DCCRN:
Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement

Yanxin Hu, Yun Liu

Interspeech 2020

2023. 03. 04. (SAT)

Younghoo Kwon

Ohoo

Content

1. Introduction
2. The DCCRN model
3. Experiments
4. Conclusions

Ohoo

1. Introduction

Ohoo

Related work

- Noisy speech can be enhanced by neural networks either time-frequency(TF) domain or directrly in time-domain.
- Time-domain approachs
  - Direct regression: 1-D conv without an explicit signal front-end
  - Adaptive front-end approach: convolution encoder-decoder(CED) or u-net taking time-domain signal in and out with STFT and iSTFT. The enhancement network is inserted between the CED.
- TF-domain approachs
  - Work on the spectrogram with the belief that fine-detailed structures of speech and noise can be separable with TF representations after STFT.
- Convolution recurrent network(CRN) is recent approach that also employs a CED structure similar to the one in the time-domain approaches but extracts high-level features for better separation by 2-D CNN from noisy speech spectrogram.
- A complex-valued spectrogram can be decomposed into magnitude and phase in polar coordinate or real and imaginary part in Cartesian coordinate

Related work

- ☐ Early studies only focus on magnitude so, there exists the upper bound of performance. Also, the neural network remains real-valued.
- ☐ Training targets defined in the TF domain mainly fall into two groups of targets.
  - masking-based targets: masks describe the time-frequency relationships between clean speech and background noise
  - Mapping based targets correspond to the spectral representations of clean speech.
- ☐ Ideal binary mask(IBM), ideal ratio mask(IRM), spectral magnitude mask(SMM) use magnitude only.
- ☐ Phase-sensitive mask(PSM), complex ratio mask(CRM) uses both of magnitude and phase or real and imaginary values
- ☐ A CRN with one encoder and two decoders for complex spectral mapping(CSM) is proposed. [24]
- ☐ CRM(complex ratio mask) and CSM(complex spectral mapping) possess the full information of the speech signal.

Ohoo

Related work

- Deep complex u-net has combined the advantages of both a deep complex network and a u-net to deal with complex-valued spectrogram.
- DCUNET is trained to estimate CRM and optimize the scale-invariant source-to-noise ratio(SI-SNR) loss.
- SI-SNR loss is calculated by transforming the output TF-domain spectrogram to a time-domain waveform by iSTFT.

Contributions

- The *deep complex convolution recurrent network*(DCCRN) is created by combining the advantages of DCUNET and CRN, using LSTM to model temporal context.
- DCCRN optimizes an SI-SNR loss.
- Various training targets are tested under DCCRN framework and the best performance can be obtained by the complex network with the complex target.
- DCCRN outperforms CRN by a large margin and achieves competitive performance with DCUNET with 1/6 computation complexivity.
- With only 3.7M parameters, DCCRN achieves the best MOS in real-time track and the second-best in non-real time track according to the P.808 subjective evaluation in the DNS challenge.

