END-TO-END SPEECH RECOGNITION WITH JOINT DEREVERBERATION OF SUB-BAND AUTOREGRESSIVE ENVELOPES

Rohit Kumar[†]*, Anurenjan Purushothaman*[†‡]*, Anirudh Sreeram*[#]*, Sriram Ganapathy*[†]

[†]Learning and Extraction of Acoustic Patterns (LEAP) lab, Indian Institute of Science, Bangalore.
[‡]College of Engineering, Thiruvanthapuram, India.
[#]University of Southern California, Los Angels, USA.
E-mail - {rohitk, anurenjanr,sriramg}@iisc.ac.in, asreeram@usc.edu

ABSTRACT

The end-to-end (E2E) automatic speech recognition (ASR) systems are often required to operate in reverberant conditions, where the long-term sub-band envelopes of the speech are temporally smeared. In this paper, we develop a feature enhancement approach using a neural model operating on sub-band temporal envelopes. The temporal envelopes are modeled using the framework of frequency domain linear prediction (FDLP). The neural enhancement model proposed in this paper performs an envelope gain based enhancement of temporal envelopes. The model architecture consists of a combination of convolutional and long short term memory (LSTM) neural network layers. Further, the envelope dereverberation, feature extraction and acoustic modeling using transformer based E2E ASR can all be jointly optimized for the speech recognition task. We perform E2E speech recognition experiments on the REVERB challenge dataset as well as on the VOiCES dataset. In these experiments, the proposed joint modeling approach yields significant improvements compared to the baseline E2E ASR system (average relative improvements of 21% on the REVERB challenge dataset and about 10% on the VOiCES dataset).

Index Terms: End-to-End automatic speech recognition, frequency domain linear prediction (FDLP), dereverberation, Joint modeling.

1. INTRODUCTION

In the present era of smart speakers, virtual assistants, and human-machine speech interfaces, the approach of end-to-end (E2E) automatic speech recognition (ASR) finds wide spread application due to the elegance in processing, computational simplicity and edge deployment possibilities. Most of these speech applications require functioning in far-field reverberant environments. The far-field recording condition smears the speech signal [1], adversely impacting the ASR performance [2], with performance degradation of up to 70% [3].

A common approach in multi-channel recording conditions is to use a weighted and delayed combination of the multiple channels using beamforming [4]. The current state-of-art approaches to beamforming use a neural mask estimator [5, 6]. The speech and noise mask estimations are used to derive the power spectral density of the source and interfering signals [7]. Further, a weighted prediction error (WPE) [8] based dereverberation is used along with beamforming. In spite of these approaches to suppress far-field effects, the temporal smearing of sub-band envelopes, causes performance degradation in ASR systems [9].

Our previous work [10, 11] explored the use of dereverberation of sub-band envelopes for hybrid speech recognition systems. The sub-band envelopes are extracted using the autoregressive modeling framework of frequency domain linear prediction [12, 13]. The deep neural enhancement model is trained to predict an envelope gain, which is multiplied with the sub-band envelopes of the reverberant speech for dereverberation.

In this paper, we extend the prior works for far-field E2E ASR systems, where a joint learning of enhancement model and the E2E ASR model is proposed. We explore the boundary equilibrium generative adversarial networks (BEGAN) based loss function [14] in the envelope dereverberation model. In various E2E ASR experiments performed on the REVERB challenge dataset [15] as well as the VOiCES dataset [16], we show that the proposed approach improves over the state-of-art E2E ASR systems based on log-mel features with generalized (GEV) beamforming. Further, we illustrate that the proposed approach yields the best published results on the REVERB challenge dataset.

2. LITERATURE REVIEW

The neural approaches to speech enhancement have witnessed considerable advances in the last decade [17]. In early efforts, Maas et. al [18] proposed a recurrent model to map the noisy features to the clean features. Santos et. al proposed a context aware recurrent neural network [19] for enhancing speech spectrogram. The mapping in the time domain signal is investigated in Pandey et. al [20], where the loss is computed in the frequency domain loss. Recently, end-to-end models with attention based modeling have also been explored on the REVERB challenge dataset [21, 22]. Previously, we had proposed a convolutional neural network model to perform dereverberation of speech [10, 11]. In the current work, we extend this prior work for E2E transformer based ASR system.

