performance of the MFCC and MHEC based subsystems as well as their fusion in terms of EER for NIST SRE-2010 extended microphone and telephone trials (conditions 1–5)*

Gender	Cond.	EER (%)		
		MFCC	MHEC	Fusion
Female	1	2.49	2.49	2.14
	2	4.47	3.99	3.50
	3	4.13	3.42	3.01
	4	2.97	2.68	2.37
	5	3.73	3.48	3.37
Male	1	1.04	0.71	0.61
	2	1.83	1.47	1.21
	3	2.78	1.97	1.92
	4	1.72	1.64	1.38
	5	2.53	2.10	1.85

Table 2: *Performance of MFCC and MHEC based subsystems as well as their fusion in terms of new (old) minDCF for NIST SRE-2010 extended microphone and telephone trials (conditions 1–5)*

Gender	Cond.	minDCF new (old)		
		MFCC	MHEC	Fusion
Female	1	0.394 (0.114)	0.370 (0.106)	0.332 (0.094)
	2	0.685 (0.234)	0.610 (0.200)	0.591 (0.178)
	3	0.616 (0.196)	0.544 (0.156)	0.498 (0.146)
	4	0.534 (0.160)	0.481 (0.129)	0.462 (0.125)
	5	0.543 (0.176)	0.495 (0.159)	0.462 (0.144)
Male	1	0.256 (0.046)	0.186 (0.035)	0.190 (0.034)
	2	0.446 (0.100)	0.340 (0.079)	0.334 (0.068)
	3	0.494 (0.132)	0.427 (0.116)	0.394 (0.100)
	4	0.325 (0.087)	0.247 (0.067)	0.243 (0.062)
	5	0.430 (0.137)	0.372 (0.104)	0.339 (0.099)

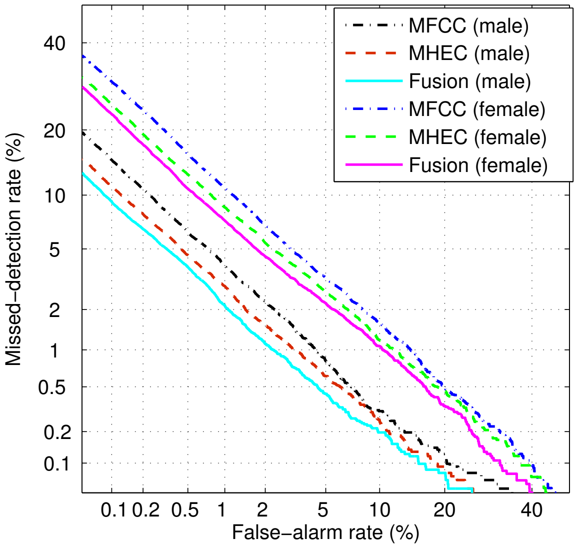

Figure 3: *DET curves for the MFCC and MHEC based subsystems as well as their fusion. Results are pooled from all extended microphone and telephone trials (conditions 1–5).*

nearly constant performance in terms of both EER and the old minDCF which reflects its robustness to this PLDA parameter.

In our next experiments, the number of columns in the eigevoice matrix Φ is fixed to 200, and the subsystems trained with individual acoustic features are evaluated using the extended microphone and telephone trials (conditions 1–5). In addition, the subsystems are fused to investigate whether or not the two front-ends are complimentary. The fusion is accomplished by adding the normalized scores of the individual subsystems. Results of these experiments are presented in Tables 1 and 2, in terms of EER as well as the new and old minDCFs, respectively. It is observed that, across all training and test conditions and for all evaluation metrics, the MHEC subsystem shows superior performance compared to the MFCC subsystem. The greatest gain in EER is seen for condition 3 in Table 1 where relative improvements of 17% and 29% are achieved for female and male genders, respectively. Condition 3 represents a challenging mismatch between training and test conditions where trials involve interview training speech and normal vocal effort conversational telephone test speech. Overall, the improvements are significant given the scale of the experiments which is reflected in the number of trials in each condition.

It can also be seen from the tables that the additive fusion of the individual subsystems yields substantial gains in performance for all conditions and with all metrics. This indicates that the two acoustic features are complimentary.

Fig. 3 shows the detection error trade-off (DET) curves for the MFCC and MHEC based subsystems along with their fusion individually for female and male genders. The curves are obtained by pooling scores of all extended microphone and telephone trails (conditions 1–5), individually for female and male genders. Consistent with our previous observations, it is seen that the MHEC based subsystem achieves superior performance across a wide range of operating points on the DET curve. Moreover, subsystem fusion can dramatically boost the performance, which confirms the complimentary nature of the two features for speaker recognition.

5. Conclusions

In this study, we have explored the effectiveness of our recently introduced MHEC features in the context of a state-of-the-art i-vector speaker recognition system with PLDA modeling. Experiments were conducted using the NIST SRE-2010 extended microphone and telephone trials for both female and male genders. The obtained results confirmed that the subsystem trained with MHECs consistently outperformed that trained using traditional MFCCs, across all core conditions available in the NIST SRE-2010 trials. In addition, it was verified that the fusion of the subsystems trained with the individual front-ends yields significant gains in speaker recognition performance. This indicates that the two acoustic representations provide complimentary information for recognizing speakers.

6. References

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus Eigenchannels in speaker recognition,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.

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

[3] S. Prince and J. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proc. IEEE Int. Conf. Computer Vision, ICCV 2007*, Rio de Janeiro, Oct. 2007, pp. 1–8.

[4] D. Garcia-Romero and C. Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems,” in *Proc. INTERSPEECH*, Florence, Italy, Sept. 2011, pp. 249–252.

[5] A. Hatch, S. Kajarekar, and A. Stolcke, “Within-class covariance normalization for SVM-based speaker recognition,” in *Proc. INTERSPEECH*, Pittsburgh, PA, Sept. 2006, pp. 1471–1474.

[6] C. Hanilci, T. Kinnunen, F. Ertas, R. Saeidi, J. Pohjalainen, and P. Alku, “Regularized all-pole models for speaker verification under noisy environments,” *IEEE Signal Process. Lett.*, vol. 19, pp. 163–166, Mar. 2012.

[7] J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O Shaughnessy, “Multi-taper MFCC features for speaker verification using i-vectors,” in *Proc. IEEE ASRU*, Hawaii, HI, Dec. 2011, pp. 547–552.

[8] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, “Linear versus mel frequency cepstral coefficients for speaker recognition,” in *Proc. IEEE ASRU*, Hawaii, HI, Dec. 2011, pp. 559–564.

[9] S. O. Sadjadi and J. H. L. Hansen, “Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions,” in *Proc. INTERSPEECH*, Makuhari, Japan, Sept. 2010, pp. 2138–2141.

[10] ——, “Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions,” in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 5448–5451.

[11] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, “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.

[12] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” in *Proc. A Speaker Odyssey*, Crete, Greece, Jun. 2001, pp. 213–218.

[13] C. Cieri, “The mixer corpus of multilingual, multichannel speaker recognition data,” DTIC Document, Tech. Rep., 2004.

[14] “The NIST year 2008 speaker recognition evaluation plan,” 2008. [Online]. Available: http://www.nist.gov

[15] “The NIST year 2010 speaker recognition evaluation plan,” 2010. [Online]. Available: http://www.nist.gov