IMPROVING ROBUSTNESS OF DEEP LEARNING BASED MONAURAL SPEECH ENHANCEMENT AGAINST PROCESSING ARTIFACTS

Ke Tan[1] *and 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
{tan.650, wang.77}@osu.edu

ABSTRACT

In voice telecommunication, the intelligibility and quality of speech signals can be severely degraded by background noise if the speaker at the transmitting end talks in a noisy environment. Therefore, a speech enhancement system is typically integrated into the transmitter device or the receiver device. Without the knowledge of whether the other end is equipped with a speech enhancer, the transmitter and receiver devices can both process a speech signal with their speech enhancers. In this study, we find that enhancing a speech signal twice can dramatically degrade the enhancement performance. This is because the downstream speech enhancer is sensitive to the processing artifacts introduced by the upstream enhancer. We analyze this problem and propose a new training scheme for the downstream deep learning based speech enhancement model. Our experimental results show that the proposed training strategy substantially elevate the robustness of speech enhancers against artifacts induced by another speech enhancer.

Index Terms— monaural speech enhancement, voice telecommunication, processing artifacts, robustness, deep learning

1. INTRODUCTION

A typical telecommunication system for voice comprises a transmitter (i.e. a microphone), a communication circuit (i.e. the physical medium that encodes and carries the speech signal) and a receiver (e.g. a mobile phone loudspeaker). If the speaker at the transmitting end is in a noisy environment, the transmitter picks up target speech as well as background noise, which can severely degrade the intelligibility and quality of the speech signal at the receiving end. In order to attenuate background noise, speech enhancement algorithms have been deployed in telecommunication devices such as mobile phones. One can perform speech enhancement in the upstream transmitter device. Alternatively, the speech enhancement system can be deployed in the downstream receiver device. Fig. 1 illustrates this situation.

Due to the variety of device makers and service providers, in voice telecommunication such as mobile communication, the receiver device typically does not have the knowledge of whether speech enhancement has been performed in the transmitter device. Similarly, the transmitter device does not have the knowledge of whether the receiver device is equipped with speech enhancement. The receiver device may choose to apply a speech enhancer to the received speech signal to cover the situation that the transmitter side lacks enhancement or its enhancement is inadequate. One would imagine that no issue should arise if noisy speech has been enhanced at the transmitting end, and further processed by a speech enhancement system at the receiving end. In this study, we find that enhancing noisy speech twice can be detrimental to the performance of speech enhancement. This occurs because the downstream speech enhancer is susceptible to the processing artifacts introduced by the upstream speech enhancer.

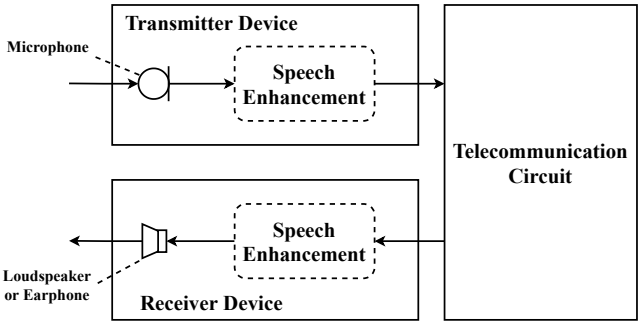

Fig. 1. Diagram of a telecommunication system equipped with a speech enhancement system. The speech enhancement system can be deployed in the transmitter device, the receiver device, or both.

