Time-Domain Speech Enhancement for Robust Automatic Speech Recognition

Yufeng Yang[1], **Ashutosh Pandey**[1], **DeLiang Wang**[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA

[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

Abstract

It has been shown that the intelligibility of noisy speech can be improved by speech enhancement algorithms. However, speech enhancement has not been established as an effective frontend for robust automatic speech recognition (ASR) in noisy conditions compared to an ASR model trained on noisy speech directly. The divide between speech enhancement and ASR impedes the progress of robust ASR systems especially as speech enhancement has made big strides in recent years. In this work, we focus on eliminating this divide with an ARN (attentive recurrent network) based time-domain enhancement model. The proposed system fully decouples speech enhancement and an acoustic model trained only on clean speech. Results on the CHiME-2 corpus show that ARN enhanced speech translates to improved ASR results. The proposed system achieves 6.28% average word error rate, outperforming the previous best by 19.3% relatively.

Index Terms:

CHiME-2; robust ASR; speech distortion; time-domain speech enhancement

1. Introduction

In real environments, acoustic interference is ubiquitous in speech communication, and negatively impacts the performance of speech-based applications such as smart home devices [1] and conference transcription systems [2]. To attenuate background noise, speech enhancement algorithms have been developed to estimate clean speech from noisy speech. These algorithms have achieved remarkable success in improving the quality and intelligibility of noisy speech [3]. However, a major disappointment is that monaurally enhanced speech does not translate to improved automatic speech recognition (ASR), and this has been attributed to the distortion introduced by monaural speech enhancement algorithms [4]. The divide between speech enhancement and ASR has persisted despite considerable research over decades to bridge enhancement and ASR [5–7]. This study represents a new effort to bridge the fields of speech enhancement and ASR.

yang.5662@osu.edu .

For monaural speech enhancement, we employ the recently proposed attentive recurrent network (ARN), which performs speech enhancement in the time domain [8]. Time-domain enhancement operates differently from spectral methods [3, 9]. It directly predicts clean speech samples from noisy speech samples, where speech magnitude and phase are enhanced simultaneously [10]. ARN incorporates a recurrent neural network (RNN) and a self-attention mechanism, and produces excellent enhancement performance [8]. We employ ARN as the frontend of ASR.

In terms of robust ASR, prevailing approaches perform acoustic modeling directly on noisy speech for noise-dependent or noise-independent models, which are proven to be effective on CHiME-2 [4], CHiME-4 [11], and CHiME-6 [12] corpora. The drawback of such approaches is that noise-dependent models do not generalize well to untrained noises and noise-independent models need an enormous amount of noisy speech for training, which is not only costly but also infeasible in many real applications. An ASR model trained on noisy speech also results in an unavoidable performance gap compared with a corresponding model trained on clean speech, when tested on clean speech. To bridge the divide between speech enhancement and ASR, attempts have been made to perform acoustic modeling on enhanced speech [4, 13] or enhanced features [14] in a distortion-independent way. Nonetheless, backends of these systems are fully dependent on the frontend. When the frontend is replaced, say by a better enhancement model, the backend has to be retrained or ASR performance degrades. Likewise, speech enhancement can also be designed to serve ASR. In [15], a perceptual loss based model was proposed to guide the training of a speech enhancement frontend using senone labels from an acoustic model. To completely combine speech enhancement and ASR, a speech enhancement model and an acoustic model can be jointly trained [16, 17]. Similar ideas are also utilized in end-to-end (E2E) systems [18, 19]. Such E2E systems often have a huge model size and are difficult to train. Furthermore, the frontend and backend inside a joint or E2E system are dependent on each other, which makes it problematic to improve either the frontend or the backend individually.

