Robust Speaker Recognition in Non-Stationary Room Environments Based on Empirical Mode Decomposition

*Taufiq Hasan and John H. L. Hansen**

Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering,
University of Texas at Dallas, Richardson, Texas, U.S.A.

Abstract

In this study, we consider the problem of speaker recognition in a non-stationary room/channel mismatched condition. In such circumstances, cepstral coefficients are affected in a way that the short-term stationarity assumption, on which conventional feature normalization methods are based on, may not be valid. We observe that the empirical mode decomposition (EMD) applied to the cepstral feature stream can partially separate out the non-stationary channel components, if present, into its residual signal and other lower order intrinsic mode functions (IMFs), which leads us to develop a filtering scheme based on this decomposition. The proposed method works in the time domain making use of the instantaneous frequency function obtained through Hilbert spectral analysis of the IMFs. Experimental evaluations on the TIMIT database with added non-stationary room channels in test demonstrate the superiority of the proposed scheme compared to conventional feature normalization schemes. Additional experiments performed on the newly released noisy robust open set speaker identification (ROSSI) and NIST SRE corpora also confirm the effectiveness of the proposed method in stationary room/channel mismatched conditions.

Index Terms: Speaker verification, non-stationary room channel, empirical mode decomposition

1. Introduction

Mismatch between training and test conditions is one of the most important problems facing speaker recognition systems today. Most state-of-the-art speaker recognition systems perform well in clean and reasonably predictable environmental mismatched conditions, but break down in adverse and unpredictable conditions. Recent NIST speaker recognition (SRE) [1] evaluations encouraged researchers to develop promising techniques for tackling channel and microphone mismatch. Approaches based on super-vectors [2] derived from speaker adapted GMMs [3] have dominated the research literature in the last decade, aided by Eigenvoice, Eigenchannel and other factor analysis based methods [4] for channel compensation. Also, low and high vocal effort, language and accent mismatch were some of the issues in the NIST SRE evaluation recently. However, since the main NIST focus has been on channel and microphone mismatch, many researchers have moved away from other important factors that can degrade speaker recognition systems in real life scenarios, such as adverse additive noise or reverberation.

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

One problematic real-life degradation of speech that is frequently overlooked is non-stationarity in the environment. This scenario is becoming quite common due to the use of cell phones. Users can walk into an office from a busy street, or into a corridor from a class-room while talking on their cell phones. Both reverberation and/or noise can be introduced in speech data in these cases. The disruption can be abrupt or slowly varying. This situation is different from session variability since multiple room/channel and noisy condition may exist in the same utterance. In this work, we consider the non-stationarity in the room acoustics only. This will create a non-stationary bias in the cepstral coefficients and may also distort the distribution of the features. Since traditional model domain channel compensation strategies assume that the mismatch is at the utterances level, this kind of mismatch needs to be normalized in the acoustic feature domain before modeling.

Most conventional feature compensation strategies are based on the assumption that the channel or microphone effect is stationary over the entire utterance. Standard cepstral mean subtraction (CMS) [5] can be used to remove stationary convolutional distortion from the cepstral feature stream, which can be applied accross the whole utterance or a sliding window. While utterance level CMS (CMS_U) cannot help in non-stationary mismatched conditions, sliding window based CMS (CMS_W) may remove speaker dependent information present in the lower frequency range [6]. In addition to introducing a shift in the mean, environmental mismatch can also distort the feature distribution which necessitates the use of normalization methods similar to cepstral variance normalization (CVN) or feature warping [7]. Another class of techniques work on applying linear filtering on the feature coefficient stream. RASTA filtering is the most widely used technique in this domain [8]. However, standard RASTA processing has been found to be less effective in speaker recognition since it potentially removes speaker related information during normalization [6, 7].

It is clear that none of the techniques discussed above specifically address the issue of non-stationarity in the channel. A variation of the linear channel will introduce a non-stationary bias in the cepstral features that cannot be removed by CMS_U. Block-wise mean removal (CMS_W) or RASTA-type filtering can help in this case, assuming that the block is small enough to assume stationarity within itself but also large enough to compute the statistics reliably. The effectiveness of feature warping will also be dependent on the sliding window duration. In this paper, we utilize the empirical mode decomposition (EMD) [9] algorithm to remove the non-stationary additive bias on the cepstral feature stream, which, unlike other techniques, does not require an *a priori* stationarity assumption. EMD adaptively decomposes a signal into a set of intrinsic mode functions (IMFs) and a residual. These IMFs contain both amplitude and frequency modulation components and yield meaningful instantaneous frequencies using the Hilbert spectral analysis [10]. This

approach of time-frequency analysis, referred to as the Hilbert-Huang transform (HHT), has recently been used extensively in a wide range of applications [9]. This study is motivated by the intuition that EMD would partially separate the non-stationarity channel effects in its lower order IMFs and the residual trend signal [9], and thus filtering in this domain can be effective in suppressing unwanted components from the cepstral feature stream.

