INVESTIGATION OF A SINGLE-CHANNEL FREQUENCY-DOMAIN SPEECH ENHANCEMENT NETWORK TO IMPROVE END-TO-END BENGALI AUTOMATIC SPEECH RECOGNITION UNDER UNSEEN NOISY CONDITIONS

Md Mahbub E Noor[†], Yen-Ju Lu[‡], Syu-Siang Wang[‡], Supratip Ghose[§], Chia-Yu Chang[‡], Ryandhimas E. Zezario[‡], Shafique Ahmed[†‡], Wei-Ho Chung[†], Yu Tsao[‡], Hsin-Min Wang[*]*

[*]Taiwan International Graduate Program in Social Network and Human-Centered Computing,
Institute of Information Science, Academia Sinica, Taiwan
E-mail: {mhb.sinica, whm}@iis.sinica.edu.tw
[†]National Tsing Hua University, Taiwan
E-mail: whchung@ee.nthu.edu.tw
[‡]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
E-mail: {neil.lu, sypdbhee, ryandhimas, shafique.khattak13, yu.tsao}@citi.sinica.edu.tw
[§]University of Information Technology & Sciences, Bangladesh
E-mail: supratip.ghose@uits.edu.bd

ABSTRACT

Due to the presence of distortion, most of the single-channel frequency-domain speech enhancement (SE) approaches are still challenging for downstream automatic speech recognition (ASR) tasks, even with satisfactory improvements in enhancing speech quality and intelligibility. Recently, transformer-based models have shown better performance in speech processing tasks. Therefore, we intend to explore a transformer-based SE model, which is fine-tuned through a two-stage training scheme. Pre-training is performed using a feature-level optimization criterion through SE loss, and then a pre-trained end-to-end ASR model is used to fine-tune the SE model using an ASR-oriented optimization criterion through SE and ASR losses. We evaluate the proposed approach on low-resourced Bengali language, which has not received as much attention as resource-rich English or Mandarin languages in both SE and ASR fields. Experimental results show that it can improve the performance of SE and ASR under severe unseen noisy conditions and its performance is reasonably good compared with other state-of-the-art SE methods.

Index Terms**—** automatic speech recognition, Bengali, speech enhancement.

1. INTRODUCTION

Despite the continuous progress on speech processing research for resource-rich languages such as English, Mandarin, and some European languages, related research for Bengali, the world's seventh largest spoken language, is still in its early stages due to various challenges, such as insufficient resources. Over the years, limited work was done to develop Bengali ASR systems [1, 2]. Hence, in this study, we first developed an end-to-end (E2E) ASR backend system, and then we explored our SE approach to estimate the performance of SE and downstream ASR tasks.

There have been much many studies on single-channel and multi-channel SE. As reported in [3, 4], multi-channel SE is helpful to build a robust ASR system, but the research on single-channel SE is still far behind, especially when SE is implemented in the frequency domain. Most single-channel SE approaches tend to introduce distortion, leading to a mismatch with the ASR back-end. Therefore, single-channel SE approaches limit their effectiveness for ASR. To solve this issue, several multi-tasking and joint-training methods of SE and ASR have been proposed [5–7]. In [5], a multi-tasking model was reported to improve the performance of SE and ASR. It has two output layers: one for SE and the other for ASR. However, they were optimized separately by a single multi-tasking model. A deep neural network (DNN)-based joint-training mechanism was proposed in [6]. The acoustic model needs to be retrained with the enhanced features before implementing the joint-training approach. Moreover, the larger the dataset, the higher is the computational cost. In [7], a method based on direct masking and a parametric Wiener filter [8] was used to reduce distortion, and thus is beneficial to ASR.

In this work, we intend to explore a multi-tasking SE model based on supervised mapping. The goal is to obtain enhanced speech with higher quality and intelligibility and good ASR accuracy without the need to retrain the ASR system with enhanced features. In this vein, we apply a transformer-based SE model that works in the frequency domain and is trained by a two-stage training scheme. The SE model is first trained through a feature level optimization criterion with the SE loss (L1 loss), and then fine-tuned through an ASR-oriented optimization criterion with the SE and ASR losses using a pre-trained E2E speech recognition model. Therefore, in this work, 1) we first built an E2E Bengali ASR system on a relatively large dataset, and then, 2) we implemented a new variant of transformer-based SE model on a smaller dataset, which aims to minimize both SE and ASR losses. The remainder of this paper is organized as follows. Section 2 presents the proposed approach. Section 3 presents the experimental setup, results and discussion. Finally, Section 4 provides the concluding remarks of this study.

