Complex Spectral Mapping with a Convolutional Recurrent Network for Monaural Speech Enhancement

THE OHIO STATE UNIVERSITY

Ke Tan and DeLiang Wang, *Perception and Neurodynamics Lab* (PNL)

Summary

Phase is important for perceptual quality in speech enhancement. However, it seems intractable to directly estimate phase spectrogram through supervised learning due to lack of spectrotemporal structure in phase spectrogram. Complex spectral mapping aims to estimate the real and imaginary spectrograms of clean speech from those of noisy speech, which simultaneously enhances magnitude and phase responses of noisy speech. In this paper, we propose a new convolutional recurrent network (CRN) for complex spectral mapping, which leads to a causal system for noise- and speaker-independent speech enhancement. In terms of STOI and PESQ, the proposed CRN significantly outperforms an existing convolutional neural network (CNN) for complex spectral mapping, as well as a strong CRN for magnitude spectral mapping.

1. Introduction

- **Phase enhancement for speech enhancement**
  - Typical speech enhancement systems enhance only the magnitude spectrogram and use the noisy phase spectrogram to reconstruct the time-domain waveform.
  - A recent study (Paliwal *et al.*, 2011) shows that considerable improvements in both objective and subjective speech quality can be achieved by accurate phase spectrum estimation.
  - Various phase enhancement algorithms for speech separation have been developed, whereas they do not address the magnitude spectrum.
- **Complex spectral mapping**
  - It was found that both real and imaginary components of the clean speech spectrogram show clear spectrotemporal structure and are thus amenable to supervised learning (Williamson *et al.*, 2016).
  - Based on this observation, complex ratio masking (cRM) was developed, which yields better objective intelligibility than ideal ratio mask (IRM) estimation.
  - In a more recent study (Fu *et al.*, 2017), a CNN was employed to estimate the clean real and imaginary spectra from the noisy ones, as known as complex spectral mapping.
- **Convolutional recurrent network**
  - Motivated by our recent work (Tan and Wang, 2018) on CRNs, we propose a new CRN architecture to perform complex spectral mapping for speech enhancement. This CRN architecture additionally incorporate a newly-developed grouping strategy to reduce the number of trainable parameters and the computational cost.
  - The proposed CRN substantially outperforms an existing CNN for complex spectral mapping in terms of STOI and PESQ. Moreover, we find that complex spectral mapping consistently outperforms magnitude spectral mapping, complex ratio masking, and complex ratio masking based signal approximation.

2. Method

- **Training targets**
  - Target complex spectrum (TCS) is used as the training target in this study, which can reconstruct clean speech.
  - In addition, we extend signal approximation (SA) (Huang *et al.*, 2014), which performs masking by minimizing the difference between the spectral magnitude of clean speech and that of estimated speech. The loss for cRM-based signal approximation (cRM-SA) is defined as $SA = |cRM \times Y - S|^2$, where Y and S denote the spectrograms of noisy speech and clean speech, respectively.

- **CRN-based complex spectral mapping**
  - We have recently developed a CRN for spectral mapping, which benefits from the feature extraction capability of CNNs and the temporal modeling capability of recurrent neural networks (RNNs) (Tan and Wang, 2018).
  - The CRN is essentially an encoder-decoder architecture, as shown in Fig. 1. Specifically, the encoder comprises five convolutional layers, and the decoder five deconvolutional layers. Between the encoder and the decoder, two long short-term memory (LSTM) layers are used to model the temporal dependencies. Additionally, skip connections are utilized to concatenate the output of each encoder layer to the input of the corresponding decoder layer.
  - In this study, we extend this architecture to perform complex spectral mapping, as illustrated in Fig. 2. The encoder and the LSTM layers are shared across the estimates of real and imaginary components, while two distinct decoder modules are employed to estimate real and imaginary spectrograms, respectively.
  - The design of such an architecture is inspired by multi-task learning, in which multiple related prediction tasks are jointly learned with information shared across the tasks. For complex spectral mapping, the estimation of the real component and that of the imaginary component can be considered as two related subtasks.

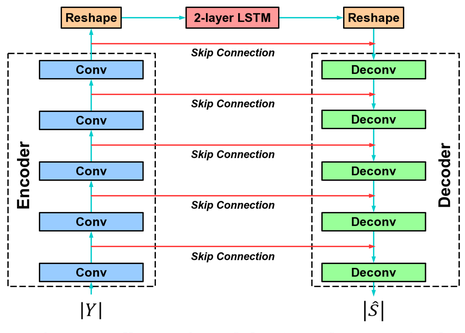


Figure 1. Illustration of the CRN for magnitude spectral mapping (Tan and Wang, 2018).