2. EMD and Hilbert spectral analysis

Unlike standard transformations (e.g. Fourier, Wavelet), EMD does not require an *a priori* basis function. The signal is directly expressed as a summation of intrinsic mode functions (IMF) and a residual. An IMF has the following properties: (a) the number of extrema and the number of zero crossings of the function are either equal or differ at most by one; and (b) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero [9]. EMD will decompose a signal $x(t)$ in the following manner:

$$x(t) = \sum_{i=1}^{n} c_i(t) + r(t), \tag{1}$$

where, $c_i(t)$ is the ith IMF and $r(t)$ is the residual. In general, $c_1(t)$ contains the highest scale (shortest period) component of the signal, while the scale gradually reduces in $c_i(t)$ for $i > 1$. Each IMF, by construction, yields a non-negative instantaneous frequency function $\omega(t)$ through the Hilbert spectral analysis [10], derived as follows. For an IMF $c(t)$ if,

$$d(t) = \mathcal{H}\left(c(t)\right) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{c(t)}{t-\tau} \mathrm{d}\tau \tag{2}$$

denotes the Hilbert transform of $c(t)$, where P indicates the Cautchy principal value, then $c(t)$ and $d(t)$ form a complex conjugate pair allowing us to form an analytic function $z(t)$ as,

$$z(t) = c(t) + jd(t) = a(t)e^{j\theta(t)}, \text{ where} \tag{3}$$

$$a(t) = \sqrt{c(t)^2 + d(t)^2} \quad \text{and} \quad \theta(t) = \arctan\left(\frac{d(t)}{c(t)}\right). \tag{4}$$

Here, $a(t)$ and $\theta(t)$ denote the instantaneous amplitude and phase, respectively. The instantaneous frequency is given by,

$$\omega(t) = \frac{\mathrm{d}\theta(t)}{\mathrm{d}t}. \tag{5}$$

Thus, from each IMF $c_i(t)$ we can compute $\omega_i(t)$ and $a_i(t)$ and the signal $x(t)$ can be represented as,

$$x(t) = \mathrm{Re}\left[\sum_{i=1}^{n} a_i(t) \exp\left(j \int \omega_i(t)\mathrm{d}t\right)\right] + r(t) \tag{6}$$

yielding a time-frequency-amplitude representation [9]. The final term $r(t)$ referred to as the residual signal, is the non-IMF portion of the decomposition and is a monotonic trend signal.

3. Proposed Method

At first the conventional Mel-frequency cepstral coefficients (MFCC) are extracted from the given test utterance. Let the kth MFCC coefficient stream at frame index t be denoted by $X_k(t)$. Application of EMD will decompose this signal into n IMFs and a residual signal as in (1):

$$X_k(t) = \sum_{i=1}^{n} c_{i,k}(t) + r_k(t). \tag{7}$$

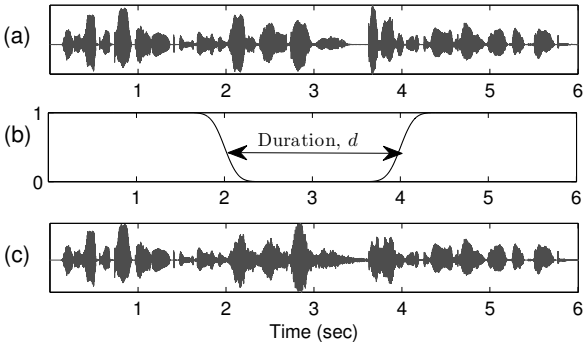

Figure 1: (a) Clean speech signal waveform $s(n)$, (b) the mixing function $m_d(n)$ used to mix $s(n)$ and $s'(n)$ to generate non-stationary room channel and (c) the degraded signal $y(n)$.

