Real-time Speech Enhancement for Mobile Communication Based on Dual-channel Complex Spectral Mapping

Ke Tan[1], Xueliang Zhang[2] and DeLiang Wang[1]
[1]The Ohio State University, United States
[2]Inner Mongolia University, China

OUTLINE

OUTLINE

- In mobile communication, speech quality and intelligibility can be severely degraded by background noise, when the far-end talker is in a noisy environment.

- Speech enhancement algorithms have been integrated into most mobile phones. In a typical dual-microphone configuration, a primary microphone is placed on the bottom of a mobile phone and a secondary microphone on the top.

Fig. 1: Illustration of a dual-microphone mobile phone.

- Real-time speech enhancement is needed for mobile communication.

- Several requirements on model design:
  - the model should use no or few future time frames;
  - the model should not have a high computational cost for the sake of processing latency;
  - memory consumption should fit the capacity of mobile phones.

Motivations

- Inspired by recent advances in complex-domain speech enhancement [1, 2, 3], we develop a new densely-connected convolutional recurrent network (DC-CRN) to perform dual-channel complex spectral mapping.

- In addition, we propose a structured pruning technique to compress the DC-CRN, which substantially reduces the model size without significantly degrading the enhancement performance.

[1] 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.
[2] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in IEEE 27th International Workshop on Machine Learning for Signal Processing. IEEE, 2017, pp. 1–6.
[3] K. Tan and D. L. 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, 2020.

