Feature Extraction Using 2-D Autoregressive Models For Speaker Recognition

Sriram Ganapathy[1], *Samuel Thomas*[1] *and Hynek Hermansky*[1,2]

[1]Dept. of ECE, Johns Hopkins University, USA
[2]Human Language Technology Center of Excellence, Johns Hopkins University, USA
{ganapathy,samuel,hynek}@jhu.edu

Abstract

The degradation in performance of a typical speaker verification system in noisy environments can be attributed to the mis-match in the features derived from clean training and noisy test conditions. The mis-match is severe in low-energy regions of the signal where noise dominates the speech signal. A robust feature extraction scheme should focus on the high-energy peaks in the time-frequency region. In this paper, we develop a signal analysis technique which attempts to model these high-energy peaks using two-dimensional (2-D) autoregressive (AR) models. The first AR model of the sub-band Hilbert envelopes is derived using frequency domain linear prediction (FDLP). Then, these all-pole envelopes from each sub-band are converted to short-term energy estimates and the energy values across various sub-bands are used as a sampled power spectral estimate for the second AR model. The output prediction coefficients from the second AR model are converted to cepstral coefficients and are used for speaker recognition. Experiments are performed using noisy versions of NIST 2010 speaker recognition evaluation (SRE) data with the state-of-art speaker recognition system. In these experiments, the proposed features provide significant improvements compared to baseline MFCC features (relative improvements of 30%). We also experiment on a large dataset of IARPA NIST 2011 speaker recognition challenge, where the 2-D AR model provides noticeable improvements (relative improvements of $15 - 20\%$).

1. Introduction

Speaker recognition in noisy environments continues to be a challenging problem mainly due to the mis-match in speech data from training and test. One common solution to overcome this mis-match is the use of multi-condition training [1] where the speaker models are trained using data from the target domain. However, in a realistic scenario it is not always possible to obtain reasonable amounts of training data from all types of noisy and reverberant environments for training the speaker models. Therefore, there is a need to attain noise robustness either at the front-end signal analysis or at the statistical speaker models. In this paper, we address the robustness issues in feature extraction.

Various techniques like spectral subtraction [2], Wiener filtering [3] and missing data reconstruction [4] have been proposed for noisy speech recognition scenarios. Feature compensation techniques have also been used in the past for speaker verification systems (for example, feature warping [5], RASTA processing [6] and cepstral mean subtraction (CMS) [7]). However, the mel frequency cepstral coefficients (MFCC) [8] with mean and variance normalization continue to represent the common front-end analysis scheme in state-of-art speaker recognition systems.

When speech is corrupted with additive noise, the valleys in the sub-band envelopes are filled with noise. Even with moderate amounts of noise, the low-energy regions are substantially modified and cause acoustic mis-match with the clean training data. Thus, a robust feature extraction scheme must rely on the high energy regions in the spectro-temporal plane. In general, an autoregressive (AR) modeling approach represents high energy regions with good modeling accuracy [9, 10]. One dimensional AR modeling of signal spectra is widely used for feature extraction of speech [11]. In the past, one dimensional AR modeling of Hilbert envelopes have also been used for speaker verification [12]. 2-D AR modeling was originally proposed for speech recognition by alternating the AR models between spectral and temporal domains [13].

In this paper, we propose a feature extraction technique based on two-dimensional (2-D) spectro-temporal AR models. The initial model is the temporal AR model based on frequency domain linear prediction [14, 15]. The FDLP model is derived by the application of linear prediction on the discrete cosine transform (DCT) of the sub-band speech signal. We use an initial sub-band decomposition of 96 sub-bands in a linear scale. The sub-band FDLP envelopes are integrated in short-term segments to obtain sub-band energy estimates. In each short-term frame, the energy values across the sub-bands form a sampled power spectral density (PSD) estimate. The inverse Fourier transform of this PSD provides autocorrelations which are used for the spectral AR model. The prediction coefficients from the second AR model are converted to cepstral coefficients using the cepstral recursion [16]. These cepstral parameters are used as features for speaker recognition.

