Learning Complex Spectral Mapping for Speech Enhancement with Improved Cross-corpus Generalization

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

pandey.99@osu.edu, dwang@cse.ohio-state.edu

Abstract

It is recently revealed that deep learning based speech enhancement systems do not generalize to untrained corpora in low signal-to-noise ratio (SNR) conditions, mainly due to the channel mismatch between trained and untrained corpora. In this study, we investigate techniques to improve cross-corpus generalization of complex spectrogram enhancement. First, we propose a long short-term memory (LSTM) network for complex spectral mapping. Evaluated on untrained noises and corpora, the proposed network substantially outperforms a state-of-the-art gated convolutional recurrent network (GCRN). Next, we examine the importance of training corpus for cross-corpus generalization. It is found that a training corpus that contains utterances with different channels can significantly improve performance on untrained corpora. Finally, we observe that using a smaller frame shift in short-time Fourier transform (STFT) is a simple but highly effective technique to improve cross-corpus generalization.

Index Terms: speech enhancement, complex spectral mapping, channel generalization, robust speech enhancement, cross-corpus generalization

1. Introduction

Background noise in a real-world environment degrades the intelligibility and quality of a speech signal for human listeners. Also, it severely degrades the performance of many speech-based applications, such as automatic speech recognition, telecommunication, and hearing aids. Speech enhancement aims at removing the background noise from a speech signal. It is used as a preprocessor in speech-based applications to improve their performance in noisy environments.

In the past few years, deep learning based supervised approaches have become the mainstream for speech enhancement [1]. Primary methods for supervised speech enhancement use time-frequency (T-F) masking or spectral mapping that enhance only the spectral magnitude of the noisy speech signal [2, 3, 4, 5, 6, 7]. This is primarily because the spectral phase was considered unimportant for speech enhancement [8], and it exhibits no clear spectro-temporal structure amenable to supervised learning [9].

A later study revealed that the phase plays an important role in the quality of enhanced speech, especially in low SNR conditions [10]. Williamson et al. [9] first studied supervised speech enhancement in the complex domain. A key insight is the use of the Cartesian representation of a complex spectrogram instead of the widely used polar representation (magnitude and phase), and the real and the imaginary components of the Cartesian representation exhibit clean spectro-temporal structure, amenable to supervised training [9]. Consequently, algorithms have been developed to jointly enhance both the magnitude and the phase [9, 11, 12, 13, 14].

There are two popular approaches to complex spectrogram enhancement: complex ratio masking [9, 13] and complex spectral mapping [11, 12, 14]. In complex ratio masking, the real and the imaginary part of a complex-valued mask is used as the training target. The complex-valued mask is derived using the noisy and the clean spectrogram. Complex spectral mapping, on the other hand, uses the real and the imaginary part of the clean spectrogram.

It is recently revealed that deep learning based speech enhancement systems do not generalize to untrained corpora [15]. A large degradation is observed in enhancement performance on untrained corpora in low SNR conditions. It is established that the main factor for degradation is the channel mismatch between training and test corpora. The authors propose several techniques to improve cross-corpus generalization, such as channel normalization, a better training corpus, and a smaller frame shift in STFT. The proposed techniques obtain significant improvements on untrained corpora for an ideal ratio mask (IRM) [2] based bidirectional LSTM (BLSTM) network [15].

For speech enhancement, complex spectral mapping is found to be superior to spectral mapping [11, 12, 14]. However, in [15], the authors find that a complex spectral mapping based convolutional recurrent network [16] fails on untrained corpora, and is worse than a spectral mapping based network. Further, channel normalization techniques based on spectral or cepstral normalization do not apply to complex spectrogram. Therefore, new techniques needs to be developed to improve cross-corpus generalization of complex spectral mapping.

In this study, we investigate techniques to improve cross-corpus generalization of complex spectrogram enhancement. We propose an LSTM network for complex spectral mapping, and develop both causal and non-causal systems. All the developed models are evaluated on untrained noises and corpora. The proposed network obtains significantly better enhancement results than a state-of-the-art GCRN [12].