In this paper, we investigate the most straightforward approach to robust ASR where the acoustic model is based on only clean speech using a Conformer-based acoustic model [11] and its input is ARN enhanced speech directly. Thus the proposed robust ASR system completely decouples the frontend and backend. In other words, the frontend is designed for speech enhancement only, and the backend acoustic model for recognizing clean speech only. Combining two modules directly to recognize noisy speech, the proposed system is demonstrated to outperform other robust ASR systems, including noise-independent, noise-dependent, and distortion-independent models. When tested on the medium vocabulary track (track 2) of the CHiME-2 corpus, our best system achieves 6.28% average word error rate (WER). To our knowledge, this WER result represents the best on this dataset to date and outperforms the previous best by 19.3% relatively. Our investigation also shows that using short-time objective intelligibility (STOI) [20] as the model selection criterion is superior for speech enhancement models in terms of ASR. ARN enhanced speech also shows consistent improvements on other baseline acoustic models, which further demonstrates that ARN eliminates the divide between speech enhancement and ASR.

The remainder of the paper is organized as follows. Section 2 describes ARN for time-domain enhancement and a Conformer-based acoustic model. Section 3 describes the experimental setup and implementation details. Evaluation results and comparisons are presented in Section 4, and Section 5 concludes the paper.

2. System Description

2.1. Problem Formulation

The monaural speech enhancement problem is formulated as follows

$$\mathbf{y} = \mathbf{s} + \mathbf{n},$$

(1)

where $\mathbf{y}$, $\mathbf{s}$, and $\mathbf{n}$ are noisy speech, clean speech, and additive noise, respectively. Speech enhancement computes an estimate of $\mathbf{s}$, $\hat{\mathbf{s}}$, from $\mathbf{y}$.

An ASR system computes the optimal word sequence $\mathbf{W}*$ given a sequence of acoustic features $\mathbf{X}$ of speech signal $\mathbf{x}$, which is formulated as a maximum *a posteriori* probability problem

$$\mathbf{W}* = \arg\max_{\mathbf{W}} \ P_{\mathscr{AM},\mathscr{LM}}(\mathbf{W} \mid \mathbf{X}),$$

(2)

where $\mathscr{AM}$ and $\mathscr{LM}$ denote an acoustic model (AM) and language model (LM), respectively. Using Bayes' theorem, Eq. 2 can be rewritten as

$$\mathbf{W}* = \arg\max_{\mathbf{W}} \ p_{\mathscr{AM}}(\mathbf{X} \mid \mathbf{W}) P_{\mathscr{LM}}(\mathbf{W}),$$

(3)

where $p_{\mathscr{AM}}$ and $P_{\mathscr{LM}}$ are AM likelihood and LM prior probability, respectively. An AM predicts the likelihood of acoustic features of a phoneme or another linguistic unit, and an LM provides a probability distribution over words or sequences of words in a speech corpus.

2.2. Attentive Recurrent Network

We employ ARN as the frontend of ASR, which is a time-domain speech enhancement model comprising RNN, self-attention, feedforward network, and layer normalization modules. Details of ARN building blocks can be found in [8]. In this work, the non-causal version of ARN is used, namely the RNN in ARN is bi-directional long short-term memory (BLSTM) and self-attention is unmasked.

The diagram of ARN is shown in Fig. 1. After an input signal is chunked into overlapping frames, all frames are projected into a latent representation of size N by a linear layer. Then the representations are processed by four consecutive ARN blocks. Finally, another linear

layer projects the output of the last ARN block back to size *L*. The enhanced speech is finally computed using the overlap-and-add (OLA) method.

2.3. Conformer-based Acoustic Model

We utilize the Conformer-based AM [11] as the backend in the proposed system. It is built upon a wide residual BLSTM network (WRBN) (see [4]). The Conformer-based AM is shown to outperform WRBN on the CHiME-4 single-channel track [11]. The system architecture of the Conformer-based AM is shown in Fig. 2, where FFN denotes a feedforward network.

