FEATURE NORMALIZATION FOR SPEAKER VERIFICATION IN ROOM REVERBERATION

Sriram Ganapathy[1], *Jason Pelecanos*[2] *and Mohamed Kamal Omar*[2]

[1]Dept. of ECE, Johns Hopkins University, USA
[2]IBM T.J Watson Research Center, USA
ganapathy@jhu.edu, {jwpeleca,mkomar}@us.ibm.com

ABSTRACT

The performance of a typical speaker verification system degrades significantly in reverberant environments. This degradation is partly due to the conventional feature extraction/compensation techniques that use analysis windows which are much shorter than typical room impulse responses. In this paper, we present a feature extraction technique which estimates long-term envelopes of speech in narrow sub-bands using frequency domain linear prediction (FDLP). When speech is corrupted by reverberation, the long-term sub-band envelopes are convolved in time with those of the room impulse response function. In a first order approximation, gain normalization of these envelopes in the FDLP model suppresses the room reverberation artifacts. Experiments are performed on the 8 core conditions of the NIST 2008 speaker recognition evaluation (SRE). In these experiments, the FDLP features provide significant improvements on the interview microphone conditions (relative improvements of 20-30%) over the corresponding baseline system with MFCC features.

Index Terms— Frequency Domain Linear Prediction (FDLP), Room Reverberation, Speaker Verification.

1. INTRODUCTION

Most state-of-the-art speaker verification systems perform well in controlled environments where speech data is collected from reasonably clean conditions. However, the performance of these systems are degraded in the presence of reverberation artifacts. This is primarily due to the temporal smearing of short-term spectra which are used for conventional features like MFCCs [1].

A number of feature compensation techniques have been proposed in the past for speaker verification systems (for example, feature warping [2], RASTA processing [3] and cepstral mean subtraction (CMS) [4]). Although these techniques (which are based on short-term spectra of speech) provide good improvements for short-term distortions like telephone channel conditions, they fail to suppress the long-term artifacts caused by room reverberation.

In reverberant environments, the speech signal that reaches the microphone is superimposed with multiple reflected versions of the original speech signal. These superpositions can be modelled by the convolution of the room impulse response, that accounts for individual reflection delays, with the original speech signal, i.e.,

$$r(t) = s(t) * h(t), \qquad (1)$$

where $s(t)$, $h(t)$ and $r(t)$ denote the original speech signal, the room impulse response and the reverberant speech respectively. The effect of reverberation on the short-time Fourier transform (STFT) of the speech signal $s(t)$ can be represented as

$$R(n, \omega_k) = S(n, \omega_k)H(n, \omega_k), \qquad (2)$$

where $S(n, \omega_k)$ and $R(n, \omega_k)$ are the STFTs of the clean speech signal $s(t)$ and reverberant speech $r(t)$ respectively. Here, $H(n, \omega_k)$ denotes the STFT of the room impulse response $h(t)$, n denotes the frame index and w_k denotes the kth frequency bin.

The amount of reverberation in speech is generally characterized by reverberation time (T_{60}) (time required for reflections of a direct sound to decay by 60dB below the level of the direct sound, typically in the range of 200-700ms). The main assumption in conventional short-term channel compensation techniques is $H(n, \omega_k) = H(\omega_k) \forall n$. While this assumption is reasonable for distortions like linear telephone channel noises, it is not valid for long-term artifacts like room reverberations. Thus, by using conventional approaches like CMS (where analysis windows for deriving cepstral features are much shorter than T_{60}), the effect of reverberation cannot be suppressed by a mean subtraction in the cepstral domain.

The use of long-term mean subtraction has been studied in the past for the suppression of room reverberation [5]. This approach involves the subtraction of a mean estimate of the log spectrum using a long-term (2s) analysis window, followed by overlap-add resynthesis. The application of gain normalization of 1s long sub-band temporal envelopes has also shown to be useful for speech recognition in room reverberations [6].