Numerous speech enhancement approaches have been developed over the last several decades. A classic speech enhancement algorithm is spectral subtraction [1], which estimates a short-term noise spectrum and then subtracts it from the noisy spectrum to produce an estimated spectrum of clean speech. In [2], Lim and Oppenheim developed an iterative Wiener filter based on an all-pole model. A noniterative Wiener filtering algorithm was proposed in [3], which computes the Wiener gain using an estimated *a priori* signal-to-noise ratio (SNR). Another classic speech enhancement algorithm is the minimum mean-square error (MMSE) estimator developed by Ephraim and Malah [4]. Following this study, they designed an MMSE log-spectral amplitude estimator (Log-MMSE) in [5], which leads to less residual noise. In [6], Cohen and Berdugo introduced a minima controlled recursive averaging (MCRA) method for noise estimation, which can be combined with the optimally modified log-spectral amplitude (OM-LSA) estimator [7] to perform speech enhancement. They presented an improved MCRA (IMCRA) in [8]. A different class of speech enhancement algorithms is signal subspace algorithms based on matrix analysis. For example, Hu and Loizou proposed a Karhunen–Loève transform (KLT) based subspace algorithm in [9]. However, these conventional speech enhancement methods make assumptions about the statistical characteristics of the speech and noise signals, e.g. stationarity of background noise. Although these assumptions hold for many

This research was supported in part by an NIDCD grant (R01 DC012048) and the Ohio Supercomputer Center.

acoustic environments, the enhancement performance is limited in dynamic environments.

Speech enhancement has been recently formulated as a supervised learning task, where discriminative patterns of clean speech and background noise are learned from training data [10]. Since deep neural networks (DNNs) were first introduced to supervised speech enhancement in 2013 [11], various deep learning based approaches have been developed [12]. The performance of supervised speech enhancement has been substantially elevated thanks to the use of deep learning. For any supervised learning task, generalization to untrained conditions is a crucial issue. In voice telecommunication, does a supervised speech enhancement model generalize to the speech signals that have been already processed by another speech enhancement algorithm?

In this study, we investigate the processing artifacts induced by monaural speech enhancement, and their effects on a succeeding speech enhancer. To alleviate performance degradation caused by the processing artifacts, we propose a new training strategy for deep learning based speech enhancement in voice telecommunication. We evaluate the proposed training technique on a commonly-used long short-term memory (LSTM) model and two newly-developed convolutional recurrent network (CRN) models. Our experimental results show that the proposed training strategy substantially improve the robustness of speech enhancement models against processing artifacts. Moreover, we find that the models trained by the proposed strategy generalize well to new speech enhancers. To our knowledge, this is the first study to examine and address the important robustness issue of deep learning based speech enhancement against processing artifacts introduced by another speech enhancement system.

The rest of this paper is organized as follows. We analyze enhancement artifacts and describe a training strategy for speech enhancement in Section 2. The experimental setup and evaluation results are provided in Section 3. Section 4 concludes this paper.

2. ALGORITHM DESCRIPTION

2.1. Analysis of artifacts induced by speech enhancement

Given a single-microphone mixture y, the goal of monaural speech enhancement is to separate target speech s from background noise n. A noisy mixture can be modeled as

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

where k is the time sample index. Taking the time-frequency (T-F) representations of both sides, we derive

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

where Y, S and N denote the T-F representations of y, s and n, respectively, and m and f index the time frame and the frequency bin, respectively. The T-F representation $\hat{S}_{m,f}$ of enhanced speech can be written as

$$\hat{S}_{m,f} = S_{m,f} + A_{m,f} + N^{(\text{res})}_{m,f}, \tag{3}$$

where A represents the processing artifact induced by speech enhancement, and $N^{(\text{res})}$ the residual noise. Typically, the artifact A is correlated with target speech S, which can result in an alteration or even a loss of speech components. For voice telecommunication, if a conventional speech enhancement method is deployed in the receiver device, such an artifact can dissatisfy the assumptions or conditions that this enhancement method is based on. For a receiver device equipped with a deep learning based speech enhancer, the alteration or the loss of speech components caused by the artifact cannot be restored by the enhancer if it is only trained with unprocessed noisy speech. The performance of such an enhancer can severely degrade on enhanced speech, due to the mismatch between the pattern of enhanced speech and that of unprocessed noisy speech. Even though the SNR of the speech signal is improved by the upstream speech enhancer, the detriment caused by processing artifacts may outweigh the benefit of enhancement in the transmitter device.

