Robust Feature Extraction Using Modulation Filtering of Autoregressive Models

Sriram Ganapathy, *Member, IEEE*, Sri Harish Mallidi, *Student Member, IEEE*, and Hynek Hermansky, *Fellow, IEEE*

***Abstract*—Speaker and language recognition in noisy and degraded channel conditions continue to be a challenging problem mainly due to the mismatch between clean training and noisy test conditions. In the presence of noise, the most reliable portions of the signal are the high energy regions which can be used for robust feature extraction. In this paper, we propose a front end processing scheme based on autoregressive (AR) models that represent the high energy regions with good accuracy followed by a modulation filtering process. The AR model of the spectrogram is derived using two separable time and frequency AR transforms. The first AR model (temporal AR model) of the sub-band Hilbert envelopes is derived using frequency domain linear prediction (FDLP). This is followed by a spectral AR model applied on the FDLP envelopes. The output 2-D AR model represents a low-pass modulation filtered spectrogram of the speech signal. The band-pass modulation filtered spectrograms can further be derived by dividing two AR models with different model orders (cut-off frequencies). The modulation filtered spectrograms are converted to cepstral coefficients and are used for a speaker recognition task in noisy and reverberant conditions. Various speaker recognition experiments are performed with clean and noisy versions of the NIST-2010 speaker recognition evaluation (SRE) database using the state-of-the-art speaker recognition system. In these experiments, the proposed front-end analysis provides substantial improvements (relative improvements of up to 25%) compared to baseline techniques. Furthermore, we also illustrate the generalizability of the proposed methods using language identification (LID) experiments on highly degraded high-frequency (HF) radio channels and speech recognition experiments on noisy data.**

***Index Terms*—Autoregressive modeling, feature extraction, modulation filtering, speaker and language recognition.**

I. Introduction

TYPICALLY, speaker and language recognition systems perform well on clean controlled environments where the background and target models match the data used for testing. However, the performance of these systems is degraded significantly when the speech data used for testing are distorted due to additive noise, reverberation or radio channel distortions. Recently, there has been initiatives from various organizations (for example IARPA [1], DARPA [2], NIST [3]) which target improved speaker and language recognition in noisy environments.

The main cause of degradation in noisy environments is the acoustic mismatch of features derived from clean training conditions and noisy test conditions. One common solution to overcome this mismatch is the use of multi-condition training [4] 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 environments. Therefore, there is a need to attain noise robustness either at the front-end signal analysis or at the statistical modeling stage. The goal of this paper is to address the robustness issues in feature extraction.

Various techniques like spectral subtraction [5], Wiener filtering [6], power bias subtraction [7] and missing data reconstruction [8] have been proposed for noisy speech recognition scenarios. These approaches assume an additive model of the noise and attempt to enhance the signal by subtracting the noise component. Feature compensation techniques like feature warping [9], RASTA processing [10] and cepstral mean subtraction (CMS) assume a convolutive noise model which is additive in the cepstral domain. By removing a fixed mean computed over the recording, the channel effects in speech are suppressed. Hence, many of the feature processing techniques perform reasonably well when the assumptions of additive or convolutive model is satisfied. However, in a realistic scenario, it is not always possible to characterize the noise model as additive or convolutive especially for non-linear channel distortions like HF radio channels [2]. In this paper, we propose to develop a robust front-end which is devoid of any noise model by merely focussing on the high energy regions of the signal.

Typically, when speech is corrupted due to various environmental distortions, 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 mismatch with the clean training data. Thus, a robust feature extraction scheme must rely on the high energy regions in the spectro-temporal plane. An autoregressive (AR) modeling approach fits the high energy regions well [11], [12]. One dimensional AR modeling of signal spectra is widely used for feature extraction of speech in the form of perceptual linear prediction (PLP) [13]. The one dimensional temporal AR model has been

Manuscript received December 12, 2013; revised April 07, 2014; accepted May 20, 2014. Date of publication June 05, 2014; date of current version June 25, 2014. This work was supported in part by Contract No. D11PC20192 and D10PC20015 DOI/NBC under the RATS program. The views expressed are those of the authors and do not reflect the official policy of the Department of Defense or the U.S. Government. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vincent Vanhoucke.

S. Ganapathy is with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: ganapath@us.ibm.com).

S. H. Mallidi and H. Hermansky are with the Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: harish@jhu.edu; hynek@jhu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2329190

proposed using frequency domain linear prediction [14], [15]. An extension of this method has also been proposed by alternating between spectral and temporal domains [16].

In the recent past, we have shown that 2-D AR modeling can generate robust speech representations which are useful for speaker verification [17], [18]. In this paper, we extend this approach to derive multiple representations of modulation filtered spectrograms. The low-pass modulation filtering is achieved by using AR modeling with a low model order in the spectral and temporal domain. The band-pass filtering is implicitly achieved by dividing a higher order AR model with a lower order AR model. These filters are separably applied in the time and frequency domains and the modulation filtered spectrograms are converted to cepstral features for speaker/language recognition in noisy conditions.

The speaker recognition experiments are performed using NIST-2010 speaker recognition evaluation (SRE) data [19]. The noise robustness is measured using the condition 2 (interview microphone training and testing) of SRE 2010 data corrupted with various additive noise types and convolutive room responses. In these experiments, the proposed modulation filtering provides significant improvements compared to the conventional features.

We perform language recognition experiments on the noisy and degraded radio channel data on the RATS database [2]. In these experiments, the emphasis is on the performance of novel channel conditions which is simulated by training on a subset of the channels and testing on rest of the channels which are not seen during training. The proposed approach shows noticeable improvements (relative improvements of about 25%) in these mismatched channel experiments. We also perform automatic speech recognition (ASR) experiments in the Aurora 4 database [20] which contains additive and channel noise artifacts. We use a deep neural network system for these experiments. The results from the LID and ASR tasks indicate that the modulation filtering method using AR models is robust to a wide range of acoustic and channel degradations.

The rest of the paper is organized as follows. In Section II, we outline the linear prediction approaches in the spectral and temporal domain. Section III details the proposed modulation filtering scheme using 2-D AR models. Section IV describes the speaker recognition experiments using the proposed front-end. In Section V, we describe our experimental setup and discuss the results of various evaluations for a language recognition task. Section VI summarizes the results on the ASR task. In Section VII, we conclude with a brief discussion of the proposed front-end.