In this paper, we extend the previous approach in [6] for the task of speaker verification in reverberant environments. The suppression of the reverberation artifacts is achieved by a gain normalization of the sub-band temporal envelopes estimated using frequency domain linear prediction (FDLP). This normalization technique assumes a constant value for the sub-band temporal envelope of the room impulse response within the analysis window. In our feature extraction, this assumption is emphasized by analyzing long temporal regions of the speech signal (10s) in narrow sub-bands (96 bands). Finally, the normalized sub-band envelopes are integrated to form mel-band energies and are converted to cepstral features similar to MFCCs.

Experiments are performed on a Gaussian Mixture Model-Universal Background Model (GMM-UBM) speaker verification system [10]. The models are evaluated using all of the 8 core conditions of the NIST 2008 SRE task. In these experiments, the proposed

This research was partly funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U.S. Government.

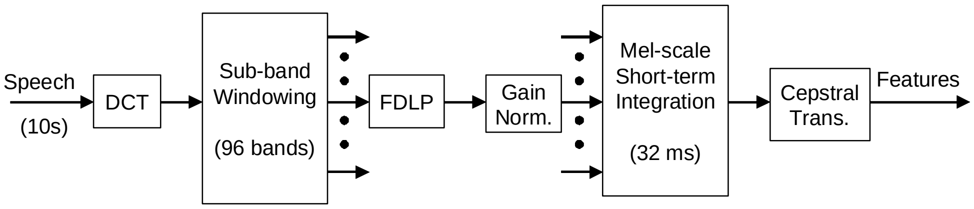

Fig. 2. Block schematic of the proposed feature extraction.

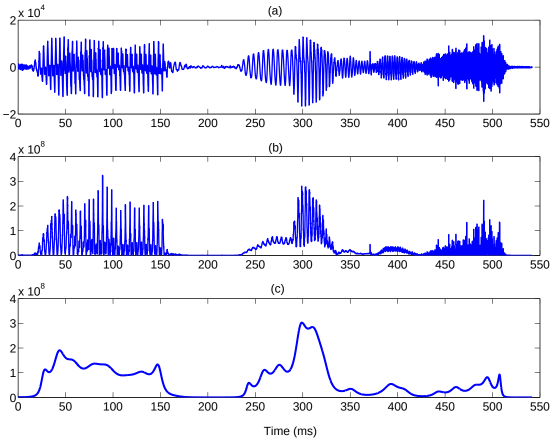

Fig. 1. Illustration of the all-pole modelling property of FDLP. (a) a portion of the speech signal, (b) its temporal envelope, and (c) all pole model obtained using FDLP.

features provide significant improvements on the interview conditions over the baseline system with MFCC features.

The rest of the paper is organized as follows. In Sec. 2, the FDLP technique for feature extraction is explained. Speaker verification experiments with the proposed features are reported in Sec. 3 and 4. In Sec. 5, we conclude with a discussion of the proposed features.

2. FEATURE EXTRACTION

2.1. Deriving sub-band envelopes using FDLP

Conventionally, linear prediction is applied to the speech signal in the time domain to obtain an autoregressive model of the power spectrum of the signal. On the other hand, linear prediction can be applied to the discrete spectral representations of the signal to provide autoregressive models of the temporal envelope[1] of the signal [7]. This technique is referred to as frequency domain linear prediction (FDLP).

In our implementation, the discrete cosine transform (DCT) is applied on long temporal segments (hundreds of ms) and linear prediction is performed on the DCT components to yield a parametric model of the Hilbert envelope of speech. Fig. 1 shows the AR modelling property of FDLP. It shows (a) a portion of the speech signal, (b) its temporal envelope and (c) an all pole approximation for the temporal envelope using FDLP.

The block schematic for the FDLP feature extraction is shown in Fig. 2. Long segments of the input speech signal (10s) are transformed using DCT. A set of rectangular overlapping windows are applied on the DCT components to yield 96 sub-band DCT components. In each sub-band, FDLP is performed by applying linear prediction on the DCT components. FDLP provides a parametric model for the sub-band envelope in the form of an all-pole polynomial (described by $\{a_0, a_1 ..., a_p\}$), where p is the FDLP model order with $a_0 = 1$ for predicting the current sample. In our experiments, we use a model order of 30 poles per sub-band for 1s of speech. The resulting sub-band envelope can be written as,

$$E(t) = \frac{G}{|\sum_{k=0}^{k=p} a_k e^{-i2\pi kt}|^2} \tag{3}$$

where $E(t)$ denotes the FDLP envelope as a function of time, t, (which approximates the sub-band temporal envelope) and G denotes the gain of the all-pole model.