2. THE PROPOSED APPROACH

2.1. ASR model

Recently, E2E ASR [9-13] systems have received considerable attention due to the elimination of the necessity of training several disjoint components (e.g., the acoustic model, pronunciation model, and language model) of the traditional style ASR, making it more suitable for the ASR systems for low-resourced languages. In [14], it is reported that the DNN-based complex modular ASR system is difficult to guide a multii-tasking SE model optimization.

Our E2E ASR system was developed based on the hybrid CTC/Attention model [11] of the Espnet toolkit [15]. A multi-objective learning criterion that combines the CTC loss $loss_{CTC}$ and the attention-based cross entropy loss $loss_{ATT}$ was used to train the model parameters:

$$Loss_{Total} = (1 - \alpha) \times loss_{ATT} + \alpha \times loss_{CTC}, \quad (1)$$

where α is a tunable parameter. It was set to 0.3 in this study. The CTC and Attention models share the same BLSTM encoder and LSTM decoder.

Let $P_{CTC}(c_t)$ and $P_{ATT}(c_t)$ be the probabilities of the output label c_t at position t for the CTC and Attention models, respectively, the combined posterior score is:

$$log(P(c_t)) = (1 - \beta) \times log(P_{ATT}(c_t)) + \beta \times log(P_{CTC}(c_t)). \quad (2)$$

In this study, β was set to 0.3. All the hyper-parameters were set following [12] except that the BLSTM encoder was formed by layers=6, layer size=320, projection layer size=320, and the LSTM decoder was formed by layers=1 and layer size =300. Besides, we set batch size=24, CTC weight=0.3, beam size=20, maximum epoch=20, patience=3, and optimizer=AdaDelta for training the ASR model. 83-dimensional fbank-pitch features extracted by the Kaldi toolkit [16] built on Espnet were used as the acoustic features. In addition, we applied the recently well-known data augmentation technique SpecAugment [17], which makes any ASR system more robust and eliminates the overfitting problem of the E2E model. For SpecAugment, we used the default parameter settings in Espnet.

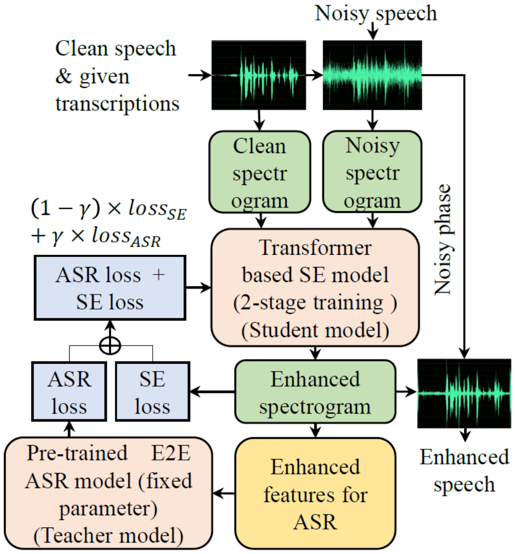
Noisy speech
Clean speech & given transcriptions
Clean spectr ogram
Noisy spectr ogram
$(1-\gamma) \times loss_{SE}$ $+ \gamma \times loss_{ASR}$
ASR loss + SE loss
Transformer based SE model (2-stage training) (Student model)
Noisy phase
ASR loss
SE loss
Enhanced spectrogram
Pre-trained E2E ASR model (fixed parameter) (Teacher model)
Enhanced features for ASR
Enhanced speech

Fig. 1 The SE model used in this study.

Moreover, we used an RNN-based language model [18], which was fused with our E2E acoustic model to generate the final recognition output. The language model was formed by an LSTM with a single hidden layer and 1000 hidden units. To train this language model, the most frequently used 53,000 Bengali words from the transcriptions of the original clean training set were selected to form the vocabulary set. The stochastic gradient descent (SGD) algorithm was used to train the language model, with the following settings: maximum epoch=20, batch size=300, and patience=3.