Experiments are performed on core conditions of NIST 2010 SRE data [17]. The speaker recognition system is based on Gaussian mixture model-universal background model (GMM-UBM). We use factor analysis methods on the GMM supervectors [18] with i-vector probabilistic linear discriminant analysis (PLDA) for score computation [19]. In order to determine the noise robustness of the speaker recognition, we use data from condition 2 (interview mic-training with interview mic-testing) of SRE 2010 data added with various noise types and signal-to-noise rations. The choice of condition 2 is motivated in part by the potential application of speaker recognition technologies on handheld devices with distant microphones in noisy environments. In these experiments, the proposed 2-D AR model provides considerable improvements compared to

This research was funded by the Intelligence Advanced Research Projects Activity (IARPA) through the Army Research Laboratory (ARL), Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20015 and the Office of the Director of National Intelligence (ODNI). The authors would like to acknowledge Brno University of Technology, Xinhui Zhou and Daniel Garcia-Romero for code fragments.

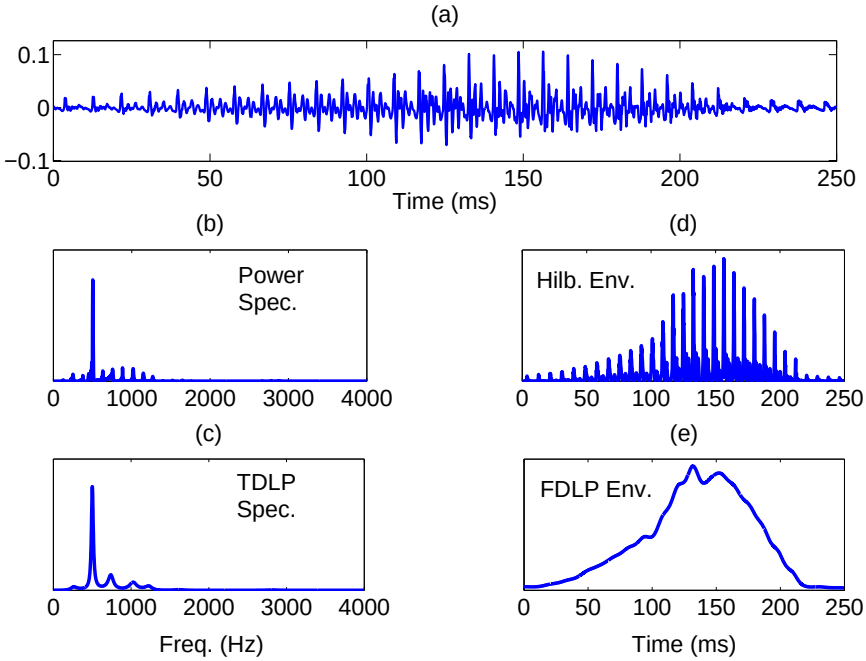

Figure 1: *Illustration of AR modeling in time and frequency domain - (a) a portion of voiced speech, (b) power spectrum, (c) AR model of power spectrum obtained from TDLP, (d) Hilbert envelope and (e) AR model of Hilbert envelope using FDLP.*

the conventional MFCC system (relative improvements of about 30%).

We also measure the performance of these speaker verification systems on a large data-set from IARPA BEST evaluation challenge 2011 [20]. The speech data in these evaluations contain a wide variety of intrinsic variabilities (within speaker variations like vocal effort), extrinsic variabilities (include differences in room acoustics, noise level, sensor differences and speech coding) and parameter variabilities (variations to different languages, aging factors etc). In these evaluations, the proposed 2-D model outperforms the MFCC system in most of the testing conditions (relative improvements of $15-20\%$).

The rest of the paper is organized as follows. In Sec. 2, we outline the linear prediction approaches in the spectral and temporal domain. Sec. 3 details the proposed feature extraction scheme using 2-D AR model. Sec. 4 describes our experimental setup used for the NIST 2010 SRE. The results of these evaluations are reported in Sec. 5. Sec. 6 describes the speaker recognition experiments using the IARPA BEST database. In Sec. 7, we conclude with a brief discussion of the proposed front-end.