The Conformer-based AM takes as input 80-dimensional mean-normalized log-Mel filterbank features extracted from the ARN output (or other forms of output), coupled with its delta and delta-delta features. First, the input is processed by a wide residual convolutional layer denoted as WRCNN, which passes the input signal through a convolution layer and uses three residual blocks to extract representations at different frequency resolutions [21]. Afterwards, an utterance-wise batch normalization and a linear layer with exponential linear unit (ELU) non-linearity are utilized to project the signal onto 320 dimensions. Then a linear layer projects the signal onto the dimension of multi-head self-attention. After two blocks of the Conformer encoder with absolute positional encoding, the signal is projected onto 1024 dimensions, followed by a ReLU (rectified linear unit) activation and dropout. Finally, a linear layer projects the signal to the final output of each frame as the posterior probability for 1965 context-dependent senone states. Then the output is sent to a decoder for final text transcripts.

3. Experimental Setup

3.1. Dataset

Our experiments are conducted on a medium vocabulary track (track 2) of the CHiME-2 corpus [22]. It is a commonly used dataset for robust ASR evaluation, which is generated by convolving Wall Street Journal (WSJ) clean speech with binaural room impulse responses (BRIRs) and mixing with non-stationary family home noise [23]. Although the speech materials are from the WSJ corpus, due to the BRIRs during the data generation process [24], the speech alignments are altered, which makes it infeasible to use WSJ anechoic-clean speech as the training target for speech enhancement. Therefore, we treat reverberant-clean speech as the training target for ARN. Two channels are averaged to produce single-channel speech.

Training data for ARN is generated by randomly mixing 7138 reverberant-clean training utterances ('reverberated') from CHiME-2 with noise segments randomly picked from 10k non-speech sounds from a sound effect library (http://www.sound-ideas.com) under signal-to-noise ratio (SNR) uniformly distributed in [−7, 0] dB and [0, 10] dB ranges, with 50% probability for each range to be selected. Validation data is generated by mixing 409 reverberant-clean validation utterances ('reverberated') from CHiME-2 with factory noise from the NOISEX-92 dataset [25] under −6 dB SNR. Once trained, ARN is tested on

reverberant-noisy test data ('isolated'), which has six SNR levels with each containing 330 utterances.

The default backend in the proposed system is trained and validated on 7138 and 409 reverberant-clean utterances ('reverberated') from CHiME-2, respectively. We also train a noise-dependent backend (i.e. dependent on the CHiME-2 noise) which is trained and validated on 7138 and 409 reverberant-noisy utterances ('isolated'), respectively. Training data for the noise-independent backend is generated by randomly mixing 7138 reverberant-clean training utterances ('reverberated') with randomly picked 10k noise under the same SNR range as for ARN training. Training data for the distortion-independent backend is generated by enhancing the training data for the noise-independent backend using ARN. The noise-independent and distortion-independent backends are trained on 157036 utterances in the same way as [4].

3.2. Implementation Details

The sampling rate for all utterances is 16 kHz. All training samples are generated randomly and dynamically for ARN. We apply root mean square normalization to noisy mixtures, and clean speech is scaled to produce a specified SNR level. During training, the number of samples for each utterance is set to 64000. Input and output frame size is set to 16 ms with a 2 ms frame shift. Dimension N for BLSTM is set to 1024. The dropout rate is set to 0.05 in feedforward blocks. ARN is trained using the PCM (phase-constrained magnitude) loss [26], which computes loss for both enhanced speech and noise in the time-frequency domain. Training over 157036 utterances is considered as one epoch, and ARN is trained for 100 epochs with batch size 16. The Adam optimizer [27] is utilized. The learning rate of the first 33 epochs is fixed to $2e^{-4}$ and then exponentially decays every epoch till the final learning rate of $2e^{-5}$. Because STOI is shown to relate to WER [28], thus we also use validation STOI as a model selection criterion in addition to validation PCM loss.