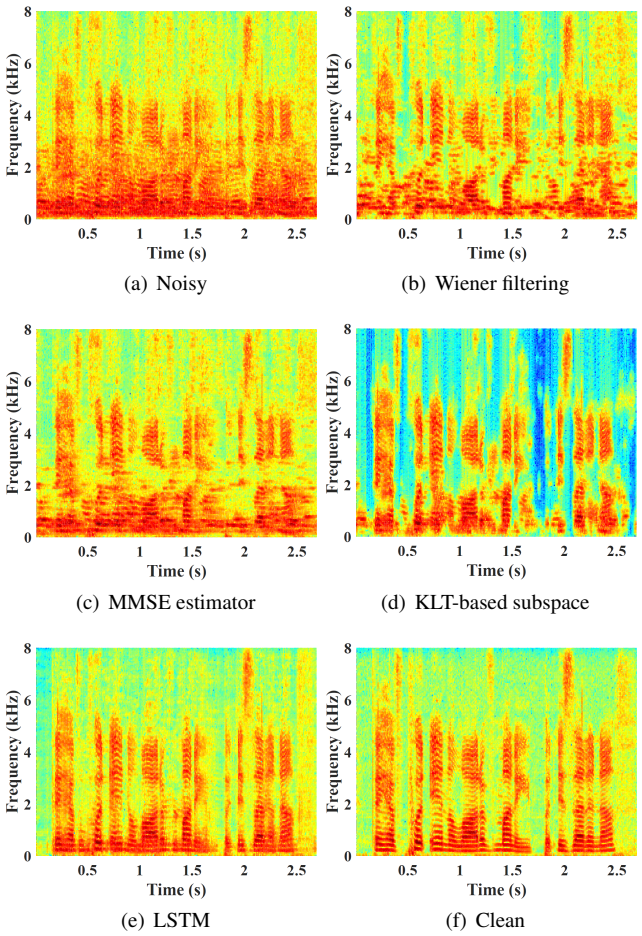

Fig. 2. (Color Online). Example of spectrograms of enhanced speech by Wiener filtering [3], an MMSE estimator [4], a KLT-based subspace method [9] and an LSTM network. The spectral magnitudes are plotted on a log scale.

Fig. 2 illustrates an example of enhanced spectrograms by Wiener filtering [3], an MMSE estimator [4], a KLT-based subspace method [9] and an LSTM network [13]. One can observe that different speech enhancement methods can exhibit very different distortion effects.

2.2. Proposed training strategy

To derive a robust speech enhancer against processing artifacts, we propose a new training strategy for deep learning based monaural speech enhancement, as summarized in Algorithm 1. Specifically, we carefully choose a set of different speech enhancers, and then process each noisy mixture in the original training set using each of these enhancers. Subsequently, we collect all these enhanced speech

Algorithm 1 Proposed training strategy

Input: A set of M different speech enhancers $E_j (1 \leqslant j \leqslant M)$, a randomly initialized speech enhancer E_{tr} to be trained, and a training set $T = \{(y_i, s_i)\}_{1 \leqslant i \leqslant K}$ that contains K pairs of unprocessed noisy speech y_i and clean speech s_i.

Output: A robust speech enhancer E'_{tr}.

