Multivariate Autoregressive Spectrogram Modeling for Noisy Speech Recognition

Sriram Ganapathy *Senior Member, IEEE*

***Abstract*—The performance of an automatic speech recognition (ASR) system is highly degraded in the presence of noise and reverberation. The autoregressive (AR) modeling approach, which preserves the high energy regions of the signal that are less susceptible to noise, presents a potential method for robust feature extraction. Secondly, there are strong correlations in the spectro-temporal domain of the speech signal which are generally absent in noise. In this paper, we propose a novel method for speech feature extraction which combines the advantages of AR approach and joint time-frequency processing using the multivariate AR modeling (MAR). Specifically, the sub-band discrete cosine transform (DCT) coefficients obtained from multiple speech bands are used in the MAR framework to derive the Riesz temporal envelopes that provide features for ASR. We perform several speech recognition experiments in the Aurora-4 database with clean and multi-condition training. In these experiments, the proposed features provide significant improvements over other noise robust feature extraction methods (relative improvements of 24 % in clean training and 14 % in multi-condition training over mel features). Furthermore, the speech recognition experiments in REVERB challenge database illustrates the extension of the MAR modeling method for suppressing reverberant artifacts.**

***Index Terms*—Multivariate Autoregressive Models, Riesz Envelopes, Feature Extraction, Speech Recognition.**

I. Introduction

Over the last decade, the architecture of automatic speech recognition (ASR) systems has witnessed a fundamental change with the use of deep neural networks [1]. While the ASR performance has seen overall improvements, the relative degradation in the presence of noise or reverberation continues to be a substantial challenge in the developing real world applications of ASR [2]. One common solution to overcome the performance degradation in noisy conditions is the use of multi-condition training [3] where the acoustic 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. Furthermore, even with multi-condition training, the performance of ASR systems are significantly worse compared to clean controlled testing conditions. The goal of this paper is to address the robustness issues in feature extraction.

In the past, several approaches have been proposed for robust feature extraction like modulation filtering (RASTA filtering [4]), spectral subtraction [5], [6], multi-step linear prediction based dereverberation [7], power normalization [8] etc. The autoregressive modeling (AR) approach to speech feature extraction relies on estimating an all-pole model where the peaks of the signal are well preserved. An example of the AR approach to model the short-term spectral envelopes is the perceptual linear prediction (PLP) technique [9] which is widely used for speech recognition. The application of AR modeling for estimating temporal envelopes [10], [11] has shown to improve robustness when used in conjunction with modulation filtering techniques [12]. However, many of these methods fail to fully exploit the two-dimensional time-frequency correlations present in the speech signal. Although a 2-D AR modeling approach was attempted in the past [13], this involved modeling spectral and temporal envelopes in an iterative and separable manner.

In this paper, we attempt to jointly model the temporal envelopes of multiple sub-bands using a time series analysis approach [14], [15]. The multivariate AR (MAR) modeling technique is the method of approximating a random time series vector as a linear combination of "past" vectors. Here, the prediction coefficients are matrices which are estimated using a generalized least squares criterion. The MAR modeling approach is widely used in econometrics for forecasting applications [16]. This work represents the first application of MAR modeling using multi-band Riesz envelope estimation to the best of our knowledge.

For application to speech processing, we use the long-term discrete cosine transform (DCT) coefficients of multiple spectral bands in the MAR framework. The MAR modeling preserves the signal peaks in the joint spectro-temporal domain and exploits the inherent 2-D structure of speech spectrograms. Given the lack of time-frequency correlations in noise, the proposed 2-D modeling allows the extraction of the long-term multi-band features representative of the underlying speech signal even in the presence of noise.

We perform several ASR experiments using the clean and multi-condition training setup in Aurora-4 database [3] as well as the REVERB challenge database [7] with a deep neural network back-end. In these experiments, the proposed front-end using MAR processing provides significant improvements compared to other noise robust feature extraction methods.

The rest of the paper is organized as follows. In Sec. II, we describe the MAR parameter estimation method. The application of the MAR model for feature extraction is discussed in Sec. III. The speech recognition experiments are described in Sec. IV followed by a summary of the paper in Sec. V.