2. AR Modeling in Time and Frequency

2.1. Spectral AR model - TDLP

Spectral AR modeling has been widely used in speech and audio signal processing for about four decades now [9, 10]. Let $x[n]$ denote the input signal for $n = 0, \ldots, N-1$. The time domain LP model is to identify the set of coefficients $a_j, j = 1, \ldots, p$ such that $\sum_{j=1}^{p} a_j x[n-j]$ approximates $x[n]$ in a least square sense [9], where p denotes the model order.

Let $\mathbf{r}_x[\tau]$ denote the autocorrelation sequence for time domain signal $x[n]$ with lag τ ranging from $-N+1, \ldots, N-1$.

$$r_x[\tau] = \frac{1}{N} \sum_{n=|\tau|}^{N-1} x[n]x[n-|\tau|] \tag{1}$$

Let $\hat{x}[n]$ denote the zero-padded signal $\hat{x}[n] = x[n], \;\; n = 0, .., N-1$ and $\hat{x}[n] = 0, \; for \; n = N, .., 2N-1$. The relation between the power spectrum of the zero-padded signal $P_x[k] = |\hat{X}[k]|^2$ and the autocorrelation $\mathbf{r}_x[\tau]$ is given by,

$$P_x[k] = \mathcal{F}\big[r_x[\tau]\big] \tag{2}$$

where $\hat{X}[k]$ is the discrete Fourier transform (DFT) of the signal $\hat{x}[n]$ for $k = 0, \ldots, 2N-1$. This relation is used in the AR modeling of the power spectrum of the signal [10]. Time domain linear prediction (TDLP) refers to the use of time domain autocorrelation sequence to solve the linear prediction problem. The optimal set of a_j along with the variance of prediction error G and $a_0 = 1$ provides an AR model of the power spectrum,

$$\hat{P}_x[k] = \frac{G}{|\sum_{j=0}^{j=p} a_j e^{-i2\pi jk}|^2} \tag{3}$$

An illustration of AR model of power spectrum obtained from TDLP is shown in Fig. 1, where we plot the original power spectrum in (b) for a 250 ms portion of speech signal in (a). The TDLP approximation of the power spectrum in shown in Fig. 1 (c). We use a model order of 40.

2.2. Temporal AR model - FDLP

Linear prediction in the spectral domain was first proposed by Kumaresan [14]. The analog signal theory is used for developing the concept and the extension of the solution for a discrete-sample case is provided. This was reformulated by Athineos

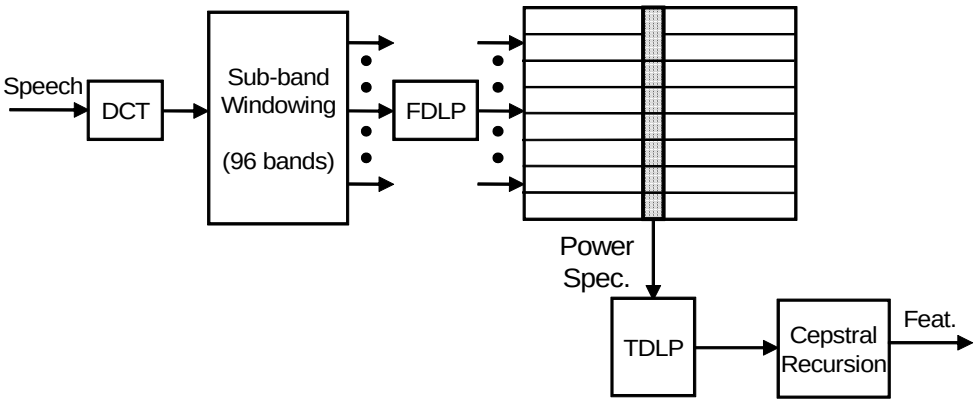

Figure 2: *Block schematic of the proposed feature extraction using 2-D AR modeling.*

and Ellis [15] using matrix notations and the connection with DCT sequence is established. In this paper, we derive the discrete-time relations underlying the FDLP model without using matrix notations. We begin with the definition of analytic signal (AS). Then, we show the Fourier transform relation between the squared magnitude of AS, a.k.a. Hilbert envelope and the autocorrelation of DCT signal. This would mean that linear prediction in DCT domain can be used for AR modeling of the Hilbert envelope of the signal.

