Deep Learning Based Dereverberation of Temporal Envelopes for Robust Speech Recognition

Anurenjan Purushothaman, Anirudh Sreeram, Rohit Kumar, Sriram Ganapathy

Learning and Extraction of Acoustic Patterns (LEAP) lab,
Electrical Engineering, Indian Institute of Science, Bangalore.
{anurenjanr, sanirudh, rohitk, sriramg}@iisc.ac.in

Abstract

Automatic speech recognition in reverberant conditions is a challenging task as the long-term envelopes of the reverberant speech are temporally smeared. In this paper, we propose a neural model for enhancement of sub-band temporal envelopes for dereverberation of speech. The temporal envelopes are derived using the autoregressive modeling framework of frequency domain linear prediction (FDLP). The neural enhancement model proposed in this paper performs an envelop gain based enhancement of temporal envelopes and it consists of a series of convolutional and recurrent neural network layers. The enhanced sub-band envelopes are used to generate features for automatic speech recognition (ASR). The ASR experiments are performed on the REVERB challenge dataset as well as the CHiME-3 dataset. In these experiments, the proposed neural enhancement approach provides significant improvements over a baseline ASR system with beamformed audio (average relative improvements of 21% on the development set and about 11% on the evaluation set in word error rates for REVERB challenge dataset).

Index Terms: Automatic speech recognition, frequency domain linear prediction (FDLP), Dereverberation, Neural speech enhancement.

1. Introduction

Automatic speech recognition (ASR) systems find widespread use in applications like human-machine interface, virtual assistants, smart speakers etc, where the input speech is often reverberant and noisy. The ASR performance has improved dramatically over the last decade with the help of deep learning models [1]. However, the degradation of the systems in presence of noise and reverberation continues to be a challenging problem due to the low signal to noise ratio [2]. For *e.g.* Peddinti *et al.,* [3] reports a 75% rel. increase in word error rate (WER) when signals from a far-field array microphone are used in place of those from headset microphones in the ASR systems, both during training and testing. This degradation could be primarily attributed to reverberation artifacts which smear the time domain envelopes of the speech signal [4, 5].

The traditional approach to multi-channel far-field ASR combines all the available channels by beamforming [6]. Recently, unsupervised DNN-mask estimator based beamforming is also proposed for generalized eigen value (GEV) based beamforming [7]. Along with the beamforming, the weighted prediction error (WPE) [8] based dereverberation is used in state-of-art ASR systems in reverberant environments. In addition, multi-condition training, where reverberation is simulated in training data is commonly employed to reduce the mis-match between training and testing [9]. However, even with these methods, the temporal smearing of sub-band envelopes, caused by the combination of the direct path and the reflected paths in reverberation, continue to degrade the ASR performance [10].

In this paper, we propose an approach for sub-band envelope enhancement which attempts to learn the mapping of the reverberated envelopes to the close-talking ones. The sub-band envelopes are derived using the autoregressive modeling framework of frequency domain linear prediction [11, 12]. A deep neural model based on convolutional and recurrent layers is trained to enhance the reverberated sub-band FDLP envelopes. Following the DNN model training, which predicts an envelope gain, the output of the model is multiplied with the sub-band envelopes of the reverberant speech to suppress the effects of reverberation. The enhanced sub-band envelopes are used for feature extraction of ASR. In various ASR experiments on the REVERB challenge dataset [5] as well as the CHiME-3 dataset [13], we show that the proposed approach improves over the state-of-art ASR systems based on log-mel features with GEV beamforming and WPE enhancement.

2. Related Prior Work

The early works by Xu et. al. [14] targeted the enhancement of signals corrupted by additive noise where a supervised neural network method was proposed to enhance speech by means of finding a mapping function between noisy and clean speech signals. In a similar manner, speech separation (the problem of separating the target speaker speech from the background interference) has seen considerable progress using neural methods with ideal ratio mask based mapping [15].

For reverberant speech, Zhao et al., proposed a LSTM model to predict late reflections in the spectrogram domain [16]. A spectral mapping approach using the log-magnitude inputs was attempted by Han et. al [17]. A mask based approach to dereverberation on the complex short-term Fourier transform domain was explored by Williamson et. al [18]. A recurrent neural network model to predict the spectral magnitudes for dereverberation of speech was also proposed by Santos et. al [19]. Speech enhancement for speech recognition based on neural networks has been explored in [20, 21, 22]. In [23] a recurrent neural network is used to map noise-corrupted input features to their corresponding clean versions.