II. AR Modeling in Time and Frequency

A. Spectral AR model–TDLP

Spectral AR modeling has been widely used in speech and audio signal processing for at least four decades now [11], [12]. Let $x[n]$ denote the input signal for $n = 0, \ldots, N-1$. The time domain LP model is formulated 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 [11], where p denotes the model order.

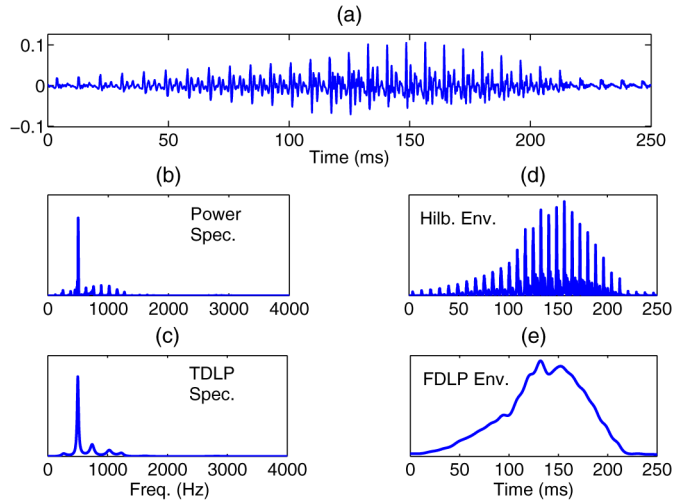

Fig. 1. Illustration of AR modeling in time and frequency domain - (a) a portion of voiced speech, (b) power spectrum, (c) AR model of power spectrum obtained from TDLP, (d) Hilbert envelope and (e) AR model of Hilbert envelope using FDLP.

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

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

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 [12]. 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 with $a_0 = 1$ provides an AR model of the power spectrum,

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

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.

B. Temporal AR model - FDLP

Linear prediction in the spectral domain was first proposed by Kumaresan [14]. This was reformulated by Athineos and Ellis [15] using matrix notations and the connection with DCT sequence is established. In this paper, we simplify the derivation without using matrix notations [21]. In a discrete-time case, an "analytic" signal (AS) $x_a[n]$ can be defined using the following procedure [22]-

1) Compute the N-point DFT sequence $X[k]$

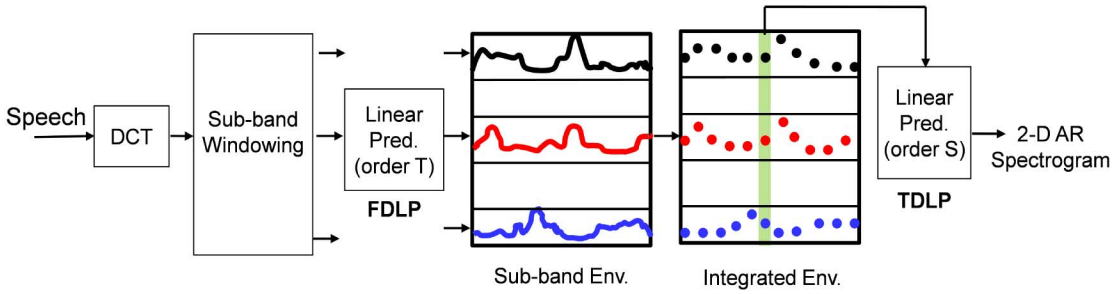

Fig. 2. Block schematic of 2-D AR model spectrogram using FDLP and TDLP.

2) Find the N-point DFT of the AS as,

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

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 $x[0] = 0$ respectively. 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 $k = 0, \ldots, N-1$ 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,

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

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

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 $k = 0, \ldots, N-1$. Let $\hat{y}$ denote the zero-padded DCT with $\hat{y}[k] = y[k]$ for $k = 0, \ldots, N-1$ and $\hat{y}[k] = 0$ for $k = N, \ldots, M-1$. 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 $k = 0, \ldots, M-1$. 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. Similar to the relation between the auto-correlation of signal $x[n]$ and the power spectrum $|\hat{X}[k]|^2$ (Eq. (2)), we can obtain a relation between the auto-correlation of the DCT sequence and the Hilbert envelope.

The auto-correlation of the DCT signal is 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 is easily shown that,

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

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 analogues 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.

III. Modulation Filtering Using AR Models

A. 2-D AR Spectrogram

The block schematic for the generation of 2-D AR spectrogram is shown in Fig. 2. Long segments of the input speech signal (of the order of few seconds of non-overlapping windows) are transformed using a discrete cosine transform [24]. The full-band DCT signal is windowed into a set of over-lapping sub-bands. In each sub-band, linear prediction is applied on the sub-band DCT components to estimate an all-pole representation of Hilbert envelope as described in Section II. This constitutes the temporal AR modeling stage. The FDLP envelopes

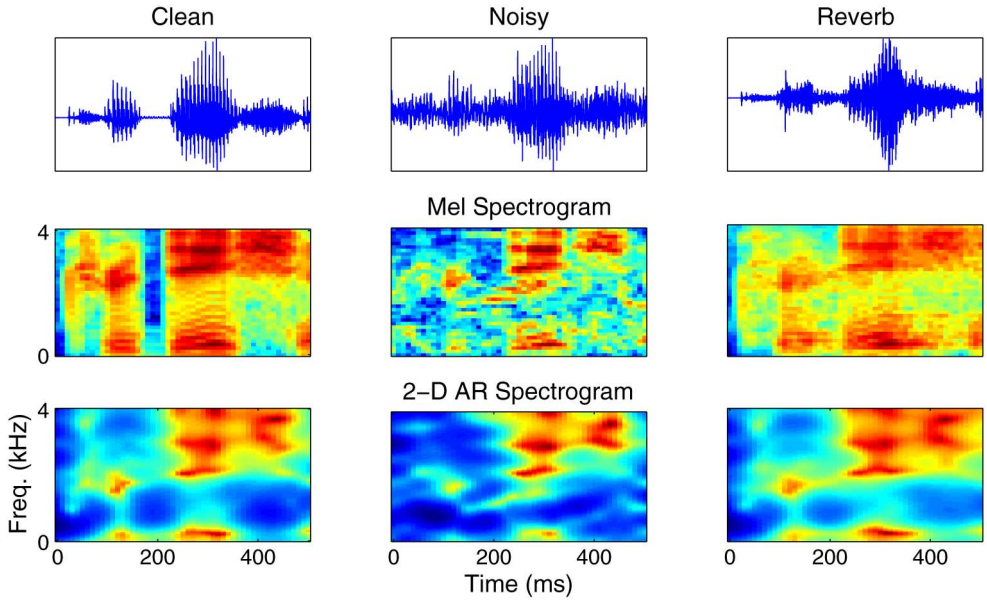

Fig. 3. Comparison of spectrographic representations obtained from clean speech, noisy speech (babble noise at 0 dB) and reverberant speech reverberation time of 300 ms for the mel-spectrogram (with 13 dimensional cepstral smoothing) and the proposed 2-D AR model spectrogram.

from various sub-bands are stacked together to obtain a two-dimensional representation as shown in Fig. 2.