In a discrete-time case, an "analytic" signal (AS) $x_a[n]$ can be defined using the following procedure [21]-

1. Compute the N-point DFT sequence $X[k]$
2. Find the N-point DFT of the AS as,

$$X_a[k] = \begin{cases} X[0] & \text{for } k = 0 \\ 2X[k] & \text{for } 1 \leq k \leq \frac{N}{2} - 1 \\ X[\frac{N}{2}] & \text{for } k = \frac{N}{2} \\ 0 & \text{for } \frac{N}{2} + 1 \leq k \leq N \end{cases} \tag{4}$$

3. Compute the inverse DFT of $X_a[k]$ to obtain $x_a[n]$

We assume that the discrete-time sequence $x[n]$ has a zero-mean property in time and frequency domains, i.e., $x[0] = 0$ and $X[0] = 0$. This assumption is made so as to give a direct correspondence between the DCT of the signal and DFT. Further, these assumptions are mild and can be easily achieved by appending a zero in the time-domain and removing the mean of the signal.

The type-I odd DCT $y[k]$ of a signal for $k = 0, \ ... \ , N-1$ is defined as [23]

$$y[k] = 4 \sum_{n=0}^{N-1} c_{n,k} x[n] \cos\left(\frac{2\pi nk}{M}\right) \tag{5}$$

where the constants $M = 2N - 1$, $c_{n,k} = 1$ for $n, k > 0$ and $c_{n,k} = \frac{1}{2}$ for $n, k = 0$ and $c_{n,k} = \frac{1}{\sqrt{2}}$ for the values of n, k, where only one of the index is 0. The DCT defined by Eq. 5 is a scaled version of the original orthogonal DCT with a factor of $2\sqrt{M}$.

We also define the even-symmetrized version $q[n]$ of the input signal,

$$q[n] = \begin{cases} x[n] & \text{for } n = 0 \ , .. , \ N-1 \\ x[M-n] & \text{for } n = N, \ ... \ , M-1 \end{cases} \tag{6}$$

A important property of $q[n]$ is that it has a real spectrum given by,

$$Q[k] = 2 \sum_{n=0}^{N-1} x[n] \cos\left(\frac{2\pi nk}{M}\right) \tag{7}$$

for $k = 0, \ ... \ , M-1$.

For signals with the zero-mean property in time and frequency domains, we can infer from Eq. 5 and Eq. 7 that,

$$y[k] = 2Q[k] \tag{8}$$

for $k = 0, \ ... \ , N-1$. Let $\hat{y}$ denote the zero-padded DCT with $\hat{y}[k] = y[k]$ for $k = 0, \ ... \ , N-1$ and $\hat{y}[k] = 0$ for $k = N, \ ... \ , M-1$. From the definition of Fourier transform of the analytic signal in Eq. 4, and using the definition of the even symmetric signal in Eq. 6, we find that,

$$Q_a[k] = \hat{y}[k] \tag{9}$$

for $k = 0, \ ... \ , M-1$. This says that the AS spectrum of the even-symmetric signal is equal to the zero-padded DCT signal. In other words, the inverse DFT of the zero-padded DCT signal is the even-symmetric AS. Since the auto-correlation of signal $x[n]$ is related to the power spectrum $|\hat{X}[k]|^2$ (Eq. 2), we can obtain a similar relation to the auto-correlation of the DCT sequence.

The auto-correlation of the DCT signal defined as (similar to Eq. 1),

$$r_y[\tau] = \frac{1}{N} \sum_{k=|\tau|}^{N-1} y[k] y[k - |\tau|] \tag{10}$$

From Eq. 9, the inverse DFT of zero-padded DCT signal $\hat{y}[k]$ is the AS of the even-symmetric signal. It can be shown that,

$$r_y[\tau] = \frac{1}{N} \sum_{n=0}^{M-1} |q_a[n]|^2 e^{-j\frac{2\pi n\tau}{M}} \tag{11}$$

