Interspeech 2018
2-6 September 2018, Hyderabad

Far-Field Speech Recognition Using Multivariate Autoregressive Models

Sriram Ganapathy[1], *Madhumita Harish*[2]

[1]Learning and Extraction of Acoustic Patterns (LEAP) Lab, Indian Institute of Science, Bangalore.
[2] Carnegie Mellon University, Pittsburgh, U.S.A.

sriramg@iisc.ac.in,mharish@andrew.cmu.edu

Abstract

Automatic speech recognition (ASR) in far-field reverberant environments is challenging even with the state-of-the-art recognition systems. The main issues are artifacts in the signal due to the long-term reverberation that results in temporal smearing. The autoregressive (AR) modeling approach to speech feature extraction involves representing the high energy regions of the signal which are less susceptible to noise. In this paper, we propose a novel method of speech feature extraction using multivariate AR modeling (MAR) of temporal envelopes. The sub-band discrete cosine transform (DCT) coefficients obtained from multiple speech bands are used in a multivariate linear prediction setting to derive features for speech recognition. For single channel far-field speech recognition, the features are derived using multi-band linear prediction. In the case of multi-channel far-field speech recognition, we use the multi-channel data in the MAR framework. We perform several speech recognition experiments in the REVERB Challenge database for single and multi-microphone settings. In these experiments, the proposed feature extraction method provides significant improvements over baseline methods (average relative improvements of 9.7 % and 3.9 % in single microphone conditions for clean and multi-conditions respectively and 6.3 % in multi-microphone conditions). The results with clean training on single microphone conditions further illustrates the effectiveness of the MAR features.

Index Terms: Far-field speech recognition, 3D CNN modeling, Multi-variate Autoregressive (MAR) modeling, Time-frequency analysis, Single and multi-Channel speech processing.

1. Introduction

The advancement of deep neural networks has established a benchmark in modeling an Automatic Speech Recognition system (ASR). Although its performance has shown to have improved over the decades, degradation due to reverberation is a notable challenge in the development of a real world application of hands free ASR. [1]. For example, Peddinti *et al.,* [2] reports a 75% rel. degradation in word error rate (WER) when far-field array microphone signals are used instead of the headset microphones in the ASR systems, both during training and testing. The main issue in reverberant environments is the temporal smearing of the received speech signal.

A conventional solution to the reverberation artifacts is to employ multi-microphone sensors to record speech signals. A delay-sum based approach called beamforming is subsequently employed for multi-channel signal based speech enhancement [3, 4]. Performance degradation in reverberated conditions can also be overcome by training on multi conditioned data [5]. It has been observed that the performance of an ASR system is substandard even after going through speech enhancement and the multi-condition training, when compared to the clean test data (without reverberations). This points to a need to attain noise robustness either at the signal analysis stage in the front-end or at the statistical modeling stage. In this paper, the issue of robustness in feature extraction has been addressed.

The use of a multivariate time series analysis for feature extraction has been explored in this paper [6, 7]. The multivariate AR (MAR) modeling is an approach where a linear combination of past vectors have been used to approximate a random time series vector. The coefficients of prediction are matrices estimated using a least squares criterion. The MAR modeling technique has been broadly used for forecasting applications in econometrics [8]. In the past, speech enhancement using autoregressive modeling has been explored for multi-channel dereverberation [9]. Applications of the MAR model for joint-time frequency modeling shows a significant promise for noisy speech recognition [10]. In this paper, we use the MAR model for feature extraction in single and multi-channel speech recognition.

The MAR approach to feature extraction uses the long-term windows of the signal (2000 ms) which are processed with a discrete cosine transform (DCT). The DCT coefficients are windowed in mel-spaced sub-bands. The mel-windowed DCT components from multi-microphone data are jointly used in the MAR framework to directly model the multi-microphone data (without any beamforming). The MAR coefficients here characterize the sub-band temporal envelopes of the multi-channel speech signal. The MAR modeling allows a representation of signal peaks in the multi-channel data and exploits the inherent 2-D structure along the time-channel space. Hence, these representations can be suitable for dealing with reverberation artifacts in the speech signal. In the case of single-channel speech recognition, the MAR features are extracted using the joint time-frequency modeling.