The sub-band envelopes are convolved with a Hamming window of 25 ms and sub-sampled at 100 Hz (one value for each sub-band in a 10 ms frame). This is equivalent to integration with a Hamming wind in each sub-band over a 25 ms window with a 10 ms shift. The output of this integration and sub-sampling provides a short-term estimate of the temporal envelope over 25 ms. The integration in time of a sub-band signal will result in a power spectrum of the same frequency resolution as the sub-band signal. The frequency resolution of this power spectrum is equal to the initial sub-band decomposition. We use an initial sub-band decomposition of 96 bands. For each 25 ms frame, these power spectral estimates are transformed to temporal autocorrelation estimates using inverse Fourier transform and used for time domain linear prediction (TDLP). This gives the 2-D AR spectrogram.

When a speech signal is corrupted by noise or reverberation, the valleys in the sub-band envelopes are dominated by noise. Even with moderate amounts of distortion, the low-energy regions are substantially modified and cause acoustic mismatch with the clean training data. Since the AR modeling tends to fit the high energy regions with good accuracy [12], the spectro-temporal AR modeling approach could be more robust to noise and reverberation artifacts. This is illustrated in Fig. 3 where we plot a portion of clean speech signal, speech with additive noise (babble noise at 0 dB SNR) and speech with artificial reverberation (reverberation time of 300 ms). The spectrographic representation obtained from mel frequency cepstral coefficients (13 dimensional cepstra) is shown in the second panel and the corresponding representation obtained from spectro-temporal AR models is shown in the bottom panel. The mel spectrogram captures a lot of details from the clean signal which are not well preserved in noise and reverberation. The representation obtained by AR modeling is relatively smooth and some detail is lost due to AR smoothing. We hypothesize that enough information is still preserved in the representation for the application at hand. Further, the information retained in the AR model at high energy regions provides a greater similarity between the clean and the noisy versions of the same signal. This is desirable and contributes to improved robustness when these features are used for speaker and language recognition in noisy environments.

B. Modulation filtering Using 2-D AR models

A temporal modulation filter is referred to as a rate filter and a spectral modulation filter is referred to as a scale filter. In the proposed feature extraction framework, the AR modeling process represents a filter impulse response, whose frequency response ("time response" in the case of the temporal AR filter) can be controlled by the model order. A lower model order implies more smoothing in a given domain (denoted by T, S for the temporal and spectral AR models respectively in Fig. 2), while the higher model captures finer details. The band pass modulation representation can be then created by dividing a higher order envelope with a lower order envelope.

The method for obtaining band pass modulation representation from the 2-D AR spectrogram is outlined in Fig. 4. As shown here, the low pass representation is obtained using AR models in temporal domain and spectral domain. The band pass representation is obtained by dividing a higher order AR model with a lower order AR model. For example, for the temporal domain, the FDLP technique is applied on the sub-band DCT components using two different model orders ($T1$,$T2$) with $T1 > T2$. The AR model obtained from these representations (Eq. (11)) are divided to obtained a band pass temporal envelope. This envelope is used by the spectral AR model to form the temporal band-pass spectral low-pass (TBP-SLP) model. Similarly, the temporal low-pass spectral band-pass (TLP-SBP) model is obtained by dividing a higher order spectral AR model with a lower order spectral AR model. The band-pass model typically

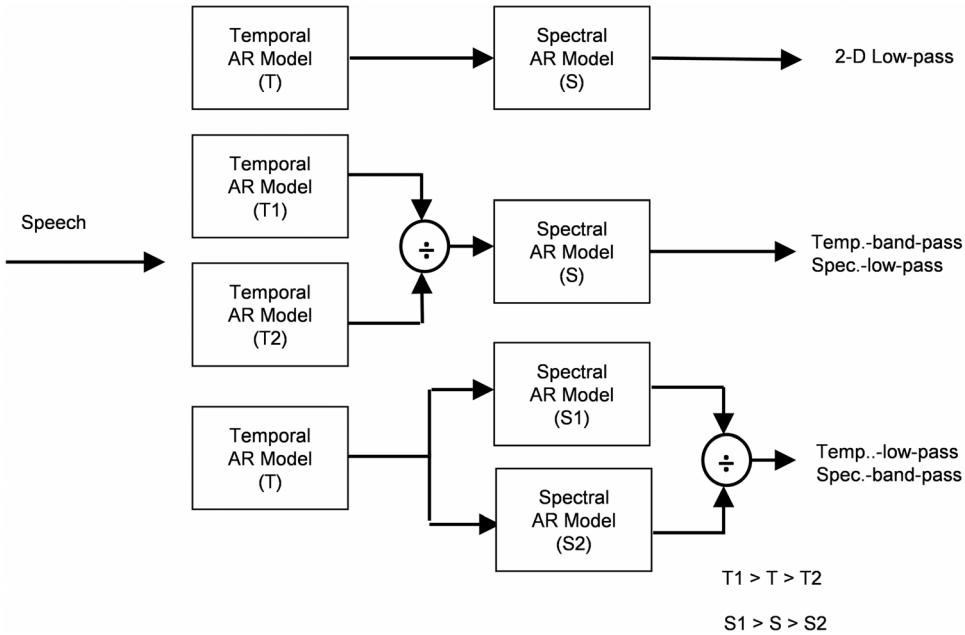

Fig. 4. Block schematic for modulation filtering using AR models. Band pass filtering is achieved by dividing a higher order model with a lower order model.

emphasizes certain range of spectral and temporal fluctuations while deemphasizing the steady state constants.