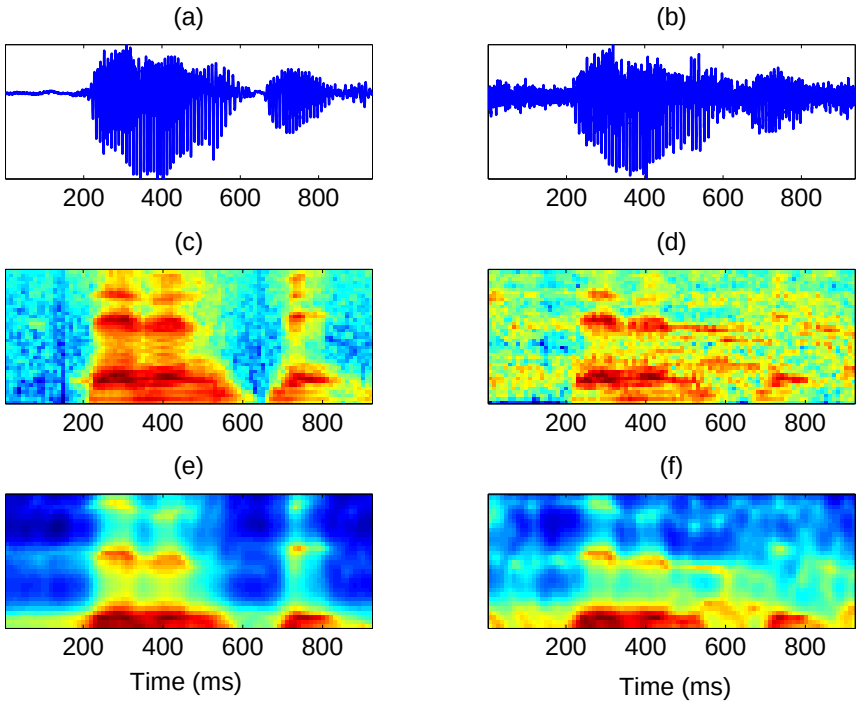

Figure 3: *Comparing mel spectrogram with 2-D AR model spectrogram - (a) a portion of clean speech, (b) a portion of noisy speech (babble noise at* 10 *dB), (c) Mel spectrogram of clean speech, (d) Mel spectrogram of noisy speech (e) 2-D AR model spectrogram of clean speech and (f) 2-D AR model spectrogram of noisy speech.*

i.e., the auto-correlation of the DCT signal and the squared magnitude of the AS (Hilbert envelope) of the even-symmetric signal are Fourier transform pairs. This is exactly dual to the relation in Eq. 2. In other words, we have established that AR modeling of Hilbert envelope can be achieved by linear prediction of DCT components. The AR modeling property of FDLP is illustrated in Fig. 1 where we plot the discrete time Hilbert envelope of the signal in (d) and the FDLP envelope in (e) using a model order of 40. As seen in this figure, the temporal AR model provided by FDLP is dual to the spectral AR model provided by TDLP.

3. 2-D AR Modeling

The block schematic for the proposed feature extraction is shown in Fig. 2. Long segments of the input speech signal (10s) are transformed to the frequency domain using a DCT [12]. The full-band DCT signal is windowed into a set of 96 linear sub-bands in the frequency range of 125-3800 Hz [22]. In each sub-band, linear prediction is applied on the sub-band DCT components to estimate an all-pole representation of Hilbert envelope. We use a model order of 30 poles per sub-band per second. At the output of this stage we obtain the temporal AR model. The FDLP envelopes in each sub-band are integrated in short-term frames (25ms with a shift of 10ms). The output of the integration process provides an estimate of the power spectrum of signal in the short-term frame level. The frequency resolution of this power spectrum is equal to the initial sub-band decomposition of 96 bands.

The power spectral estimates from the short-term integration are inverse Fourier transformed to obtain an autocorrelation sequence. This autocorrelation sequence is used for TDLP with a model order of 12. The TDLP model provides an all-pole approximation of the 96 point short-term power spectrum. The output LP parameters of this AR model are transformed to 13 dimensional cepstral coefficients using the standard cepstral recursion [16]. Delta and acceleration coefficients are extracted to obtain 39 dimensional features which are used for speaker recognition.