OUTLINE

Densely-connected CRN

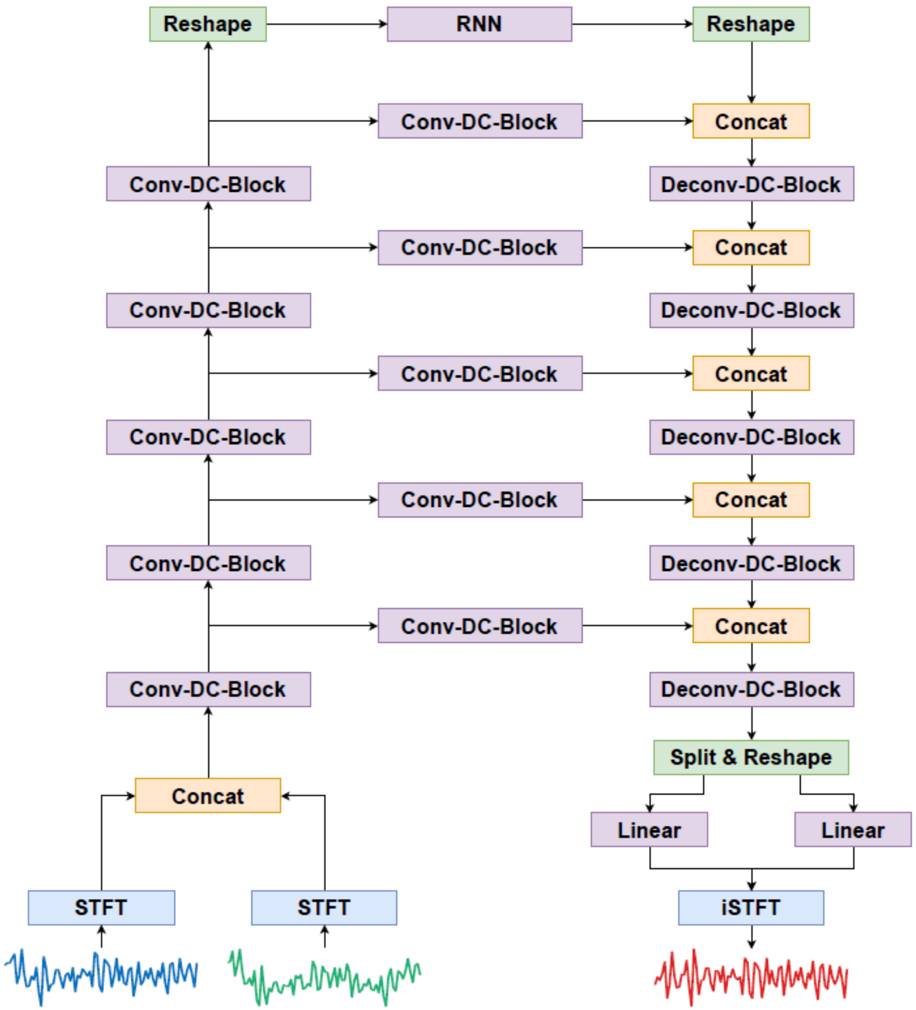


Fig. 2. Diagram of the DC-CRN.

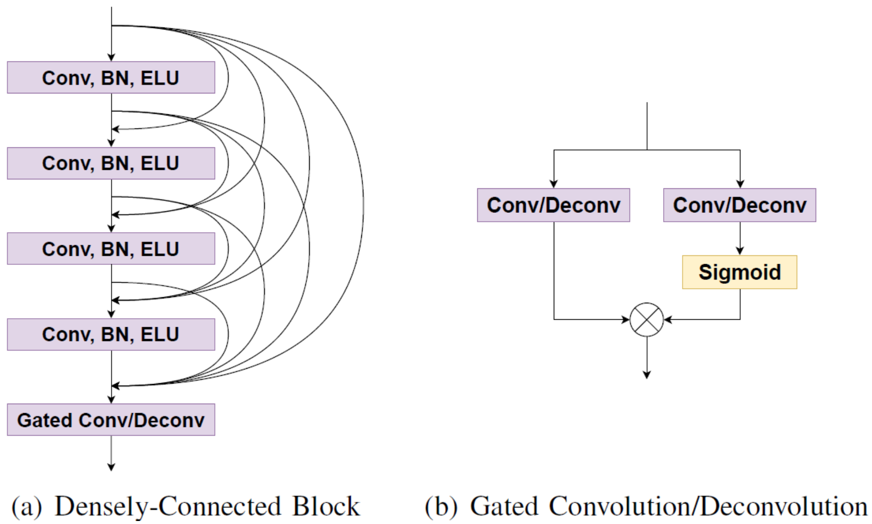


Fig. 3. Diagrams of the densely-connected block (a) and the gated convolution/deconvolution (b).

- We train the DC-CRN to perform dual-channel complex spectral mapping with a loss function as follows:

$$\begin{aligned}\mathcal{L}_{\text{RI+Mag}} =& \frac{1}{M \cdot F} \sum_{m,f} \left| \hat{S}_1^{(r)}(m,f) - S_1^{(r)}(m,f) \right| \\ &+ \left| \hat{S}_1^{(i)}(m,f) - S_1^{(i)}(m,f) \right| \\ &+ \left| |\hat{S}_1(m,f)| - |S_1(m,f)| \right|,\end{aligned}$$

$$\left| \hat{S}_1(m,f) \right| = \sqrt{\hat{S}_1^{(r)}(m,f)^2 + \hat{S}_1^{(i)}(m,f)^2}$$

- Noncausal DC-CRN:
  - a reasonably large number of trainable parameters (~8M)
  - using bidirectional LSTM for sequential modeling

- Causal DC-CRN:
  - a relatively small number of trainable parameters (~290K)
  - using unidirectional LSTM for sequential modeling

- The causal DC-CRN is still not amenable to the capacity of most mobile phones.

- We propose a structured pruning technique to compress the causal DC-CRN, without significantly sacrificing the enhancement performance.
- Structured pruning is a class of coarse-grained parameter pruning techniques, and it leads to more regular sparsity patterns than unstructured pruning. For example, structured pruning can remove an entire column of a weight matrix, unlike unstructured pruning that prunes individual weights.
- The regularity of sparse structure makes it easier to apply hardware acceleration.
- To achieve a high compression rate, we adopt a group sparse regularization [4] technique to impose the group-level sparsity of the weight matrices or tensors.