S. Ganapathy is with the Learning and Extraction of Acoustic Patterns (LEAP) labs, Electrical Engineering, Indian Institute of Science, Bangalore, India. (phone: +91-(80) 2293 2433; fax +91-(80)-23600444 ; e-mail: sriram@ee.iisc.ernet.in)

II. Multivariate Autoregressive Modeling

Given a D dimensional vector process $\mathbf{y}$ of sequential data indexed by $q = 1...Q$, a multivariate AR model of order p is given by [15],

$$\mathbf{y}_q = \sum_{k=1}^{p} \mathbf{A}_k \mathbf{y}_{q-k} + \mathbf{u}_q, \tag{1}$$

where $\mathbf{u}$ is a D dimensional white noise random process with a covariance matrix $\mathbf{\Sigma}_u$ and the MAR coefficients $\mathbf{A}_k$ are square matrices of size D which characterize the model. In the following subsection, we provide the generalized least squares (GLS) estimation of the parameters $[\{\mathbf{A}_k\}_{k=1}^{p}, \mathbf{\Sigma_u}]$ of the MAR model.

In order to succinctly represent the MAR model in Eq. (1), let us define, $\mathbf{Y} := [\mathbf{y}_1\ \mathbf{y}_2\ ...\ \mathbf{y}_Q]$ of dimension $D \times Q$, $\mathbf{B} := [\mathbf{A}_1\ \mathbf{A}_2\ ...\ \mathbf{A}_p]$ of dimension $D \times Dp$, $\mathbf{Z}_q := [\mathbf{y}_q^T\ \mathbf{y}_{q-1}^T\ ...\ \mathbf{y}_{q-p+1}^T]^T$ of dimension $Dp \times 1$ and $\mathbf{U} := [\mathbf{u}_1\ \mathbf{u}_2...\ \mathbf{u}_Q]$ of dimension $D \times Q$. With these definitions, the MAR model of Eq. (1) for $q = 1..Q$ is given by,

$$\mathbf{Y} = \mathbf{BZ} + \mathbf{U} \tag{2}$$

where $\mathbf{Z} := [\mathbf{Z}_0, ..., \mathbf{Z}_{Q-1}]$ of dimension $Dp \times Q$ and we have assumed the availability of presample data observations $\mathbf{y}_{-p+1}, .., \mathbf{y}_0$. Let vec denote the splicing operator which converts a matrix of size $m \times n$ into a vector of size $mn \times 1$ by stacking the columns of the matrix one below the other. If $\mathbf{u} := vec(\mathbf{U})$ (of dimension $DQ \times 1$), $\boldsymbol{\beta} := vec(\mathbf{B})$ (of dimension $D^2p \times 1$) and $\mathfrak{y} := vec(\mathbf{Y})$ (of dimension $D^2p \times 1$), then,

$$\begin{aligned} \mathfrak{y} &= vec(\mathbf{BZ}) + \mathbf{u} \\ &= (\mathbf{Z}^T \otimes \mathbf{I}_D)\boldsymbol{\beta} + \mathbf{u} \end{aligned} \tag{3}$$

where $\otimes$ is the Kronecker product and $\mathbf{I}_D$ is the identity matrix of size D. The covariance matrix of $\mathbf{u}$ is $\mathbf{\Sigma_u} = \mathbf{I}_Q \otimes \mathbf{\Sigma}_u$. The GLS estimator minimizes the cost $S(\boldsymbol{\beta})$ given by [17],

$$\begin{aligned} S(\boldsymbol{\beta}) &= \mathbf{u}^T {\mathbf{\Sigma_u}}^{-1} \mathbf{u} \\ &= [\mathfrak{y} - (\mathbf{Z}^T \otimes \mathbf{I}_D)\boldsymbol{\beta}]^T {\mathbf{\Sigma_u}}^{-1} [\mathfrak{y} - (\mathbf{Z}^T \otimes \mathbf{I}_D)\boldsymbol{\beta}] \\ &= \boldsymbol{\beta}^T(\mathbf{ZZ}^T \otimes \mathbf{\Sigma}_u^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}^T(\mathbf{Z} \otimes \mathbf{\Sigma}_u^{-1}) + C \end{aligned} \tag{4}$$

where the inverse of Kronecker product[1] and the commutative property[2] are invoked and C is a constant independent of $\boldsymbol{\beta}$.