2.2. SE model

We adopted the transformer-based SE model in [19, 20] to implement our SE system. Unlike previous work, the objective function used to train our SE model consisted of two parts, and the training process was divided into two stages. As shown in Fig. 1, our SE system was trained with the SE loss $loss_{SE}$ and the ASR loss $loss_{ASR}$

$$Loss_{Total} = (1 - \gamma) \times loss_{SE} + \gamma \times loss_{ASR}, \quad (3)$$

where γ is a tunable parameter. Note that if γ is set to 0, the SE model is trained in the conventional manner with only the SE loss.

Transformer was originally proposed for machine translation [21], and some research has also been conducted in the SE field, such as [19, 20, 22, 23]. For sequence-to-sequence learning, the transformer comprises an encoder and a decoder. When applied to SE, since the input sequence (noisy speech) and the output sequence (enhanced speech) have the same length, decoder learning is omitted. The transformer used in this study consists of four convolutional layers for encoding the input spectrogram with its location information and eight attention blocks, each consisting of eight heads (64 neurons for each head) and two fully connected feedforward layers. Both sublayers contain residual connection and layer normalization [24]. Leaky ReLU was used as the activation function. To reduce the size of training parameters, we trained our SE system based on a teacher-student scheme [25], where the ASR model acts as a larger teacher model, and the SE model as a smaller student model.

2.2.1 Two-stage training

In this study, all SE and ASR models were trained on a single 12GB GPU in a Linux environment. As our SE model training is completed through a two-stage scheme, we call the first stage SE-part and the second stage as ASR-part. In SE-part, the SE loss was optimized at the feature level for pre-training. After that, in ASR-part, ASR-oriented optimization was conducted by the weighted sum of the SE loss and the ASR loss given by the pre-trained E2E ASR model. The pre-trained E2E ASR model was trained using clean training data with transcripts in a fully E2E manner. It should be noted that during SE-part, the ASR loss was ignored by setting $\gamma = 0.0$, i.e., it had no effect on model optimization. Even so, we found that the ASR loss was reduced to a certain extent. In ASR-part, the ASR loss began to affect the total loss through a certain weight factor, which can balance between the ASR and SE losses. During training, we fixed the parameters of the ASR model and only updated the parameters of the SE model, because our goal is to get better enhanced data, aiming to obtain better SE and ASR evaluation performance. This approach is different from that in [5], which updated the parameters based on two losses of two output layers. On the contrary, we only focused on the SE model, thus fixing the parameters of the ASR model. We performed a total of 150 epochs in the training, including 70 epochs in SE-part and 80 epochs in ASR-part. The batch size was set to 2. In our experiments, we found that $\gamma = 0.000009$ in ASR-part could strike a good balance between the SE and ASR losses to yield better performance. Generally, the ASR model and the SE model are in completely different domains, and different types of features are used in the training process. It has been observed in our experiments that the ASR loss has a much higher impact on the total loss function than the SE loss. For this reason, we used such a low weight value for the ASR loss. However, this setting may need to be adjusted with the dataset, because different datasets have different ASR performance.

In our experiments, we also evaluated single-stage training. The SE model was trained by 150 epochs using both SE and ASR losses with $\gamma = 0.000009$ from the beginning.

Table 1: The division of the dataset in this paper.

Split	No. of utterances	No. of speakers
Train	213,496	495
Development	3,000	8
Test	2,206	5(3 male, 2 female)
Total	218,702	508

3. EXPERIMENTS

3.1. Experimental setup

We evaluated the proposed approach on the Bengali set in the OpenSlr corpus [26]. Since the dataset does not have a standard train/development/test split, we divided it by ourselves. The details are shown in Table 1. We picked up the first 3,000 utterances after sorting by the Espnet toolkit, including 8 speakers (6 males and 2 females), to form the development set, 2,206 utterances from 5 speakers (3 males and 2 females) to form the test set, and the rest were used as the training data. We carried out such a split to reasonably match the previous state-of-the-art work on this dataset [2]. However, in [2], the identities of the test speakers were not revealed, so we randomly selected 5 speakers to obtain a similar number of test utterances.

To train our SE models, we randomly selected 10,000 utterances (5 to 15 seconds in length) from the training data of the ASR model. This subset covers 404 speakers. Then, we synthesized the corresponding 10,000 noisy utterances by artificially adding noise to the clean training utterances. For the training data, 100 types of environmental noises in [27] were used, and 14 signal-to-noise ratio (SNR) levels ranging from –6 to 20 dB, with an interval of 2dB, were applied. For the evaluation data, the 2206 noisy test utterances were synthesized from the corresponding same clean utterances used for ASR evaluation. Four unseen types of noise were used, namely car, siren, street, and cafeteria babble, from the other resource [28] at low SNR levels (–7 to 5dB), with an interval of 2dB. The speakers, noise types and SNR levels between the training and test sets do not overlap. It should be noted that the original speech utterances in the corpus are not completely clean, which makes our mapping-based SE systems more challenging.