This work was partly funded by the project grants from Samsung Research India, Bangalore.

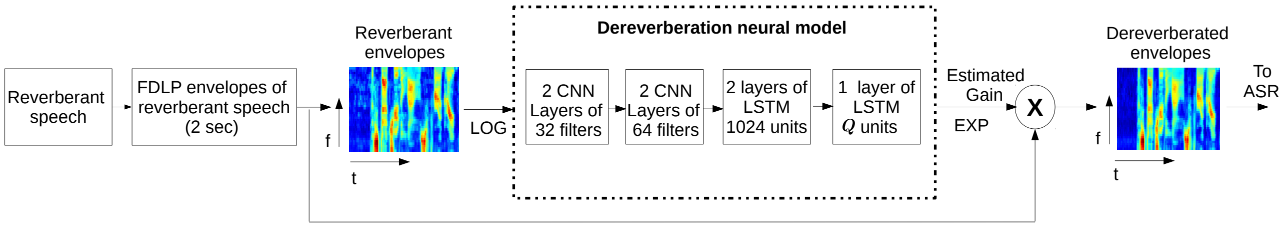
Reverberant speech
FDLP envelopes of reverberant speech (2 sec)
Reverberant envelopes
f
t
LOG
Dereverberation neural model
2 CNN Layers of 32 filters
2 CNN Layers of 64 filters
2 layers of LSTM 1024 units
1 layer of LSTM Q units
Estimated Gain
EXP
X
Dereverberated envelopes
f
t
To ASR

Figure 1: *Block schematic of envelope dereverberation model*

3. Proposed Approach

3.1. Signal model

When speech is recorded in far-field reverberant environment, the data collected in the microphone can be expressed as

$$r(t) = x(t) * h(t), \tag{1}$$

where $x(t)$, $h(t)$ and $r(t)$ denote the clean speech signal, the room impulse response and the reverberant speech respectively. The room response function $h(t) = h_e(t) + h_l(t)$, where $h_e(t)$ and $h_l(t)$ represent the early and late reflection components.

Let $x_q(t)$, $h_q(t)$ and $r_q(t)$ denote the sub-band clean speech, room-response and the reverberant speech respectively and $q = 1, \dots, Q$ denotes the sub-band index. Assuming an ideal band-pass filtering we can write (using Eq. 1),

$$r_q(t) = x_q(t) * h_q(t). \tag{2}$$

Now, the analytic signal $r_{aq}(t) = r_q(t) + \mathcal{H}[r_q(t)]$ can be shown to be [11, 24],

$$r_{aq}(t) = \frac{1}{2}[x_{aq}(t) * h_{aq}(t)], \tag{3}$$

For band-pass filters with small band-width, applying magnitude on both sides, we get the following approximation between the sub-band envelope (defined as the magnitude of the analytic signal) components of the reverberant signal and those of the clean speech signal.

$$m_{rq}(t) \simeq \frac{1}{2} m_{xq}(t) * m_{hq}(t), \tag{4}$$

where $m_{rq}(t)$, $m_{xq}(t)$, $m_{hq}(t)$ denote the sub-band envelopes of reverberant speech, clean speech and room response respectively. With this model of reverberation in the envelope domain, we can further split the envelope into early and late reflection coefficients.

$$m_{rq}(t) = m_{rqe}(t) + m_{rql}(t), \tag{5}$$

In this work, the envelopes are also estimated using the autoregressive modeling framework of frequency domain linear prediction (FDLP). Specifically, the discrete cosine transform (DCT) of sub-band signals $r_q(t)$ is computed and a linear prediction (LP) is applied on the DCT components. The LP envelope estimated using the prediction on the DCT components provides an all-pole model of the sub-band envelopes $m_{rq}(t)$ [24].