A. Model parameter estimation

The parameters ($\boldsymbol{\beta}$) can then be estimated by setting $\frac{\partial S}{\partial \boldsymbol{\beta}} = \mathbf{0}$. This estimate can be shown to be

$$\hat{\boldsymbol{\beta}} = ((\mathbf{ZZ}^T)^{-1}\mathbf{Z} \otimes \mathbf{I}_K)\mathfrak{y} \tag{5}$$

In this case, the Hessian matrix $\frac{\partial^2 S}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = 2(\mathbf{ZZ}^T \otimes \mathbf{\Sigma}_u^{-1})$ is indeed positive definite which guarantees a minimum estimate. Note that, with $D = 1$, the above formulation simplifies to the normal equations in a conventional AR model. The estimator given in Eq. (5) can be shown to be consistent and asymptotically normal [15]. The estimate of the covariance matrix of the noise $\hat{\mathbf{\Sigma}}_u$ can be obtained as,

$$\hat{\mathbf{\Sigma}}_u = \frac{1}{Q}\sum_{q=1}^{Q} \mathbf{u}_q \mathbf{u}_q^T = \frac{1}{Q}\mathbf{Y}(\mathbf{I}_Q - \mathbf{Z}^T(\mathbf{ZZ}^T)^{-1}\mathbf{Z})\mathbf{Y}^T \tag{6}$$

Thus, given the observations $\mathbf{y}_q$ for the vector random process, the parameters of the MAR model $[\{\mathbf{A}_k\}_{k=1}^{p}, \mathbf{\Sigma_u}]$ can be estimated using Eq. (5),(6).

[1]If $\mathbf{A}$, $\mathbf{B}$ are two matrices, $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$.

[2]If $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ are matrices, $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$.

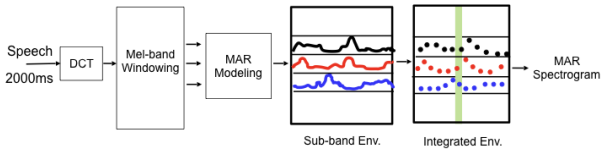

Fig. 1. Block schematic of the MAR spectrogram model.

B. Envelope Estimation

Unlike the application of MAR modeling in forecasting [16], this paper uses the MAR model for temporal envelope estimation. For the one dimensional AR modeling case in the time domain, the spectral envelope of the sequence y_q (q is the sample index in time domain) with $\mathbf{A}_k = a_k$ is given by,

$$\hat{s}_y[f] = \frac{\sigma_u^2}{|1 - \sum_{k=1}^{p} a_k e^{-i2\pi kf}|^2} \tag{7}$$

where $s_y[f]$ is the power spectral density, f is the normalized discrete frequency index and σ_u^2 is the prediction gain [18]. In the case of AR estimation of temporal envelopes [11], [10], the DCT sequence y_q is used as the input and the AR envelope $\hat{s}_y[n]$ denotes the Hilbert envelope of the signal (n denotes the discrete time index).

In this paper, the input $\mathbf{y}_q$ denotes DCT coefficients indexed by q for multiple speech sub-bands and the sub-band Riesz envelopes are estimated using MAR modeling. The following multidimensional z-transform (in temporal AR modeling, z represents complex time domain variable [11]) filter expression can be written for the model described in Eq. (1),

$$\Big[\mathbf{I}_D - \sum_{k=1}^{p} \mathbf{A}_k z^{-k}\Big]\mathbf{y}_q = \mathbf{u}_q \tag{8}$$

Let $\mathbf{H}(z) = \mathbf{I}_D - \sum_{k=1}^{p} \mathbf{A}_k z^{-k}$. If $\mathbf{s_y}[n]$ denotes the Riesz envelope (extension of Hilbert envelope to 2-D signals) of speech sub-bands, then the MAR estimate of the Riesz envelope is given by (for $\mathbf{H}[n] = \mathbf{H}[z]|_{z=e^{-j2\pi n}}$),

$$\hat{\mathbf{s}}_\mathbf{y}[n] = diag\big(\mathbf{H}[n]^{-1}\hat{\mathbf{\Sigma}}_u\mathbf{H}[n]^{-1}\big) \tag{9}$$

III. Feature Extraction using MAR

The block schematic of the proposed approach for feature extraction is shown in Fig. 1. Long segments of the input speech signal (2000ms of non-overlapping windows) are transformed using DCT. The full-band DCT signal is windowed into a set of 39 over-lapping sub-bands using Gaussian shaped windows with center frequencies chosen uniformly along the mel scale. The DCT sequences of multiple sub-bands are