To examine the importance of training corpus, we train LSTM and GCRN using two different corpora: WSJ-SI-84 (WSJ) [17] and LibriSpeech [18]. WSJ contains utterances with similar channels, whereas LibriSpeech contains utterances with different channels. Therefore, LibriSpeech should aid cross-corpus generalization. We find LibriSpeech to be substantially better than WSJ for both LSTM and GCRN.

Finally, we compare different frame shifts in STFT for LSTM and GCRN. We find that using a smaller frame shift in STFT is a simple and highly effective technique to improve cross-corpus generalization for complex spectral mapping.

This research was supported in part by two NIDCD (R01 DC012048 and R01 DC015521) grants and the Ohio Supercomputer Center.

The rest of this paper is organized as follows. We describe complex spectral mapping in Section 2. The techniques to improve cross-corpus generalization are given in Section 3. Experiments are discussed in Section 4. Section 5 concludes this paper.

2. Complex Spectral Mapping for Speech Enhancement

Given a clean speech signal $\boldsymbol{s}$ and a noise signal $\boldsymbol{n}$, the noisy speech signal $\boldsymbol{y}$ is modeled as

$$y[k] = s[k] + n[k] \tag{1}$$

where $\{\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{n}\} \in \mathbb{R}^{M\times 1}$, M represents the number of samples in the signal, and k is the time sample index. The goal of speech enhancement is to get a close estimate $\widehat{\boldsymbol{s}}$ of $\boldsymbol{s}$ given $\boldsymbol{y}$. Taking STFT on both sides in Eq. (1), we get

$$Y_{t,f} = S_{t,f} + N_{t,f} \tag{2}$$

where $\{\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{N}\} \in \mathbb{R}^{T\times F}$. $\boldsymbol{Y}$, $\boldsymbol{S}$ and $\boldsymbol{N}$ respectively represent the STFTs of $\boldsymbol{y}$, $\boldsymbol{s}$, and $\boldsymbol{n}$. t and f denote the frame and the frequency index. T is the number of frames, and F is the number of frequency bins. In polar coordinates, Eq. (2) becomes

$$|Y_{t,f}|e^{i\theta_{Y_{t,f}}} = |S_{t,f}|e^{i\theta_{S_{t,f}}} + |N_{t,f}|e^{i\theta_{N_{t,f}}} \tag{3}$$

where $|z|$ and θ_z, respectively, denote the magnitude and the phase of a complex variable z. Generally, speech enhancement is formulated as a magnitude enhancement problem in which $|\boldsymbol{Y}|$ is used to get a close estimate, $|\widehat{\boldsymbol{S}}|$, of $|\boldsymbol{S}|$. The enhanced magnitude is combined with the noisy phase to get the enhanced STFT.

$$\widehat{S}_{t,f} = |\widehat{S}_{t,f}|e^{i\theta_{Y_{t,f}}} \tag{4}$$

Finally, inverse STFT (ISTFT) is used to obtain the enhanced waveform $\widehat{\boldsymbol{s}}$.

$$\widehat{\boldsymbol{s}} = \text{ISTFT}(\widehat{\boldsymbol{S}}) \tag{5}$$

Complex spectrogram enhancement, on the other hand, aims at enhancing both the magnitude and the phase. Phase enhancement is considered difficult as it does not exhibit a clear spectro-temporal structure amenable to supervised learning [9]. However, the magnitude and the phase can be jointly enhanced by exploiting the Cartesian representation of STFT. In Cartesian coordinates, a complex variable z is represented as

$$z = z^r + iz^i \tag{6}$$

where z^r and z^i, respectively, denote the real and the imaginary part of z. Using Eq. (6) in Eq. (2), we obtain

$$Y^r_{t,f} + iY^i_{t,f} = S^r_{t,f} + N^r_{t,f} + i(S^i_{t,f} + N^i_{t,f}) \tag{7}$$