3. PROPOSED APPROACH

3.1. Signal model

The speech data recorded by a far-field microphone is modeled as,

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

where $x(n)$, $h(n)$ and $r(n)$ are the source speech signal, the room impulse response function and the far-field speech respectively. The

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

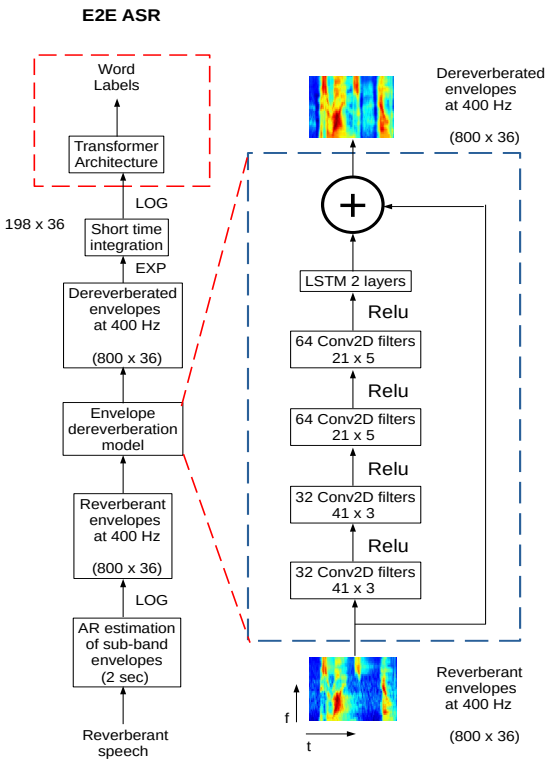

Fig. 1: Block schematic of envelope dereverberation model, the feature extraction module and the E2E ASR model.

room response function can be further expanded as $h(n) = h_e(n) + h_l(n)$, where $h_e(n)$ and $h_l(n)$ represent the early and late reflection components.

Let $x_q(n)$, $h_q(n)$ and $r_q(n)$ denote the sub-band clean speech, room-response and the reverberant speech respectively for the q^{th} sub-band. The sub-band envelopes of far-field speech $m_{rq}(n)$, extracted using frequency domain linear prediction (FDLP) [23, 1], can be approximated as, [10]

$$m_{rq}(n) \approx \frac{1}{2} m_{xq}(n) * m_{hq}(n) \tag{2}$$

where, $m_{xq}(n)$, $m_{hq}(n)$ denote the sub-band envelope of the clean source signal and the room impulse response function respectively. Given this envelope convolution model, we can further split the far-field speech envelope into early and late reflection components.

$$m_{rq}(n) = m_{rqe}(n) + m_{rql}(n) \tag{3}$$

where $m_{rqe}(n)$ and $m_{rql}(n)$ denote the sub-band envelopes of early and late reflection parts.

3.2. Envelope Dereverberation and E2E ASR

In this section, we describe the proposed approach to envelope dereverberation (Figure 1).

3.2.1. Envelope dereverberation model

As seen in Eq. (3), the FDLP envelope of reverberant speech can be written as the sum of the direct component (early reflection) and those with the late reflection. The envelope dereverberation model tries to subtract the late reflection components $m_{rql}(n)$ from reverberant sub-band temporal envelope $m_{rq}(n)$. Similar to the popular Wiener filtering technique for dealing with additive noise in speech [24], the dereverberation task is posed as envelope gain estimation problem. The sub-band envelope residual (targets for the neural model) is the log-difference of the sub-band envelope for the direct components and the sub-band envelope of the reverberant sub-band signal. The neural model is trained with reverberant sub-band envelopes ($log\ (m_{rq}(n))$ as input and model outputs the gain (in the log domain this is $log\ \frac{m_{rq}(n)}{m_{xq}(n)}$), which when multiplied with the reverberant envelopes (additive in log domain), generates the estimate of source signal envelope ($log\ (\hat{m}_{xq}(n))$.

Figure 1 shows the block schematic of the proposed envelope dereverberation model. For training the dereverberation model, we use the FDLP envelope of the close talking microphone to compute the target residual envelope. Thus, the model behaves like a residual network for dereverberation.