[4] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini, "Group sparse regularization for deep neural networks," Neurocomputing, vol. 241, pp. 81–89, 2017.

Iterative Structured Pruning

- To determine the pruning ratio for each layer, we perform a per-layer sensitivity analysis.

- Subsequently, we perform group-level pruning as per layer-wise pruning ratios, and then fine-tune the pruned model.

- This procedure is repeated until the number of pruned weights becomes trivial in an iteration or a significant degradation of STOI or PESQ is observed on a validation set.

OUTLINE

1. Background and Motivations

2. Algorithm Description

3. Experiments

4. Conclusion

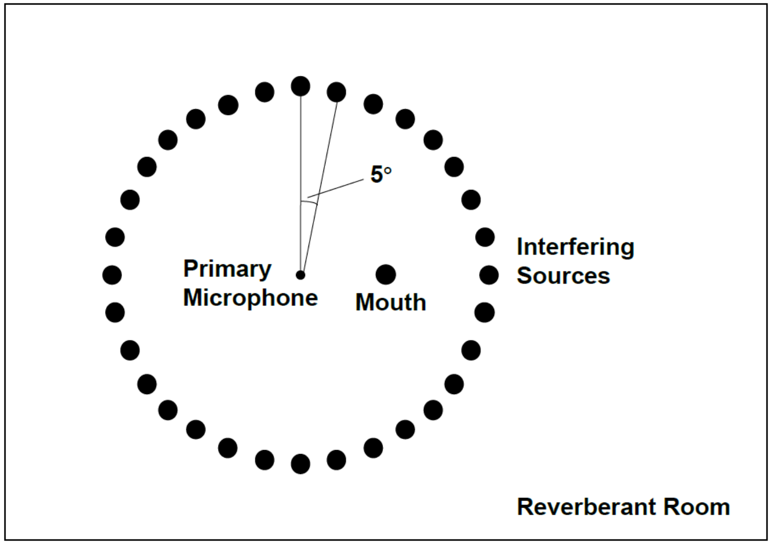
- Speech corpus: training set of WSJ0, including 12776 utterances from 101 (= 89 + 6 + 6) speakers.

- We simulate a rectangular room with a size of $10 \times 7 \times 3$ m^3 using the image method. The target source (mouth) is at the center of the room. The primary microphone is placed on a sphere centered at the target source, with a radius randomly sampled between 0.01 m and 0.15 m.

- The distance between microphones is fixed to 0.1 m. Thus the location of the secondary microphone is randomly chosen on a sphere with a radius of 0.1 m, centered at the primary microphone.

- The reverberation time (T_{60}) is randomly sampled between 0.2 s and 0.5 s.

- We simulate a diffuse babble noise in the following way.
  - concatenate the utterances spoken by each of the 630 speakers in the TIMIT corpus, and then split them into 480 and 150 speakers for training and testing.
  - randomly select 72 speech clips from 72 randomly chosen speakers, and place them on a horizontal circle centered at and with the same height as the primary microphone, where the azimuths range from 0 to 355 degrees with a step of 5 degrees.



The distance between the primary microphone and each of the interfering sources is 2 m.

- In order to mimic the head shadow effect, we downscale the amplitude of the speech signal at the secondary channel prior to mixing, where the downscaling ratio is randomly sampled between -10 and 0 dB.

- For both training and validation data, the SNR is randomly sampled between -5 and 0 dB, where the SNR is with respect to the reverberant speech signal and the reverberant noise signal at the primary channel. We create a test set consisting of 846 mixtures for each of four SNRs, i.e. -5, 0, 5 and 10 dB.

Results

Table 1. Comparisons of alternative models in STOI and PESQ. Here ✓ indicates causal model, and ✗ indicates noncausal model.