Experiments are performed on the REVERB challenge dataset using both single and multi-microphone conditions. In the case of single microphone conditions, the MAR framework has been used to model the joint time frequency by performing MAR on multi-band data (similar to previous work in [10]). We further compare the proposed approach to other noise robust feature extraction methods as well as the beamforming approach for multi-channel speech enhancement. In these experiments, the proposed MAR method provides significant improvements for the single and multi-microphone conditions. We also illustrate the benefits of the proposed approach for mismatched conditions in the REVERB challenge data using clean training conditions as well.

The rest of the paper is organized as follows. In Sec. 2, we describe the MAR model and the estimation method. The application of MAR model for speech feature extraction in single microphone conditions is discussed in Sec. 3 and for multi-microphone conditions in Sec. 3.2. The speech recognition experiments and results are described in Sec. 4. In Sec. 5, we

10.21437/Interspeech.2018-2003

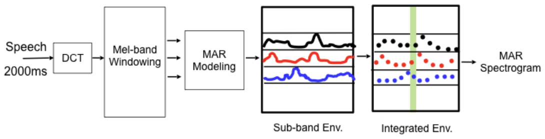

Figure 1: *Block schematic of the MAR spectrogram model for single channel speech.*

conclude with a summary of the proposed features.

2. Multivariate Autoregressive Modeling

A multivariate AR model of order p is given by [7],

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

where $\mathbf{y}$ is a vector process of a sequential data with index q, varying from $1, .., Q$ and of dimension D. $\boldsymbol{\nu}$ is the D dimensional mean vector, $\mathbf{u}$ is a white noise random process having covariance $\boldsymbol{\Sigma}_u$ and dimension D. The MAR coefficients $\mathbf{A}_k$ are D dimension square matrices characterizing the model. The generalized least squares estimation (GLS) of the MAR parameters has been presented in the next subsection of the paper.

To illustrate the MAR Model described by Eq. 1, we define a $D \times Q$ matrix, $\mathbf{Y} := [\mathbf{y}_1, ..., \mathbf{y}_Q]$, a $D \times (Dp + 1)$ matrix, $\mathbf{B} := [\boldsymbol{\nu}, \mathbf{A}_1 ..., \mathbf{A}_p]$, a $D \times Q$ dimension matrix, $\mathbf{U} := [\mathbf{u}_1, ..., \mathbf{u}_Q]$ and a vector, $\mathbf{Z}_t := [1 \ \mathbf{y}_q^T \ \mathbf{y}_{q-1}^T \ ... \ \mathbf{y}_{q-p+1}^T]^T$ of dimension $(Dp + 1) \times 1$. Eq. 1 can thus be re-written as

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

where $\mathbf{Z} := [\mathbf{Z}_0, ..., \mathbf{Z}_{Q-1}]$ with dimension $(Dp + 1) \times Q$. Here, presample observations $\mathbf{y}_{-p+1}, .., \mathbf{y}_0$ are assumed to be available. Let vec denote a stacking operator which converts an $m \times n$ matrix to an $mn \times 1$ vector by stacking the columns of the matrix one below the other. Defining $\mathfrak{u} := vec(\mathbf{U})$ $(DQ \times 1)$, $\boldsymbol{\beta} := vec(\mathbf{B})$ $((D^2p + D) \times 1)$ and $\mathfrak{y} = vec(\mathbf{Y})$ $((D^2p + D) \times 1)$, we obtain,

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

where the Kronecker product is denoted by $\otimes$ and $\mathbf{I}_D$ is a D dimensional identity matrix. The covariance matrix of $\mathfrak{u}$ is thus given by $\mathbf{I}_Q \otimes \boldsymbol{\Sigma}_u$. Here, the cost function $S(\boldsymbol{\beta})$[11] is minimized by the GLS estimator. Using matrix operations such as the inverse of Kronecker product[1] and the commutative property [2], we get,

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

where C is a constant, independent of $\boldsymbol{\beta}$.