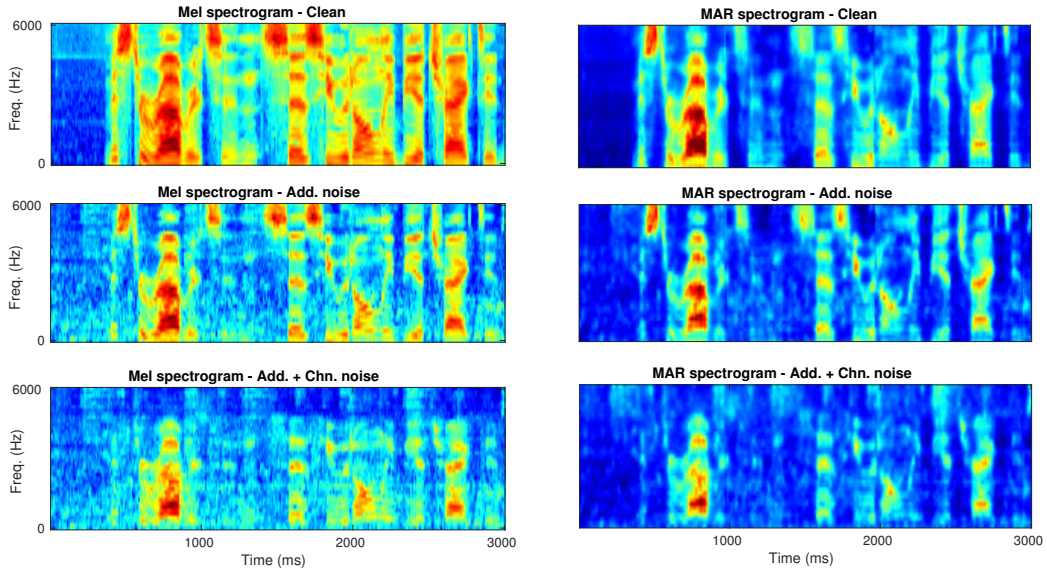

Fig. 2. Comparison of spectrogram estimation using MAR modeling with conventional mel spectrogram for clean and noisy speech recordings from Aurora-4 database.

stacked together to form vector series data $\mathbf{y}_q$ (q denotes the DCT coefficient index) of Eq. (1). The estimation procedure of the MAR model is applied and model parameters are estimated (Eq. 5). We use a fixed model order of $p = 160$ for the MAR estimation of 2000ms of speech. The sub-band temporal envelopes are then computed using Eq. (9). In this paper, the DCT coefficients of 3 mel bands are jointly used in the MAR modeling (i.e, $D = 3$).

The sub-band MAR envelopes are integrated with a Hamming window over a 25 ms window with a 10 ms shift. The integration in time of the sub-band envelopes yields an estimate of the MAR spectrogram of the input speech signal. In Fig. 2, we compare the spectrographic representation from MAR modeling and the conventional mel spectrogram. As seen here, the MAR modeling results in a smooth representation which emphasizes only the high energy regions of the signal. The joint estimation of the envelopes obtained by the two-dimensional spectro-temporal modeling also allows the model to focus primarily on time-frequency correlations of the underlying speech signal while suppressing the effects of noise as illustrated by the representations obtained for the noisy signal (babble noise at 10 dB SNR as well as in the presence of channel noise). These properties of the MAR model improve the noise robustness in the representations derived from this approach. For the ASR feature extraction, the integrated sub-band temporal envelope for a duration of 200 ms (centered around a 10 ms frame) are transformed to 14 DCT coefficients for each sub-band. The MAR features are also appended with spectral delta features yielding features of dimension of 1092.

IV. Experiments and Results

A. Noisy Speech Recognition

The WSJ Aurora-4 corpus is used for conducting ASR experiments in noisy speech [3]. This database consists of continuous read speech recordings recorded under clean and noisy conditions (street, train, car, babble, restaurant, and airport) at $10-20$ dB SNR. The training data has two sets of 7138 clean and multi condition recordings from 84 speakers. The validation data has two sets of 1206 recordings (14 speakers) for clean and multi condition and the test data has 330 recordings (8 speakers), each of the 14 clean and noise conditions. The test data is classified into group A - clean data, B - noisy data, C - clean data with channel distortion, and D - noisy data with channel distortion.

TABLE I
Word error rate (%) in Aurora-4 database for clean training condition with various feature extraction schemes.