1: **for** j in $\{1, 2, \ldots, M\}$ **do**
2: **for** i in $\{1, 2, \ldots, K\}$ **do**
3: Process y_i with E_j to produce enhanced speech $y_i^{(j)}$;
4: Make a new pair of signals $(y_i^{(j)}, s_i)$;
5: **end for**
6: Collect $(y_i^{(j)}, s_i)$ for all i's into a new training set $T^{(j)} = \{(y_i^{(j)}, s_i)\}_{1 \leqslant i \leqslant K}$;
7: **end for**
8: Let $T' = T \cup T^{(1)} \cup T^{(2)} \cup \cdots \cup T^{(M)}$;
9: Train E_{tr} on the comprehensive training set T' to obtain a robust speech enhancer E'_{tr};
10: **return** E'_{tr}

signals as well as the original unprocessed noisy speech signals to form a new comprehensive training set, which is used to train the deep learning based speech enhancer. Note that an unprocessed noisy signal and its multiple enhanced versions correspond to the same target speech signal. In other words, the speech enhancer tends to learn a many-to-one mapping. In this study, we choose a set of five representative traditional speech enhancement algorithms and a commonly-used feedforward DNN as E_j's: (1) E_1: spectral subtraction [1]; (2) E_2: a Wiener filter based on *a priori* SNR estimation [3]; (3) E_3: an MMSE estimator [4]; (4) E_4: the IMCRA method [8]; (5) E_5: a KLT-based subspace algorithm with embedded pre-whitening [9]; (6) E_6: a feedforward DNN that has four hidden layers with 1024 units in each layer, where the output layer performs a spectral mapping in the magnitude domain.

The selection of such a speech enhancer set is motivated by the fact that these methods fall into different classes of speech enhancement algorithms, which are based on different principles or assumptions [14]. The spectral-subtractive algorithms exploit the fact that noise is additive and one can heuristically derive an estimate of the clean speech signal spectrum simply by subtracting the noise spectrum from the noisy speech spectrum. Wiener filtering obtains the enhanced speech signal by optimizing a mathematically tractable error criterion, i.e. the mean squared error (MSE). Statistical model based methods (e.g. the MMSE estimator and the IMCRA method) use nonlinear estimators of the spectral magnitude, which employ various statistical models and optimization criteria. Subspace algorithms are based on the principle that the clean speech signal might be confined to a subspace of the noisy Euclidean space. Supervised speech enhancement methods train a model to learn discriminative patterns from training data. Therefore, the speech signals enhanced by these different approaches may include a relatively wide range of distortion effects. Thus, training with these processed speech signals can improve the robustness of a speech enhancer against enhancement artifacts.

3. EVALUATION AND ANALYSIS

3.1. Experimental setup

In our experiments, we use the WSJ0 SI-84 training set [15] which includes 7138 utterances from 83 speakers (42 males and 41 females). Of these speakers, we set aside six (3 males and 3 females) as untrained speakers for testing. The models are trained with the 77 remaining speakers. We use 10,000 noises from a sound effect library (available at https://www.sound-ideas.com) as the training noises, of which the total duration is about 126 hours. For testing, we use two highly nonstationary noises (babble and cafeteria) from an Auditec CD (available at http://www.auditec.com).

Table 1. Evaluation of LSTM models on different speech enhancers.

Metrics	STOI (in %)			PESQ		
SNR	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
Unprocessed	57.84	69.80	81.06	1.49	1.79	2.12
LSTM1	72.82	84.98	91.57	1.88	2.39	2.80
LSTM2	**73.80**	**85.28**	**91.67**	**1.92**	**2.39**	**2.79**
Spectral subtraction [1]	56.14	70.43	82.77	1.61	1.96	2.33
Spectral subtraction - LSTM1	60.14	76.42	88.24	1.44	2.09	2.73
Spectral subtraction - LSTM2	**72.84**	**84.89**	**91.55**	**1.90**	**2.41**	**2.82**
Wiener filtering [3]	54.63	68.96	81.29	1.52	1.89	2.26
Wiener filtering - LSTM1	57.48	74.46	86.51	1.35	2.02	2.64
Wiener filtering - LSTM2	**72.50**	**84.82**	**91.57**	**1.90**	**2.40**	**2.82**
MMSE estimator [4]	54.19	67.21	79.26	1.61	1.96	2.31
MMSE estimator - LSTM1	55.55	70.27	83.27	1.41	1.96	2.57
MMSE estimator - LSTM2	**71.63**	**84.32**	**91.30**	**1.86**	**2.37**	**2.80**
IMCRA method [8]	55.33	69.50	81.56	1.54	1.90	2.27
IMCRA method - LSTM1	56.11	73.07	85.92	1.29	1.95	2.60
IMCRA method - LSTM2	**73.00**	**85.02**	**91.50**	**1.89**	**2.41**	**2.82**
KLT-based subspace [9]	55.72	71.32	83.24	1.20	1.68	2.11
KLT-based subspace - LSTM1	50.20	70.38	85.65	0.91	1.65	2.39
KLT-based subspace - LSTM2	**71.70**	**84.29**	**91.17**	**1.87**	**2.37**	**2.77**
DNN mapping	68.09	81.29	89.21	1.73	2.21	2.60
DNN mapping - LSTM1	68.78	82.37	89.76	1.69	2.26	2.69
DNN mapping - LSTM2	**71.70**	**84.29**	**91.17**	**1.87**	**2.37**	**2.77**