2.1. Model parameter estimation

The parameter estimates of the model can then be obtained by setting $\frac{\partial S}{\partial \boldsymbol{\beta}} = \mathbf{0}$. The estimate $\hat{\boldsymbol{\beta}}$ can be written as,

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

The Hessian matrix $\frac{\partial^2 S}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = 2(\mathbf{ZZ}^T \otimes \Sigma_u^{-1})$ is positive definite which indicates a minima. The above formulation reduces to the normal equations in a traditional AR model if $D = 1$. The estimator in Eq. (5) is consistent and asymptotically normal [7]. The estimate $\hat{\boldsymbol{\Sigma}}_u$ can be obtained as follows,

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

2.2. Envelope Estimation

MAR modeling has widely seen applications in forecasting [8]. However, in this paper, we use MAR models to estimate temporal envelopes. For a 1-D time domain AR model, the spectral envelope of the sequence y_q with coefficients $\mathbf{A}_k = a_k$ is given by,

$$\hat{s_y}[f] = \frac{\sigma_u^2}{|\sum_{k=0}^{p} a_k e^{-i2\pi k f}|^2} \tag{7}$$

where $s_y[f]$ is the power spectral density where f is the normalized frequency index. σ_u^2 is the prediction gain [12]. In the case of MAR, computation of spectral envelope is more involved. If $\mathbf{y}_q$ denotes a mean removed time series data ($\boldsymbol{\nu} = \mathbf{0}$) indexed by q, the following multidimensional z-transform filter expression can be written for the model described in Eq. 1,

$$[\mathbf{I}_D - \sum_{k=0}^{p} \mathbf{A}_k z^{-k}]\mathbf{y}_q = \mathbf{u}_q \tag{8}$$

$\mathbf{H} = \mathbf{I}_D - \sum_{k=0}^{p} \mathbf{A}_k z^{-k}$ estimated at $z = e^{-j2\pi f}$. If $\mathbf{s}_\mathbf{y}[f]$ denotes the vector power spectral density for the process $\mathbf{y}_q$, then the MAR estimate of the spectral envelope is given by,

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

In our model estimation, we have also found that the model covariance matrix $\hat{\boldsymbol{\Sigma}}_u$ captures the significant variations when there is a mis-match. By setting $\hat{\boldsymbol{\Sigma}}_u = \boldsymbol{I}$, we can achieve gain normalization of the temporal envelopes, a property that we will later use in the model for mis-match train/test conditions.

3. Feature Extraction using MAR

3.1. Single Channel Speech

The one dimensional autoregressive modeling applied to the data in the time domain yields an all-pole estimate of power spectrum of the signal [12]. In a dual manner, the autoregressive modeling of discrete cosine transform (DCT) coefficients of a signal, yields the all-pole estimate of the Hilbert envelope of the signal [13, 14]. In the latter case, the AR modeling is applied on the DCT coefficients of each sub-band individually. In this paper, the DCT coefficients of multiple sub-bands are modeled using the MAR approach. Thus, the Hilbert envelope of multiple sub-bands are jointly estimated using Eq. 9.

The block schematic of the proposed approach for feature extraction in single channel case is shown in Fig. 1. Here, long input speech segments (2000ms of non-overlapping windows) are transformed using DCT. The full-band DCT signal is windowed into a set of 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 stacked together to form vector series data $\mathbf{y}_q$ of Eq. (1).

[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, $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$.