3.2. Envelope dereverberation model

As seen in Eq. (5), the FDLP envelope of reverberant speech can be expressed as sum of the direct component (early reflection) and those with the late reflection. In the envelope dereverberation model, our aim is to input the envelope of the reverberant sub-band temporal envelope $m_{rq}(t)$ to predict the late reflection components $m_{rql}(t)$. Once this prediction is achieved, the late reflection component can be subtracted from the sub-band envelope to suppress the artifacts of reverberation. A similar analogy to this envelope subtraction approach is the spectral subtraction model where the noise and clean power spectral density (PSD) gets added in noisy speech PSD. If Gaussian assumptions are made for PSD components [25], the Wiener filtering approach to noisy speech enhancement provides the minimum mean squared error, where the noisy PSD is multipled by the gain of the filter. In a similar manner, we pose the dereverberation problem as an envelope gain estimation problem. The sub-band envelope gain in this case is the ratio of the sub-band envelope for the direct components to the sub-band envelope of the reverberant sub-band signal. This sub-band envelope gain estimation is achieved using a deep neural network model in the proposed work. Following the model training, the dereverberation is achieved by multiplying the estimated sub-band envelope gain with the sub-band envelope of reverberant speech.

3.3. Implementation of the envelope dereverberation

The block schematic of the envelope dereverberation model is shown in Figure 1. The input to the dereverberation model is the FDLP sub-band envelope of the reverberant speech. The model is trained to learn the sub-band envelope gain which is the ratio of the clean envelopes (direct component) with the reverberant envelopes. We use the FDLP envelope of the close talking microphone as an estimate of the direct component. As the envelopes and the gain parameters are positive in nature, the model implementation in the neural architecture uses a logarithmic transform at the input and the estimated gain is followed by an exponential operation. This implementation in the log envelope domain makes the model behave like a residual network based dereverberation architecture. It is also noteworthy that the entire model developed in Section 3.1 is applicable only on long analysis windows (which are typically greater than the T60 of the room response function). Hence, unlike previous models for dereverberation, the proposed approach operates on long temporal envelopes of the order of 2 sec. duration. In the neural model, we also predict the envelope gain of all sub-bands jointly to exploit the sub-band correlations that exist in speech.

From the reverberant speech and the corresponding clean speech, the FDLP sub-band envelopes corresponding to 2sec. non-overlapping segments are extracted. If the input sampling rate is 16 kHz, a 2sec. segment will correspond to 32,000 samples. FDLP envelopes are extracted at a down sampled rate of

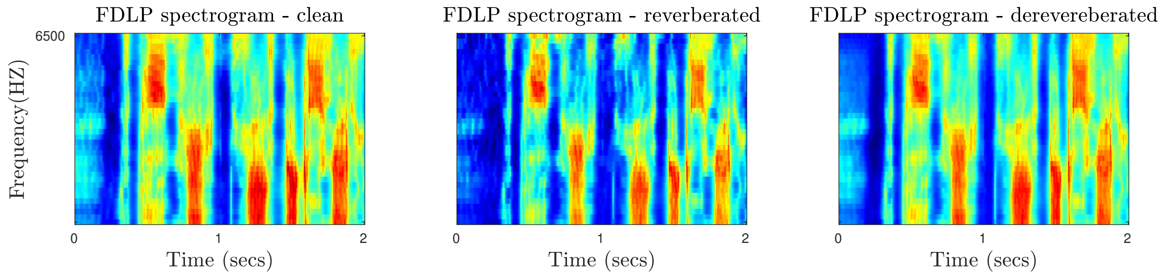

Figure 2: *Comparison of spectrograms, FDLP spectrogram for clean (near-room), reverberant speech (far-room) and far-room after the proposed dereverberation, recordings from the REVERB Challenge dataset.*

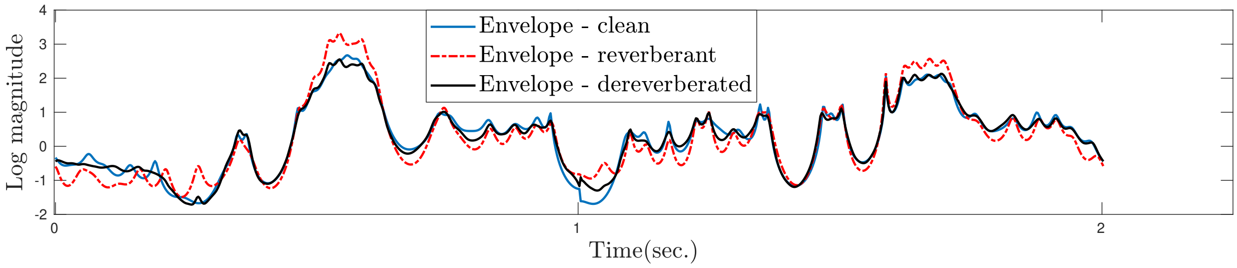

Figure 3: *Comparison of temporal envelopes, FDLP envelopes for clean (near-room), reverberant speech (far-room) and far-room after the proposed dereverberation, recordings from the REVERB Challenge dataset.*