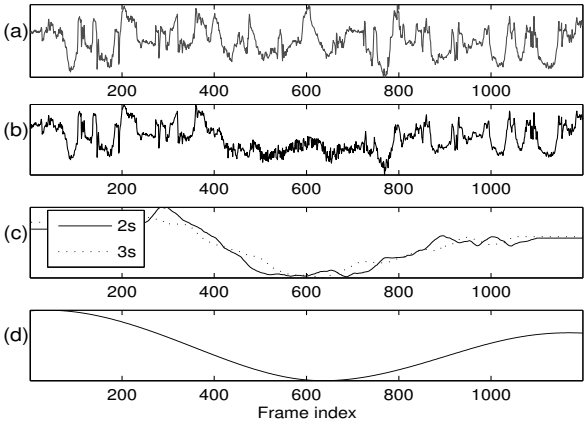

Figure 2: Time trajectory of the 2nd MFCC cepstral coefficient (C_2) for: (a) the clean speech, (b) degraded speech, (c) the block-wise mean computed using 2s and 3s sliding windows from the degraded speech C_2 (d) summation of the residual and the last IMF generated from EMD computed from the degraded speech C_2.

Here $c_{i,k}(t)$ and $r_k(t)$ denote the ith IMF and the residual trend, respectively. At first, we investigate the speaker relevant information contained in the residual. Next, we analyze the IMFs in non-stationary conditions.

3.1. Frequency analysis of the residual

Let there be a total N feature frames in an utterance sampled at skip rate T_s. For the kth cepstrum trajectory $X_k(t)$ we have the EMD residual $r_k(t)$ as in (7). Since the monotonic residual signal is not an IMF, it can only be a signal with at most one extrema or a DC signal. Thus the maximum frequency of the residual signal is theoretically bounded by $F_{\max} = 1/NT_s$. This can occur only if the residual has the form of a raised cosine waveform given by,

$$r(t) = \alpha\left[1 - \cos\left(\frac{2\pi t}{NT_s}\right)\right], \text{where } \alpha \in \mathcal{R}. \tag{8}$$

For a 1 minute duration utterance and $T_s = 0.01$ this corresponds to $F_{\max} = 0.0167$Hz. Since this is below 0.125Hz, we can assume that there is no speaker information below this frequency [6]. From a cepstrum point of view, the residual signal is nothing but the mean trend of the data that can be compara-

ble to a sliding window based mean envelope which is removed during CMS_W. If the channel is stationary, removing the EMD residual is the same as performing CMS_U, since in this case EMD would ideally produce a DC residual signal. Thus we can conclude that discarding the residual signal from each cepstrum trajectory $X_k(t)$ should improve speaker recognition performance, and at the very least, not degrade it.

3.2. EMD domain filtering

We expect that, EMD will be able to separate the additive channel bias due to the non-stationarity in the environment into a set of IMFs. To validate this assumption we perform an experiment on a clean TIMIT utterance, $s(n)$. First we simulate an office room channel degraded signal, $s'(n)$, using an impulse response from the AIR database [11]. The clean signal $s(n)$ and the degraded signal $s'(n)$ was then combined using a mixing function $m_d(n)$ [1] shown in Fig. 1(b) as,

$$y(n) = s(n)m_d(n) + s'(n)(1 - m_d(n)). \qquad (10)$$

The objective here is to generate a non-stationary environment effect for a duration of $d = 2$ seconds by adding signals in two different stationary conditions (clean and office room) using $m_d(n)$. This is an attempt to simulate the effect of changing a room acoustic condition during the speech. Fig. 2 shows the effect of this distortion on the second MFCC coefficient, C_2, over time. From Fig. 2(b) it is seen that the change in the acoustic condition distorted both the mean and the variance of C_2 in frames $400 - 800$. Fig. 2(c) shows the mean trend estimated using a sliding window (as done in CMS_W), and Fig. 2(d) shows the sum of the residual and the final IMF $(c_{n,k}(t) + r_k(t))$ generated by computing EMD on C_2. From these figures we observe that, while CMS_W does estimate the non-stationary trend of C_2, it also shows some high frequency ripples that may contain speaker identity information. On the contrary, the EMD estimated mean-trend is much smoother and thus is a better candidate for the non-stationary trend removal. From this analysis, we conclude that besides the residual signal, the lower order IMFs may also contain components due to the non-stationary environment. Thus we aim at suppressing unimportant low frequency components [6] of the modulation spectrum from the IMFs of the feature stream based on their instantaneous frequency values. Using (6), $X_k(t)$ is given by,

$$X_k(t) = \text{Re}\left[\sum_{i=1}^{n} a_{i,k}(t) \exp\left(j \int \omega_{i,k}(t)\text{d}t\right)\right] + r_k(t).$$

Applying a transfer function $H(\omega)$ on the instantaneous amplitudes $a_{i,k}(t)$ and removing $r_k(t)$ we obtain its filtered version,

$$\begin{aligned}\hat{X}_k(t) &= \text{Re}\left[\sum_{i=1}^{n} a_{i,k}(t)H(\omega_{i,k}(t)) \exp\left(j \int \omega_{i,k}(t)\text{d}t\right)\right] \\ &= \sum_{i=1}^{n} c_{i,k}(t)H(\omega_{i,k}(t)). \qquad (11)\end{aligned}$$