For the backend, the configuration of WRCNN is kept the same as in [11]. The attention dimension is set to 256. The kernel size of the 1-D depthwise convolution is set to 16. We use a learning rate schedule from [29], with 2k warm-up steps and a learning rate factor k of 100. The Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1e^{-9}$ is utilized for model training. The batch size is set to 3 and short utterances are padded with zeros to match the length of the longest utterance in each batch. The dropout rate is set to 0.15 for network and attention weights. All backends are trained for 25 epochs and model selection is based on the cross-entropy loss on the validation set. Training labels are 1965 senones in our experiments, and they are generated as in [16]. For log-Mel feature extraction, the approach in [4] is applied, and pre-emphasizing, dithering, and direct current offset removal steps are skipped. A Hamming window is applied to the input waveform for STFT. Then a small value of e^{-40} is added to prevent the underflow of logarithmic operation. For the noise-dependent, noise-independent, and distortion-independent backends, factor k and the number of warm-up steps pairs are empirically set to $\{1e^4, 5k\}$, $\{5e^3, 80k\}$, and $\{5e^3, 80k\}$ respectively for better convergence and validation cross-entropy. Two distortion-independent backends are trained on the enhanced speech by ARN with minimum PCM validation

and maximum STOI validation, denoted as distortion-independent PCM backend and STOI backend, respectively.

The decoder used in this work is the same as in [16]. AM outputs are first substracted by the log priors, and then fed to the decoder, which is based on the CMU pronunciation dictionary and the official 5k close-vocabulary tri-gram language model. The decoding beamwidth is set to 13, and lattice beamwidth is 8. The number of active tokens ranges from 200 to 700. Language model weights ranging from 4 to 25 are utilized.

4. Evaluation and Comparison Results

4.1. Results on Speech Enhancement

Table 1 presents the speech enhancement results of ARN on the CHiME-2 corpus in terms of STOI [20] and perceptual evaluation of speech quality (PESQ) [30]. ARN with minimum validation PCM loss and ARN with maximum validation STOI are both evaluated. Under −6 dB SNR, ARN with minimum validation PCM and ARN with maximum validation STOI improve STOI by 19.64% and 19.57%, and PESQ by 1.27 and 1.29, respectively. ARN with minimum validation PCM outperforms ARN with maximum validation STOI at all SNR levels except for PESQ under −6 dB SNR, although the performance differences between the two models are small. Enhancement results on CHiME-2 from other baseline models are not available, but the enhancement comparisons between ARN and other models are available in [8].

4.2. Results of Acoustic Models Trained on Noisy Speech

Evaluations of ASR performance are presented in Table 2, where *Enhancement Model* denotes the frontend enhancement model of the system, and *AM Type* denotes the type of training data for the backend. When tested on CHiME-2 reverberant-clean test utterances ('scaled'), the WERs for the default backend, noise-dependent backend, noise-independent backend, distortion-independent PCM backend, and distortion-independent STOI backend are 2.90%, 4.35%, 3.48%, 4.62%, and 4.33%, respectively.

We first evaluate and compare AMs trained on noisy speech. The noise-independent and noise-dependent WRBN in Table 2 are trained on CHiME-2 clean speech mixed with 10k noise and CHiME-2 noisy speech, respectively [4]. Compared with noise-independent WRBN, our noise-independent backend achieves reasonably good performance. The noise-dependent backend outperforms the noise-dependent WRBN by 2.6%, and is the best-performing model. The results demonstrate the effectiveness of the Conformer-based AM baselines.

4.3. Results of the Proposed System

We next evaluate the proposed system and compare it with other systems that incorporate speech enhancement. One system is WRBN trained on GRN (gated residual network) [31] enhanced speech [4], and the other is WRBN trained on GRN enhanced magnitude spectra [14]. A perceptual loss based model [15] is also included, and it employs a Wide ResNet trained with a perceptual loss as the frontend and its backend uses an off-the-shelf Kaldi

CHiME-2 recipe. The proposed system with the ARN selected using minimum validation PCM achieves 7.84% WER on average, which is close to the previous best 7.78% from distortion-independent WRBN with enhanced features [14]. The proposed system with the ARN selected using maximum validation STOI achieves 6.28% WER, which outperforms the previous best by 19.3% relatively. Switching the model selection criterion for ARN reduces WER from 7.84% to 6.28%, corresponding to a relative improvement of 19.9%. This suggests that maximum validation STOI is a strong option for speech enhancement model selection when tested on downstream ASR tasks. Results from the proposed system demonstrate that speech enhancement advances translate to improved ASR results. In this way, the divide between speech enhancement and ASR is eliminated by the use of ARN speech enhancement.