Table 2. Evaluation of a KLT-based subspace method on an MMSE estimator.

Metrics	STOI (in %)			PESQ		
SNR	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
Unprocessed	57.84	69.80	81.06	1.49	1.79	2.12
MMSE estimator [4]	54.19	67.21	79.26	1.61	1.96	2.31
KLT-based subspace [9]	55.72	71.32	83.24	1.20	1.68	2.11
MMSE - KLT	50.34	67.02	80.33	1.04	1.54	2.03

We simulate a training set by mixing randomly selected training utterance with random cuts from the 10,000 training noises. The SNR of a mixture is randomly sampled from {-8, -7, -6, -5, -4, -3, -2, -1, 0, 4, 8, 12, 16, 20} dB. We denote this training set as "training set 1", which contains 80,000 training examples. Following the procedures 1-8 in Algorithm 1, we process each mixture in training set 1 using each of the six speech enhancers, i.e. spectral subtraction, Wiener filtering, MMSE, IMCRA, KLT-based subspace and a four-layer DNN. These procedures yield a comprehensive training set (denoted as "training set 2"), which comprises 560,000 (=80,000×(1+6)) training examples. Note that the DNN enhancer is trained with a different training set including 320,000 mixtures, which is created following the same procedure that generates training set 1. Moreover, we simulate a test set including 150×3 mixtures, which are created from 25×6 utterances of 6 untrained speakers. Three different SNRs are used for the test set, i.e. -5, 0 and 5 dB.

In this study, all signals are sampled at 16 kHz. A 20-ms Hamming window is employed to segment the signals into a set of time frames, with a 50% overlap between adjacent time frames. We use 161-dimensional spectra, which corresponds to a 320-point short-time Fourier transform (STFT) (16 kHz×20 ms). We train all models using the AMSGrad optimizer [16] with a learning rate of 0.001. The MSE is used as the objective function. The minibatch size is set to 8 at the utterance level. Within a minibatch, all training examples are zero-padded to have the same number of time steps as the longest

Table 3. STOI and PESQ evaluations on two unseen conventional speech enhancers.

