A Causal U-net based Neural Beamforming Network for Real-Time Multi-Channel Speech Enhancement

Xinlei Ren, Xu Zhang, Lianwu Chen, Xiguang Zheng, Chen Zhang, Liang Guo, Bing Yu
Kuai Shou Technology Co. Beijing, China
renxinlei@kuaishou.com

Abstract

People are meeting through video conferencing more often. While single channel speech enhancement techniques are useful for the individual participants, the speech quality will be significantly degraded in large meeting rooms where the far-field and reverberation are increased. Approaches based on microphone array signal processing are proposed to explore the inter-channel correlation among the individual microphone channels. In this work, a new causal U-net based multiple-in-multiple-out structure is proposed for real-time multi-channel speech enhancement. The proposed method incorporates the traditional beamforming structure with the multi-channel causal U-net by explicitly adding a beamforming operation at the end of the neural beamformer. The proposed method has entered the INTERSPEECH Far-field Multi-Channel Speech Enhancement Challenge for Video Conferencing. With 1.97M model parameters and 0.25 real-time factor on Intel Core i7 (2.6GHz) CPU, the proposed method has outperformed the baseline system of this challenge on PESQ, Si-SNR and STOI metrics.

Index Terms: multi-channel speech enhancement, U-NET, encoder-decoder, deep learning

1. Introduction

Video conferencing has become a new normal as people collaborate from geographically disjoint offices more often than before. For large conference rooms, circular and linear microphone arrays are usually employed to achieve better speech capture and enhancement. Traditional signal processing based beamforming techniques [1][2] have been proposed, where the multi-channel optimal filters are estimated to only boost the signals coming from the desired target direction while attenuating the interferences from the other directions. In recent years, deep learning based speech enhancement (SE) approaches have achieved significant improvement over the signal processing based methods, especially for single-channel SE [3][4][5]. Motivated by this success, the multi-channel deep learning based speech enhancement methods are proposed. Many of these methods incorporate the deep neural network (DNN) with the traditional beamforming broadly known as the neural beamformers. In [6][7], a single channel mask is produced by the single channel deep noise suppression network for each channel. The multi-channel spatial covariances of the noise signals are calculated for MVDR beamforming using these single-channel masks. [8] employs deep learning to train the first network to exploit the inter-channel phase and level pattern features for the two-channel inputs. These features are served as the additional directional features to the second multi-channel neural source separation network. [9] proposes an all deep learning MVDR (ADL-MVDR) beamformer that trains the Conv-Tasnet[10] variant with the interaural phase difference (IPD) alongside with the log-power spectra (LPS) as the input features. These methods either require additional spatial features such as the inter-channel phase and level patterns or only employ the DNN to estimate a single channel mask.

More recently, the U-net structure, previously achieved state-of-the-art music separation [11] performance, is employed for single and multi-channel speech enhancement. Methods described in [12] employ the Wave-U-net [3] structure to estimate a single channel speech from the input multi-channel noisy signal. Time-frequency domain U-net is also proposed in [13] where the Channel-Attention is placed between the encoder and the decoder to produce the estimated single-channel speech and noise. Comparing to the neural beamformers, these U-net approaches generally do not require explicit spatial features, where the output single channel clean speech estimation is created directly from the input multi-channel (MISO).

In this paper, a causal multiple-in-multiple-out (MIMO) U-net neural beamformer is proposed to combine the MISO U-net with the beamforming structure. The contribution of this paper can be summarized as follows. First, the proposed method extends the causal U-net[14][15] structure previously proposed for single channel speech enhancement to produce multi-channel time-frequency domain complex beamforming filter. Second, the proposed method incorporates the traditional beamforming structure with the multi-channel causal U-net by explicitly adding a beamforming operation at the end of the neural beamformer. Comparing to the existing neural beamformers, the proposed method does not require explicit spatial feature such as the IPDs. In comparison to the existing MISO U-nets, the proposed method directly outputs the complex beamforming filters to work with the beamforming layer. The evaluation results show improved PESQ, Si-SNR and STOI scores over the MISO U-net using the same dataset and U-net configurations. The proposed method has also entered the INTERSPEECH Far-field Multi-Channel Speech Enhancement Challenge for Video Conferencing(ConferenceSpeech 2021 Challenge). With 1.97M model parameters and 0.25 real-time factor on Intel Core i7 (2.6GHz) CPU, the proposed method has outperformed the baseline system of the ConferenceSpeech 2021 Challenge on PESQ, Si-SNR and STOI metrics.