4.4.   Results of ARN on Other Acoustic Models

We finally test ARN enhanced speech on the noise-independent, noise-dependent, and distortion-independent backends for comprehensive comparisons. The WER results are given in the bottom parts of Table 2. Tested on ARN enhanced speech, WERs of noise-independent and noise-dependent backends are consistently reduced in comparison to those tested on noisy speech directly. With ARN selected using maximum validation STOI, the WER results for the noise-independent and noise-dependent backends are improved by 42.6% and 19.2% respectively over the corresponding results on noisy speech (see Table 2). The improvements show that ARN enhanced speech can benefit not only AM trained on clean speech only but also AM trained on noisy speech. With ARN enhanced speech as the input, even though ASR performances get improved for AMs trained on noisy speech, the proposed system with AM trained on clean speech still outperforms both baselines. Compared with distortion-independent backends, the proposed system achieves comparable performance with the 6.16% WER from STOI backend tested on matched ARN enhanced speech, and outperforms at SNR higher than 0 dB. This further demonstrates the utility of the proposed system that fully decouples the enhancement frontend and the backend trained on clean speech only.

5.   Concluding Remarks

This study aims to eliminate the divide between speech enhancement and ASR. The time-domain enhancement model of ARN is employed as the frontend to a Conformer-based AM trained on clean speech only. The proposed system fully decouples speech enhancement and ASR. Results on the CHiME-2 corpus show that better speech enhancement translates to improved ASR results. The proposed system achieves 6.28% WER on the CHiME-2 corpus, outperforming the previous best by 19.3% relatively. Future work includes conducting more experiments on other corpora, applying ARN to other ASR tasks such as continuous speech separation, and extending to multi-channel robust ASR tasks.

Acknowledgements

References

[1]. Heymann J, Bacchiani M, and Sainath TN, "Performance of mask based statistical beamforming in a smart home scenario," in Proc. IEEE ICASSP, 2018, pp. 6722–6726.

[2]. Fu Y, Cheng L, Lv S et al. , "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," in Proc. INTERSPEECH, 2021, pp. 3665–3669.

[3]. Wang DL and Chen J, "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. [PubMed: 31223631]

[4]. Wang P, Tan K, and Wang DL, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 39–48, 2019.

[5]. Cooke M, Green P, Josifovski L, and Vizinho A, "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Communication, vol. 34, no. 3, pp. 267–285, 2001.

[6]. Raj B, Seltzer ML, and Stern RM, "Reconstruction of missing features for robust speech recognition," Speech Communication, vol. 43, no. 4, pp. 275–296, 2004.

[7]. Narayanan A and Wang DL, "Investigation of speech separation as a front-end for noise robust speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 826–835, 2014.

[8]. Pandey A and Wang DL, "Self-attending RNN for speech enhancement to improve cross-corpus generalization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1374–1385, 2022. [PubMed: 36245814]

[9]. Wang Z-Q, Wang P, and Wang DL, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1778–1787, 2020. [PubMed: 33748326]

[10]. Fu S-W, Tsao Y, Lu X, and Kawai H, "Raw waveform-based speech enhancement by fully convolutional networks," in Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf, 2017, pp. 006–012.

[11]. Yang Y, Wang P, and Wang DL, "A Conformer based acoustic model for robust automatic speech recognition," arXiv:2203.00725, 2022.