Metrics	STOI (in %)			PESQ		
SNR	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
Unprocessed	57.84	69.80	81.06	1.49	1.79	2.12
LSTM1	72.82	84.98	91.57	1.88	2.39	2.80
LSTM2	**73.80**	**85.28**	**91.67**	**1.92**	**2.39**	**2.79**
CRN1 [17]	73.66	84.92	91.53	1.90	2.36	2.76
CRN2 [17]	**73.74**	**85.30**	**91.81**	**1.91**	**2.39**	**2.80**
Bayesian estimator [18]	53.16	66.45	78.56	1.58	1.95	2.33
Bayesian estimator - LSTM1	43.13	55.61	73.13	1.17	1.65	2.33
Bayesian estimator - LSTM2	**68.72**	**81.40**	**89.35**	**1.80**	**2.36**	**2.82**
Bayesian estimator - CRN1	48.81	60.68	75.14	1.05	1.44	2.08
Bayesian estimator - CRN2	**69.97**	**82.36**	**90.04**	**1.81**	**2.38**	**2.86**
Log-MMSE estimator [5]	53.75	66.98	79.09	1.52	1.89	2.26
Log-MMSE estimator - LSTM1	49.77	63.29	78.74	1.35	2.02	2.64
Log-MMSE estimator - LSTM2	**71.05**	**83.60**	**90.76**	**1.87**	**2.40**	**2.84**
Log-MMSE estimator - CRN1	53.31	65.52	79.23	1.25	1.69	2.31
Log-MMSE estimator - CRN2	**71.39**	**83.93**	**91.21**	**1.85**	**2.41**	**2.86**

example.

3.2. Experimental results and analysis

We first investigate the performance of two LSTM models on both original unprocessed noisy speech and enhanced speech produced by another speech enhancer. Specifically, both LSTM models have four stacked LSTM hidden layers with 1024 units in each layer, and a fully connected layer is used to estimate the spectral magnitude of clean speech, with a softplus activation function [19]. These two LSTMs with the same architecture are trained on training sets 1 and 2, respectively, which are denoted as "LSTM1" and "LSTM2".

Table 1 presents short-time objective intelligibility (STOI) [20] and perceptual evaluation of speech quality (PESQ) [21] results for different approaches at different SNR levels. The numbers represent the averages over all test examples on the two test noises at each SNR. Note that "Method 1 - Method 2" indicates that the unprocessed mixtures are successively processed by "Method 1" and "Method 2". It can be observed that the five conventional speech enhancers severely degrade the enhancement performance of LSTM1, particularly in low-SNR conditions. For example, "LSTM1" improves STOI from 57.84% to 72.82% and PESQ from 1.49 to 1.88 over the unprocessed mixtures at -5 dB. If the noisy mixtures are first processed by an MMSE estimator prior to being enhanced by LSTM1, LSTM1 yields significantly lower STOI and PESQ, i.e. 55.55% and 1.41, which are even lower than those of unprocessed mixtures. It should be pointed out that most conventional speech enhancement algorithms perform poorly on nonstationary noises such as the babble and cafeteria noises used in this study, especially when the SNR is low (e.g. -5 dB). In addition, Table 1 shows that LSTM1 is sensitive to artifacts introduced by another DNN-based speech enhancer. At -5 dB, for example, "DNN mapping - LSTM1" produces a STOI score of 68.78% and a PESQ score of 1.69, which are significantly lower than those yielded by "LSTM1" (i.e. 72.82% and 1.88). Moreover, we find that conventional speech enhancement algorithms can be susceptible to processing artifacts as well. For example, "MMSE-KLT" underperforms both "MMSE estimator" and "KLT-based subspace", as shown in Table 2.

In addition, the LSTM trained by our proposed training strategy, i.e. LSTM2, is far more robust than LSTM1 against artifacts introduced by the six speech enhancers. In other words, the enhancement artifacts induced by a preceding speech enhancer leads to far slighter or no performance degradation for LSTM2. For example, "Wiener filtering - LSTM2" yields a STOI score of 72.50% and a PESQ score of 1.90 at -5 dB, which are marginally lower than those produced by "LSTM2", i.e. 73.80% and 1.92. It should be noted that LSTM2 performs even slightly better than LSTM1 on unprocessed mixtures. A possible interpretation is that training with enhanced speech produced by different speech enhancers tends to regularize the LSTM model, which improves its generalization capability to the untrained speakers and the untrained noises for testing.

Table 4. STOI and PESQ evaluations on an unseen deep learning based speech enhancer.