An illustration of the band pass filtering using AR models is shown in Fig. 5. The low-pass 2-D AR spectrogram captures the broad spectral and temporal variations using temporal and spectral AR models. The band-pass representations enhances the changes in the temporal or spectral domain and deemphasizes the constant regions. This enhancement of changes has been observed in the human auditory processing and various auditory models have been developed in the past for describing this phenomenon [10], [25]. In the proposed approach, we derive band pass modulation representations within the framework of AR models. As seen here, other streams can also be derived (for example, band-pass representations in both domains). For the scope of this study, we focus on the extraction of only three streams - 2-D AR LP, TBP-SLP and TLP-SBP. Although one could also derive representation with band-pass in both temporal and spectral domain, we found this noisy and less useful for representing speech in our applications. Following the generation of these spectrographic representations, the steps involved in converting them to features are similar to conventional mel frequency features [26]. The cepstral coefficients are derived by DCT on the 96 band log-spectrogram. We use the first 13 DCT components and the delta and acceleration coefficients are extracted to obtain 39 dimensional features.

IV. Speaker Recognition Experiments

A. Database

The proposed features are used for speaker recognition using the core conditions of the NIST 2010 speaker recognition evaluation (SRE) [19]. We use a GMM-UBM based speaker verification system [27]. The input speech features are feature warped [9] which forms a normalization of the mean, variance and higher order moments. 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.

Following the training of the UBM, the zeroth order and first order statistics of the Gaussian mixture components are derived. We use the i-vector based factor analysis technique [28] on these statistics 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 subspace training set. Gender specific i-vectors of 450 dimensions are extracted and these are used to train a PLDA system [29]. The output scores are obtained using a 250 dimensional PLDA sub-space for each gender.

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, resulting in speech at 5, 10, 15 and 20 dB SNR. These noises are added using the FaNT tool [30]. For simulating reverberant recording conditions, we also convolve the test data for Cond.-2 with three artificial room responses [31] with reverberation time of 100, 300 and 600 ms. In our experiments, the enrollment data consists of "clean" speech data present in NIST 2010 and the test data may be clean or noisy data. The voice-activity decisions provided by NIST which are obtained from the clean speech are used in these experiments. The GMM-UBM, i-vector and the PLDA sub-spaces trained from the development data are used without any modification. The performance metric used is the EER (%) and the false-alarm rate at a miss-rate of 10% (Miss10). In our experiments, we do not have any score normalization. All the front-ends considered

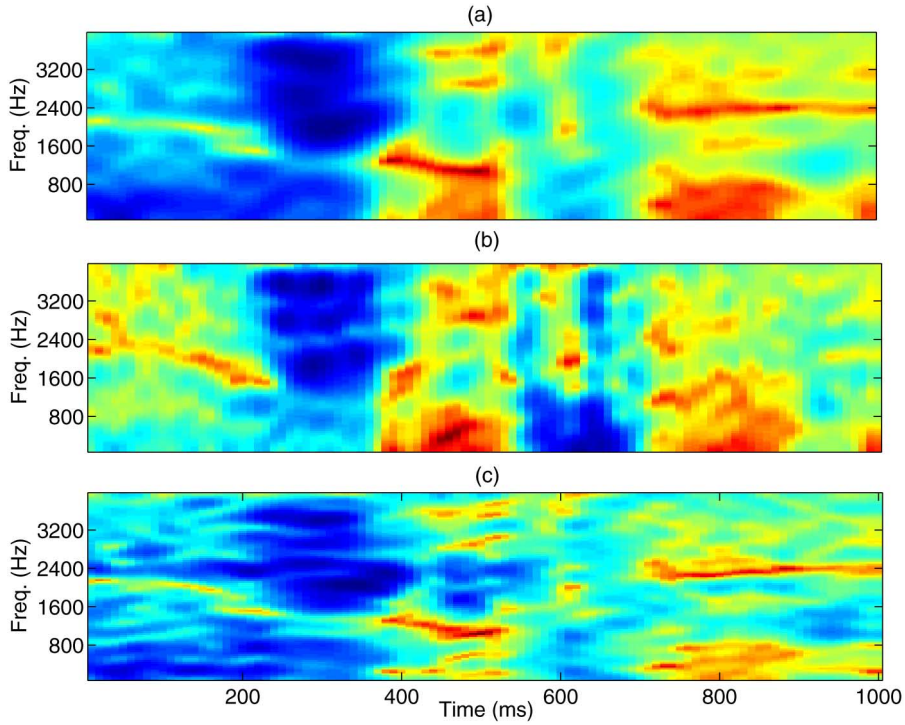

Fig. 5. Illustration of modulation filtering using AR models. (a) Low pass 2-D AR spectrogram, (b) Temporal band-pass spectral low-pass (TBP-SLP) spectrogram, (c) Temporal low-pass spectral band-pass (TLP-SBP) spectrogram.

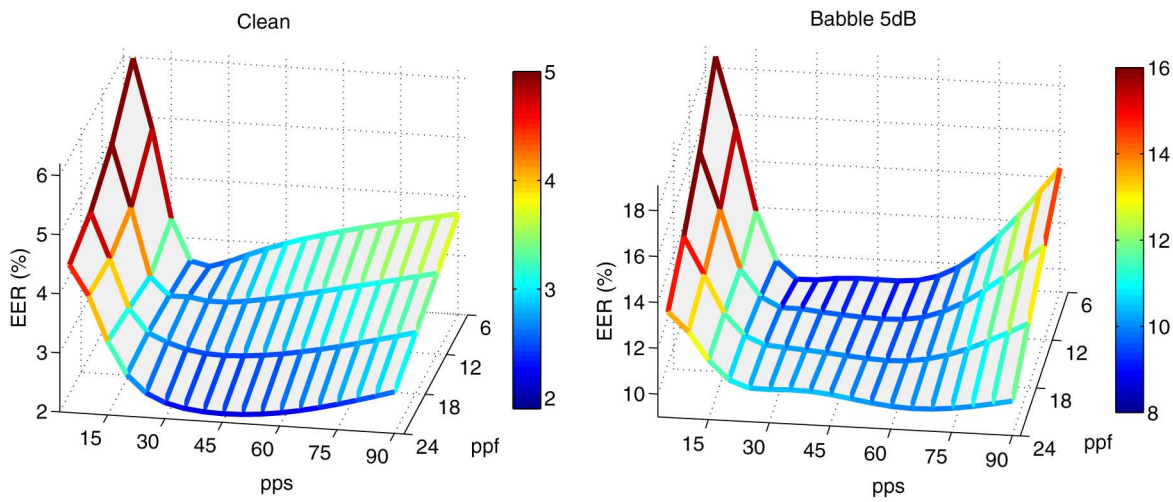

Fig. 6. EER (%) clean and noisy version (babble at 5 dB SNR for Cond.-2 of NIST 2010 SRE for 2-D AR LP features for various choices of model order, the temporal AR model (T) in terms of poles per sec (pps) and the spectral AR model (S) in terms of poles per frame (ppf).

here use the same processing pipeline without any threshold selection or tuning for any noise condition.