In this representation, speech enhancement is formulated as a problem of enhancing the real and the imaginary part of the noisy STFT. In other words, $\boldsymbol{Y}$ is used to get the close estimates, $\widehat{\boldsymbol{S}}^r$ and $\widehat{\boldsymbol{S}}^i$, of $\boldsymbol{S}^r$ and $\boldsymbol{S}^i$, respectively. The enhanced STFT is obtained using the following equation.

$$\widehat{S}_{tf} = \widehat{S}^r_{t,f} + i\widehat{S}^i_{t,f} \tag{8}$$

Finally, Eq. (5) is used to obtain the enhanced waveform $\widehat{\boldsymbol{s}}$.

In complex spectral mapping, a DNN is used to jointly compute $\widehat{\boldsymbol{S}}^r$ and $\widehat{\boldsymbol{S}}^i$ using $\boldsymbol{Y}$ as the input. The input to the DNN is formed by concatenating $\boldsymbol{Y}^r$ and $\boldsymbol{Y}^i$. Similarly, the output of the DNN is a concatenation of $\widehat{\boldsymbol{S}}^r$ and $\widehat{\boldsymbol{S}}^i$. In this study, we employ an LSTM network for complex spectral mapping. The proposed framework for complex spectral mapping is shown in Fig. (1).

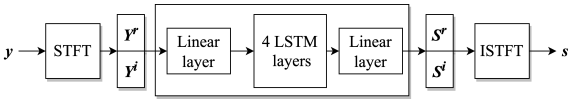

Figure 1: *The proposed framework for complex spectral mapping.*

3. Improving Cross-corpus Generalization

We have proposed the following techniques to improve cross-corpus generalization for complex spectral mapping.

3.1. Model Architecture

Model architecture can have a significant impact on cross-corpus generalization. A model with better performance on a trained corpus does not necessarily obtain a better performance on an untrained corpus [15]. For example, it has been shown in [15] that the architectures such as CRN [16], AECNN [19], and TCNN [20], that are superior to LSTM on a trained corpus, are worse than LSTM on untrained corpora. We investigate the following two architectures for complex spectral mapping.

3.1.1. GCRN

GCRN is a recently proposed architecture for complex spectral mapping and represents state-of-the-art on WSJ [12]. It is an encoder-decoder based architecture with skip connections. A network with two LSTM layers is inserted between the encoder and the decoder for a recurrent context aggregation. The decoder has two branches; one for the real part and the other for the imaginary part. The encoder and decoder consist of convolutions with gated linear units [21]. The input to the network is $\boldsymbol{Y}$ arranged as a 4D signal of shape $[BatchSize, 2, T, F]$. $\boldsymbol{Y}^r$ and $\boldsymbol{Y}^i$ are stacked along the second dimension representing the channels in the input. The LSTMs in the GCRN are replaced with bidirectional LSTMs (BLSTMs) to get a non-causal GCRN [12]. We denote non-causal GCRN as NC-GCRN in our results.

3.1.2. LSTM

We propose to use an LSTM network for complex spectral mapping. The network consists of four LSTM layers with one linear layer at the input and the output. An illustrative diagram of the proposed LSTM network is shown in Fig. (1). The input to the LSTM network is $\boldsymbol{Y}$ arrranged as a 3D signal of shape $[BatchSize, T, 2 \cdot F]$. $\boldsymbol{Y}^r$ and $\boldsymbol{Y}^i$ are concatenated along the third dimension. The output layer has $2 \cdot F$ units to output $\widehat{\boldsymbol{S}}^r$ and $\widehat{\boldsymbol{S}}^i$ together. We replace LSTMs with BLSTMs to get a non-causal speech enhancement system.

