Paper ID: 1775

Uformer: A Unet Based Dilated Complex & Real Dual-path Conformer Network for Simultaneous Speech Enhancement and Dereverberation

Yihui Fu[1], Yun Liu[2], Jingdong Li[2], Dawei Luo[2], Shubo Lv[1], Yukai Jv[1], Lei Xie[2]

[1] *Audio, Speech and Language Processing Group (ASLP@NPU), Northwestern Polytechnical University, Xi'an, China*

[2] *AI Interaction Division, Sogou Inc., Beijing, China*

1. Contribution

We propose Uformer to deliver single-channel speech enhancement and dereverberation in both complex and magnitude domain.

- **Dilated complex & real dual-path conformer**
  This module is applied on the bottle-neck feature between encoder and decoder. It includes feed forward (FF), time attention (TA), frequency attention (FA) and dilated convolution (DC) layers.
- **Hybrid encoder and decoder**
  This module aims to model complex spectrum and magnitude simultaneously. The rationale is that superb magnitude estimation can profit better recovery for phase and vice versa.
- **Encoder decoder attention**
  This model estimate attention mask to reveal the relevance between the corresponding hybrid encoder and decoder layers.

2. Complex Self Attention

Given the complex input $\mathbf{X}$ and learnable linear transformation $\mathbf{W}_Q$, the complex valued $\mathbf{Q}$ is calculated by:

$$
\begin{aligned}
\mathbf{Q}^{\Re} &= \mathbf{X}^{\Re}\mathbf{W}_Q^{\Re} - \mathbf{X}^{\Im}\mathbf{W}_Q^{\Im}, \\
\mathbf{Q}^{\Im} &= \mathbf{X}^{\Re}\mathbf{W}_Q^{\Im} + \mathbf{X}^{\Im}\mathbf{W}_Q^{\Re},
\end{aligned} \tag{1}
$$

where $\Re$ and $\Im$ indicate the real and imaginary parts, respectively. [1] $\mathbf{K}$ and $\mathbf{V}$ are calculated in the same way. Thus, the complex self attention is calculated by:

$$
\begin{aligned}
&\mathrm{ComplexAttention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \\
&(\mathrm{Attention}(\mathbf{Q}^{\Re},\mathbf{K}^{\Re},\mathbf{V}^{\Re}) - \mathrm{Attention}(\mathbf{Q}^{\Re},\mathbf{K}^{\Im},\mathbf{V}^{\Im}) - \\
&\mathrm{Attention}(\mathbf{Q}^{\Im},\mathbf{K}^{\Re},\mathbf{V}^{\Im}) - \mathrm{Attention}(\mathbf{Q}^{\Im},\mathbf{K}^{\Im},\mathbf{V}^{\Re})) + \\
&i(\mathrm{Attention}(\mathbf{Q}^{\Re},\mathbf{K}^{\Re},\mathbf{V}^{\Im}) + \mathrm{Attention}(\mathbf{Q}^{\Re},\mathbf{K}^{\Im},\mathbf{V}^{\Re}) + \\
&\mathrm{Attention}(\mathbf{Q}^{\Im},\mathbf{K}^{\Re},\mathbf{V}^{\Re}) - \mathrm{Attention}(\mathbf{Q}^{\Im},\mathbf{K}^{\Im},\mathbf{V}^{\Im})).
\end{aligned} \tag{2}
$$

3. Proposed Uformer

The overall architecture of Uformer is shown in Figure 1.