B. Model Order Selection

The initial set of experiments discuss the selection of model order for the low-pass filtering scheme with AR models using the clean data for Cond.-2 as well as validation data from babble noise at 5 dB SNR. This choice of validation data was not optimized in any manner and the performance on other types of noise and SNR levels relates to the generalization of the parameter selection process. In these experiments, we use a temporal analysis window of 10 s. The results for various choices of model order (described in terms of number of poles per second for temporal model or number poles per frame for the spectral AR model) is shown in Fig. 6. In these model order selection experiments, we use 8 choices of spectral and temporal AR model order. The EER results from these experiments are interpolated for graphical illustration in Fig. 6.

As seen here, the experiments on clean conditions show that a higher model order is generally better as more details are preserved. However, in the noisy case (babble noise at 5 dB SNR), the performance degrades when a higher order is used as the detailed model may capture the characteristics of the noise instead of the speech signal. Based on the results provided in Fig. 6, we select a model order of $T = 30$ and $S = 12$). On the average, this would approximately correspond to 0–15 Hz of modulations in the temporal rate axis [10] and about 0–1.2 octaves per cycle in the spectral scale axis [32].

TABLE I
EER (%) FOR CORE EVALUATION CONDITIONS IN NIST 2010 SRE. THE DESCRIPTION FOR SRE10 CORE CONDITIONS ARE IN ENROLL CONDITION-TEST CONDITION MANNER. HERE, $Int.mic$ REPRESENTS INTERVIEW MICROPHONE, tel REPRESENTS TELEPHONE, $phn.call - mic$ REPRESENTS CONVERSATIONAL TELEPHONE SPEECH RECORDED OVER A ROOM MICROPHONE, $phn.call - tel$ REPRESENTS CONVERSATION TELEPHONE SPEECH RECORDED OVER A TELEPHONE MICROPHONE, lv,hv REPRESENTS LOW AND HIGH VOCAL EFFORT RESPECTIVELY. THE ROOM MICROPHONE (1,2,4,7,9) HAS MORE FAR-FIELD EFFECTS ON THE SPEECH SIGNAL AND TELEPHONE MICROPHONE (3,5,6,8) IS NEAR-FIELD

Cond.	MFCC	FDLP	PNCC	ETSI	TLP-SBP	TBP-SLP	2-D AR LP
1. Int.mic - Int.mic-same.	2.0	2.1	2.3	2.4	3.2	2.5	1.9
2. Int.mic - Int.mic-diff.	3.0	2.9	3.4	3.5	3.9	3.5	2.7
3. Int.mic - Phn.call-tel	3.9	3.6	4.7	4.6	5.7	5.8	3.8
4. Int.mic - Phn.call-mic	3.3	2.8	3.4	3.9	3.9	4.0	2.9
5. Phn.call - Phn.call-tel	2.9	2.9	2.9	3.0	5.4	6.0	3.8
6. Phn.call - Phn.call-tel-hv	4.3	5.1	4.8	4.6	7.4	7.2	5.1
7. Phn.call - Phn.call-mic-hv	7.6	5.8	8.1	8.1	6.1	6.2	4.7
8. Phn.call - Phn.call-tel-lv	2.1	2.6	2.5	2.0	4.0	4.9	2.8
9. Phn.call - Phn.call-mic-lv	2.1	2.1	3.1	2.5	3.2	2.8	1.8
Avg.	3.4	3.3	3.9	3.8	4.7	4.8	3.3

TABLE II
PERFORMANCE IN TERMS OF EER (%) FOR THREE NOISE TYPES (BABBLE, EXHALL AND RESTAURANT) WITH FOUR SNR VALUES (5,10,15,20) DB AS WELL AS REVERBERANT CONDITION WITH REVERBERATION TIME OF 100,300,600 MS

Cond.	MFCC	FDLP	PNCC	ETSI	TLP-SBP	TBP-SLP	2-D AR LP
Clean	3.0	2.9	3.4	3.5	3.9	3.5	2.7
Babble-5	12.5	12.6	11.3	11.1	12.6	12.2	9.8
Babble-10	7.5	7.3	6.7	6.8	7.4	7.2	5.8
Babble-15	4.8	4.7	5.3	5.0	5.7	5.1	4.0
Babble-20	3.8	3.7	4.1	4.2	4.8	4.3	3.3
Exhall-5	9.4	9.5	8.8	9.1	9.3	8.7	8.2
Exhall-10	5.5	5.5	5.7	6.1	6.1	5.9	4.9
Exhall-15	4.1	4.0	4.4	4.6	4.8	4.7	3.7
Exhall-20	3.6	3.4	3.8	4.1	4.2	4.1	3.2
Rest.-5	9.1	8.7	8.9	8.8	10.1	9.9	7.6
Rest.-10	5.9	5.9	5.9	5.8	6.3	6.4	5.1
Rest.-15	4.2	4.2	4.5	4.6	4.8	4.8	3.8
Rest.-20	3.6	3.4	3.9	4.1	4.2	4.2	3.2
Avg.	6.2	6.1	6.1	6.2	6.7	6.5	5.2
Revb.-100	3.9	4.0	4.7	5.1	5.0	4.6	3.4
Revb.-300	4.9	5.2	5.8	5.9	5.9	5.4	4.2
Revb.-600	9.6	10.3	11.2	11.5	10.8	10.3	9.0
Avg.	6.1	6.5	7.3	7.5	7.2	6.8	5.5

C. Results

We compare the performance of the proposed features with other robust feature extraction techniques like power normalized cepstral coefficients (PNCC) [7] and Advanced ETSI features [6]. The other baseline features evaluated in this framework are MFCC features [26] as well as the FDLP features which involves one dimensional temporal AR model [24]. For the proposed modulation filter based features, we experiment with the following three configurations -

- Temporal low-pass spectral band-pass (TLP-SBP) obtained by dividing a spectral AR model of $S1 = 24$ with $S1 = 2$ with $T = 60$.
- Temporal band-pass spectral low-pass (TBP-SLP) obtained by dividing a temporal AR model $T1 = 60$ with $T2 = 4$ with a spectral AR model $S = 12$.
- Low pass 2-D AR spectrogram with $T = 30$ and $S = 12$.