In Fig. 3, we show the spectrographic representation of clean and noisy speech (babble noise at 10 dB) using the mel-spectrogram as well as the 2-D AR model based spectrogram. As shown in this figure, the conventional mel-spectrogram is modified significantly due to the presence of additive noise (Fig. 3 (c) and (d)) which will cause a mis-match between the clean training and noisy test conditions. The 2-D AR model spectrogram is relatively more robust compared to Mel spectrogram ((Fig. 3 (e) and (f)). When features are derived from 2-D AR model, the mis-match between clean and noisy conditions is reduced.

4. Experimental Setup

We use a GMM-UBM based speaker verification system [24]. The input speech features are feature warped [5] and gender dependent GMMs with 1024 mixture components are trained on the development data. The development data set consists of a combination of audio from the NIST 2004 speaker recognition database, the Switchboard II Phase 3 corpora, the NIST 2006 speaker recognition database, and the NIST08 interview development set. There are 4324 male recordings and 5461 female recordings in development set.

Table 1: EER (%) and False Alarm (%) at 10% Miss Rate (Miss10) in parantheses for core evaluation conditions in NIST 2010 SRE.

Cond.	MFCC-baseline	2-D AR Feat.
1. Int.mic - Int.mic-same-mic.	2.1 (0.1)	1.8 (0.1)
2. Int.mic - Int.mic-diff.-mic.	3.0 (0.5)	2.7 (0.3)
3. Int.mic - Phn.call-tel	3.8 (0.9)	3.8 (0.9)
4. Int.mic - Phn.call-mic	3.4 (0.5)	2.9 (0.3)
5. Phn.call - Phn.call-diff.-tel	2.9 (0.5)	3.6 (0.9)
6. Phn-call - Phn.call-high-vocal-effort-tel	4.5 (1.5)	5.3 (2.5)
7. Phn-call - Phn.call-high-vocal-effort-mic	7.6 (4.9)	4.6 (1.9)
8. Phn-call - Phn.call-low-vocal-effort-tel	1.9 (0.2)	2.9 (0.6)
9. Phn-call - Phn.call-low-vocal-effort-mic	1.8 (0.1)	1.5 (0.1)

Table 2: EER (%) and False Alarm (%) at 10% Miss Rate (Miss10) in parantheses for condition 2.

Noise	SNR (dB)	MFCC-baseline	2-D AR Feat.
Babble	20	3.8 (0.8)	3.3 (0.5)
	15	4.8 (1.6)	4.0 (0.8)
	10	7.2 (4.5)	5.9 (2.6)
	5	12.0 (15.2)	10.3 (10.6)
Exhall	20	3.7 (0.8)	3.1 (0.5)
	15	4.3 (1.3)	3.7 (0.7)
	10	5.9 (2.9)	5.1 (1.6)
	5	9.4 (8.7)	7.9 (5.7)
Restaurant	20	3.6 (0.8)	3.2 (0.5)
	15	4.3 (1.3)	3.8 (0.8)
	10	6.0 (2.9)	5.2 (1.9)
	5	9.4 (8.8)	8.4 (6.5)

Once the UBM is trained, the mixture component means are MAP adapted and concatenated to form supervectors. We use the i-vector based factor analysis technique [18] on these supervectors in a gender dependent manner. For the factor analysis training, we use the development data from Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2, NIST04-05 and extended NIST08 far-field data. There are 17130 male recordings and 21320 female recordings in this sub-space training set. Gender specific i-vectors of 450 dimensions are extracted and these are used to train a PLDA system [19]. The output scores are obtained using a 250 dimensional PLDA subspace for each gender.

5. Results on NIST 2010 SRE

The proposed features are used to evaluate the core conditions of the NIST 2010 speaker recognition evaluation (SRE) [17]. There are 9 conditions in the NIST 2010 which are described in Table 1. The baseline features consist of 39 dimensional MFCC features [8] containing 13 cepstral coefficients, their delta and acceleration components. These features are computed on 25ms frames of speech signal with a shift of 10ms. We use 37 Mel-filters in the frequency range of 125-3800 Hz for the baseline features.