To implement our SE model, first, the entire original clean training set was used to pre-train our Teacher ASR model. Then, the Student SE model was trained though the two-stage training process described in Section 2.2.1. Here, short-time Fourier transform (STFT) was performed with a Hamming window size of 25ms and a hop size of 10ms to extract the spectral features.

For comparison, we also implemented several conventional SE models, such as Karhunen-Loeve transform (KLT) [29], deep denoising auto encoder (DDAE) [30], Wave-U-Net [31], and fully convolutional neural network (FCNN). KLT is a filtering-based traditional SE method that does not need any model training, while the latter three are

DNN-based models. In our preliminary experiments, we observed that less contextual frames did not work well for our FCNN model, because Bengali is spoken differently from English, and the clean data in this dataset are not purely clean. After we used wider contextual information from the neighboring frames in the input features and deeper hidden layers, a considerable improvement was obtained. We used the same number of contextual frames in DDAE. Five contextual frames were used in the original DDAE, so the size of the input features was 257*(2*2+1) =1,285. In this study, nine contextual frames were used, resulting in the 257*(4*2+1)=2,313-dimensional features. The FCNN model consists of 15 convolutional layers. After each layer, batch normalization was used to avoid the overfitting problem during training. Each convolutional layer contained five channels {16, 32, 64, 128, 256}. Each channel used three types of strides {1, 1, 3}. The source codes are available[1]. Wave-U-Net is a time-domain SE model, which was first proposed in image processing and later applied to speech enhancement. For DDAE and FCNN, we performed 200 epochs, and for Wave-U-Net, we performed 1000 epochs.

We also implemented a more robust ASR model by combining SpecAugment augmentation. In this work, we built several E2E ASR systems to evaluate the performance of our SE model in the ASR domain. The ASR models were trained using only clean training data or using clean training data and augmented data by SpecAugment. In addition, the ASR systems were evaluated with language model rescoring or without any language model.

3.2. Experimental results

We used four standardized objective metrics to evaluate the SE performance, including perceptual evaluation of speech quality (PESQ) [32], short-time objective intelligibility (STOI) [33], speech distortion index (SDI) [34], and segmental signal-to-noise ratio improvement (SSNRI) [35]. PESQ, with a score ranging from -0.5 to 4.5, is used to evaluate the quality of processed speech. STOI, with a score ranging from 0 to 1, is used to evaluate the intelligibility of processed speech. The higher the PESQ, STOI, and SSNRI scores, the better is the SE performance. On the contrary, the lower the SDI score, the less distortion occurs in the enhanced speech. For ASR performance evaluation, character error rate (CER) and word error rate (WER) were used.

We first intend to investigate the influence of the weight of the ASR loss in Eq. 3 on the performance of our proposed SE model ($\mathrm{SE}^{(\mathrm{T\_SE+ASR})}$). The results are shown in Table 2. The results show that γ=0.000009 achieves the best performance. Using higher weights, the training process seems to be overfitting, resulting in a drop in SE performance. When the weight is increased to 0.001, although the SSNRI score is good, the quality of the enhanced speech deteriorates to the point that the evaluation tool cannot measure the PESQ value. From Table 2, we can see that our proposed SE model achieves better performance than the SE model trained with only SE loss (cf. 0.0 (using only SE loss)) and the SE model trained by the single-stage training scheme (cf. 0.000009 (single-stage training)).

Table 2: SE performance (PESQ, STOI, SDI and SSNRI) with respect to different weights of the ASR loss in Eq.3.

γ	**PESQ↑**	**STOI↑**	**SDI↓**	**SSNRI↑**
0.001	NaN	0.217	1.392	**14.919**
0.0001	2.379	0.601	0.429	12.361
0.00001	2.395	0.614	0.414	12.382
0.000009	**2.412**	**0.618**	**0.375**	12.677
0.000009 (single-stage training)	2.386	0.613	0.426	12.292
0.000001	2.305	0.605	0.421	12.210
0.0 (using only SE loss)	2.367	0.612	0.415	12.486

Table 3: SE performance (PESQ, STOI, SDI and SSNRI) of different SE models.