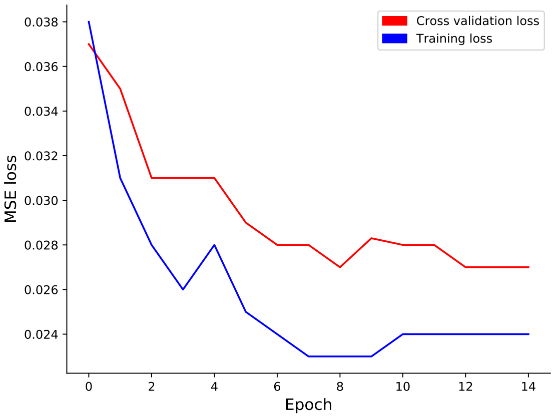

Figure 4: *MSE loss for REVERB challenge dataset*

400 Hz. Thus every 2sec. segment of audio corresponds to 800 samples of FDLP envelope for each sub-band. We use a 36 band mel decomposition. This makes the representation at the input of the enhancement model of size 800 × 36. The target signal for the enhancement model in Figure 1 is the ratio of the close talking (clean) FDLP envelopes with those of reverberant envelopes.

The architecture of the neural model is based on convolutional long short term memory (CLSTM) networks (Figure 1). The input 2-D data of sub-band envelopes are fed to a set of convolutional layers where the first two layers have 32 filters each with kernels of size of 41 × 5. The next two CNN layers have 64 filters with 21 × 3 kernel size. All the CNN layer outputs with ReLU activations are zero padded to preserve the input size and no pooling operation is performed. The output of the CNN layers are reshaped to perform time domain recurrence using 3 layers of LSTM cells. The first two LSTM layers have 1024 cells while the last layer has 36 cells corresponding to the size of the target signal (envelope gain). The training criteria is based on the mean square error between the target and predicted output. The model is trained with stochastic gradient descent using Adam optimizer. The dereverberated envelopes are integrated into 25ms windows with a shift of 10 ms and these are log transformed and used as features for ASR [26].

An analysis of dereverberation training loss variation as a function of the epoch is shown in Figure 4. The training loss and validation loss show consistent reduction during the training process. We run the dereverberation model for about 10 epochs.

An illustration of the envelope enhancement is shown in Figure 2. Here, we plot the FDLP spectrogram (integrated envelopes) for clean signal, reverberated signal and enhanced signal (using the dereveberation model). As seen in Figure 2, the dereverberation model improves the spectogram visibly and makes it closer to the clean FDLP spectrogram.

A more careful visualization of the dereverberation can be achieved using the plot of the sub-band envelope of one single sub-band (10 th mel-band) as shown in Figure 3. The sub-band envelopes of reverberant signal deviate from their clean signal counterparts (as explained in Sec. 3.1). Using the dereverberation model proposed in this paper, we find that the FDLP envelopes are more closely matched with the clean signal envelopes.

4. Experiments and results

The experiments are performed on REVERB challenge and CHiME-3 datasets. For the baseline model, we use WPE enhancement along with unsupervised GEV beamforming. This signal is processed with filter-bank energy features (denoted as BF-FBANK). The FBANK features are 36 band log-mel spectrogram with frequency range from 200 Hz to 6500 Hz. This is the same frequency decomposition used in the FDLP and FDLP-dereverberation experiments. The acoustic model corresponds to 2-D CLSTM network described in [27], consisting of 4 layers of CNN, a layer of LSTM with 1024 units performing recurrence over frequency and 3 fully connected layers with batch normalization.

Table 1: *Word Error Rate (%) in REVERB dataset for different features and proposed dereverberation method.*

Model Features	Dev			Eval		
	Real	**Simu**	**Avg**	**Real**	**Simu**	**Avg**
BF-FBANK	19.1	6.1	12.6	14.7	**6.5**	10.6
BF-FDLP	17.8	6.8	12.3	14.0	7.0	10.5
BF-FBANK + derevb.	17.3	5.5	11.4	13.1	6.9	10.0
BF-FDLP + derevb.	**14.4**	**5.3**	**9.9**	**12.0**	6.8	**9.4**

4.1. ASR framework

We used Kaldi toolkit [28] for deriving the senone alignments used in the PyTorch deep learning framework for acoustic modeling. A hidden Markov model - Gaussian mixture model (HMM-GMM) system is trained with MFCC (Mel Frequency Cepstral Coefficients) features [29] to generate the alignments for training the CLSTM model. A tri-gram language model [30] is used in the ASR decoding and the best language model weight obtained from development set is used for the evaluation set.