2.2. Gain normalization

When a speech signal is corrupted by room reverberation, the sub-band temporal envelope of the reverberant speech (in narrow sub-bands of long analysis windows) is a convolution of the temporal envelope of clean speech with that of the room impulse response function [8]. For a first-order approximation, the temporal envelope of room impulse response function in narrow sub-bands is assumed to be a constant [6]. Thus, gain normalization of the sub-band envelopes (setting $G = 1$) provides reasonable suppression of the reverberant artifacts in speech.

2.3. Cepstral features

The gain normalized sub-band envelopes are integrated into short-time frames (32ms with a shift of 10ms) using a Hamming window. The frequency axis of the 96 linear sub-bands is warped according to the mel-scale. We use 37 Mel bands in the frequency range of 125-3800 Hz. The output of the integration process provides a gain normalized mel-scale energy representation of speech similar to the mel-spectrogram obtained in conventional MFCC feature extraction [1]. These mel-band energies are converted to cepstral coefficients by using a log operation followed by a DCT (with the matlab package in [9]). We use 13 cepstral coefficients along with derivative and acceleration components yielding 39 dimensional features.

2.4. CMS versus Gain normalization

Cepstral mean subtraction (CMS) tries to suppress the effect of short-term convolutions in speech (like telephone channel distortions) by subtracting the mean of the cepstral features. Generally, the mean is computed over a sliding window (of more than 1s) or over the entire recording. However, if the convolutive effect is spread over

[1] We use the term temporal envelope to denote the Hilbert envelope of the signal, which is the square magnitude of the analytic signal.

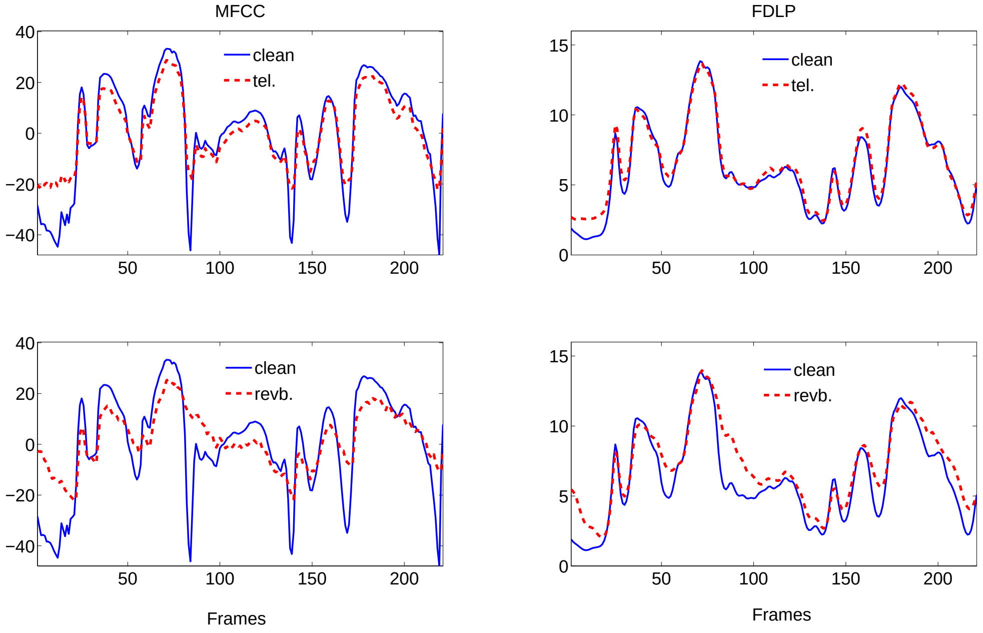

Fig. 3. Comparison of CMS for MFCC and gain normalization for FDLP.

long regions of the speech signal (more than frame duration) such as with room reverberation, CMS is unable to suppress the artifacts. For these distortions, the gain normalization technique used for FDLP features is more effective as the normalization is performed over long segments (10s). Furthermore, the gain normalization technique does not involve a mean computation or a rolling window operation.