Cond	Mel.	PN [8]	ET [6]	MH [19]	RA [4]	MAR
A. Clean with same Mic						
Clean	3.4	3.3	3.2	3.5	3.5	**3.1**
B: Noisy with same Mic						
Airport	22.5	18.3	15.0	19.5	19.3	13.3
Babble	19.2	16.0	15.5	17.7	19.9	13.0
Car	7.8	6.2	9.8	7.9	7.9	5.1
Rest.	25.6	22.9	20.5	23.2	23.0	17.9
Street	20.5	17.8	19.5	18.1	18.7	13.5
Train	20.0	16.3	17.4	17.9	19.4	13.9
Avg.	19.3	16.2	16.3	17.4	18.0	**12.8**
C: Clean with diff. Mic						
Clean	15.3	**11.7**	14.5	14.6	16.0	11.8
D: Noisy with diff. Mic						
Airport	39.4	36.4	31.4	39.2	38.7	31.4
Babble	36.0	34.2	32.1	36.8	38.5	30.9
Car	24.3	21.5	24.9	25.9	24.8	19.2
Rest.	39.8	39.0	35.4	39.3	39.1	32.8
Street	35.9	34.1	35.0	35.8	35.8	29.0
Train	35.6	31.8	33.2	35.9	36.4	29.4
Avg.	35.2	32.8	32.0	35.4	35.6	**28.8**
Avg. of all conditions						
Avg.	24.7	22.1	21.9	23.9	24.4	**18.9**

We compare the ASR performance of the proposed MAR approach with traditional mel filter bank energy (MF) features, power normalized filter bank energy (PN) features [8], advanced ETSI front-end (ET) [6], mean Hilbert envelope coefficients (MH) features [19] and RASTA features (RA) [4]. All the features are processed with utterance level mean and variance normalization.

The speech recognition Kaldi toolkit [20] is used for training

TABLE II
WORD ERROR RATE (%) IN AURORA-4 DATABASE FOR MULTI CONDITION TRAINING WITH VARIOUS FEATURE EXTRACTION SCHEMES.

Cond	Mel.	PN [8]	ET [6]	MH [19]	RA [4]	MAR
A. Clean with same Mic						
Clean	4.2	4.1	4.5	4.1	4.6	**3.7**
B: Noisy with same Mic						
Airport	9.1	7.9	8.0	8.2	8.1	7.1
Babble	8.7	7.9	7.9	8.6	8.7	7.1
Car	5.1	4.9	5.6	4.9	5.0	4.4
Rest.	10.7	10.2	11.0	11.1	11.0	9.4
Street	9.2	8.8	10.0	8.8	9.0	7.5
Train	9.3	8.3	9.3	8.4	9.1	8.0
Avg.	8.7	8.0	8.6	8.3	8.5	**7.3**
C: Clean with diff. Mic						
Clean	8.6	**7.8**	8.0	8.1	9.7	**7.8**
D: Noisy with diff. Mic						
Airport	20.7	20.9	18.5	20.8	20.1	18.2
Babble	20.6	20.9	19.3	21.4	20.0	17.9
Car	12.8	13.1	14.1	12.8	12.5	10.5
Rest.	22.8	23.7	21.8	23.1	23.1	20.0
Street	19.7	20.0	19.4	20.5	18.9	17.6
Train	18.7	19.6	19.6	18.9	19.9	17.7
Avg.	19.2	19.7	18.8	19.6	19.1	**17.0**
Avg. of all conditions						
Avg.	12.9	12.7	12.7	12.8	12.8	**11.2**

the ASR system which is based on hybrid Hidden Markov Model-Deep Neural Network (HMM-DNN) framework. A deep belief network- deep neural network (DBN-DNN) with 4 hidden layers each having 2048 hidden units is used for acoustic modeling. All the baseline input features are processed with a input temporal context of 31 acoustic frames with 40 bands (yielding an input dimension of 1240) and the DNN is trained with sigmoid nonlinearity and soft-max output layer. A trigram language model is used in the ASR decoding.

The ASR performance in clean training condition is reported in Table I. From this table, it can be observed that PN and ET features provide better performance compared to the Mel. and RA features. The proposed approach of using MAR spectrogram improves the baseline performance significantly in various noise and channel distortion conditions (average relative improvements of 24 % over mel features and about 14 % over the ET features).

In the multi condition training scenario (reported in Table II), most of the noise robust front-ends show only minor improvements over the baseline mel features. However, the proposed MAR features perform better than all other noise robust front-ends (average relative improvements of 12 % over ET features). Even when the training and test conditions are matched, the proposed features improve the ASR performance as the MAR model focuses on the 2-D structure of the spectrogram. Since the noise and other channel artifacts present in the signal are typically less correlated and lack an inherent time-frequency structure, the MAR model extracts more speech related information from the signal compared to other robust front-ends. This is evident from improvements obtained for a variety of noise types and channel distortions (Table I, II).