4.2. REVERB Challenge ASR

The REVERB challenge dataset [31] for ASR consists of 8 channel recordings with real and simulated reverberation conditions. The simulated data is comprised of reverberant utterances generated (from the WSJCAM0 corpus [32]) obtained by artificially convolving clean WSJCAM0 recordings with the measured room impulse responses (RIRs) and adding noise at an SNR of 20 dB. The simulated data has six different reverberation conditions. The real data, which is comprised of utterances from the MC-WSJ-AV corpus [33], consists of utterances spoken by human speakers in a noisy reverberant room. The training set consists of 7861 utterances from the clean WSJCAM0 training data by convolving with 24 measured RIRs.

4.2.1. Discussion

Table 1 shows the WER results for experiments on REVERB challenge dataset. The WPE applied unsupervised GEV beamformed signal is used for the FDLP baseline (denoted as BF-FDLP). The BF-FDLP baseline by itself is better than the BF-FBANK baseline (average relative improvements of 2% on the development set and about 1% on the evaluation set). For a fair comparision of the proposed approach, we have applied a similar dereverbaration method on BF-FBANK baseline. Here, we have trained the neural model with log-mel features corresponding to 2 sec. duration with all the 36 mel-bands jointly. This approach is denoted as BF-FBANK + dereverberation. Average relative improvements of 10% on the development set and about 6% on the evaluation set is achieved compared to the BF-FBANK baseline.

Finally applying the proposed neural model based dereverberation on BF-FDLP baseline (denoted as BF-FDLP + dereverberation) yields average relative improvements of 21% on the development set and about 11% on the evaluation set, compared to the BF-FBANK baseline. The improvement in real condition is much more than that of simulated data. Average relative improvements of 25% on the real development set and about 18% on the real evaluation set, compared to the BF-FBANK baseline is achieved by the proposed method. This suggests that, even though the neural model is trained only with simulated reverberations, it generalizes well on unseen real data.

Table 2: *Word Error Rate (%) in CHiME-3 dataset for different features and proposed dereverberation method.*

Model Features	Dev			Eval		
	Real	**Simu**	**Avg**	**Real**	**Simu**	**Avg**
BF-FBANK	7.8	8.0	8.0	14.0	9.7	11.8
BF-FDLP	7.0	8.1	7.5	12.0	10.0	11.0
BF-FBANK + derevb.	7.2	8.3	7.7	12.9	9.8	11.4
BF-FDLP + derevb.	7.2	**7.9**	7.5	13	**9.6**	11.3
+ reg.	**6.9**	8.0	**7.4**	**11.8**	9.8	**10.8**

Table 3: *Word Error Rate (%) in CHiME-3 dataset for different features and proposed dereverberation method with RNN-LM*

Model Features	Dev			Eval		
	Real	**Simu**	**Avg**	**Real**	**Simu**	**Avg**
BF-FBANK	5.8	6.2	6.0	10.8	7.2	9.0
BF-FDLP	5.1	6.1	5.6	**9.2**	7.5	**8.4**
BF-FBANK + derevb.	5.0	6.3	5.7	10.0	7.6	8.8
BF-FDLP + derevb.	5.0	6.2	5.6	9.9	**7.2**	8.6
+ reg.	**5.0**	**6.0**	**5.5**	9.5	7.6	8.6

4.3. CHiME-3 ASR

The CHiME-3 dataset [13] for the ASR has multiple microphone tablet device recording in four different environments, namely, public transport (BUS), cafe (CAF), street junction (STR) and pedestrian area (PED). For each of the above environments real and simulated data are present. The real data consists of 6 channel recordings from WSJ0 corpus sampled at 16 kHz spoken in the four varied environments. The simulated data was constructed by mixing clean utterances with the environment noise. The training dataset consists of 1600 (real) noisy recordings and 7138 (simulated) noisy recordings from 83 speakers.