The results for the 9 core conditions in NIST 2010 SRE are reported in Table I. The band pass filtering based feature processing (TBP-SLP and TLP-SBP) results in increased error rates as the information removed by dividing with the low-pass AR model may be useful in clean conditions. From these results, it can be seen that the proposed 2-D AR LP features provide good improvements in 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). The degradation in performance for Cond. 5, 6 and 8 seen in the 2-D AR LP features may be attributed to the reduced resolution caused by the AR modeling.

The comparison of the performance for various noisy and reverberant versions of the core condition-2 data is shown in Table II. The additive noise conditions are reported for three different types of noise - Babble, Exhall and Restaurant with four different SNR values 5,10,15,20 dB. The reverberant conditions include three different room responses with reverberation time of 100, 300 and 600 ms. In these experiments, the noise robust features like PNCC and ETSI provide good improvements over the MFCC and FDLP features at 5 dB SNR conditions. However, these features do not perform well in the presence of reverberation. This may be due to the additive model of noise assumed in these feature extraction techniques. For the band-pass filtering schemes (TBP-SLP and TLP-SBP), the results are slightly worse compared to PNCC/ETSI features in additive noise at low SNR levels. However, these features provide improved results in reverberation condition.

TABLE III
PERFORMANCE IN TERMS OF FALSE ALARM (%) AT 10% MISS RATE (MISS10) FOR EVALUATION CONDITIONS IN IARPA BEST 2011 TASK. HERE, $Int.mic$ REPRESENTS INTERVIEW MICROPHONE, tel REPRESENTS TELEPHONE, $noisy$ REPRESENTS BOTH NOISY AND REVERBERANT CONDITIONS

Cond.	MFCC	FDLP	2-D AR LP
Clean			
Int.mic - Phn.call-mic	3.7	3.5	2.8
Int.mic - Phn.call-tel	3.3	2.1	2.8
Phn-call-mic - Phn.call-mic	7.4	8.1	6.7
Phn.call-mic - Phn.call-tel	7.5	5.7	6.3
Phn.call-tel - Phn.call-tel	1.3	1.4	1.8
Avg.	4.6	4.2	4.1
Noisy			
Int.mic - Int.mic-noisy.	15.5	13.3	11.3

The proposed 2-D AR LP features provide the best performance in both the additive noise and reverberation conditions. In comparison with PNCC features, the 2-D AR LP features, provide average relative improvements of about 15% in additive noise conditions and 25% in reverberant conditions. The improvements obtained for the proposed 2-D AR LP features over the FDLP features shows that AR modeling in spectro-temporal domain provides better robustness compared with AR modeling in the temporal domain alone.

In the final set of experiments, we measure the speaker verification performance using the IARPA BEST 2011 data [1]. 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 III. Condition 1 contains majority of the trials (20M trials) recorded using interview microphone data with varying amounts of additive noise and artificial reverberation. In these experiments, only the MFCC, FDLP and 2-D AR LP features are compared as the evaluation allowed for submission of limited number of systems. We use the background and factor analysis models trained for NIST-SRE 2010 for these experiments.

The performance (Miss10) (EER results were not provided by the evaluation agency) for the baseline MFCC system is compared with proposed features in Table III. In the clean conditions, the proposed features are better than baseline features (MFCC, FDLP) for interview microphone conditions in training and test. However, the baseline features provide better results on telephone channel conditions. On the average, for clean conditions, the proposed features provide 12% relative improvement compared to MFCC features while being similar to FDLP features.

The noisy condition contains noisy only in the test data and the enrollment data is relatively clean. In the presence of noise or reverberation, the proposed features provide noticeable improvements as shown in the last row of Table III. We obtain relative improvements of 27% and 13% compared to MFCC and FDLP features respectively. The improvements seen on this large IARPA database are also consistent with those obtained for the noisy NIST-2010 SRE evaluations reported in Table II.

TABLE IV
DESCRIPTION OF RATS RADIO COMMUNICATION CHANNELS USED IN OUR LID EXPERIMENTS [2]

Channel	Characteristic
A	Receiver 50kHz offset
C	Receiver 3kHz offset
D	Frequency Shift
F	Spread Spectrum

V. LANGUAGE RECOGNITION EXPERIMENTS

A. Database

The development and test data for the LID experiments use the LDC releases of phase-I RATS LID evaluation [2]. This consists of speech recordings from previous NIST-LRE clean recordings as well as other RATS clean recordings passed through noisy radio communication channels. Each channel induces a degradation mode to the audio signal based on its device non-linearities, carrier modulation types, network parameter settings etc. In the RATS initiative, a set of eight channels is used with specific parameter settings and carrier modulations. The description of the eight channels is shown in Table IV. The five target languages are Levantine-Arabic, Farsi, Dari, Pashto and Urdu. In addition to this, the database consists of several other imposter languages. In order to investigate the effects of an unseen communication channel (not seen in training), we divide the eight channels to two groups - channels B,E,G,H used in the training and the channels A,C,D,F used in testing. This division of channels is done to target the realistic application of these systems where the noise and channel characteristics of the test data are not available during training.

The training data consist of 24,123 recordings with 270 hours of data from each of the four noisy communication channels (B,E,G,H) and the test set consists of 7,164 recordings with about 15 hours of data from each of the four target channels of interest (A,C,D,F). The training and test recordings contain 120 seconds of speech.

B. LID System

The speech signal is downsampled to 8 kHz and used for feature extraction. The features are processed with feature warping [9] and are used to train a Gaussian mixture model-Universal background model (GMM-UBM) with 1024 mixture components. Then, an i-vector projection model of 300 dimensions is trained [28].

The back-end classifier is a multi-layer perceptron (MLP) trained with the i-vectors as the input and the corresponding language labels as the targets [33]. The MLP has 2000 hidden units and 6 output neurons corresponding to the five target languages plus the imposter language. The model is trained with a cross-entropy cost function. For testing, the i-vectors corresponding to test utterance are forward passed through the MLP to obtain 6 dimensional scores. A common threshold value is applied on the scores and this threshold is varied to obtain the detection error trade-off (DET) curve. The performance of the LID system is measured in terms of equal error rate (EER).