The model developed in Section 3.1 is applicable only for envelopes extracted from long analysis windows (greater than the T60 of room impulse response). Further, the neural model also predicts the envelope residual of all sub-bands jointly to effectively utilize the sub-band correlations that exist in speech. For training the neural dereverberation model, the FDLP sub-band envelopes, corresponding to 2 sec. non-overlapping segments of both the reverberant speech and clean speech are extracted. With a 2 sec. segment of FDLP envelopes, sampled at 400 Hz, and a mel decomposition of 36 bands, the input representation to the neural dereverberation model is of dimension 800×36.

The target signal for the dereverberation model in Figure 1 is the log-difference between the envelope of the close talking (clean) FDLP envelopes and those of the reverberant speech. The final 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 (two layers having 32 filters each with kernels of size 41×5 followed by two layers of 64 filters with size 21×3). The CNN layers do not have pooling and include zero padding to preserve the input size. The output of the CNN layers are input to 3 layers of LSTM cells with $1024, 1024$, and 36 units respectively. The last layer size is matched with that of the target signal (log-difference). The training criteria is based on the mean square error between the target and predicted model output. The model is trained with stochastic gradient descent using Adam optimizer. We also experiment with the learning of the model with boundary equilibrium generative adversarial network (BEGAN) loss [14]. The BEGAN [14] is an energy based GAN. The model attempts to match the distribution of loss function using an auto-encoder (AE) architecture. The model uses the equilibrium of AE loss using the hyper-parameter $\gamma \in [0, 1]$, termed as the diversity ratio. We use the BEGAN discriminator to introduce the adversarial setting in the training process.

The predicted sub-band envelope from the dereverberation model ($\hat{m}_{xq}(n)$) for each band is integrated in short Hamming shaped windows of size 25 ms with a shift of 10 ms [1]. A log compression is applied to limit the dynamic range of values. The integrated sub-band envelopes are input to the ASR as 2-D time frequency features of the audio signal. It is noteworthy that the feature extraction steps (Hamming window based integration and log compression) can be represented as fixed neural layers consisting of a 1-D CNN layer. Hence, the entire set of steps, starting from FDLP envelope extraction to the E2E ASR transcript generation, can be

Table 1: WER (%) in the REVERB dataset for different dereverberation model architectures with BF-FDLP features. All the features are used with transformer based E2E acoustic model with separate training of the dereverberation module and E2E ASR module.

Model Architecture	Derevb. Model Parameters (in Million)	Loss Function	DEV			EVAL		
			Real	Sim	Avg	Real	Sim	Avg
2 CNN + Transformer	54.7	MSE	15.1	10.7	12.9	12.2	9.9	11.1
4 CNN + Transformer	137.4	MSE	14.8	9.8	12.3	11.3	9.7	10.5
Linear Layer + Transformer	19.8	MSE	14.3	9.5	11.9	11.5	9.3	10.4
4 CNN + 2 LSTM	14.5	MSE	11.4	7.7	9.5	9.5	7.0	8.2
4 CNN + 2 LSTM	14.5	MSE + 0.2*BEGAN	**11.2**	8.2	9.7	12.5	9.3	10.9
4 CNN + 2 LSTM	14.5	MSE + 0.1*BEGAN	11.3	**7.5**	**9.4**	**8.7**	**6.6**	**7.6**

Table 2: Word Error Rate (%) in REVERB dataset for different end-to-end architectures.

Model features	Model architecture	Dev avg	Eval avg
BF-FBANK	VGG (E2E)	14.9	11.7
BF-FDLP	VGG (E2E)	12.1	9.7
BF-FBANK	Transformer (E2E)	12.9	10.4
BF-FDLP	Transformer (E2E)	10.4	8.3

realized using differentiable neural operations[1].