3.2. Training Corpus

Corpus channel is one of the major factors for performance degradation on untrained corpora [15]. The channel of an utterance is defined as the stationary component acquired due to fixed recording conditions, such as microphones and room acoustics. A corpus such as WSJ is recorded in a controlled environment, and hence contains utterances with similar channels. LibriSpeech, on the other hand, contains audios that are recorded by many volunteers across the globe, implying that it contains utterances with different channels. Therefore, WSJ is more likely to overfit on corpus compared to LibriSpeech. In [15], the authors find LibriSpeech to be highly effective for

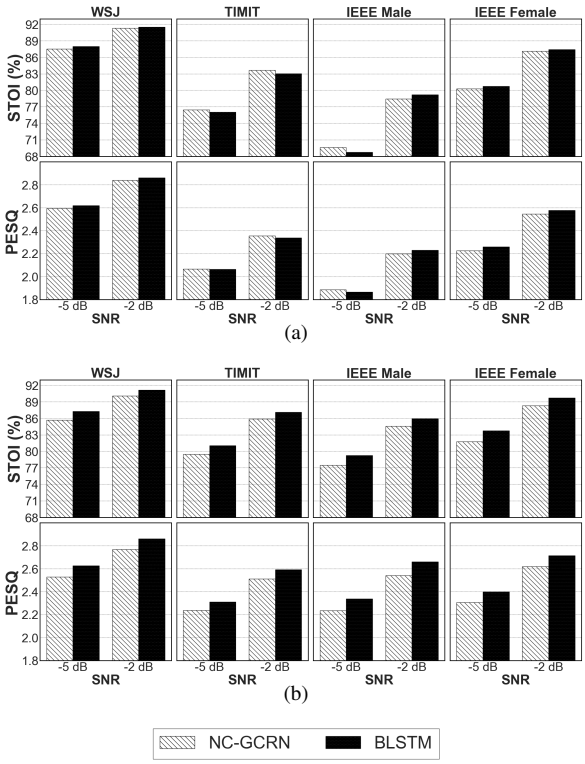

Figure 2: *STOI and PESQ comparisons between NC-GCRN and BLSTM. a) Trained on WSJ, b) Trained on LibsriSpeech.*

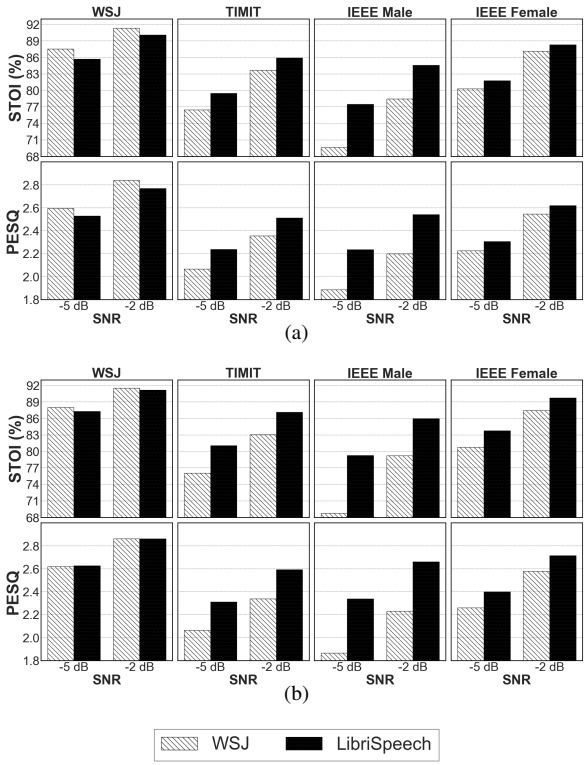

Figure 3: *STOI and PESQ comparisons between WSJ and LibriSpeech. a) NC-GCRN, b) BLSTM.*

cross-corpus generalization. In this study, we compare LibriSpeech and WSJ, which is a widely utilized corpus for speech enhancement and other related tasks.