2. Formulation of the problem

The signal recorded by the $m^{th}$ microphone array can be represented by:

$$y_m(t) = x(t)*h_m(t) + n_m(t), m = 0, ..., M-2, M-1 \quad (1)$$

where $m$ represents microphone index, $M$ represents the number of microphones, $y_m(t)$ represents the signal recorded by the $m^{th}$ microphone, $x(t)$ represents the clean speech signal, $h_m(t)$ represents the transfer function from the clean speech to the $m^{th}$ microphone, $n_m(t)$ represents the noise signal recorded by the $m^{th}$ microphone, and $*$ denotes the convolution operator.
operator. In the time-frequency domain transformed by Short-Time Fourier Transform (STFT), (1) can be expressed as:

$$Y_m(l, f) = X_m^{early}(l, f) + X_m^{late}(l, f) + N_m(l, f)$$

where

$$X_m^{early}(l, f) = X(l, f) \cdot H_m^{early}(l, f)$$

$$X_m^{late}(l, f) = X(l, f) \cdot H_m^{late}(l, f)$$

of the network is $W \in \mathbb{C}^{2F \times L \times M}$, transposing and reshaping it, a complex mask matrix $W \in \mathbb{C}^{M \times L \times F}$ is obtained.

Due to the success of the traditional signal beamforming for multi-channel speech enhancement, the proposed approach performs enhancement just like that of the beamforming.

### 3.2. Model structure

The causal U-net\[15\] is used as our neural network, which is a well-known encoder-decoder architecture and use no future information. The encoder consists of 8 Conv2d blocks for extracting high-level features gradually, and the decoder consists of 8 Conv2dTranspose blocks for reconstructing the size of input features from the output of the encoder. Between the encoder and decoder, skip connections are used to concatenate each layer in the encoder with its corresponding layer in the decoder. Figure 2 shows an example of causal convolutions U-net\[15\] with kernel size 2 and stride 1, which only utilizes the current and history information.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Filter number</th>
<th>Kernel</th>
<th>Stride</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv2d$^{1st}$</td>
<td>32</td>
<td>(6, 2)</td>
<td>(2, 1)</td>
</tr>
<tr>
<td>Conv2d$^{2nd}$</td>
<td>32</td>
<td>(6, 2)</td>
<td>(2, 1)</td>
</tr>
<tr>
<td>Conv2d$^{3rd}$</td>
<td>64</td>
<td>(7, 2)</td>
<td>(2, 1)</td>
</tr>
<tr>
<td>Conv2d$^{4th}$</td>
<td>64</td>
<td>(6, 2)</td>
<td>(2, 1)</td>
</tr>
<tr>
<td>Conv2d$^{5th}$</td>
<td>96</td>
<td>(6, 2)</td>
<td>(2, 1)</td>
</tr>
<tr>
<td>Conv2d$^{6th}$</td>
<td>96</td>
<td>(6, 2)</td>
<td>(2, 1)</td>
</tr>
<tr>
<td>Conv2d$^{7th}$</td>
<td>128</td>
<td>(2, 2)</td>
<td>(2, 1)</td>
</tr>
<tr>
<td>Conv2d$^{8th}$</td>
<td>256</td>
<td>(2, 2)</td>
<td>(1, 1)</td>
</tr>
</tbody>
</table>

The configuration of each Conv2d layer in the encoder is presented in Table 1. The encoder has 8 Conv2d blocks. Each block is consisted by a Conv2d layer followed by a batch normalization layer, a dropout layer and a LeakyReLU activation function. The dropout rate is set to 0.5. The corresponding decoder also has 8 blocks. Each of the decoder block is identical to the encoder block except for replacing the Conv2d layer with the Conv2dTranspose layer. The input shape of each layer in the encoder-decoder architecture is specified in [BatchSize, Frequency, Frame, Channel] format. The causal U-net is followed...
by a dense layer with 512 units to estimate the complex mask. In the implementation of causal U-net, $K$ zeros frames are padded in front of input feature, and the last $K$ frames are discarded for the output mask, the frame number $K$ is decided by the kernel size over the time dimension, which is set to 8 in the proposed model.

3.3. Beamforming

After estimating $W \in \mathbb{C}^{M \times L \times F}$, the enhance process is performed as that of the traditional beamforming. The complex spectrogram of noisy signal $Y \in \mathbb{C}^{M \times L \times F}$ is multiplied by $W \in \mathbb{C}^{M \times L \times F}$ with element-wise multiplication, then the result is summed over the channel axis to estimate a single-channel enhanced complex spectrogram $\hat{X} \in \mathbb{C}^{L \times F}$. A time domain enhanced signal $\hat{x}$ is obtained after a iSTFT operation on $\hat{X} \in \mathbb{C}^{L \times F}$. From the perspective of this process, $W \in \mathbb{C}^{M \times L \times F^2}$ can be understood as the weights of a filter-and-sum beamformer.