B. Reverberant speech recognition

The ASR experiments on reverberant speech data are performed using the REVERB challenge [7] data which uses the WSJCAM0 database for training. This database consists of 7861 recordings from 92 training speakers, 1488 recordings from 20 development test (dt) speakers and 2178 recordings from two sets of 14 evaluation test (et) speakers, with each speaker providing about 90 utterances. These recordings were carried out with two sets of microphone- head mounted as well as desk microphone positioned about half meter from the speaker's head. The database consists of three subsets: training data set (Train) - for both clean and multi condition training using simulated reverb data, a simulated test dataset (Sim) and a naturally reverberant recording of the test dataset (Real). The Kaldi toolkit [20] is used for training and the acoustic model configuration is similar to the one used for Aurora-4. The average performance on 6 simulated room environments and 2 real recordings for each of the (dt) and (et) speaker set is reported in Table III.

TABLE III
WORD ERROR RATE (%) IN REVERB CHALLENGE DATABASE FOR CLEAN AND MULTI-CONDITION TRAINING.

Cond	Mel	PN	ET	MAR	Mel	PN	ET	MAR
	Clean training				Sim. Reverb training			
Sim-dt	37.2	36.3	25.9	**23.8**	11.9	11.3	12.3	**11.0**
Sim-et	35.8	35.2	25.0	**22.4**	12.2	11.5	12.0	**10.7**
Real-dt	70	73.3	57.6	**51.1**	25.9	25.7	36.4	**25.1**
Real-et	73.1	77	59.7	**57.6**	30.9	30.7	34.1	**28.5**
Avg.	54.0	55.5	42.1	**38.7**	20.2	19.8	23.7	**18.8**

It can be observed that the proposed features perform better than other features under all the test conditions with clean and multi-condition training data. For the clean training, there is an average relative improvement of 28 % over MF features and the results with the proposed MAR front-end are better than the best published results in REVERB Challenge [7]. For the multi condition reverb training (simulated), there is an average relative improvement of 7 % over MF features.

V. SUMMARY AND DISCUSSION

In this paper, a novel method of multivariate autoregressive (MAR) modeling is proposed for speech spectrogram estimation. The MAR model uses the linear prediction model on the multi band DCT components to predict the long-term Riesz transform envelopes of speech. With the joint modeling of multiple sub-bands, the MAR model can capture the 2-D structure of the speech spectrogram in the time-frequency domain and the AR modeling attempts to emphasize the high energy regions in the resulting envelopes. The MAR spectrogram, used for speech recognition experiments with noisy and reverberant speech, shows significant improvements over various other noise robust front-ends. These experiments also highlight that robustness in ASR can be achieved by focussing primarily on speech characteristics without explicitly modeling noise or reverberation artifacts. Hence, the improved robustness in the proposed approach does not lead to any degradation in clean or matched test conditions. The proposed modeling framework still uses only localized sub-band structure (only 3 bands in the MAR modeling). In future, we plan to extend the approach to efficiently utilize the entire spectral range of the sub-bands.

REFERENCES

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[3] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, 2013, pp. 7398–7402.

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[5] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[6] E. ETSI, "202 050 v1. 1.1 stq; distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES*, vol. 202, no. 050, p. v1, 2002.

[7] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, 2009.

[8] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4101–4104.

[9] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[10] M. Athineos and D. P. Ellis, "Autoregressive modeling of temporal envelopes," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5237–5245, 2007.

[11] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1912–1924, 1999.

[12] S. Ganapathy, S. H. Mallidi, and H. Hermansky, "Robust feature extraction using modulation filtering of autoregressive models," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 8, pp. 1285–1295, 2014.

[13] M. Athineos, H. Hermansky, and D. P. Ellis, "PLP2: Autoregressive modeling of auditory-like 2-D spectro-temporal patterns," in *Workshop on Statistical and Perceptual Audio Processing (SAPA)*. EPFL-CONF-83126, 2004.

[14] A. Neumaier and T. Schneider, "Estimation of parameters and eigenmodes of multivariate autoregressive models," *ACM Transactions on Mathematical Software (TOMS)*, vol. 27, no. 1, pp. 27–57, 2001.

[15] H. Lütkepohl, *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[16] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[17] A. Zellner, "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias," *Journal of the American statistical Association*, vol. 57, no. 298, pp. 348–368, 1962.

[18] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[19] S. O. Sadjadi and J. H. Hansen, "Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification," *Speech Communication*, vol. 72, pp. 138–148, 2015.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.