3.2.2. E2E ASR framework

We use the ESPnet toolkit [25] to perform all the end-to-end speech recognition experiments, with the Pytorch backend [26]. We experimented with two end to end model architectures, (i) VGG based model and (ii) transformer architecture [27]. The first E2E model architecture uses 3-layer VGG-BLSTM based encoder with 1024 units, and 1-layer of decoder with 1024 units. In the transformer architecture, the encoder used is a 12-layer transformer network with 2048 units in the projection layer. The encoder-decoder attention is used where the decoder network is a 6-layer transformer architecture with 2048 units in the projection layer. During training, a combination of multiple cost functions is used [27] which consists of connectionist temporal cost (CTC) loss and the attention based cross entropy (CE) loss. The CTC-weight is fixed at 0.3 and during decoding the beam-size is fixed at 10. The model is trained for several epochs till the loss function saturates on the validation data, with the patience factor of 2 epochs for REVERB dataset and 3 epochs for VOiCES dataset.

A recurrent neural network based language model (RNN-LM) with 1 layer of 1000 LSTM cells is employed. The stochastic gradient descent (SGD) optimizer with a batch of 32 is used to train the model. The language model is incorporated in the end-to-end system and we have augmented the training data with clean Wall Street Journal (WSJ) data.

3.2.3. Joint learning

The joint learning of the envelope dereverberation module and the E2E ASR architecture is achieved by constructing the single neural model (as shown in Figure 1). Given the deep structure consisting of convolutions, LSTMs and transformer based layers, we initialize the modules with isolated training of each component. Specifically, the envelope dereverberation model is trained using MSE+BEGAN loss and the E2E architecture is separately trained on the acoustic features from the envelope dereverberation model. The final model is jointly optimized using the E2E ASR loss function (combined CTC and CE loss). The audio signal is divided into non overlapping segments of 2 sec. length and passed through the envelope dereverberation model. The feature vectors for the 2 sec audio chunk are passed through the E2E architecture to predict the acoustic model targets.

[1]The implementation of the work can be found in `https://github.com/iiscleap/Joint_FDLP_envelope_dereverberation_E2E_ASR/tree/master`

4. EXPERIMENTS AND RESULTS

For all the models, we use WPE enhancement [8] along with unsupervised generalized eigen value (GEV) beamforming [7]. The baseline features are the filter-bank energy features (denoted as BF-FBANK). Since VOiCES is a single channel dataset, we use only the WPE enhancement. The FBANK features are 36 band log-mel spectrogram with frequency range from 200 Hz to 6500 Hz (similar to the sub-band decomposition in the FDLP feature extraction).

4.1. REVERB Challenge ASR

The REVERB challenge dataset [15] for ASR consists of 8 channel recordings with real and simulated reverberation conditions. The simulated data is comprised of reverberant utterances (from the WSJCAM0 corpus [28]) 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 [29], 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 convolved with 24 measured RIRs.

The first experiments report the performance of the two E2E architectures explored in this paper. We experiment with baseline features (BF-FBANK) and the FDLP features without dereverberation for these experiments. These results are reported in Table 2. The transformer based E2E architecture shows significant improvements over the VGG based model. The rest of the experiments reported in the paper use the transformer architecture for the E2E model training. Further, the BF-FDLP results are observed to perform consistently better than the BF-FBANK baseline (average absolute improvements of 2.5% on the development set and about 2.1% on the evaluation set). These improvements may be attributed to the advantages of autoregressive modeling of sub-band envelopes, where the signal peaks are given more prominence [1].

The next set of experiments report the performance for various envelope dereverberation model architectures. In these experiments, we perform a separate dereverberation and speech recognition E2E

Table 3: WER (%) in REVERB dataset for separate learning of the dereverberation and E2E models as well as the joint learning.

Model Config.	Dev			Eval		
	Real	**Sim**	**Avg**	**Real**	**Sim**	**Avg**
BF-FBANK (baseline)	15.3	10.5	12.9	11.5	9.2	10.4
BF-FDLP	14.1	6.7	10.4	10.1	6.5	8.3
- + derevb. [MSE]	11.4	7.7	9.5	9.5	7.0	8.2
- + derevb. [MSE+BEGAN]	11.3	7.5	9.4	8.7	6.6	7.6
- + joint. [MSE]	10.3	6.3	8.3	**7.1**	**5.6**	**6.3**
- + joint. [MSE+BEGAN]	9.3	6.1	**7.7**	7.7	5.9	6.8