Test set	**PESQ↑**	**STOI↑**	**SDI↓**	**SSNRI↑**
Noisy	2.005	0.545	0.711	0.000
$\mathrm{SE}^{(\mathrm{KLT})}$	1.649	0.542	0.391	**14.741**
$\mathrm{SE}^{(\mathrm{DDAE})}$	2.150	0.549	0.461	13.344
$\mathrm{SE}^{(\mathrm{Wave\text{-}U\text{-}net})}$	2.233	0.581	0.430	12.580
$\mathrm{SE}^{(\mathrm{FCNN})}$	2.293	0.614	0.474	13.989
$\mathrm{SE}^{(\mathrm{T\_SE+ASR})}$	**2.412**	**0.618**	**0.375**	12.677

Table 3 shows the performance of different SE models. Obviously, our proposed $\mathrm{SE}^{(\mathrm{T\_SE+ASR})}$ model outperforms all other SE models in terms of PESQ, STOI, and SDI, although its SSNRI score is worse than that of some models. Interestingly, the conventional KLT model achieves the best SSNRI score, although it performs worse than all other neural network-based models in the PESQ and STOI metrics. In Fig. 2, the spectrograms of different versions of a sample speech are depicted. It can be seen from the figure that the KLT and Wave-U-Net SE models can remove a reasonable amount of original background noise, which is prominent in the clean spectrogram. However, they failed to remove the artificially additive noise. This is why their performance is not good in Table 3. Among all SE models, it is clear that the spectrogram of the enhanced speech by our model $\mathrm{SE}^{(\mathrm{T\_SE+ASR})}$ is the closest to that of the clean speech. It is worth noting that the clean utterance is not completely clean, because this is a crowdsourced data, which makes this dataset far more challenging for mapping-based speech enhancement. However, our SE model can still achieve satisfactory improvements in SE and ASR performance, which is a remarkable contribution of this work. In the future, we will study SE techniques that do not require any clean ground truth, because it is difficult to collect completely clean speech data, especially for low-resourced languages.

[1] https://github.com/mahbubnoor/E2E

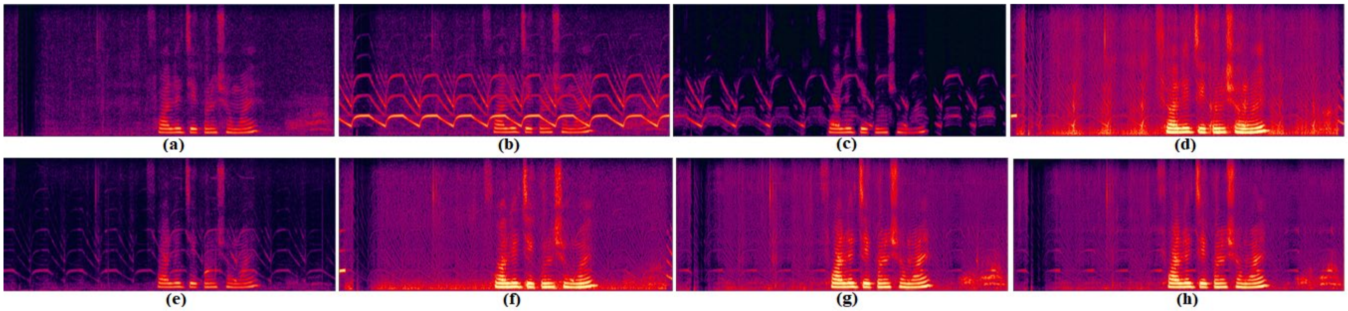

Fig. 2 Spectrograms of different versions of a sample utterance in the synthetic test set: (a) clean speech, (b) noisy speech (fast siren wail noise at -5dB), (c) enhanced speech by KLT, (d) enhanced speech by DDAE, (e) enhanced speech by Wave-U-Net, (f) enhanced speech by FCNN, (g) enhanced speech by transformer-based model trained with only SE loss, (h) enhanced speech by our model $SE^{(T\_SE+ASR)}$.

Table 4: ASR results of different test scenarios (e.g., clean, noisy, enhanced by different SE models).