[12]. Chan W, Park D, Lee C et al. , "SpeechStew: Simply mix all available speech recognition data to train one large neural network," arXiv:2104.02133, 2021.

[13]. Meng Z, Li J, Gong Y, and Juang B-H, "Adversarial feature-mapping for speech enhancement," in Proc. INTERSPEECH, 2018, pp. 3259–3263.

[14]. Wang P and Wang DL, "Enhanced spectral features for distortion-independent acoustic modeling," in Proc. INTERSPEECH, 2019, pp. 476–480.

[15]. Plantinga P, Bagchi D, and Fosler-Lussier E, "Perceptual loss with recognition model for single-channel enhancement and robust ASR," arXiv:2112.06068, 2021.

[16]. Wang Z-Q and Wang DL, "A joint training framework for robust automatic speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 4, pp. 796–806, 2016.

[17]. Zhu Q-S, Zhang J, Zhang Z-Q, and Dai L-R, "Joint training of speech enhancement and self-supervised model for noise-robust ASR," arXiv:2205.13293, 2022.

[18]. Shi J, Chang X, Watanabe S, and Xu B, "Train from scratch: Single-stage joint training of speech separation and recognition," Comput. Speech Lang, vol. 76, p. 101387, 2022.

[19]. Chang X, Maekaku T, Fujita Y, and Watanabe S, "End-to-end integration of speech recognition, speech enhancement, and self-supervised learning representation," in Proc. INTERSPEECH, 2022, pp. 3819–3823.

[20]. Taal CH, Hendriks RC, Heusdens R, and Jensen J, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125–2136, 2011.

[21]. Zagoruyko S and Komodakis N, "Wide residual networks," arXiv:1605.07146, 2016.

[22]. Vincent E, Barker J, Watanabe S et al. , "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in Proc. IEEE ICASSP, 2013, pp. 126–130.

[23]. Barker J, Vincent E, Ma N, Christensen H, and Green P, "The PASCAL CHiME speech separation and recognition challenge," Comput. Speech Lang, vol. 27, no. 3, pp. 621–633, 2013.

[24]. Christensen H, Barker J, Ma N, and Green PD, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in Proc. INTERSPEECH, 2010.

[25]. Varga A and Steeneken HJ, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Communication, vol. 12, no. 3, pp. 247–251, 1993.

[26]. Pandey A and Wang DL, "Dense CNN with self-attention for time-domain speech enhancement," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1270–1279, 2021. [PubMed: 33997107]

[27]. Kingma DP and Ba J, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Representations, 2015.

[28]. Moore AH, Parada PP, and Naylor PA, "Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures," Comput. Speech Lang, vol. 46, pp. 574–584, 2017.

[29]. Vaswani A, Shazeer N, Parmar N et al. , "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst, 2017, pp. 5998–6008.

[30]. Rix AW, Beerends JG, Hollier MP, and Hekstra AP, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in Proc. IEEE ICASSP, vol. 2, 2001, pp. 749–752.