model training. Table 1 shows the results for the different models that are used to perform the envelope dereverberation. The 2 CNN+transformer model employs two CNN layers with 32 filters each with the kernel size of 41×5, and four layers of transformer encoder architecture with 8 attention heads. The 4 CNN+transformer model, uses the same transformer architecture, and the same initial two layers of CNN with the only difference being two additional CNN layers. Here, the 2 additional layers of CNN contain 64 filters each and with kernel size of 21×3. In the model defined as linear+transformer, we use the same transformer configuration, but use a simple linear layer to project the feature matrix, which is passed through the transformer. Further, the architecture with 4 CNN and 2 LSTM layers gave the best performance (similar to the previous findings on hybrid ASR model [10]).

The last two rows of Table 1 highlight the performance with regularized loss function (MSE + BEGAN loss). As seen here, the regularization with 0.1 parameter for BEGAN loss results in the best performance reported in this Table. In particular, on the real evaluation data, the model trained with MSE and BEGAN loss improves over the MSE alone training. We observe a 27% relative improvement in the development and evaluation dataset compared to the baseline BF-FBANK (Table 2).

The results using the joint learning of the dereverberation network and the E2E ASR model are reported in Table 3. The joint training model is initialized using the deverberation model and the E2E model trained separately. In these experiments, we use the deverberation network with 4 CNN layers and 2 LSTM layers, termed as CLSTM network. The proposed joint training of the model yields average absolute improvements of 4.6% and 4.1% on the development set and evaluation set respectively over the baseline system. The improvement in real condition is more than those observed in the simulated data. The joint model, initialized with the dereverberation model trained with BEGAN loss regularization, improved over the one without the BEGAN loss regularization, in the development data. However, this did not show consistent improvement in the evaluation data.

For the joint model initialized with the MSE loss based CLSTM dereverberation network, we observe average relative improvements of 36% on the development set and about 39% on the evaluation set, compared to the BF-FBANK baseline. The joint training is also shown to improve over the set up of having separate networks for dereverberation and E2E ASR. These results show that the joint learning of the two modules and the application of autoregressive modelling of sub-band envelopes can yield considerable benefits for E2E ASR. The comparison of the results from prior works reported on the REVERB challenge dataset is given in Table 4. The table includes results from end-to-end ASR systems [21, 22, 30] as well as the joint enhancement and ASR modeling work reported in [31]. To the best of our knowledge, the results from the proposed algorithm

Table 4: Comparison of the results with other works reported on REVERB challenge dataset.

System	Eval-sim.	Eval-real	Avg.
Subramanian et. al. [21]	6.6	10.6	8.6
Heymann et. al. [31]	-	10.8	-
Fujita et. al. [30]	**4.9**	9.8	7.4
Zhang et. al. [22]	-	10.0	-
This work	5.6	**7.1**	**6.3**

Table 5: Performance (WER %) on the VOiCES dataset.

Model Config.	**Dev**	**Eval**
FBANK	42.9	52.5
FDLP	40.2	49.9
FDLP + CLSTM derevb.	39.2	48.6
+ joint. (prop)	**38.1**	**47.6**

achieves the best average performance on the REVERB challenge evaluation dataset (relative improvements of 15% over the recent work by Fujita et. al. [30]).

4.2. VOiCES ASR

The training set of the VOiCES corpus [16] consists of 80-hour subset of the clean LibriSpeech corpus. The training set has close talking microphone recordings from 427 speakers recorded in clean environments. The development and evaluation set consists of 19 hours and 20 hours of far-field microphone recordings from diverse room dimensions, environment and noise conditions. There are no common speakers between the training set and the development set or the evaluation set. The significant mismatch the training set and development/evaluation set allows the testing of the robustness of the trained models. We have used the same transformer based E2E ASR system that was developed for the REVERB challenge dataset. Further, the current experiments do not perform any data augmentation in the ASR model training.

The WER results for VOiCES corpus is given in Table 5. As seen, the FDLP features show better WER compared to the FBANK features. This is improved with the dereverberation of the FDLP envelopes, achieved using the CLSTM network. The same architecture of the dereverbation network used in the REVERB dataset is also explored for the VOiCES dataset. In this case, the dereverberation network was trained with simulated reverberation.