The effectiveness of the proposed approach is illustrated in Fig. 3, where we plot $C0$ for MFCC features and FDLP features. In these plots, MFCC features are processed with CMS and the FDLP features are derived from gain normalized sub-band envelopes. The FDLP features provide better invariance to telephone distortions as well as reverberant artifacts compared to MFCC features.

3. SPEAKER VERIFICATION SYSTEM

3.1. Baseline features

The baseline features consist of 39 dimensional MFCC features [1] containing 13 cepstral coefficients, their delta and acceleration components. These features are computed on 32ms frames of speech signal with a shift of 10ms. As with the case of FDLP features, we use 37 Mel-filters in the frequency range of 125-3800 Hz for the baseline features.

3.2. Experimental set-up

The proposed features as well as the baseline features are used in a GMM-UBM based speaker verification system [10]. The input speech features are feature warped [2] and a 512 component GMM is trained on the development data. Once the UBM is trained, the mixture component means are MAP adapted and concatenated to form supervectors [11]. Nuisance attribute projection (NAP) is applied on the supervectors to remove directions which correspond to large intra speaker variability (like session variability). In our system, we remove 64 nuisance directions based on the principal components extracted from the within-class covariance matrix [12].

For the task of verification, scores are computed as

$$s = \Phi_e^T \mathrm{K} \Phi_v \tag{4}$$

where Φ_e, Φ_v are the supervectors corresponding to enrollment and verification recordings respectively, K is the NAP projection matrix and s is the score for this pair of conversation sides. These scores are further normalized using the ZT score normalization procedure [13].

The proposed features are evaluated on the core conditions of the NIST 2008 speaker recognition evaluation (SRE) [14]. The description of the 8 core evaluation conditions is given in Table. 1. The development data set consists of a combination of audio from the NIST 2004 speaker recognition database, the Switchboard II Phase III corpora, the NIST 2006 speaker recognition database, and the NIST08 interview development set. The collection contains 13770 recordings. There are 1769 speakers in the development data: 988 female speakers and 781 male speakers. The development set was used to estimate the UBM parameters, the expected within-class covariance matrix over all speakers for NAP compensation, as well as for gender-dependent ZT score normalization.

4. RESULTS

The baseline features are 39 dimensional MFCC features as described in Sec. 3.1. The FDLP features are used in 3 configurations. All configurations use the gain normalization technique on the FDLP envelopes. FDLP-MEL-1s corresponds to features derived from temporal envelopes directly on the mel-bands (37 bands instead of 96 bands). These features use a temporal analysis window of $1s$ on the input speech similar to [6] (and hence, a $1s$ window for the gain normalization as well). FDLP-MEL-10s also uses mel-band temporal envelopes obtained from an input analysis window of 10s. FDLP-96bands-10s features use a 10s analysis window

Table 1. Core evaluation conditions for the NIST 2008 SRE task.

Cond.	Task
1.	Interview speech in training and test.
2.	Interview speech from the same microphone type in training and test.
3.	Interview speech from different microphones types in training and test.
4.	Interview training speech and telephone test speech.
5.	Telephone training speech and non-interview microphone test speech.
6.	Telephone speech in training and test from multiple languages.
7.	English language telephone speech in training and test.
8.	English language telephone speech spoken by a native U.S. English speaker in training and test.

Table 2. Performance of various features in terms of min DCF ($\times 10^3$) and EER (%) in parentheses.