2. The DCCRN model

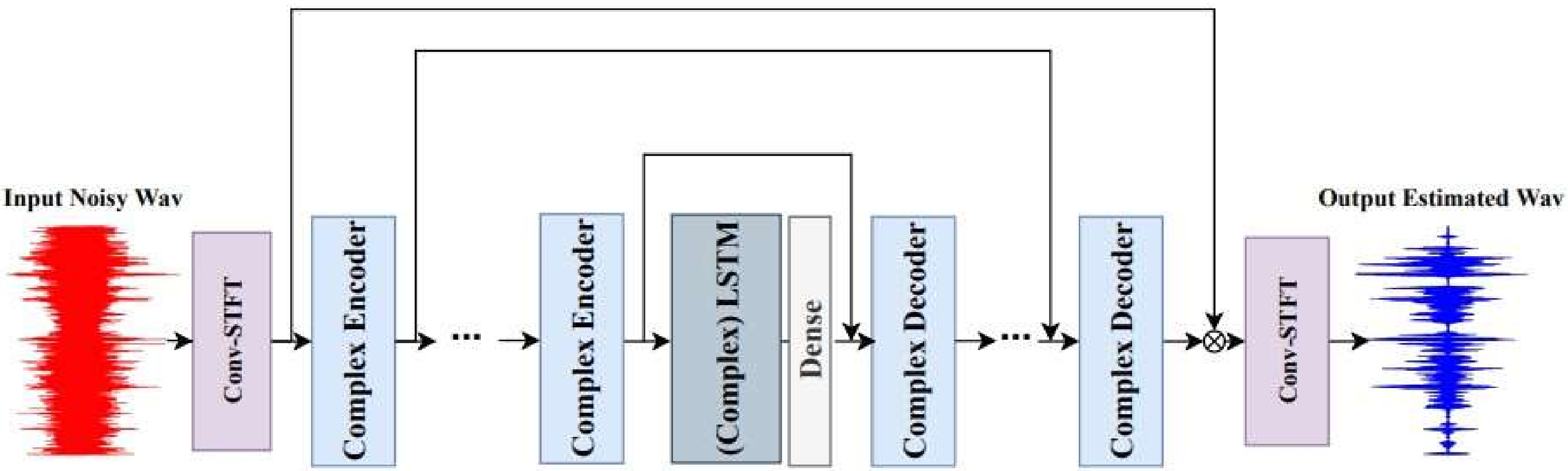
Ohoo

Convolution recurrent network architecture

- The convolution recurrent network(CRN) [14] is an essentially causal CED architecture with two LSTM layers between the encoder and decoder.
  - The encoder consists of five Conv2d block aiming extracting high-level features from the input features, or reducing the resolution.
  - The decoder reconstructs the low-resolution features to the original size of input.
  - The CED is composed of convolution/deconvolution layer followed by batch normalization and activation function in a symmetric design.
  - The LSTM is specifically used to model the temporal dependencies.
- The complex spectral mapping [24] does not model only magnitude but the real and imaginary parts of complex STFT spectrogram from the input mixture to the clean speech with one encoder and two decoders.
- However, it treats real and imaginary parts as two input channels
  - It only applies real-valued convolution operation with one shared real-valued convolution filter.

Convolution recurrent network architecture

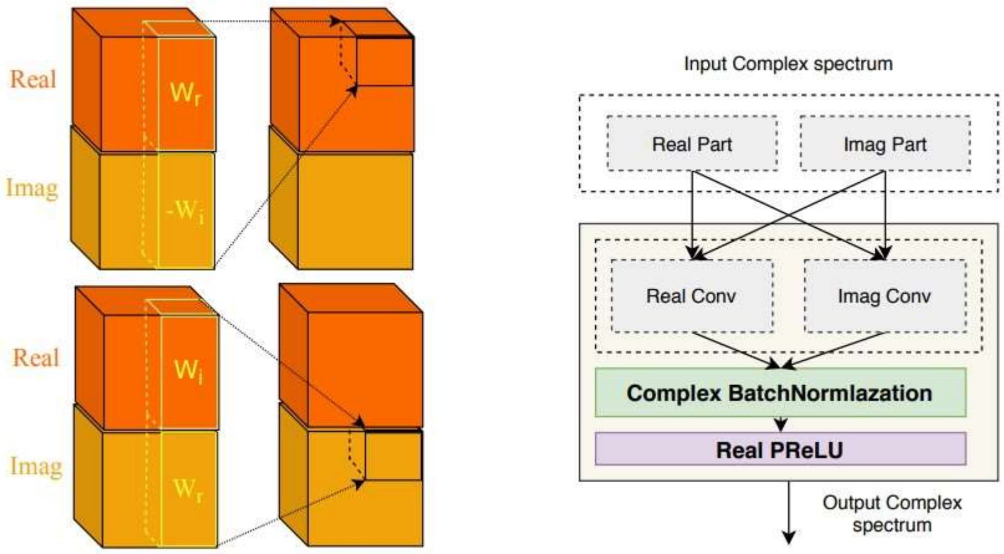
- DCCRN modifies CRN substantially with complex CNN and complex BN in CED and complex LSTM with the prior knowledge of complex multiplication. This models the correlation between magnitude and phase.