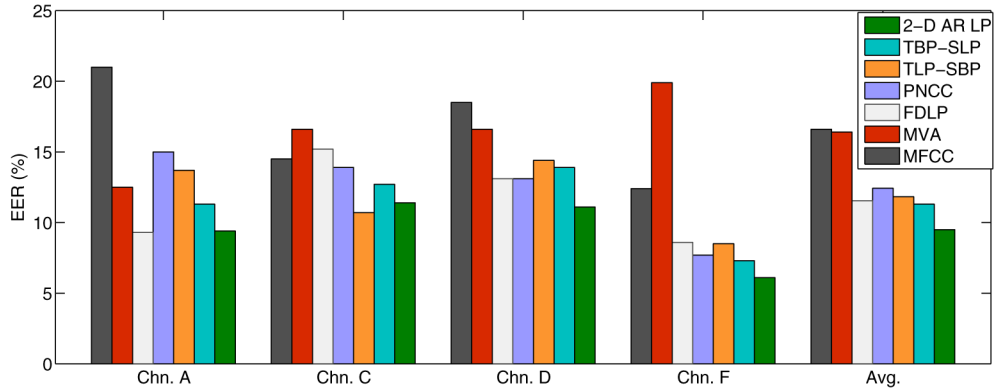

Fig. 7. LID Performance (EER %) for various extraction methods with different channels in RATS database (A,C,D,F).

C. Results

We experiment with various feature extraction schemes in the LID system like - MFCC features, MVA features [34], PNCC features [7] and FDLP features [24]. For the proposed modulation filter based feature representations, we use a mel-scale sub-band decomposition. We experiment with the following configurations,

- Temporal low-pass spectral band-pass (TLP-SBP) obtained by dividing a spectral AR model of $S1 = 18$ with $S2 = 2$ using a temporal AR model $T = 80$.
- Temporal band-pass spectral low-pass (TBP-SLP) obtained by dividing a temporal AR model $T1 = 80$ with $T2 = 2$ and a spectral AR model of $S = 12$.
- Low pass 2-D AR spectrogram with $T = 80$ and $S = 12$.

The LID results for various feature extraction techniques are shown in Fig. 7. As seen in this figure, the performance of the features vary depending on the type of non-linearity involved in the test data. The FDLP features, which involves one dimensional AR modeling [24] provide good performances on channel A, but results in a degradation in performance for channel C. Among the proposed features, the 2-D AR LP features provide significant improvements compared to other feature streams. In comparison with the PNCC baseline, the average relative improvements for the TLP-SBP, TLP-SBP and 2-D AR LP features are 5%, 9% and 24% respectively.

In order to illustrate the complimentary nature of the proposed features, we also perform a system combination experiment where we compare the three way fusion of baseline features (MFCC-MVA-PNCC) with the three way fusion of the proposed streams (TBPSLP-TLPSBP-ARLP). The system combination is achieved by computing the geometric mean of the posteriors obtained from the MLP outputs. The results for the system combination experiment are shown in Table V. As seen here, the system combination using the proposed streams achieves significant improvements across all the channels except channel D where a linear frequency shift is involved. On the average, the performance improvement provided by proposed streams is about 25%.

TABLE V
EER (%) FOR TWO DIFFERENT FUSION SCHEMES USING MFCC-MVA-PNCC FEATURES AS WELL AS THE PROPOSED TBPSLP-TLPSBP-ARLP FEATURES

Cond.	MFCC-MVA-PNCC	TBPSLP-TLPSBP-ARLP
Channel A	14.0	9.6
Channel C	12.3	9.2
Channel D	10.3	10.8
Channel F	9.8	5.1
Average	11.6	8.7

TABLE VI
WORD ERROR RATE (%) FOR VARIOUS FEATURES ON AURORA 4

Feat.	Clean	Chn.	Add.	Add.+Chn.	Avg.
MFBE	3.6	10.1	27.2	37.9	25.7
PNCC	3.5	11.3	17.5	34.6	21.1
ETSI	3.0	14.5	16.8	31.6	21.0
2-D AR LP	3.7	13.2	14.0	30.5	19.2

VI. SPEECH RECOGNITION EXPERIMENTS

We also perform a set of automatic speech recognition experiments in the Aurora4 database [20] using a deep neural network (DNN) hybrid system [8]. We use the clean training setup which contains 7308 clean recordings (14 h) for training the acoustic models. The system uses a tri-gram language model with 5 k vocabulary size. The test data consist of 330 recordings each from 14 conditions which include one clean condition, one channel noise condition ($Chn.$), 6 additive noise conditions ($Add.$) and 6 conditions with the combination of additive and channel noise ($Add. + Chn.$). We experiment with various feature extraction methods for the DNN-ASR system namely - mel filter bank energies (MFBE), PNCC and ETSI. All these features use a 21 frame context with utterance based mean variance normalization. For the proposed features (2-D AR LP with $T = 50$ and $S = 24$), we use 14 modulation components from each mel-band obtained by a DCT on 200 ms windows of sub-band envelopes. The modulation components are spliced with their frequency derivatives to form the input features for the DNN. We use a DNN with 4 hidden layers of 1024 activations and uses context dependent phoneme targets obtained from an initial alignment using a hidden-Markov-model-GMM system. The DNN training and ASR setup are obtained from the Kaldi toolkit [35]. The performance of the ASR system for various feature processing schemes is shown in Table VI. The proposed features provide noticeable improvements in the mismatched conditions of $Add.$ and $Add. + Chn.$. The results of the ASR experiments are also consistent with the trends reported for speaker and language recognition experiments.

VII. Summary

In this paper, we have proposed a feature processing scheme which uses two dimensional modulation filtering based on autoregressive models. The proposed scheme uses a frequency domain linear prediction based temporal AR model and a time domain linear prediction based spectral AR model. The band pass modulation filtering is achieved by dividing the AR model of higher order with a lower order. We perform several speaker recognition, language identification and speech recognition experiments with mismatched conditions. In these experiments, the proposed features provide significant improvements compared to various other noise robust front-ends. The generalization of the robustness even to non-linear distortions shows that the proposed scheme of low-pass modulation filtering using AR models is robust to a wide variety of acoustic and channel distortions.