Feature.	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Cond. 5	Cond. 6	Cond. 7	Cond. 8
MFCC (Baseline)	28.8 (5.3)	3.2 (0.8)	29.7 (5.4)	35.5 (7.8)	32.1 (7.9)	41.1 (7.6)	15.5 (3.3)	15.0 (3.5)
FDLP-MEL (1s)	28.4 (5.2)	3.1 (0.7)	29.2 (5.3)	36.1 (8.8)	29.1 (7.6)	44.2 (8.1)	14.0 (3.1)	15.1 (3.4)
FDLP-MEL (10s)	24.4 (4.8)	2.2 (0.8)	24.9 (4.9)	32.8 (7.5)	26.0 (6.2)	42.2 (7.7)	12.9 (3.0)	13.4 (3.5)
FDLP-96 bands (10s)	19.7 (3.6)	1.5 (0.3)	20.5 (3.7)	27.1 (6.4)	24.3 (6.8)	45.8 (8.2)	14.6 (3.4)	13.5 (3.2)

and derive temporal envelopes in 96 sub-bands. Gain normalization is applied and the sub-band envelopes are warped back to mel-scale as described in Sec. 2.

The speaker verification results for the various feature extraction techniques are reported in Table 2. FDLP-MEL-1s features provide performances similar to the baseline MFCC features. When the analysis window is increased to 10s, there is a relative performance improvement of about 15% on almost all the conditions. Furthermore, applying an initial sub-band analysis of 96 bands provides significant improvements for the interview mic conditions (relatively about 20-30% over the baseline system). This is due to the application of gain normalization on longer analysis windows in narrow sub-bands which validates the first order approximation made in the technique. A drop in performance is observed for Cond. 6 which may be attributed to the use of different languages in training and test conditions (where the use of longer context degrades the performance).

5. SUMMARY

In this paper, we have proposed a feature normalization technique for speaker verification in reverberant conditions. The normalization procedure is applied on FDLP features derived from long temporal segments of speech in narrow sub-bands. The application of the gain normalization is followed by the integration of the sub-band envelopes to provide short-term mel-band energies similar to the mel-spectrogram in MFCC feature extraction. In this way, the proposed technique can be viewed as a pre-processing mechanism for the MFCC features to improve robustness in reverberant environments.

6. REFERENCES

[1] Davis, B. and Mermelstein, P., “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 28, pp. 357-366, 1980.

[2] Pelecanos, J. and Sridharan, S., “Feature warping for robust speaker verification”, *Proc. Speaker Odyssey 2001 Speaker Recognition Workshop*, Greece, pp. 213-218, 2001.

[3] Hermansky, H. and Morgan, N., “RASTA processing of speech,” *IEEE Trans. on Speech and Audio Process.*, Vol. 2, pp. 578-589, 1994.

[4] Furui, S., “Cepstral analysis technique for automatic speaker verification”, *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 29, pp. 254-272.

[5] Gelbart, D. and Morgan, N., “Double the trouble: handling noise and reverberation in far-field automatic speech recognition,” *Proc. ICSLP*, Colorado, USA, pp. 2185-2188, 2002.

[6] Thomas, S., Ganapathy, S. and Hermansky, H., “Recognition of reverberant speech using frequency domain linear prediction,” *IEEE Signal Proc. Letters*, Vol. 15, pp. 681-684, 2008.

[7] Athineos, M. and Ellis, D., “Autoregressive modelling of temporal envelopes,” *IEEE Tran. Signal Proc.*, Vol. 55, pp. 5237-5245, 2007.

[8] Mourjopoulos, J. and Hammond, K., “Modelling and enhancement of reverberant speech using an envelope convolution method,” *Proc. ICASSP*, Boston, USA, pp. 1144-1147, 1983.

[9] Ellis, D., *http://labrosa.ee.columbia.edu/matlab/rastamat/*.

[10] Reynolds, D., Quatieri, T. and Dunn, R., “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Process.*, Vol. 10 (1-3), pp. 19-41, 2000.

[11] Campbell, W., Sturim, D., Reynolds, D. and Solomonoff, A., “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation”, *Proc. ICASSP*, France, pp. 97-100, 2006.

[12] Hatch, A., Kajarekar, S. and Stolcke, A., “Within-class covariance normalization for SVM-based speaker recognition,” *Proc of Interspeech*, Pennsylvania, USA, pp. 1471-1474, 2006.

[13] Auckenthaler, R., Carey, M. and Lloyd-Thomas, H., “Score normalization for text-independent speaker verification systems,” *Digital Signal Process.*, Vol. 10, pp. 42-54, 2000.

[14] “National Institute of Standards and Technology (NIST),” speech group website, *http://www.nist.gov/speech*, 2008.