DCCRN network

Encoder and decoder with complex network

- The complex encoder block includes complex Conv2d, complex batch normalization [26] and real-valued PReLU [28].
  - Complex Conv2d block is from the one in DCUNET [25] and consists of four traditional Conv2d operation.



(a) complex convolution (b) complex encoder

Encoder and decoder with complex network

☐ The complex-valued convolutional filter W is defined as $W = W_r + jW_i$ where the real-valued matrics W_r and W_i represent the real and imaginary part of complex convolution kernel, respectively.

☐ The complex output Y from the complex convolution operation $X \circledast W$ with the input matrics $X = X_r + jX_i$:

$$F_{out} = (X_r * W_r - X_i * W_i) + j(X_r * W_i + X_i * W_r)$$

☐ The complex output of complex LSTM with the complex input X_r and X_i can be defined as:

$$\begin{aligned} F_{rr} &= \mathrm{LSTM}_r(X_r); \quad F_{ir} = \mathrm{LSTM}_i(X_r) \\ F_{ri} &= \mathrm{LSTM}_i(X_r); \quad F_{ii} = \mathrm{LSTM}_i(X_i) \\ F_{out} &= (F_{rr} - F_{ii}) + j(F_{ri} + F_{ir}) \end{aligned}$$

☐ F_{out} denotes the output feature of one convolution layer.

Training target

- DCCRN estimates complex ratio mask(CRM) and is optimized by signal approximation(SA).
- Given the complex-valued STFT spectrogram of clean speech S and noisy speech Y, CRM can be defined as:

$$\mathrm{CRM} = \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2} + j\frac{Y_r S_i - Y_i S_r}{Y_r^2 + Y_i^2}$$

- For comparison, magnitude target(SMM) can be considered. $(\mathrm{SMM} = |S|/|Y|)$
- Singal approximation(SA) directly minimizes the difference between magnitude or complex spectrogram of clean speech and that of noisy speech applied with mask.
  - CRM-based SA: $\mathrm{CSA} = Loss(\widetilde{M} \cdot Y, S)$
  - SMM-based SA: $MSA = Loss(|\widetilde{M}| \cdot |Y|, |S|)$
- The Cartesian coordinate representation of mask $\widetilde{M} = \widetilde{M_r} + j\widetilde{M_i}$ can also be expressed in polar coordinates:

$$\widetilde{M}_{\mathrm{mag}} = \sqrt{\widetilde{M}_r^2 + \widetilde{M}_i^2}$$
$$\widetilde{M}_{\mathrm{phase}} = \mathrm{arctan2}(\widetilde{M}_i, \widetilde{M}_r)$$

Training target

- Three multiplicative patterns for DCCRN are proposed like below:

$$\mathrm{DCCRN-R:}\ \tilde{S} = \left(Y_r \cdot \widetilde{M}_r\right) + j\left(Y_i \cdot \widetilde{M}_i\right)$$

$$\mathrm{DCCRN-C:}\ \tilde{S} = \left(Y_r \cdot \widetilde{M}_r - Y_i \cdot \widetilde{M}_i\right) + j\left(Y_r \cdot \widetilde{M}_i + Y_i \cdot \widetilde{M}_r\right)$$

$$\mathrm{DCCRN-E:}\ \tilde{S} = Y_{\mathrm{mag}} \cdot \widetilde{M}_{\mathrm{mag}} \cdot e^{Y_{\mathrm{phase}} + \widetilde{M}_{\mathrm{phase}}}$$

- DCCRN-R estimates the mask of the real and imaginary parts of $\tilde{Y}$, respectively.
- DCCRN-C obtains $\tilde{S}$ in the manner of CSA.