Test utterance condition	**ASR without any language model**				**ASR with a language model**			
	Clean		**Clean+SpecAug**		**Clean**		**Clean+SpecAug**	
	CER	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**	**WER**
Clean	10.7	32.9	10.4	31.5	**5.6**	**14.6**	7.2	16.5
Noisy	48.7	77.3	41.5	68.3	44.8	60.6	42.1	57.8
$SE^{(KLT)}$	64.2	92.0	52.6	83.8	55.0	73.1	50.8	69.2
$SE^{(DDAE)}$	51.9	86.5	45.7	76.5	46.6	67.3	45.3	63.9
$SE^{(Wave\text{-}U\text{-}net)}$	99.0	100.0	94.1	99.7	96.5	99.6	94.4	99.9
$SE^{(FCNN)}$	41.3	73.1	36.2	64.2	36.2	53.1	35.0	50.5
$SE^{(T\_SE)}$	40.3	71.9	35.2	63.3	35.2	52.0	33.5	48.4
$SE^{(T\_SE+ASR)}$ Single-stage	41.0	72.3	34.5	62.6	35.6	51.9	33.1	48.7
$SE^{(T\_SE+ASR)}$ Two-stage	**39.0**	**70.3**	**33.7**	**61.4**	**33.5**	**50.0**	**32.2**	**47.5**

Table 4 shows the downstream ASR results of different ASR models (ASR model with/without a language model trained with only clean training data or clean training data plus augmented data by SpecAugment) evaluated under different test utterance conditions (such as clean, noisy, and enhanced speech by different SE models). The best CER of 5.6% and WER of 14.6% on the clean test speech can be obtained by the ASR model that is trained with only clean training data and equipped with language model rescoring. It is worth noting that that the CER and WER of the noisy test speech are very high, even when the ASR model was trained with clean training data and augmented data by SpecAugment. Because the original test speech is not completely clean, added noise to it at low SNR levels makes the test set even more challenging. Among all SE models, we can see that the enhanced speech by $SE^{(FCNN)}$, $SE^{(T\_SE)}$ (only SE loss), $SE^{(T\_SE+ASR)}$ (single-stage), and $SE^{(T\_SE+ASR)}$ (two-stage) always yields better ASR performances than the unprocessed noisy speech. On the other hand, $SE^{(KLT)}$ and $SE^{(Wave\text{-}U\text{-}net)}$ show worse ASR performances. It was obvious because they removed the original background noises which were prominent in clean data and thus feature level mismatch occurs with the ASR models. Obviously, the proposed $SE^{(T\_SE+ASR)}$ (two-stage) model outperforms all other models. The CER was reduced from 42.1% (noisy) to 32.2%, and the WER was reduced from 57.8% (noisy) to 47.5%.

4. CONCLUSION

In this paper, we have confirmed that a transformer-based SE model trained through a two-stage training scheme can improve both SE and downstream ASR performance. The contribution of this paper is four-fold: First, we confirmed the effectiveness of the proposed SE model in speech enhancement evaluation metrics and downstream ASR evaluation metrics. Second, we showed that the two-stage training scheme for the SE model is more effective than the single-stage training method. Third, we showed that ASR performance can be improved by a single-channel frequency-domain SE system without the need to retrain the ASR system with the enhanced features by the SE system. Fourth, we benchmarked new ASR performance on the clean and noisy test speech in the Bengali set of the OpenSlr corpus. In our future work, we will explore multi-condition training to include different types of training data, such as noisy speech with various noises and enhanced speech by different representative SE models, to improve the robustness of the ASR system. In addition, we will also apply our approach to other languages.

5. REFERENCES

[1] B. Das, S. Mandal, and P. Mitra, “Bengali speech corpus for continuous automatic speech recognition system,” in *Proc. Oriental COCOSDA,* 2011, pp. 51-55.

[2] N. Sadeq, N. T. Chowdhury, F. T. Utshaw, S. Ahmed, and M. A. Adnan, “Improving End-to-End Bangla Speech Recognition with Semi-supervised Training,” in *Proc. EMNLP,* 2020, pp. 1875-1883.

[3] M. Fujimoto, and H. Kawai, “One-Pass Single-Channel Noisy Speech Recognition Using a Combination of Noisy and Enhanced Features,” in *Proc. INTERSPEECH*, 2019, pp. 486-490.

[4] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, “Improving noise robust automatic speech recognition with single-channel time-domain enhancement network,” in *Proc. ICASSP,* 2020, pp. 7009-7013.

[5] Z. Chen, S. Watanabe, H. Erdogan and J. R. Hershey, “Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks,” in *Proc. INTERSPEECH,* 2015, pp. 3274-3278.