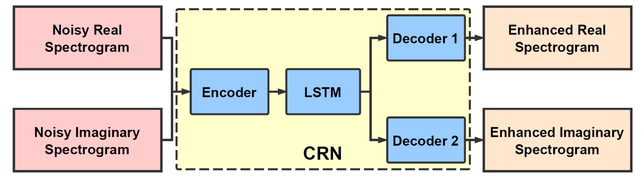
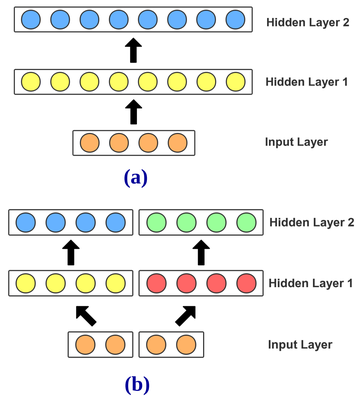


Figure 2. Network architecture of our proposed CRN for complex spectral mapping.

- **Model complexity reduction via a grouping strategy**
  - We adopt a newly-developed grouping strategy to improve the efficiency of recurrent layers while maintaining their performance (Gao *et al.*, 2018). This grouping strategy is illustrated in Fig. 3.
  - In a recurrent layer, both input features and hidden states are split into disjoint groups, and intra-group features are learned separately within each group. Thus the model complexity is substantially reduced by the grouping operation.
  - The inter-group dependency, however, cannot be captured. In other words, an output only depends on the input in the corresponding feature group, which significantly degrades the representation power.
  - To mitigate this problem, a parameter-free representation rearrangement layer between two consecutive recurrent layers is used to rearrange the features and hidden states, so that the inter-group correlations are recovered.



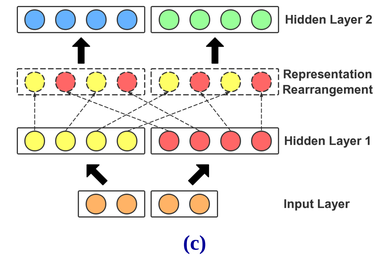


Figure 3. Illustration of the grouping strategy for RNNs.

3. Experimental results & comparisons

- **Corpus**
  - WSJ0 SI-84 training set including 7138 utterances from 83 speakers. Among these speakers, 6 speakers (3 males and 3 females) are treated as untrained speakers. Hence, we train the models with the 77 remaining speakers.
  - 10,000 training noises from a sound effect library. Two testing noises (babble and cafeteria) from an Auditec CD.
  - We create a training set including 320,000 mixtures with a total duration of about 500 hours, as well as a testing set for each noise using 6 untrained speakers.
- **Evaluation metrics: short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ)**
- **Comparisons of approaches**

metrics	STOI (in %)			PESQ		
SNR	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
unprocessed	57.94	70.01	81.20	1.51	1.80	2.12
LSTM + TMS	74.84	85.66	91.57	1.97	2.43	2.81
CRN + TMS [14]	76.28	86.15	91.92	2.02	2.46	2.83
CRN + cIRM (K=2)	74.83	86.05	91.99	1.94	2.44	2.85
CRN + cRM-SA (K=2)	77.73	88.44	93.56	2.03	2.56	2.96
CNN + TCS [11]	66.42	80.39	87.91	1.64	2.11	2.47
CRN + TCS (K=1)	80.12	89.68	94.03	**2.19**	**2.70**	**3.07**
CRN + TCS (K=2)	**80.14**	**89.84**	94.15	2.17	2.68	3.05
CRN + TCS (K=4)	80.01	89.78	**94.21**	2.18	2.69	**3.07**
CRN + TCS (K=8)	78.63	89.06	93.83	2.15	2.67	3.05

Table 1. Comparisons of different models and training targets in STOI and PESQ metrics on untrained noises and untrained speakers. Note that K denotes the group number in LSTM layers.

metrics	STOI (in %)			PESQ		
SNR	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
unprocessed	57.94	70.01	81.20	1.51	1.80	2.12
noisy phase	76.28	86.15	91.92	2.02	2.46	2.83
estimated phase	78.49	88.72	93.73	2.14	2.65	3.02
clean phase	80.83	90.34	94.86	2.35	2.85	3.22

Table 2. Evaluation of phase estimation provided by complex spectral mapping.

4. Conclusion

- As shown in Table 1, our proposed CRN model significantly outperforms an existing CNN for complex spectral mapping. Additionally, complex spectral mapping consistently outperforms magnitude spectral mapping, as well as complex ratio masking and complex ratio masking based signal approximation.
- Moreover, our complex spectral mapping provides an effective phase estimate, as shown in Table 2. Thus it avoids the use of the noisy phase.