References

[1] "IARPA BEST Speaker Recognition Challenge 2011," [Online]. Available: http://www.nist.gov/itl/iad/mig/best.cfm 2011

[2] K. Walker and S. Strassel, "The RATS Radio traffic collection system," in *Proc. Odyssey Speaker Lang. Recognition Workshop*, 2012.

[3] "NIST Speaker Recognition Evaluation 2012," [Online]. Available: http://www.nist.gov/itl/iad/mig/sre12.cfm 2012

[4] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1711–1723, Jul. 2007.

[5] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[6] "ETSI ES 202 050 v1.1.1 STQ: Distributed speech recognition: Advanced Front-End: Compression algorithms," 2002.

[7] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *Proc. INTERSPEECH*, 2009, pp. 28–31.

[8] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7398–7402.

[9] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey, Speaker Recogn. Workshop*, 2001.

[10] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[11] B. Atal and S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 47, pp. 637–655, 1970.

[12] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[13] H. Hermansky, "Perceptual linear predictive analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.

[14] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *J. Acoust. Soc. Amer.*, vol. 105, pp. 1912–1920, 1999.

[15] M. Athineos and D. Ellis, "Autoregressive modeling of temporal envelopes," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5237–5245, Nov. 2007.

[16] M. Athineos, H. Hermansky, and D. Ellis, "PLP-2 autoregressive modeling of auditory-like 2-D spectro-temporal patterns," 2004.

[17] S. Ganapathy, S. Thomas, and H. Hermansky, "Feature extraction using 2-D autoregressive models for speaker recognition," in *Proc. ISCA Speaker Odyssey*, 2012.

[18] H. Mallidi, S. Ganapathy, and H. Hermansky, "Robust speaker recognition using spectro-temporal autoregressive models," in *Proc. INTERSPEECH*, 2013.

[19] "National Institute of Standards and Technology (NIST)," [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/2010/

[20] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ASR2000-Autom. Speech Recognit.: Challenges for the New Millenium ISCA Tutorial and Research Workshop*, 2000.

[21] S. Ganapathy, "Signal analysis using autoregressive models of amplitude modulation," Ph.D. dissertation, Johns Hopkins Univ., Baltimore, MD, USA, 2012.

[22] S. Marple, Jr., *Digital spectral analysis with applications*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1987, vol. 1.

[23] S. Martucci, "Symmetric convolution and the discrete sine and cosine transforms," *IEEE Trans. Signal Process.*, vol. 42, no. 5, pp. 1038–1051, May 1994.

[24] S. Ganapathy, J. Pelecanos, and M. Omar, "Feature normalization for speaker verification in room reverberation," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 4836–4839.

[25] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 106, p. 2040, 1999.

[26] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

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

[28] 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, Aug. 2011.

[29] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. 12fth Annu. Conf. Int. Speech Commun. Assoc.*, 2011.

[30] "FaNT Filtering and Noise Adding Tool," [Online]. Available: http://dnt.kr.hsnr.de/download.html 2002

[31] D. Gelbart and N. Morgan, "Evaluating long-term spectral subtraction for reverberant ASR," in *Proc. IEEE Workshop Autom. Speech Recogn. Understand.*, 2001, pp. 103–106.

[32] S. Nemala, K. Patil, and M. Elhilali, "A multistream feature framework based on bandpass modulation filtering for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 416–426, Feb. 2013.

[33] P. Matejka *et al.*, "Patrol team language identification system for DARPA RATS p1 evaluation," in *Proc. INTERSPEECH*, 2012.

[34] C. P. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 257–270, Jan. 2007.

[35] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, 2011, pp. 1–4.

Sriram Ganapathy is a research staff member at the IBM T. J. Watson Research Center, Yorktown Heights, USA, where he is working on signal analysis methods for radio communication speech in highly degraded environments. He received his Doctor of Philosophy from the Center for Language and Speech Processing, Johns Hopkins University in 2011 with Prof. Hynek Hermansky. Prior to this, he obtained his Bachelor of Technology from College of Engineering, Trivandrum, India in 2004 and Master of Engineering from the Indian Institute of Science, Bangalore in 2006. He has also worked as a Research Assistant at the Idiap Research Institute, Switzerland, from 2006 to 2008 contributing to various speech and audio projects. His research interests include signal processing, machine learning and robust methodologies for speech and speaker recognition. He has obtained over 50 publications in leading international journals and conferences in the area of speech and audio processing along with several patents.

Sri Harish Mallidi received his B.Tech (2008) and M.S. (2010) in Electronics and Communications from International Institute of Information Technology, Hyderabad (IIIT-H), India. Currently, he is a Ph.D. student at the Center for Language and Speech Processing, Dept. of ECE, Johns Hopkins University, USA. His research interests include signal processing for robust speech applications and machine learning. He has been a student member of the IEEE since 2013.

Hynek Hermansky (S'78–M'83–SM'92–F'01) received the Dr. Eng. degree from the University of Tokyo, and the Dipl. Ing. Degree from Brno University of Technology, Czech Republic. He is the Julian S. Smith Professor of Electrical Engineering and the Director of the Center for Language and Speech Processing at the Johns Hopkins University in Baltimore, Maryland. He is also a Professor at the Brno University of Technology, Czech Republic, and an External Fellow at the International Computer Science Institute at Berkeley, California. He has been working in speech processing for over 30 years, previously as a Director of Research at the IDIAP Research Institute, Martigny and a Titullary Professor at the Swiss Federal Institute of Technology in Lausanne, Switzerland, a Professor and Director of the Center for Information Processing at OHSU Portland, Oregon, a Senior Member of Research Staff at U S WEST Advanced Technologies in Boulder, Colorado, a Research Engineer at Panasonic Technologies in Santa Barbara, California, and a Research Fellow at the University of Tokyo. His main research interests are in acoustic processing for speech recognition. He is a Fellow of IEEE, and of the International Speech Communication Association (ISCA), is the General Chair of the 2013 IEEE Automatic Speech Recognition and Understanding Workshop, was in charge of plenary sessions at the 2011 ICASSP in Prague, was the Technical Chair at the 1998 ICASSP in Seattle and an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He is also an Member of the Editorial Board of Speech Communication, an elected Member of the Board of ISCA, a Distinguished Lecturer for ISCA, and the recipient of the 2013 ISCA Medal for Scientific Achievement.