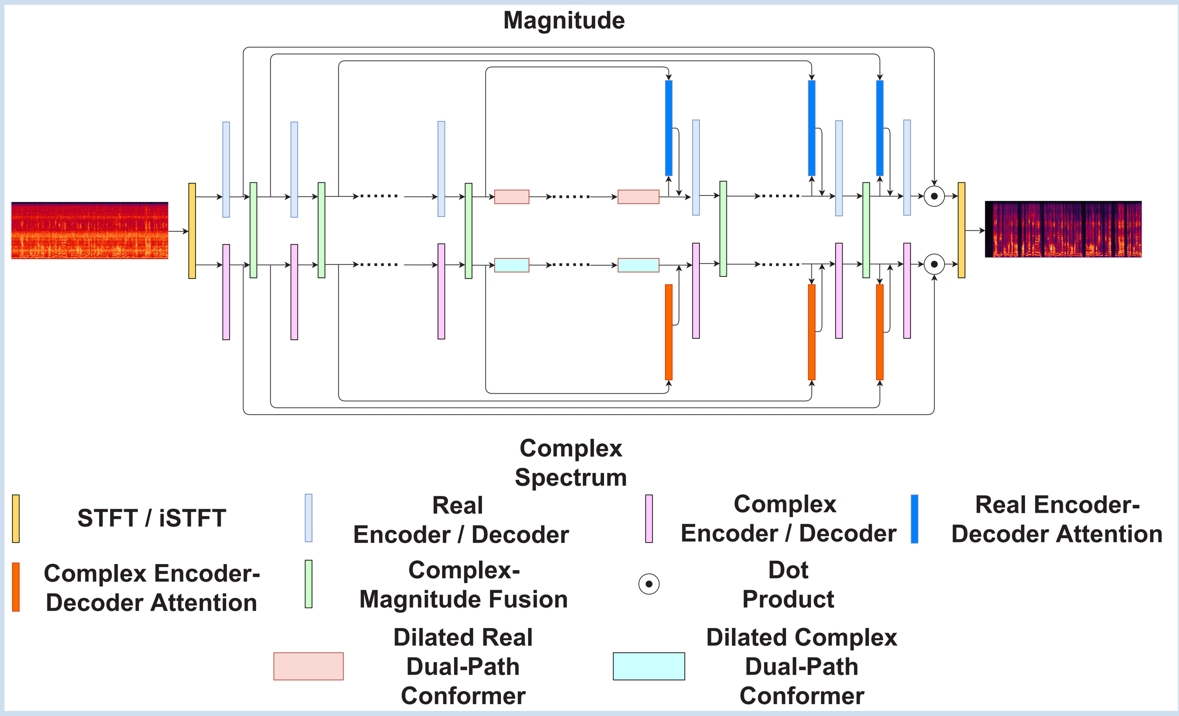

Figure 1: The overall architecture of Uformer.

- **Dilated Dual-path Conformer**

Time attention (TA): TA aims to model the local temporal relevance information. It uses the current frame, together with its frame expansion to deliver self attention along time axis.

Frequency attention (FA): FA lays different importance on different frequency bands. It delivers self attention on frequency axis.

Dilated convolution (DC): DC aims to better capture long range sequence dependencies. Gated D-Conv2d is applied with opposite dilation in two D-Conv2ds to get different scale of receptive field.

- **Hybrid Encoder and Decoder**

Both encoder and decoder model the complex spectrum and magnitude simutaneously:

$$
\begin{aligned}
\hat{\mathbf{C}}_i^{\Re} &= \mathbf{C}_i^{\Re} + \sigma(\mathbf{M}_i), \\
\hat{\mathbf{C}}_i^{\Im} &= \mathbf{C}_i^{\Im} + \sigma(\mathbf{M}_i), \\
\hat{\mathbf{M}}_i &= \mathbf{M}_i + \sigma(\sqrt{\mathbf{C}_i^{\Re^2} + \mathbf{C}_i^{\Im^2}}).
\end{aligned} \tag{3}
$$

where $\mathbf{C}_i$ and $\mathbf{M}_i$ denotes the complex spectrum and magnitude output of encoder/decoder layer i respectively.

- **Encoder Decoder Attention**

After getting the output of encoder and decoder layers, two Conv2ds are applied to generate high dimensional features $\mathbf{G}_i$:

$$\mathbf{G}_i = \sigma(\mathbf{W}_i^E * \mathbf{E}_i + \mathbf{W}_i^D * \mathbf{D}_i), \tag{4}$$

A third Conv2d is applied and sigmoid attention mask is estimated to represent the relevance between encoder and decoder:

$$\hat{\mathbf{D}}_i = \sigma(\mathbf{W}_i^A * \mathbf{G}_i) \odot \mathbf{D}_i, \tag{5}$$

Finally, we concatenate $\mathbf{D}_i$ and $\hat{\mathbf{D}}_i$ along channel axis as the input of the next decoder layer.

- **Loss Function**

We use hybrid time and frequency domain loss as the optimization function:

$$\mathcal{L} = \alpha\mathcal{L}_{\textbf{SI-SNR}} + \beta\mathcal{L}_{\textbf{L1}}^{\textbf{T}} + \gamma\mathcal{L}_{\textbf{L2}}^{\textbf{C}} + \zeta\mathcal{L}_{\textbf{L2}}^{\textbf{M}}, \tag{6}$$

$\mathcal{L}_{\textbf{SI-SNR}}$, $\mathcal{L}_{\textbf{L1}}^{\textbf{T}}$, $\mathcal{L}_{\textbf{L2}}^{\textbf{C}}$ and $\mathcal{L}_{\textbf{L2}}^{\textbf{M}}$ denote SI-SNR loss in time domain, L1 loss in time domain, complex spectrum L2 loss and magnitude L2 loss, respectively. α, β, γ and ζ denote the weight of four losses.