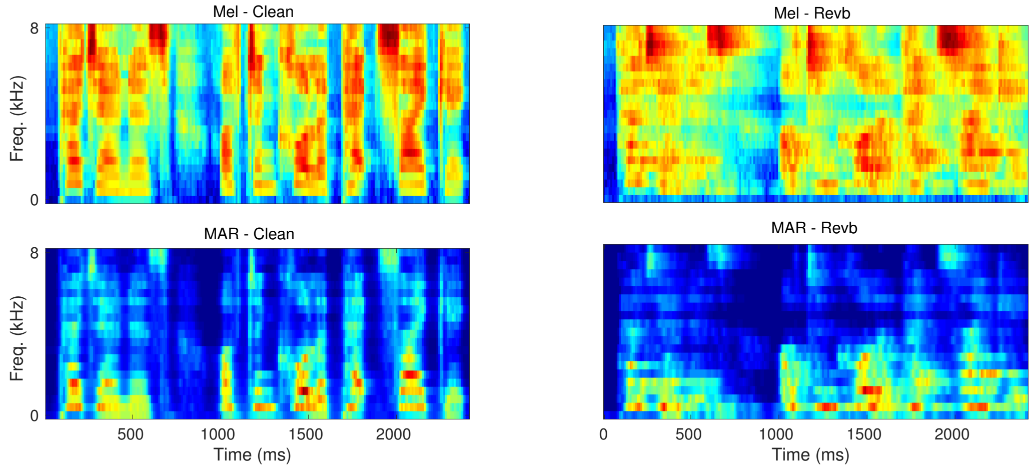

Figure 2: *Spectrogram of a portion of speech (2.5s) in clean and simulated reverberant conditions for conventional mel processing and the proposed MAR model.*

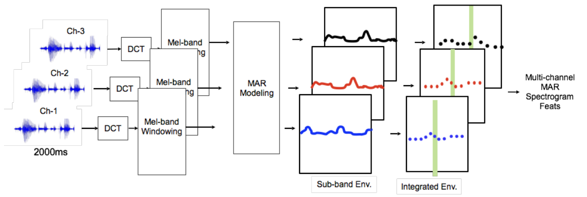

Figure 3: *Block schematic of the MAR spectrogram model for multi-channel speech.*

In this case, q corresponds to the sub-band DCT coefficient index. The estimation procedure of the MAR model is applied and model parameters are estimated (Eq. 1. We use a fixed model order of 160 for the MAR estimation of 2000ms of speech. The sub-band temporal envelopes are then computed using Eq. 9.

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 envelope yields an estimate of the short-term power spectrum. This gives an estimate of the MAR spectrogram of the input speech signal. In Fig.2, we compare the spectrogram 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 on relatively high signal-to-noise (SNR) regions of the speech signal as illustrated by the spectrogram representations obtained for the reverberated signal (simulated reverb condition). The envelope estimation in the proposed MAR model enhances the changes in the signal energy while suppressing the constant regions of the signal. These properties of the MAR model provide robustness in the representations derived from this approach.

3.2. Multi-Channel Speech

In the case of multi-channel speech, we use the speech data from all the channels in deriving the MAR model. In this paper, we use 3 parallel channel recordings form the microphone array. The MAR feature extraction for multi-channel speech is illustrated in Fig. 3. Here, the DCT coefficients of the same sub-band from all the recording (in the multi-channel setting) are used jointly in the MAR framework. Thus, q in Eq. 1 corresponds to the channel index. The main difference in the multi-channel and single channel processing is that the multi-channel processing used MAR to model sub-band envelopes of multiple microphone signals for each sub-band whereas the single channel MAR framework uses multi-band modeling. While it is possible to model the multi-band and multi-channel signal jointly in the MAR framework, we have not explored this option for the current work.

4. Experimental Setup

4.1. Database

The ASR experiments on reverberant speech data are performed using WSJCAM0 corpus in a single channel scenario, released as a part of REVERB challenge [15]. 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 and a 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). For single channel ASR experiments, we use the clean condition and multi-condition setting and for the multi-channel ASR, we use 3 channels from the array microphone recordings.