The performance metric used is the EER (%) and the false-alarm rate at a miss-rate of 10 % (Miss10). The Miss10 is an useful metric for variety of applications in which a low false-alarm rate is desired. The speaker recognition results for the baseline system as well as the proposed 2-D AR features is shown in Table 1. From these results, it can be seen that the proposed 2-D features provides good improvements in mismatched far-field microphone conditions like Cond. 1,2 7 and 9). In these conditions the modeling of high-energy regions in time-frequency domain is beneficial. However, the baseline MFCC system performs well in telephone channel matched conditions (Cond. 5, 6 and 8.)

For evaluating the robustness of these features in noisy conditions, the test data for Cond-2 is corrupted using (a) babble noise, (b) exhibition hall noise, and (c) restaurant noise from the NOISEX-92 database, each resulting in speech at 5, 10, 15 and 20 dB SNR. These noises are added at various SNRs using the FaNT tool [25]. The generation of the noisy version of the test data is done using the setup described in [26]. The choice of condition-2 is motivated in part by speaker recognition applications in far-field noisy environments. Further, the IARPA BEST evaluation [20] also targets noisy data recorded using interview microphone. Condition-2 has the highest number of trials in the NIST 2010 SRE evaluation with 2.8M trials and it contains 2402 enrollment recordings and 7203 test recordings. Enrollment data is the NIST 2010 clean speech data and voice-activity decisions provided by NIST are used in these experiments. For these noisy speaker recognition experiments, the GMM-UBM, i-vector and the PLDA sub-spaces trained from the development data are used without any modification.

The results of noisy speaker recognition experiments is shown in Table. 2. The results of the proposed features are consistently better than the baseline feature for all noise types and signal-to-noise-ratios. On the average, the proposed features provide about 35 % relative Miss10 improvement over the baseline MFCC system. These improvements are mainly due to the robust representation of the high energy regions by two dimensional AR modeling. When the signal is distorted by noise, these peaks are relatively well preserved and therefore the speaker recognition system based on these features outperforms the MFCC baseline system.

6. Results on BEST 2011 Challenge

The speaker verification systems outlined in the previous section are used for a speaker verification task using the IARPA BEST 2011 data [20]. The database contains 83198 recordings (25822 enrollment utterances and 57376 test utterances) with a wide-variety of intrinsic and extrinsic variabilities like language, age, noise and reverberation. There are 38M trials which are split into various conditions as shown in Table 3. Condition 1 contains majority of the trials (20M trials) recorded using interview microphone data with varying amounts of additive noise and artificial reverberation. We use the GMM-UBM and factor analysis models trained using the development data

Table 3: False Alarm (%) at 10% Miss Rate (Miss10) for evaluation conditions in IARPA BEST 2011 task.

Cond.	MFCC-baseline	2-D AR Feat.
1. Int.mic - Int.mic-noisy.	15.5	11.3
2. Int.mic - Phn-call-mic	3.7	2.8
3. Int.mic - Phn.call-tel	3.3	2.8
4. Phn-call-mic - Phn.call-mic	7.4	6.7
5. Phn.call-mic - Phn.call-tel	7.5	6.3
6. Phn.call-tel - Phn.call-tel	1.3	1.8

(Sec. 4) for these experiments. For these speaker recognition experiments, we use the automatic voice activity decision obtained using multi-layer-perceptrons [27].

The performance (Miss10)[1] for the baseline MFCC system is compared with proposed features in Table 3. In these experiments, the proposed features provide noticeable improvements for all conditions except the matched telephone scenario (Cond. 6). On the average, the proposed features provide improvements of about 18% in the Miss10 metric relative to the baseline system.

7. Summary

In this paper, we have proposed a two-dimensional autoregressive model for robust speaker recognition. An initial temporal AR model is derived from long segments of the speech signal. This model provides Hilbert envelopes of sub-band speech which are integrated in short-term frames to obtain power spectral estimates. These estimates are used for a spectral AR modeling process and the output prediction coefficients are converted to cepstral parameters for speaker recognition. Various experiments are performed with noisy test data on NIST 2010 SRE where the proposed features provide significant improvements. These results are also validated using a large speaker recognition dataset from BEST. The results are promising and encourage us to pursue the problem of joint 2-D AR modeling instead of a separable time and frequency linear prediction schemes adopted in this paper.