For a low-pass filter in this domain, we define $H(\omega)$ as,

$$H(\omega) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\omega - \omega_L}{\sigma_L}\right)\right], \qquad (12)$$

[1] For an utterance of M seconds duration the mixing function is,

$$m_d(n) = 1 - \frac{1}{2}\left[\text{erf}\left(\frac{n - M/3}{\sigma_m}\right) - \text{erf}\left(\frac{n - M/3 - d}{\sigma_m}\right)\right], \quad (9)$$

where σ_m controls the abruptness and d controls the duration of non-stationarity. In Fig. 1(b), $m_d(n)$ changes from 1 to 0 for a duration $d = 2$ seconds starting from $M/3 = 2$ seconds. σ_m is set to 0.12.

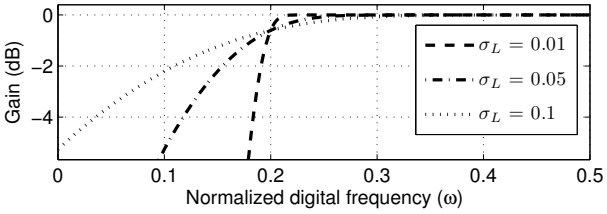

Figure 3: Variation of the transfer function $H(\omega)$ of an EMD domain low-pass filter with an $\omega_L = 0.2$ and various σ_L values. $\omega = 0.5$ is equivalent to 50Hz in the modulation spectrum.

where ω_L denote the lower cut-off frequency and σ controls the cutoff steepness of the filter. An example $H(\omega)$ for a low-pass filter with an $\omega_L = 0.2$ is shown in Fig. 3. In general, components below 0.5Hz in the modulation spectrum ($\omega < 0.005$) can be considered unimportant for speaker identity [6].

4. Experiments

Since there are no commercially available speech database that contains speech utterances in a changing environment in the same session, we generate this effect using the TIMIT database for this evaluation. We are currently in progress of collecting such data from speakers using a mobile recording framework. To test the algorithm on a real speech corpus, we use the NIST SRE database and also use the newly developed ROSSI database. Further details are given in the following subsections.

4.1. System Description

Since this work deals with acoustic feature normalization alone, a classical GMM-UBM system [3] is used for the evaluation. For the front-end we extract MFCC features with 12 coefficients ($\Delta + \Delta\Delta$ was used for NIST and ROSSI data only). We use a 25 ms analysis window with 10 ms shift. A phone recognizer based voice activity detector (VAD) is utilized [12]. Next, we performed one of the following: (a) utterance level CMS (CMS_U), (b) sliding-window based CMS (CMS_W), (c) RASTA filtering, and (d) proposed EMD based compensation (EMDC) Finally, feature warping (FW) [7] is applied using a 3-s sliding window. The system with only FW is referred to as the baseline system. For the proposed EMDC scheme, a low pass filter is applied using (12). For UBM training, the number of mixtures used is 1024 for NIST SRE and 512 for other experiments. UBM is trained using maximum likelihood (ML) criterion and later adapted to each enrollment speaker using classical MAP adaptation [3] with one iteration and a relevance factor of 19.

4.2. System setup for non-stationary environment

In order to evaluate the effectiveness of the proposed feature compensation strategy we generate non-stationary environments in the test data using the AIR database [11]. For the evaluation, we select 50 male and 50 female speakers from the TIMIT database for enrollment testing. 8 conversations (approximately 24 seconds) of data per speaker is used for training and the remaining 2 conversation (approximately 6 seconds) is used for test. From the remaining data, 120 male and 120 female speakers' data is used for UBM training. Each enrollment speaker is tested 1 target and 99 non-target test speakers. To generate non-stationary room environment, we select the following 7 room impulse responses from the AIR database [11]: (a) bathroom, (b) corridor, (c) kitchen, (d) lecture hall, (e) meeting room, (f) office and (g) stairway. We also corrupt

Table 1: EER performance in stationary and non-stationary room channel mismatched conditions in different datasets