3.3. Frame Shift

In an STFT computation for speech enhancement, a frame shift equal to the half of the frame size typically is used, and overlap-and-add (OLA) is used for final reconstruction in the time domain. However, using a smaller frame shift leads to multiple predictions (> 2) for a given T-F unit, and these multiple predictions are averaged in OLA stgae. The simple idea of using a smaller frame shift can lead to a significant improvement in cross-corpus generalization [15]. Further, using a smaller frame shift leads to better performance on trained corpus as well [22]. In this study, we compare two different frame shifts: half of the frame size and quarter of the frame size, and illustrate that using a smaller frame shift is a simple and effective technique to improve cross-corpus generalization.

4. Experiments

4.1. Datasets

We train all the models on WSJ and LibriSpeech. The WSJ training set consists of 6385 utterances of 77 speakers. The LibriSpeech training set consists of 252702 utterances of 2087 speakers. Noisy utterances are generated during training by randomly adding noise to all the utterances in a batch. For training noises, we use 10000 non-speech sounds from a sound effect library (`www.sound-ideas.com`). A random noise segment is added to an utterance at a random SNR in $\{-5$ dB, -4 dB, -3 dB, -2 dB, -1 dB, 0 dB$\}$.

All the models are evaluated on WSJ, TIMIT [23], and IEEE [24]. A random male and female speaker is used from IEEE to create IEEE Male and IEEE Female corpus. The WSJ test set consists of 150 utterances of 6 speakers not included in WSJ training. The TIMIT test set consists of 192 utterances from the core test set. IEEE Male and IEEE Female each contain 144 randomly selected utterances. Test set is generated using 4 different noises: babble, cafeteria, factory and engine, at SNRs of -5 dB and -2 dB. The babble and cafeteria noises are from Auditec CD (available at `http://www.auditec.com`). Factory and engine noises are from Noisex [25].

4.2. Experimental Settings

All the utterances are resampled to 16 kHz. All the noisy utterances are normalized to the range of $[-1, 1]$, and corresponding clean utterances are scaled accordingly to maintain an SNR. A frame size of 20 ms is used for GCRN, similar to the original paper [12]. BLSTM uses a frame size of 16 ms. Hamming window is used in STFT.

The Adam optimizer [26] is used with a learning rate schedule as in [15]. LSTM is trained with a batch size of 32 utterances, whereas GCRN is trained with a batch size of 8 utterances. An ISTFT layer is used at the output to compute enhanced waveform, and an utterance level mean squared error loss in the time domain is used for training.

4.3. Evaluation Metrics

In our experiments, models are compared using short-time objective intelligibility (STOI) [27] and perceptual evaluation of speech quality (PESQ) [28] scores. STOI has a typical value range from 0 to 1, which can be roughly interpreted as percent

Table 1: *STOI and PESQ comparisons between different models on babble and cafeteria noise. a) Non-causal systems, b) Causal systems.*