Test SNR	-5 dB		0 dB		5 dB		10 dB		# Param.	Causal
Metric	STOI (%)	PESQ	STOI (%)	PESQ	STOI (%)	PESQ	STOI (%)	PESQ		
Unprocessed	58.71	1.49	72.08	1.73	83.53	2.04	91.41	2.38	-	-
NC-CRN-PSM	85.48	2.20	90.79	2.60	93.82	2.93	95.47	3.17	12.99 M	✗
NC-DC-CRN-RI	**92.77**	**3.07**	**96.09**	**3.41**	**97.66**	**3.63**	**98.45**	**3.78**	8.36 M	✗
IRM	92.02	2.83	94.21	3.10	96.24	3.39	97.74	3.68	-	-
PSM	94.08	3.16	96.26	3.40	97.87	3.66	98.87	3.88	-	-
C-CRN-PSM	78.77	1.76	86.80	2.18	91.53	2.56	94.05	2.88	73.15 K	✓
C-DC-CRN-RI	**87.57**	**2.56**	**93.36**	**2.99**	**96.35**	**3.30**	**97.74**	**3.53**	290.44 K	✓
C-DC-CRN-RI-P1	86.88	2.54	93.08	2.97	96.16	3.26	97.63	3.46	124.96 K	✓
C-DC-CRN-RI-P2	87.13	2.56	93.10	2.98	96.14	3.27	97.62	3.47	113.68 K	✓
C-DC-CRN-RI-P3	86.64	2.52	92.89	2.95	96.07	3.26	97.61	3.47	108.77 K	✓
C-DC-CRN-RI-P4	86.63	2.49	92.85	2.91	96.03	3.22	97.59	3.44	106.21 K	✓
C-DC-CRN-RI-P5	86.63	2.48	92.86	2.90	96.07	3.20	97.65	3.43	104.76 K	✓
C-DC-CRN-RI-P6	86.45	2.51	92.64	2.94	95.88	3.27	97.47	3.51	103.07 K	✓

[5] K. Tan, X. Zhang, and D. L. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2019, pp. 5751–5755.

Table 2. Effects of dense connectivity at -5 dB SNR.

Test SNR	-5 dB			# Param.
Metric	STOI (%)	PESQ	SNR (dB)	
Unprocessed	58.71	1.49	-5.03	-
C-DC-CRN-RI	**87.57**	**2.56**	**8.61**	290.44 K
− DC_{Skip} (i)	87.23	2.53	8.49	253.32 K
− DC_{ED} (ii)	86.26	2.42	8.02	218.69 K
− DC_{Skip} − DC_{ED} (iii)	82.77	2.10	6.37	181.57 K

Table 3. Investigation of inter-channel features for magnitude- and complex-domain approaches. "ICFs" represent the inter-channel features.

Test SNR	-5 dB			Domain
Metric	STOI (%)	PESQ	SNR (dB)	
Unprocessed	58.71	1.49	-5.03	-
C-CRN-PSM w/ ICFs	78.77	1.76	5.13	Magnitude
C-CRN-PSM w/o ICFs	76.14	1.67	4.56	Magnitude
C-DC-CRN-RI w/ ICFs	87.64	2.56	8.44	Complex
C-DC-CRN-RI w/o ICFs	87.44	2.56	8.61	Complex

Demos

Noncausal:

Unprocessed (-5 dB):

NC-CRN-PSM:

NC-DC-CRN-RI:

IRM:

Clean:

Causal:

Unprocessed (-5 dB):

C-CRN-PSM:

C-DC-CRN-RI-P6:

IRM:

Clean:

OUTLINE

1. Background and Motivations

2. Model Description

3. Experiments

4. Conclusion

Conclusion

- In this study, we have proposed a novel framework for dual-channel speech enhancement on mobile phones, which employs a new causal DC-CRN to perform dual-channel complex spectral mapping.

- By applying an iterative structured pruning technique, we derive a low-latency and memory-efficient enhancement system, which is amenable to real-time processing on mobile phones.

- Evaluation results demonstrate that the proposed approach significantly outperforms a previous method for dual-channel speech enhancement.