Metrics	STOI (in %)			PESQ		
SNR	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
Unprocessed	57.84	69.80	81.06	1.49	1.79	2.12
CRN1 [17]	73.66	84.92	91.53	1.90	2.36	2.76
CRN2 [17]	**73.74**	**85.30**	**91.81**	**1.91**	**2.39**	**2.80**
RI-CRN1 [22]	76.82	87.26	93.20	2.00	2.52	2.95
RI-CRN2 [22]	**77.13**	**88.09**	**93.50**	**2.04**	**2.56**	**2.96**
LSTM masking	71.37	82.60	89.81	1.84	2.48	2.89
LSTM masking - CRN1	72.14	84.29	91.09	1.86	2.39	2.79
LSTM masking - CRN2	**72.80**	**85.13**	**91.66**	**1.86**	**2.43**	**2.85**
LSTM masking - RI-CRN1	72.88	85.67	91.97	1.84	2.48	2.89
LSTM masking - RI-CRN2	**76.72**	**87.81**	**93.14**	**2.00**	**2.58**	**2.98**

We now investigate the generalization to unseen speech enhancers. We additionally train a newly-developed CRN in [17] for evaluation. Akin to the LSTM models, "CRN1" is trained on training set 1 and "CRN2" on training set 2. The LSTM and CRN models are evaluated on enhanced speech produced by two unseen speech enhancers, i.e. a Bayesian estimator based on weighted Euclidean distortion measure [18] and a Log-MMSE estimator [5]. The STOI and PESQ comparisons are shown in Table 3. We can observe that the performance of both LSTM1 and CRN1 is severely degraded by the two preceding speech enhancers. This performance degradation can be considerably reduced by our proposed training strategy, as shown in Table 3. Take, for example, the CRN models. "Bayesian estimator - CRN1" yields a STOI score of 48.81% and a PESQ score of 1.05, which are far lower than those produced by "CRN1". Going from "Bayesian estimator - CRN1" to "Bayesian estimator - CRN2" improves STOI by 21.16% and PESQ by 0.76.

We further investigate the generalization to an unseen deep learning based speech enhancer. Aside from the CRN in [17], we use another CRN (denoted as "RI-CRN") that learns a complex spectral mapping, which is developed in [22]. We train a four-layer LSTM that estimates the ideal ratio mask (IRM). Note that this LSTM is trained on a training set different from training sets 1 and 2, akin to "DNN mapping" in Table 1. Both CRN and RI-CRN are evaluated on enhanced speech by this LSTM speech enhancer (denoted as "LSTM masking"). As shown in Table 4, our proposed training strategy consistently improves the robustness of CRN and RI-CRN against the processing artifacts introduced by the LSTM enhancer. At -5 dB SNR, for example, "LSTM masking - RI-CRN2" improves STOI by 3.84% and PESQ by 0.16 over "LSTM masking - RI-CRN1".

4. CONCLUSION

In voice telecommunication, the performance of speech enhancement can severely degrade if we enhance the speech signal twice. In this study, we have examined this problem and proposed a new training strategy for the downstream speech enhancer in the receiver device. Our experimental results show that the proposed training strategy substantially elevate the robustness of deep learning based speech enhancement systems against processing artifacts induced by another speech enhancer. In addition, we find that the models trained by the proposed strategy generalize well to two new conventional speech enhancers and a new deep learning based speech enhancer.

5. REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 1979, vol. 4, pp. 208–211.

[2] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[3] P. Scalart and J. Vieira Filho, "Speech enhancement based on a priori signal to noise estimation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 1996, vol. 2, pp. 629–632.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[6] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, 2002.

[7] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[8] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[9] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[10] D. L. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE press, 2006.

[11] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[12] 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, no. 10, pp. 1702–1726, 2018.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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

[15] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the Workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[16] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *International Conference on Learning Representations*, 2018.

[17] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Interspeech*, 2018, pp. 3229–3233.

[18] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, 2005.

[19] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[20] 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.

[21] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2001, vol. 2, pp. 749–752.

[22] K. Tan and D. L. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6865–6869.