The acoustic model is built using Kaldi ASR setup [16] where an initial HMM-GMM model is trained to obtain the alignments. A tri-gram language model is used in our ASR experiments. The senone alignments are then used with a Keras engine [17] to build the deep neural network based acoustic model. All the models are trained using frame-level cross entropy criterion. For single channel ASR experiments, we use a deep neural network (DNN) or a convolutional neural network (CNN) with 2-D convolutions. For multi-channel ASR experiments, we also experiment with the recently proposed CNN-3D architecture [18].

Table 1: *Word error rate (%) in single channel multi-condition setting for simulated (S) and naturally reverberant conditions (R) on development (dt) and evaluation (et) datasets. All the models use log-mel features.*

Model	S-dt	S-et	R-dt	R-et	Avg.
DNN	12.7	13.6	31.8	37.5	23.9
LSTM	11.1	11.6	28.3	31.2	20.5
CNN2D	11.3	11.4	26.8	29.6	19.8
CNN2D-Dropout	**10.2**	**10.8**	**25.5**	**27.7**	**18.6**

Table 2: *Word error rate (%) in single channel multi-condition setting with CNN-2D-Dropout model.*

Feat.	S-dt	S-et	R-dt	R-et	Avg.
Mel	10.2	10.8	25.5	27.7	18.6
PNFBE	**9.9**	10.7	**24.3**	29.4	18.6
MAR-3Band	10.1	**10.3**	24.4	**26.7**	**17.9**

4.2. Single Channel Multi Condition Training

The initial experiments reported in Table 1 are performed with 23 dimensional log-mel filterbank energies which are mean and variance normalized. A context of 21 frames is used as the context for the DNN model or for generating the time-frequency representation used at the input of the CNN. The DNN model consists of 4 layers of 1024 dimensions. The LSTM model has 3 layers of 256 units each. The CNN model has 4 convolutional layers (2 layers of 256 kernels and 2 layers of 128 kernels with all kernel sizes set to 3×3) and two feed-forward layers of 1024 dimensions. A frequency max-pooling was also applied after the second and fourth convolutional layers. We also experiment with the use of dropout in CNN model [19]. A dropout factor of 0.2 is used in all the layers. As seen from the experiments in Table 1, the CNN2D models with dropout gives the best performance. The rest of experiments reported in this paper use the CNN architecture.

The next set of ASR experiments reported in Table 2 compares the performance of various feature extraction schemes using the CNN2D model with dropout. The noise-robust feature extraction scheme using power normalized filter bank energy (PNFBE) features [20] (40 dimensional) are compared in these experiments with the proposed MAR modeling approach (21 dimensional). All the features are processed with utterance level mean and variance normalization. As seen in these experiments, the proposed MAR approach using a multi-band framework for processing the signal provides significant improvements compared to other feature extraction methods (average relative improvements 3.9 % over the baseline log-mel features). These improvements can be attributed to the peak modeling property of the AR estimation as well as the joint multi-band modeling provided by the MAR framework.

4.3. Multi-Channel Multi-Condition Training

Here, we experiment with two different ways of acoustic modeling in a multi-microphone setting (with 3 parallel microphone channels). In the first case, the microphone recordings are beamformed (BF) and the enhanced signal is used for feature extraction and acoustic modeling. In the second approach, the features for the 3 channels are extracted separately and a CNN-3D model is used to jointly model the time-frequency-channel space [18]. The results are reported in Table 3. In the case of log-mel features, the beamforming model approach improves the performance over the single channel conditions. The model with dropout training further improves the ASR accuracies. While the CNN3D model has previously shown improvements for multi-speaker case [18], the BF based signal enhancement provided the best ASR results in the REVERB Challenge case (as there is only a single source (speaker) in these recordings).