4.3.1. Discussion

The WER results for experiments on CHiME-3 dataset are shown in Table 3. The FDLP baseline, denoted as BF-FDLP is better than the FBANK baseline (BF-FBANK). We observe average relative improvements of 8% on the development set and about 12% on the evaluation set when comparing BF-FDLP and BF-FBANK baseline systems. It can also be seen from Table 3 that the proposed dereverberation method improves the FBANK-baseline system.

In the CHiME-3 dataset, we observed that the significant cause of degradation in the signal quality came from the additive noise sources. Hence, the application of the dereverberation model degraded the performance on the BF-FDLP system (which showing improvements in the BF-FBANK system). On further investigation, we found that the dereverberation model also resulted in smoothing of the spectral variations in the FDLP spectrogram. In order to circumvent this issue, we regularized the MSE loss with a term that encouraged the spectral channels to be uncorrelated. The regularization parameter was kept at 0.05. Using this regularized MSE loss, we improved the BF-FDLP-Dereverberation system results over the dereverberation approach with MSE loss alone. These experiments suggest that even when the audio data does not have significant late reflection components (like CHiME-3 dataset), the proposed approach improves significantly over the baseline method (average relative improvements of 8.5 % over the baseline BF-FBANK system in the eval. condition).

5. Summary

In this paper, we propose a new neural model for dereverberation of temporal envelopes. Using the proposed neural dereverberation approach, we perform speech recognition experiments on the REVERB challenge dataset as well as on the CHiME-3 dataset. These experiments indicate that the proposed neural dereverberation approach generalizes well on unseen reverberant data. The analysis of results also highlight the incremental benefits achieved for application of the proposed approach in other features like, log-mel filter bank features.

6. References

[1] D. Yu and L. Deng, *Automatic Speech Recognition.* Springer, 2016.

[2] T. Hain, L. Burget, J. Dines, P. N. Garner, F. Grézl, A. El Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the amida systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 486–498, 2012.

[3] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and lstms," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2018.

[4] T. Yoshioka *et al.*, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[5] K. Kinoshita *et al.*, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE WASPAA*, 2013, pp. 1–4.

[6] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[7] R. Kumar, A. Sreeram, A. Purushothaman, and S. Ganapathy, "Unsupervised neural mask estimator for generalized eigen-value beamforming based asr," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7494–7498.

[8] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[9] 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*, 2013, pp. 7398–7402.

[10] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, 2017.

[11] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Processing Letters*, vol. 15, pp. 681–684, 2008.

[12] S. Ganapathy and M. Harish, "Far-field speech recognition using multivariate autoregressive models." in *Interspeech*, 2018, pp. 3023–3027.

[13] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.

[14] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[15] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[16] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Late reverberation suppression using recurrent neural networks with long short-term memory," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5434–5438.

[17] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4628–4632.

[18] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[19] J. F. Santos and T. H. Falk, "Speech dereverberation with context-aware recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1236–1246, 2018.

[20] M. Wöllmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional lstm networks for conversational speech recognition in highly non-stationary noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6822–6826.

[21] Z. Chen, S. Watanabe, H. Erdogan, and J. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," *Nuclear Physics A*, vol. 2015-January, pp. 3274–3278, 2015.

[22] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[23] A. L. Maas, T. M. O'Neil, A. Y. Hannun, and A. Y. Ng, "Recurrent neural network feature enhancement: The 2nd chime challenge," in *Proceedings The 2nd CHiME Workshop on Machine Listening in Multisource Environments held in conjunction with ICASSP*, 2013, pp. 79–80.

[24] S. Ganapathy and V. Peddinti, "3-d cnn models for far-field multi-channel speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5499–5503.

[25] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE transactions on speech and audio processing*, vol. 13, no. 5, pp. 845–856, 2005.

[26] S. Ganapathy, "Signal analysis using autoregressive models of amplitude modulation," Ph.D. dissertation, Johns Hopkins University, 2012.

[27] A. Purushothaman, A. Sreeram, and S. Ganapathy, "3-d acoustic modeling for far-field multi-channel speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6964–6968.

[28] 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 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[29] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling." in *ISMIR*, vol. 270, 2000, pp. 1–11.

[30] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[31] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[32] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcamo: a british english speech corpus for large vocabulary continuous speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1995, pp. 81–84.

[33] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (mc-wsj-av): Specification and initial experiments," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, 2005, pp. 357–362.