[31]. Tan K, Chen J, and Wang D, "Gated residual networks with dilated convolutions for monaural speech enhancement," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 1, pp. 189–198, 2018. [PubMed: 31355300]
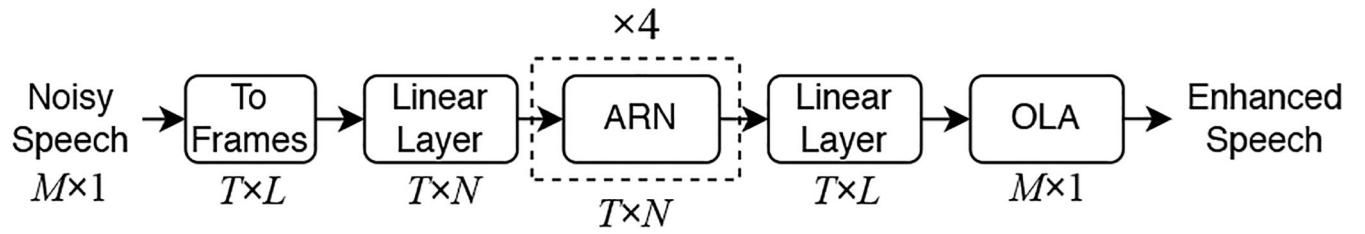
Figure 1:
Diagram of ARN. M, L, and T denote the number of samples of an input signal, frame length, and the total number of frames, respectively.
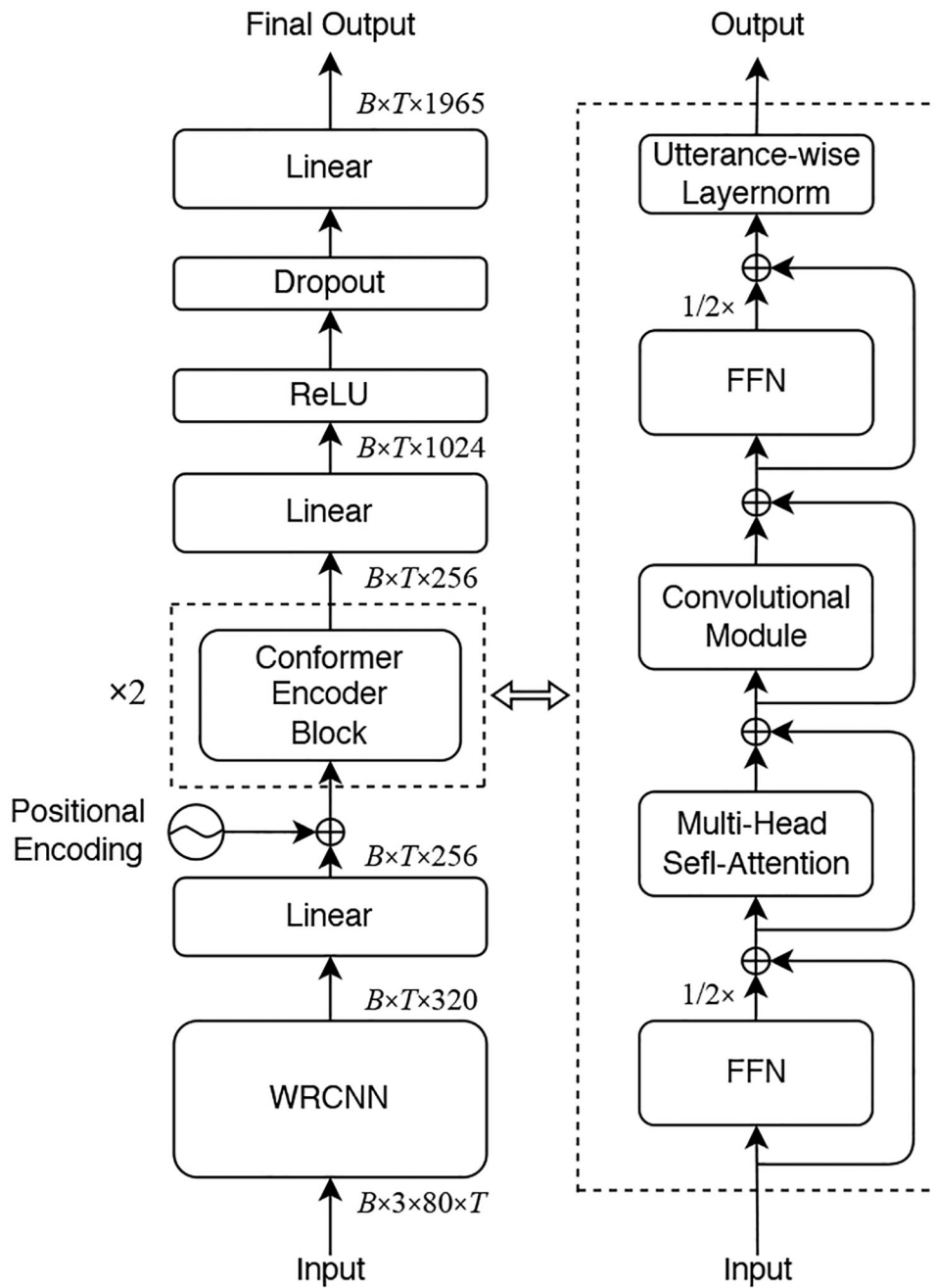
Figure 2:
System architecture of Conformer-based AM. B denotes the batch size, and T denotes the number of time frames of the longest utterance in a batch.