$$\begin{aligned}
\tilde{S} &= \left(Y_r \cdot \widetilde{M}_r - Y_i \cdot \widetilde{M}_i\right) + j\left(Y_r \cdot \widetilde{M}_i + Y_i \cdot \widetilde{M}_r\right) \\
&= \left(Y_r \cdot \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2} - Y_i \cdot \frac{Y_r S_i - Y_i S_r}{Y_r^2 + Y_i^2}\right) + j\left(Y_r \cdot \frac{Y_r S_i - Y_i S_r}{Y_r^2 + Y_i^2} + Y_i \cdot \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2}\right) \\
&= S_r + jS_i
\end{aligned}$$

- DCCRN-E is mathematically similar to DCCRN-C but the only difference is that it uses $tanh$ activation function to limit the mask magnitude to 0 to 1.

Ohoo

Loss function

- The loss function of model training is SI-SNR. It is defined as:

$$\boldsymbol{s}_{\text{target}} := (< \tilde{\boldsymbol{s}}, \boldsymbol{s} > \cdot\ \boldsymbol{s}) / ||\boldsymbol{s}||_2^2$$
$$\boldsymbol{e}_{\text{noise}} := \tilde{\boldsymbol{s}} - \boldsymbol{s}_{\text{target}}$$
$$\text{SI} - \text{SNR} := 10\log10\left(\frac{||\boldsymbol{s}_{\text{target}}||_2^2}{||\boldsymbol{e}_{\text{noise}}||_2^2}\right)$$

- $<\cdot,\cdot>$ denotes the dot product between two vectors and $||\cdot||_2$ is Euclidean norm(L2 norm).

Ohoo

3. Experiments

Ohoo

Datasets

- The first datasets: WSJ0 [30] for speech and MUSAN [31] for noise
  - 20K, 3K, and 1.5K utterances for train, validation, evaluation is selected from WSJ0.
  - There exists 131 speakers (66 males and 65 females).
  - MUSAN os 42.6 hours music for training and validation and 7 hours for evaluation.
  - The speech-noise mixture in training and validation is generated by randomly selecting utterances.
  - The mixing SNR is randomly selected between -5 dB and 20 dB.
  - The evaluation set is gernerated at 5 typical SNRs (0 dB, 5 dB, 10 dB, 15 dB, 20 dB).
- The second datasets: DNS for speech and noise
  - 180 hours DNS challenge noise set includes 150 classes and 65,000 noise clips.
  - Clean speech set includes over 500 hours of clip from 2,150 speakers.
  - The speech-noise mixture is mixed with dynamic mixing during model training.
  - At each training epoch, speech and noise are rst convolved with a room impulse response(RIR) randomly-selected from a simulated 3000-RIR set by the image method [32]
  - The speech-noise mixtures are generated dynamically by mixing reverb speech and noise at random SNR between -5 dB and 20 dB.

Ohoo

Training setup and baselines

- ☐ The window length and hop size are 25 ms and 6.25 ms, and FFT length is 512.
- ☐ The optimizer is Adam.
- ☐ The initial learning rate is set to 0.001, and it will decay 0.5 when the validation loss goes up.
- ☐ All the waveforms are resampled at 16 kHz.
- ☐ The models are selected by early stopping.

- ☐ The experiments are processed with LSTM, CRN, DCCRN, and DCUNET.
- ☐ The four target patterns of DCCRN are also used (DCCRN-R, DCCRN-C, DCCRN-E, DCCRN-CL).
- ☐ The number of channel for the first three DCCRN is {32,64,128,128,256,256} and one of the last one is {32,64,128,256,256,256}.
- ☐ The kernel size and stride are set to (5,2) and (2,1) respectively.
- ☐ The real LSTMs of the first three DCCRN are two layers with 256 units and DCCRN-CL uses complex LSTM with 128 units for the real part and imaginary part, respectively.
- ☐ A dense layer with 1024 units is after the last LSTM.

Ohoo

Training setup and baselines

- Semi-causal convolution has only two differences with commonly used causal convolution in practice.
  - Zeros are padded in front of the time dimension at each Conv2ds in the encoder.
  - For decoder, one frame is looked ahead in each convolution layer.
  - This eventually leads to 6 frames look-ahead, totally $6 \times 6.25 = 37.5\ \mathrm{ms}$, confined with the DNS challenge limit – 40ms

Experimental results and discussion

- The model performance is first accessed by PESQ on the simulated WSJ0 dataset.