		Test Noise	**Babble**								**Cafeteria**							
		Test Corpus	WSJ		TIMIT		IEEE Male		IEEE Female		WSJ		TIMIT		IEEE Male		IEEE Female	
		Test SNR	-5 dB	-2 dB	-5 dB	-2 dB	-5 dB	-2 dB	-5 dB	-2 dB	-5 dB	-2 dB	-5 dB	-2 dB	-5 dB	-2 dB	-5 dB	-2 dB
STOI (%)		Mixture	58.6	65.5	54.0	60.9	55.0	62.3	55.5	62.9	57.4	64.5	53.1	60.1	54.8	60.9	55.1	62.0
	a)	Baseline	82.4	87.3	75.1	82.1	74.3	83.2	74.8	84.3	80.3	85.5	74.5	80.7	73.4	80.4	77.6	84.0
		NC-GCRN	85.7	90.4	76.8	85.0	74.0	83.8	76.6	86.8	83.7	88.9	77.8	84.7	74.8	82.6	80.4	87.0
		BLSTM	**88.0**	**91.9**	**79.2**	**86.9**	**76.2**	**85.8**	**79.5**	**89.1**	**85.4**	**89.9**	**79.4**	**86.0**	**76.7**	**83.8**	**82.3**	**88.5**
	b)	GCRN	81.8	88.0	71.5	81.3	68.7	79.9	70.0	82.5	79.6	86.3	72.8	81.3	69.5	78.4	74.7	83.7
		LSTM	**84.4**	**89.6**	**74.9**	**83.8**	**72.0**	**82.3**	**73.0**	**84.5**	**81.5**	**87.5**	**74.8**	**83.0**	**71.6**	**79.9**	**76.2**	**84.8**
PESQ		Mixture	1.54	1.69	1.46	1.63	1.46	1.63	1.12	1.32	1.44	1.64	1.33	1.52	1.37	1.54	1.01	1.20
	a)	Baseline	2.43	2.70	2.20	2.52	2.11	2.47	1.94	2.41	2.41	2.66	2.25	2.51	2.15	2.44	2.16	2.47
		NC-GCRN	2.48	2.75	2.10	2.43	2.05	2.44	1.98	2.46	2.44	2.70	2.18	2.45	2.17	2.49	2.24	2.57
		BLSTM	**2.64**	**2.90**	**2.21**	**2.55**	**2.16**	**2.59**	**2.12**	**2.62**	**2.52**	**2.76**	**2.24**	**2.52**	**2.24**	**2.58**	**2.32**	**2.64**
	b)	GCRN	2.18	2.48	1.87	2.21	1.84	2.20	1.70	2.17	2.14	2.44	1.95	2.23	1.92	2.25	1.98	2.33
		LSTM	**2.29**	**2.59**	**1.95**	**2.32**	**1.96**	**2.36**	**1.80**	**2.29**	**2.20**	**2.48**	**1.98**	**2.27**	**2.01**	**2.32**	**2.01**	**2.35**

correct. PESQ values range from -0.5 to 4.5.

4.4. Experimental Results

First, we compare NC-GCRN (non-causal GCRN) and BLSTM trained on WSJ and LibriSpeech. STOI and PESQ scores averaged over 4 test noises are plotted in Fig. 2. We observe that when using WSJ as the training corpus, there is no clear winner between NC-GCRN and BLSTM. BLSTM is similar to NC-GCRN on WSJ, better on IEEE Female and IEEE Male -2 dB, and worse in other cases. However, when LibriSpeech is used for training, BLSTM is significantly better in all the cases. This behavior can be attributed to the fact that BLSTM has far more parameters compared to NC-GCRN, and hence can learn underlying larger variations in LibriSpeech.

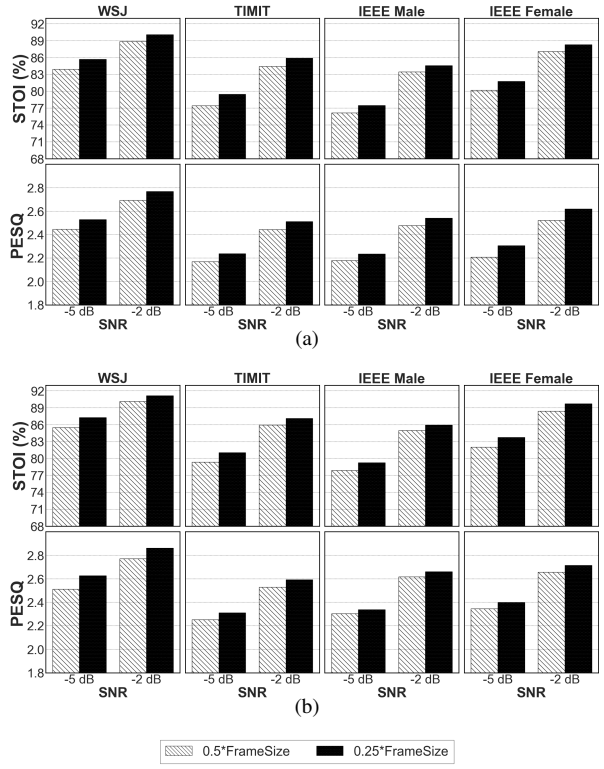

Figure 4: *STOI and PESQ comparisons between two frame shifts. a) NC-GCRN, b) BLSTM.*