Table 3: *Word error rate (%) in multi-channel multi-condition setting. Here beamforming is denoted as BF.*

Feat+Model	S-dt	S-et	R-dt	R-et
BF-Mel-CNN2D	9.7	10.0	24.8	26.4
BF-Mel-CNN2D-Dropout	9.0	9.5	23.8	25.3
Mel-CNN3D	9.8	10.3	26.7	28.4
Mel-CNN3D-Dropout	9.1	9.8	24.6	25.8
MAR-1Ch-CNN2D-Dropout	9.9	10.1	23.4	26.4
MAR-CNN3D-Dropout	10.1	10.3	23.2	27.0
BF-MAR-CNN2D-Dropout	**8.7**	**8.9**	**22.9**	**24.3**

Table 4: *Word error rate (%) in single channel clean-condition setting with the CNN-2D-Dropout model.*

Feat.	S-dt	S-et	R-dt	R-et	Avg.
Mel	37.7	35.6	79.6	79.8	58.2
PNFBE	39.5	37.9	77.9	79.6	58.7
MAR-3Band	**33.9**	**31.7**	**70.3**	**74.0**	**52.5**

For the proposed MAR features, the ASR results are further improved over the log-mel features. These results are further encouraging as the beamforming already provided significant benefits by suppressing the reverberation and enhancing the quality of the signal. In the case of multi-microphone experiments, the MAR features provide 6.3 % relative improvements over the log-mel baseline.

4.4. Single Channel Clean Condition Training

The performance of same set of features using the CNN2D model is also compared in a clean training condition (Table 4). All the features have significant drop in performance owing to the mis-match in training and test conditions. Even in the mis-matched conditions, the proposed MAR features provide significant benefits of far-field speech recognition. In the case of single-microphone clean conditioned experiments, the MAR features provide 9.7 % relative improvements over the log-mel baseline.

5. Summary and Future Work

A novel method for feature extraction of speech using the multivariate AR modeling of the temporal envelopes has been proposed in this paper. A multivariate linear prediction is performed on the sub-band DCT components from multiple speech bands for a robust feature extraction in the ASR setup. The MAR features extracted using the joint time frequency modeling deals with the reverberation artifacts by emphasizing on the speech characteristics which shows a significant improvement in the performance of the ASR. With several experiments on matched conditions (single and multi-microphone) and in mismatched training conditions, we have shown the effectiveness of the proposed features. In future, we plan to extend the MAR framework by deriving multi-band and multi-channel features jointly.

6. References

[1] Jinyu Li, Li Deng, Yifan Gong, and Reinhold 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.

[2] Vijayaditya Peddinti, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, 2017.

[3] Matthias Wölfel and John McDonough, *Distant speech recognition*, John Wiley & Sons, 2009.

[4] Marc Delcroix et al., "Strategies for distant speech recognition in reverberant environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 60, 2015.

[5] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7398–7402.

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

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

[8] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung, *Time series analysis: forecasting and control*, John Wiley & Sons, 2015.

[9] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 85–88.

[10] Sriram Ganapathy, "Multivariate autoregressive spectrogram modeling for noisy speech recognition," *IEEE signal processing letters*, vol. 24, no. 9, pp. 1373–1377, 2017.

[11] Arnold 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.

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

[13] Marios Athineos and Daniel PW Ellis, "Autoregressive modeling of temporal envelopes," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5237–5245, 2007.

[14] Sriram Ganapathy, "Signal analysis using autoregressive models of amplitude modulation," *PhD Thesis, Johns Hopkins University*, 2012.

[15] Keisuke Kinoshita, Marc Delcroix, Tomohiro Nakatani, and Masato 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.

[16] Daniel Povey et al., "The kaldi speech recognition toolkit," in *IEEE ASRU*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[17] François Chollet et al., "Keras: Deep learning library for theano and tensorflow," *URL: https://keras. io/k*, 2015.

[18] Sriram Ganapathy and Vijayaditya Peddinti, "3-d cnn models for far-field multi-channel speech recognition," *Proceedings of ICASSP*, 2018.

[19] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[20] Chanwoo Kim and Richard 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.