[6] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, “Joint training of frontend and back-end deep neural networks for robust speech recognition,” in *Proc. ICASSP*, 2015, pp. 4375-4379.

[7] T. Menne, R. Schlüter, and H. Ney, “Investigation into joint optimization of single channel speech enhancement and acoustic modeling for robust ASR,” in *Proc. ICASSP,* 2019, pp. 6660-6664.

[8] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer handbook of speech processing*, SpringerVerlag Berlin Heidelberg, 2008.

[9] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in *Proc. NIPS*, 2015, pp. 577-585.

[10] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *Proc. ICASSP,* 2016, pp. 4960-4964.

[11] S. Watanabe, T. Hori, S. Kim, J. R., Hershey and T. Hayashi, “Hybrid CTC/attention architecture for end-to-end speech recognition,” *IEEE Journal of Selected Topics in Signal Processing,* vol. 11, no. 8, pp. 1240-1253, 2017.

[12] A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. ICASSP,* 2013, pp. 6645-6649.

[13] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang et al., “A comparative study on transformer vs RNN in speech applications,” in *Proc. ASRU*, 2019, pp. 449-456.

[14] L. Chai, J. Du, Q.-F. Liu, and C.-H. Lee, “A Cross-Entropy-Guided Measure (CEGM) for Assessing Speech Recognition Performance and Optimizing DNN-Based Speech Enhancement,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 106-117, 2020.

[15] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno et al. “Espnet: End-to-end speech processing toolkit,” in *Proc. INTERSPEECH*, 2018, pp. 2207-2211.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel et al. “The Kaldi speech recognition toolkit,” in *Proc. ASRU*, 2011.

[17] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk *et al.*, “SpecAugment: A simple data augmentation method for automatic speech recognition,” in *Proc. INTERSPEECH*, 2019, pp. 2613-2617.

[18] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, “Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM.” in *Proc. INTERSPEECH*, 2017, pp. 949–953.

[19] Y.-J. Lu, C.-F. Liao, X. Lu, J.-W. Hung, and Y. Tsao, “Incorporating broad phonetic information for speech enhancement,” in *Proc. INTERSPEECH*, 2020, pp. 2417-2421.

[20] S.-W. Fu, C.-F. Liao, T.-A. Hsieh, K.-H. Hung, S.-S. Wang, C. Yu et al. “Boosting Objective Scores of a Speech Enhancement Model by MetricGAN Post-processing,” in *Proc. APSIPA ASC*, 2020, pp. 455-459.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez *et al.*, “Attention is all you need,” in *Proc. NeurIPS*, 2017, pp. 1-11.

[22] X. Hao, C. Shan, Y. Xu, S. Sun, and L. Xie, “An attention-based neural network approach for single channel speech enhancement,” in *Proc. ICASSP,* 2019, pp. 6895-6899.

[23] Y. Koizumi, K. Yaiabe, M. Delcroix, Y. Maxuxama, and D. Takeuchi, “Speech enhancement using self-adaptation and multi-head selfattention,” in *Proc. ICASSP*, 2020, pp. 181-185.

[24] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.

[25] S. Abbasi, M. Hajabdollahi, N. Karimi, and S. Samavi, “Modeling teacher-student techniques in deep neural networks for knowledge distillation,” in *Proc. MVIP*, 2020, pp. 1-6.

[26] O. Kjartansson, S. Sarin, K. Pipatsrisawat, M. Jansche and L. Ha, “Crowd-sourced speech corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali,” in *Proc. SLTU*, 2018, pp. 52–55.

[27] G. Hu and D. Wang, “A tandem algorithm for pitch estimation and voiced speech segregation,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[28] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[29] A. Rezayee, and S. Gazor. “An adaptive KLT approach for speech enhancement,” *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 2, pp. 87–95, 2001.

[30] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Proc. INTERSPEECH*, 2013, pp. 436–440.

[31] C. Macartney, and T. Weyde, “Improved speech enhancement with the wave-u-net,” *arXiv preprint arXiv:1811.11307*, 2018.

[32] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *Proc. ICASSP,* 2001, pp. 749–752.

[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[34] J. Chen, J. Benesty, Y. Huang, and S. Doclo, “New insights into the noise reduction Wiener filter,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 14, pp. 1218–1234, 2006.

[35] J. Chen, *Fundamentals of Noise Reduction in Spring Handbook of Speech Processing*, Chapter 43, Springer, 2008.