The best results reported in Table 5 is for the model with joint learning of the dereverberation network and the E2E model. For training the joint E2E model, we initialize the weights of encoder and decoder with weights of CLSTM E2E model, and then train the joint model for 5 epochs. The final WER shows an average relative WER improvement of 11.2% in development set and 9.3% on the evaluation set over the baseline FBANK system.

5. SUMMARY

In this paper, we propose a feature enhancement model for E2E ASR systems using frequency domain linear prediction based subband envelopes. Using the joint learning of the neural dereverberation approach and the E2E ASR model, we perform several speech recognition experiments on the REVERB challenge dataset as well as on the VOiCES dataset. These results shows that the proposed approach improves over the state of art E2E ASR systems based on log mel features. Further, ablation studies show the justification for the choice of the dereverberation network architecture and the choice of loss functions in training the models.

6. REFERENCES

[1] S. Ganapathy, *Signal analysis using autoregressive models of amplitude modulation*, Ph.D. thesis, Johns Hopkins University, 2012.

[2] D. Yu and Li Deng, *Automatic Speech Recognition.*, Springer, 2016.

[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] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE TASLP*, vol. 15, no. 7, pp. 2011–2022, 2007.

[5] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[6] R. Kumar, A. Sreeram, A. Purushothaman, and S. Ganapathy, "Unsupervised neural mask estimator for generalized eigenvalue beamforming based ASR," in *IEEE ICASSP*, 2020, pp. 7494–7498.

[7] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[8] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE TASLP*, vol. 18, no. 7, pp. 1717–1731, 2010.

[9] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters, vol. 25, issue 3, pp. 373-377*, 2017.

[10] Anurenjan Purushothaman, Anirudh Sreeram, Rohit Kumar, and Sriram Ganapathy, "Deep Learning Based Dereverberation of Temporal Envelopes for Robust Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 1688–1692.

[11] Anurenjan Purushothaman, Anirudh Sreeram, Rohit Kumar, and Sriram Ganapathy, "Dereverberation of autoregressive envelopes for far-field speech recognition," *Computer Speech & Language*, vol. 72, pp. 101277, 2022.

[12] 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.

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

[14] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.

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

[16] C. Richey, M. Barrios, et al., "VOiCES obscured in complex environmental settings (voices) corpus," *arXiv preprint arXiv:1804.05053*, 2018.

[17] 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 ICASSP*, 2013, pp. 6822–6826.

[18] A. Maas, T. O'Neil, A. Hannun, and A. 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.

[19] J. Santos and T. 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] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[21] A. Subramanian, X. Wang, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita, "An investigation of end-to-end multichannel speech recognition for reverberant and mismatch conditions," *arXiv preprint arXiv:1904.09049*, 2019.

[22] W. Zhang, A. Subramanian, X. Chang, S. Watanabe, and Y. Qian, "End-to-end far-field speech recognition with unified dereverberation and beamforming," *arXiv preprint arXiv:2005.10479*, 2020.

[23] S. Ganapathy and V. Peddinti, "3-d CNN models for far-field multi-channel speech recognition," in *IEEE ICASSP*, 2018, pp. 5499–5503.

[24] 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.

[25] S. Watanabe, T. Hori, et al., "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[26] A. Paszke, S. Gross, et al., "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[27] S. Karita, N. Chen, et al., "A comparative study on transformer vs RNN in speech applications," in *2019 IEEE ASRU*. IEEE, 2019, pp. 449–456.

[28] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAMO: A British english speech corpus for large vocabulary continuous speech recognition," in *IEEE ICASSP*, 1995, vol. 1, pp. 81–84.

[29] M. Lincoln, I. McCowan, J. Vepa, and H. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," in *IEEE ASRU*, 2005, pp. 357–362.

[30] Y. Fujita, A. Subramanian, M. Omachi, and S. Watanabe, "Attention-based asr with lightweight and dynamic convolutions," in *ICASSP*. IEEE, 2020, pp. 7034–7038.

[31] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Joint optimization of neural network-based wpe dereverberation and acoustic model for robust online ASR," in *ICASSP*. IEEE, 2019, pp. 6655–6659.