System	Non-stationary			Stationary		
	d=1s	d=2s	d=3s	TIMIT	NIST	ROSSI
Baseline	7.99	12.00	10.99	4.12	12.18	2.79
CMS_U+FW	8.01	11.99	11.00	4.19	12.19	2.86
CMS_W+FW	8.98	10.99	10.87	5.89	11.57	3.00
RASTA+FW	7.00	10.01	**9.00**	6.05	17.21	4.98
EMDC+FW	**6.22**	**8.83**	9.03	**3.77**	**11.44**	**2.11**

the waveforms using the FaNT [13] tool that contains the four ITU defined standard telephone channels: (a) G.712, (b) IRS, (c) MIRS and (d) P.341. If, $s(n)$, $h_{\text{air}}(n)$ and $h_{\text{tel}}(n)$ denote the clean signal, room impulse response, and telephone channel impulse response, respectively, the degraded test utterance $y(n)$ with non-stationary room channel and a stationary telephone channel is generated as,

$$y(n) = [s(n)m_d(n) + s(n) * h_{\text{air}}(n)(1 - m_d(n))] * h_{\text{tel}}(n).$$

Here the mixing function $m_d(n)$ given in (9) was used. The parameter d specifies the time duration when $h_{\text{air}}(n)$ is active. With 4 types of $h_{\text{tel}}(n)$, and seven types of $h_{\text{air}}(n)$, a total of 28 different channel conditions were created. Additional four test conditions were generated as follows: three non-stationary conditions with d =1, 2, 3 seconds in $m_d(n)$, and one stationary condition ($d = 0$) with only telephone channel mismatch. For the enrollment and UBM utterances, only $h_{\text{tel}}(n)$ is applied.

4.3. Stationary environment condition

The ROSSI and NIST SRE database is used to evaluate the proposed scheme in standard stationary channel mismatched conditions. The ROSSI database is especially designed for testing and evaluating automatic speaker recognition systems in real environments. Each evaluation set contains 100 in-set speakers (data for both training and testing) and 100 out-of-set speakers. Each utterance is about 2 minutes in duration. 600 additional development speakers' data is also included. These are used for UBM training. Audio data in ROSSI is recorded under various environmental conditions including telephone channels and noise. We utilize the ROSSI set-1 which uses table-mic (far-field) for train and lapel-mic (close talk) for test. Other ROSSI train/test conditions include, cell phone in public, vehicle, office and roadside, and land-line phone in office. For NIST SRE'08 experiments, 5min tel train-interview mic test trials are used for evaluation and the SRE'04 and 05 data for UBM training.

5. Results

The EER performance of the system in different conditions is summarized in Table 1. For the non-stationary room channel cases d =1, 2 and 3s, it is clear that CMS_U was not able to compensate for the mismatch. CMS_W helps in general but degrades the system performance for $d = 1$. This is expected since a $3s$ sliding window is used here. RASTA filtering performed reasonably well. However, the proposed EMDC method has outperformed the other techniques for d =1 and 2. In case of $d = 3$ RASTA performed slightly better. This demonstrates the ability of the proposed method to remove non-stationarity from cepstral trajectory even if it is of small duration.

It is interesting to note that the proposed technique also performs well in stationary mismatched conditions. While, CMS_U, CMS_W and RASTA degrade system performance, the proposed scheme was still able to achieve improvement over baseline system performance. This, we believe, is due to the fact that EMD was effectively able to separate some of the unimportant components in its lower order IMFs that were successfully removed by filtering. Other model domain channel compensation may be used to further improve system performance.

6. Conclusion

In this study we considered the problem of speaker recognition in non-stationary room channel mismatch during test. We have demonstrated that the short time sliding window based feature normalization methods are not fully able to mitigate the mismatch introduced by this type of distortion, and subsequently a new method based on the empirical model decomposition of the feature stream was presented. A new cepstral filtering scheme was presented based on the instantaneous frequencies of the feature stream computed using EMD and Hilbert spectral analysis. Experimental results on synthetic and real-life data demonstrated the effectiveness of the approach in speaker verification.

7. References

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

[2] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, May 2006, pp. 97–100.

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

[4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, July 2008.

[5] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 2003.

[6] S. Van Vuuren and H. Hermansky, "On the importance of components of the modulation spectrum for speaker verification," *Proc. ICSLP*, vol. 2, 1998.

[7] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. A Speaker Odyssey*, 2001, pp. 213–218.

[8] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 2002.

[9] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and Hilbert spectrum for non-linear and non-stationary time series analysis," *Proc. Roy. Soc. London A*, vol. 454, pp. 903–995, 1998.

[10] A. Oppenheim and R. Schafer, *Discrete-time signal processing*. Prentice Hall Signal Processing, 1999.

[11] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. ICDSP 2009*, Santorini, Greece, 2009, pp. 1–5.

[12] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. IEEE ICASSP*, vol. 1, May 2006.

[13] H. G. Hirsch, "FaNT-filtering and noise adding tool." [Online]. Available: http://dnt.kr.hs-niederrhein.de/