Table 1:

Speech enhancement results of different ARN models on CHiME-2. STOI results are in percentage.

	Test SNR	−6 dB	−3 dB	0 dB	3 dB	6 dB	9 dB	Avg.
STOI	Mixture	73.68	77.80	81.34	85.16	88.12	90.93	82.84
	Min PCM ARN	**93.32**	**94.44**	**95.17**	**96.01**	**96.50**	**97.13**	**95.43**
	Max STOI ARN	93.25	94.17	94.71	95.35	95.75	96.24	94.91
PESQ	Mixture	2.13	2.33	2.49	2.66	2.85	3.04	2.58
	Min PCM ARN	3.40	**3.50**	**3.59**	**3.68**	**3.75**	**3.81**	**3.62**
	Max STOI ARN	**3.42**	**3.50**	3.56	3.64	3.70	3.76	3.60

Table 2:

ASR (% WER) results of the proposed system and comparison systems on CHiME-2. Enh. denotes 'enhanced'.

Model Name	Enhancement Model	AM Type	SNR						Avg.
			-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	
Mixture	-	Clean	73.81	64.88	57.59	45.10	36.05	28.54	51.00
Noise-independent WRBN [4]	-	Noisy	17.45	13.06	10.69	8.82	7.72	6.63	10.73
Noise-independent backend	-	Noisy	19.82	13.30	10.91	9.25	7.25	6.63	11.19
Noise-dependent WRBN [4]	-	Noisy	14.83	9.98	8.95	6.78	6.26	5.49	8.72
Noise-dependent backend	-	Noisy	14.44	10.33	7.92	6.73	6.03	5.49	8.49
Distortion-independent WRBN [4]	GRN	Enh. Speech	15.45	11.04	9.70	7.10	6.54	5.51	9.22
Distortion-independent WRBN [14]	GRN	Enh. Feature	13.11	9.43	7.92	6.20	5.45	4.54	7.78
Perceptual loss based [15]	Wide ResNet	Clean	15.2	10.9	8.3	6.7	5.8	5.2	8.7
Proposed	Min PCM ARN	Clean	13.30	9.66	7.83	6.41	5.25	4.60	7.84
Proposed	Max STOI ARN	Clean	9.94	7.04	6.50	5.45	4.54	4.22	6.28
Noise-independent backend	Min PCM ARN	Noisy	12.93	9.14	8.28	6.65	5.53	4.86	7.90
Noise-independent backend	Max STOI ARN	Noisy	9.62	7.34	6.78	5.45	4.84	4.46	6.42
Noise-dependent backend	Min PCM ARN	Noisy	12.54	9.25	8.29	6.73	6.15	5.87	8.14
Noise-dependent backend	Max STOI ARN	Noisy	9.73	7.40	7.23	6.41	5.16	5.23	6.86
Distortion-independent PCM backend	Min PCM ARN	Enh. Speech	10.14	7.98	7.29	6.28	5.94	5.51	7.19
Distortion-independent PCM backend	Max STOI ARN	Enh. Speech	8.84	6.89	6.50	6.13	5.57	5.29	6.54
Distortion-independent STOI backend	Min PCM ARN	Enh. Speech	10.61	8.22	7.36	6.35	5.98	5.32	7.31
Distortion-independent STOI backend	Max STOI ARN	Enh. Speech	8.05	6.63	6.39	5.87	5.21	4.78	6.16