Next, we compare WSJ and LibriSpeech for cross-corpus generalization of complex spectrogram enhancement. STOI and PESQ averaged over four test noises are plotted in Fig. (3). We can observe that LibriSpeech outperforms WSJ for all the test corpora except WSJ. WSJ outperforms LibriSpeech only when evaluated on WSJ, i.e., in a matched condition. Further, LibriSpeech obtains better enhancement results for both NC-GCRN and BLSTM. These results indicate that LibriSpeech is a better corpus than WSJ for robust speech enhancement. Therefore, WSJ should be replaced with LibriSpeech for future speech enhancement research.

Further, we demonstrate the effectiveness of a smaller frame shift for cross-corpus generalization. NC-GCRN and BLSTM are trained with two frame shifts: half of the frame size and quarter of the frame size. STOI and PESQ averaged over four test noises are plotted in Fig. (4). We observe that reducing the frame-shift from half of the frame size to quarter significantly improves the objective scores for all the test corpora and at all test SNRs. A similar performance trend is observed for both NC-GCRN and BLSTM. This implies that using a smaller frame shift is a simple and highly-effective technique to improve cross-corpus generalization of complex-spectrogram enhancement.

Finally, we compare best performing NC-GCRN and BLSTM with an IRM based BLSTM network proposed in [15]. STOI and PESQ scores for babble and cafeteria noises are given in Table 1. A comparison between causal enhancement systems is also presented. We observe that LSTM network is consistently and significantly better than baseline and NC-GCRN. A similar behavior is observed for factory and engine noises (not reported here). Experimental comparisons in this study indicate that the proposed LSTM network trained on LibriSpeech with a small frame shift is an effective approach for complex spectral mapping with improved cross-corpus generalization.

5. Conclusions

We have proposed techniques to improve cross-corpus generalization of complex spectral mapping based speech enhancement. A combination of better training corpus, smaller frame shift, and LSTM network has obtained superior enhancement over existing state-of-the-art models. We have also demonstrated the effectiveness of proposed techniques for causal speech enhancement. Future works include developing a computationally efficient causal speech enhancement system suitable for real-time implementation, and performing listening tests of corpus-independent enhancement.

6. References

[1] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1702–1726, 2018.

[2] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[3] 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, 2015.

[4] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *INTERSPEECH*, 2017, pp. 1993–1997.

[5] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[6] K. Tan, J. Chen, and D. Wang, "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.

[7] A. Pandey and D. L. Wang, "On adversarial training and loss functions for speech enhancement," in *ICASSP*, 2018, pp. 5414–5418.

[8] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[9] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[10] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

[11] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *International Workshop on Machine Learning for Signal Processing*. IEEE, 2017, pp. 1–6.

[12] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[13] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," *arXiv preprint arXiv:1903.03107*, 2019.

[14] A. Pandey and D. Wang, "Exploring deep complex networks for complex spectrogram enhancement," in *ICASSP*, 2019, pp. 6885–6889.

[15] ——, "On cross-corpus generalization of deep learning based speech enhancement," *arXiv preprint arXiv:2002.04027*, 2020.

[16] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *ICASSP*, 2019, pp. 6865–6869.

[17] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Workshop on Speech and Natural Language*, 1992, pp. 357–362.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[19] 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 (TASLP)*, vol. 27, no. 7, pp. 1179–1188, 2019.

[20] ——, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP*, 2019, pp. 6875–6879.

[21] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International Conference on Machine Learning*. JMLR. org, 2017, pp. 933–941.

[22] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[24] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[25] A. Varga and H. J. Steeneken, "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] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "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.

[28] 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 *ICASSP*, 2001, pp. 749–752.