8. References

[1] Ming, J., Hazen, T.J., Glass, J.R. and Reynolds, D.A., "Robust Speaker Recognition in Noisy Conditions", *IEEE Tran. on Audio Speech Lang. Proc.*, Vol 15 (5), 2007, pp. 1711 - 1723.

[2] Boll, S.F.,"Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 27 (2), Apr. 1979, pp. 113-120.

[3] "ETSI ES 202 050 v1.1.1 STQ; Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms", 2002.

[4] Cooke, M., Morris, A., Green, P., "Missing data techniques for robust speech recognition", *Proc. ICASSP*, 1997, pp. 863-866.

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

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

[7] Rosenberg, A.E., Lee, C. and Soong, F.K., "Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification," in *Proc. ICSLP*, pp. 1835-1838, 1994.

[8] Davis, S. and Mermelstein, R., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 28 (4), Aug. 1980, pp. 357-366.

[9] Atal, B.S., Hanauer, L.S, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *J. Acoust. America*, Vol 50 (28), 1971, pp. 637-655.

[10] Makhoul, J., "Linear Prediction: A Tutorial Review",in *Proc. of the IEEE*, Vol 63(4), pp. 561-580, 1975.

[11] Hermansky, H., "Perceptual Linear Predictive (PLP) Analysis of Speech," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738-1752, 1990.

[12] Ganapathy, S., Pelecanos, J. and Omar, M.K., "Feature Normalization for Speaker Verification in Room Reverberation", *Proc. ICASSP*, 2011, pp. 4836-4839.

[13] Athineos, M. and Hermansky, H. and Ellis, D., "PLP2 Autoregressive modeling of auditory-like 2-D spectro-temporal patterns", *Proc. ISCA Tutorial Research Workshop Statistical and Perceptual Audio Processing SAPA04*, pp. 3742, 2004.

[14] Kumerasan, R. and Rao, A., "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *Journal of Acoustical Society of America*, Vol. 105, no 3, pp. 1912-1924, Mar. 1999.

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

[16] Atal, B.S., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *J. Acoust. America*, Vol 55 (6), 1974, pp. 1304-1312.

[17] "National Institute of Standards and Technology (NIST)," speech group website, *http://www.nist.gov/speech*, 2010.

[18] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P and Ouellet, P., "Front-End Factor Analysis for Speaker Verification", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 19(4), pp. 788-798, 2011.

[19] Romero, D. and Espy-Wilson, C.Y., "Analysis of i-vector Length Normalization in Speaker Recognition Systems", *Proc. Interspeech*, 2011.

[1] At this moment, the key files for these experiments are not available and thus the EERs are not reported here.

[20] “IARPA BEST Speaker Recognition Challenge 2011”, *http://www.nist.gov/itl/iad/mig/best.cfm* , 2011

[21] Marple, L.S., “Computing the Discrete-Time Analytic Signal via FFT”, *IEEE Trans. on Acoust., Speech and Sig. Proc.*, Vol. 47, pp. 2600-2603, 1999.

[22] Thomas, S., Ganapathy, S. and Hermansky, H. “Recognition of Reverberant Speech Using Frequency Domain Linear Prediction,” *IEEE Signal Proc. Letters*, Vol. 15, Dec. 2008, pp. 681-684.

[23] Martucci, S.A., “Symmetric convolution and the discrete sine and cosine transforms”, *IEEE Tran. Signal Proc.*, Vol. 42(5), 1994 pp. 1038-1051.

[24] Reynolds, D., “Speaker Identification and Verification using Gaussian Mixture Speaker Models,” *Speech Comm.* Vol. 17, Aug. 1995, pp. 91-108.

[25] Hirsch, H.G., “FaNT: Filtering and Noise Adding Tool”, *http://dnt.kr.hsnr.de/download.html*.

[26] Gelbart, D. “Ensemble Feature Selection for Multi-Stream Automatic Speech Recognition”, *Ph. D. Thesis*, University of California, Berkeley, 2008.

[27] Ganapathy, S., Rajan, P. and Hermansky, H., “Multi-layer Perceptron Based Speech Activity Detection for Speaker Verification”, *IEEE Workshop on Application of Signal Proc. to Audio and Acoustics*, 2011, pp. 321-324.