4. Experiments and Results

- **Experimental Setup**

Source speech data: LibriTTS, AISHELL-3, speech data of DNS challeng and the vocal part of MUSDB. 1050 h in total.

Source noise data: MUSAN, noise data of DNS challenge, the music part of MUSDB, MS-SNSD and collected pure music data including classical and pop music. 260 h in total.

RIR: simulated by image method with [0,2, 1.2]s of RT60. Early reflection within 50 ms is used as the dereverberation training target.

Sampling rate: 16 kHz.

Signal to noise ratio (SNR): -5 dB to 15 dB.

Model evaluation: Three SNR ranges namely [-5, 0], [0, 5] and [5, 10] dB. The blind test set of Interspeech 2021 DNS Challenge is also selected as another evaluation dataset.

- **Results**

Uformer gives the best performance in both objective and subjective evaluation. while the result of the causal version doesn't degrade generally.

Compared with the time domain models, better performance is achieved for complex domain approaches, which indicates that it is more suitable to perform simultaneous enhancement and dereverberation in the T-F domain than the waveform level.

Uformer reaches 3.6032 on DNSMOS, which is superior to all complex domain neural network based models and has relatively equal ability with the SDD-Net with post processing.

The idea of using dilated convolution layer shows great ability of modeling long range sequence dependencies. In addition, FA, encoder-decoder attention also contribute to a great extend.

Table 1: Results on different models in terms of PESQ, eSTOI, DNSMOS and MOS, where PESQ and eSTOI are calculated on simulated test set while DNSMOS and MOS are calculated on Interspeech2021 DNS challenge blind test set. Note that SDD-Net and DCCRN+ here are the results submitted to the challenge.

Model	Cau.	#Param. (M)	PESQ			eSTOI			DNSMOS	MOS
SNR (dB)	-	-	[-5,0]	[0,5]	[5,10]	[-5,0]	[0,5]	[5,10]	-	-
Noisy	-	-	1.4710	1.7616	1.9904	43.50	53.54	60.96	2.4139	1.8545
UFormer	×	9.46	**2.4501**	**2.7472**	**2.9511**	**64.63**	**74.33**	**79.62**	**3.6032**	**3.3545**
UFormer	✓	9.46	2.4023	2.7265	2.9250	64.22	74.29	79.46	3.5890	3.3523
- FA	×	9.02	2.4207	2.7273	2.9306	64.19	74.11	79.37	3.5801	-
- DC	×	9.31	2.3374	2.6689	2.8883	62.86	73.03	78.52	3.5654	-
- encoder-decoder attention	×	5.33	2.4218	2.7217	2.9177	64.27	74.10	79.37	3.5381	-
dilated conformer → LSTM	×	9.47	2.4106	2.7243	2.9258	64.11	73.90	79.31	3.5839	-
- real-valued sub-modules	×	7.26	2.4266	2.7352	2.9402	64.28	74.26	79.58	3.5751	-
- complex-valued sub-modules	×	3.85	2.4039	2.7025	2.9095	63.55	73.37	78.78	3.5265	-
DCCRN	✓	8.99	2.3652	2.6674	2.8676	62.25	72.55	77.97	3.4915	3.2773
GCRN	×	30.83	2.2672	2.5768	2.7883	61.43	71.87	77.70	3.3452	-
PHASEN	×	8.41	2.3203	2.6170	2.8072	62.76	72.73	78.12	3.4518	-
SDD-Net	✓	6.38	-	-	-	-	-	-	3.36/3.47/ 3.56/**3.60**	3.3432
DCCRN+	✓	4.71	-	-	-	-	-	-	3.4260	3.0682
TasNet	×	8.69	2.2671	2.5649	2.7808	61.11	71.30	77.50	3.3832	-
DPRNN	×	2.60	2.2758	2.5723	2.7752	61.30	71.61	77.10	3.2524	-

5. Conclusion

- We propose Uformer for simultaneous speech enhancement and dereverberation in both magnitude and complex domains.
- UFormer reaches 3.6032 DNSMOS on the blind test set of Interspeech 2021 DNS Challenge.
- All proposed sub-modules are proved to be effective.

6. Key References

[1] Muqiao Yang, Martin Q Ma, Dongyu Li, Yao-Hung Hubert Tsai, and Ruslan Salakhutdinov. Complex transformer: A framework for modeling complex-valued sequence. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4232–4236. IEEE, 2020.