Model	Para.(M)	0dB	5dB	10dB	15dB	20dB	Ave.
Noisy	-	2.062	2.388	2.719	3.049	3.370	2.518
LSTM	9.6	2.783	3.103	3.371	3.593	3.781	3.326
CRN	6.1	2.850	3.143	3.374	3.561	3.717	3.329
DCCRN-R	3.7	2.832	3.192	3.488	3.717	3.891	3.424
DCCRN-C	3.7	2.832	3.187	3.477	3.707	3.840	3.409
DCCRN-E	3.7	2.859	3.203	3.492	3.718	3.891	3.433
DCCRN-CL	3.7	**2.972**	**3.301**	**3.559**	3.755	3.901	3.498
DCUNET	3.6	2.971	3.297	3.556	**3.760**	**3.916**	**3.500**

- DCCRN-CL achieves better performance than other DCCRNs. Complex LSTM is also beneficial to complex target training.
- The full-complex-value network DCCRN and DCUNET are similar in PESQ but computational complexity of DCUNET is almost 6 times than that of DCCRN-CL according to our run-time test.

Experimental results and discussion

- ☐ In the DNS chanllenge, the two best DCCRN models and DCUNET with the DNS dataset are evaluated.

Model	Para. (M)	look-ahead (ms)	no reverb	reverb	Ave.
Noisy	-	-	2.454	2.752	2.603
NSNet (Baseline) [34]	1.3	0	2.683	2.453	2.568
DCCRN-E [T1]	3.7	37.5	**3.266**	3.077	3.171
DCCRN-E-Aug [T2]	3.7	37.5	3.209	**3.219**	**3.214**
DCCRN-CL [T2]	3.7	37.5	3.262	3.101	3.181
DCUNET [T2]	3.6	37.5	3.223	2.796	3.001

- ☐ DCCRN-CL achieves a little bit better PESQ than DCCRN-E in general.
  - ☐ But, after internal subject listening, DCCRN-CL may over-suppress the speech signal on some clips.
- ☐ DCUNET obtains relatively good PESQ on synthetic non-reverb set, but its PESQ will drop significantly on the synthetic reverb set.
- ☐ Subjective listening is very critical when the objective scores are close for different systems so, DCCRN-E was finally chosen for the real-time track.
- ☐ DCCRN-E-Aug is the model which improves the performance on the reverb set.

Experimental results and discussion

- The final P.808 subjective evaluation results for several top systems in both track.

Model		Para.(M)	no reverb	reverb	realrec	Ave.
Noisy		-	3.13	2.64	2.83	2.85
NSNet (Baseline) [34]		1.3	3.49	2.64	3.00	3.03
Track 1	DCCRN-E	3.7	**4.00**	2.94	**3.37**	**3.42**
	Team 9	UNK	3.87	**2.97**	3.28	3.39
	Team 17	UNK	3.83	3.05	3.27	3.34
Track 2	Team 9	UNK	**4.07**	**3.19**	**3.40**	**3.52**
	DCCRN-E-Aug	3.7	3.90	2.96	3.34	3.38
	Team 17	UNK	3.83	3.15	3.28	3.38

- The MOS of DCCRN-E-Aug has a small improvement of 0.02 on the reverb set.
- DCCRN-E achieves an average MOS of 3.42 on all sets and 4.00 on the non-reverb set.

4. Conclusions

Conclusions

- ☐ The DCCRN model utilizes a complex network for complex-valued spectrum modeling.
- ☐ With the complex multiply rule constraint, DCCRN can achieve better performance than others in terms of PESQ and MOS in the similar configuration of model parameters.
- ☐ In the future, DCCRN in low computational scenarios will be tried and DCCRN improved noise suppression ability in reverberation conditions also can be tried.

Ohoo

END

Ohoo