3.4. Loss function

The mean absolute error (MAE) defined in time domain is used as loss function in the proposed system. Comparing to the time domain mean square error (MSE) used in [16], the MAE pays more attention to small signals and have better noise suppression performance. In addition to the speech component, the noise component is also added to the loss function, which is beneficial for improving the noise robustness of proposed model. And these two items have the same weight:

$$\text{loss}_{\text{mae}} = |x - \hat{x}| + |n - \hat{n}| \tag{6}$$

where $x$ and $\hat{x}$ are clean speech and estimated clean speech respectively, $n$ and $\hat{n}$ are noise and estimated noise respectively. They satisfy the following relationship:

$$x + n = \hat{x} + \hat{n} = y \tag{7}$$

where $y$ is noisy speech.

3.5. Post-filter

To further suppress the residual noise, we apply a Wiener filter[17] with the noise estimation algorithm based on minimum tracking as the post-filtering strategy, which is commonly employed in the traditional speech enhancement system. The window length of the minimum tracking algorithm is set to 4 seconds, and the coefficients of the Wiener filter is calculated by:

$$G(l, f) = \frac{\lambda_n(l, f) - \lambda_s(l, f)}{\lambda_s(l, f)} \tag{8}$$

where $\lambda_s(l, f)$ is power spectral density of the noisy signal and $\lambda_n(l, f)$ is the estimated noise power spectral density. $l$ and $f$ denote the frame index and the frequency index respectively.

4. Experiments and Results

4.1. Datasets

The clean speech dataset includes four open source speech databases: AISHELL-1[18], AISHELL-3[19], VCTK[20] and LibriSpeech[train-clean-360][21]. The speech utterances with SNR larger than 15dB are selected for training. The total duration of clean training speech is around 550 hours. The noise dataset is composed of MUSAN[22] and Audioset[23], the total duration is around 120 hours. Besides these two open source databases, 98 real meeting room noise files recorded by high fidelity devices are also used. The speech and noise files selected by ConferencingSpeech 2021 challenge [24] from these databases are used for data augmentation.

We generate 20000 multi-channel room impulse responses (RIRs) based on the configuration of the microphone array with the image method[25], and the range of RT60 is between 0.1 and 1.2 s. The room size ranges from $3 \times 3 \times 3 m^3$ to $8 \times 8 \times 3 m^3$. The microphone array is randomly placed in the room with height ranges from 1.0 to 1.5 m. The sound source, including speech and noise, comes from any position in the room with height ranges from 1.2 to 1.9 m. The angle between two sources are wider than 20°. The distance between sound source and microphone array ranges from 0.5 to 5.0 m.

Based on the speech, noise and RIR datasets, a total of 1000 hours of multi-channel noisy data are generated, and 70% for training while 30% for validating. 2000 extra clips are generated for testing. These data are generated with following augmentations:

- **REVERBERATION:** The single-channel speech and noise are convolved with the RIRs to generate multi-channel data. The target speech preserves 50ms early reverberation, this is because the reflections arriving within 50 ms after the direct sound is actually beneficial for intelligibility. The reflections, which arrives 50 ms after the direct sound, degrades the speech intelligibility[26][27].
- **SCALING:** The amplitude of the training data is randomly selected within the range of [-50, -0.87]dB.
- **EQ:** We filter the data with various filters for simulating EQ, these filters include low-pass filter, de-emphasis filter and so on.
- **SNR:** We mix the speech and noise with the random SNR between -3 and 25 dB.

4.2. Experiments setup

Figure 3 shows the configuration of the microphone array(MA) provided by ConferencingSpeech 2021 challenge. It is a linear microphone array with non-uniformly distributed 8 microphones (marked as MA No.1).

![Figure 3: The information of MA No.1. The unit in this figure is centimeter](image)

There are two tasks for ConferencingSpeech 2021 challenge:

- **Task1:** Multi-channel speech enhancement with single microphone array. This task is focusing on processing speech from the MA No.1. It needs to meet real-time requirements: no future information can be used; the processing time of the test clip should be less and equal than the time of the test clip and frame length should be less than or equal to 40ms.
- **Task2:** Multi-channel speech enhancement with multiple distributed microphone arrays. Five microphone arrays are provided for this task, and there is no any constraints, any algorithms can be explored with these microphone arrays.
For task1 and task2, we only develop our system on the data of MA No.1, and the system aims to estimate the clean speech of the 3rd microphone. We use the Adam optimizer to train the model. The initial learning rate is set to 0.001, and it will decay 0.5 when the validation loss does not decrease for 5 epochs.

4.3. Results

First, several causal U-net models with different inputs and outputs are evaluated. The details model configuration are shown as followed:

- **SISO-U-net**: It inputs single-channel noisy spectrogram and outputs single-channel complex mask. The single-channel mask is multiplied with noisy spectrogram to generate enhanced speech spectrogram.
- **SISO+IPD-U-net**: It inputs single-channel noisy spectrogram along with the interchannel phase difference(IPD) features[8], which are extracted from four microphone-pairs((0,4),(1,5),(2,6),(3,7)). The output is single-channel complex mask.
- **MISO-U-net**: It inputs multi-channel noisy spectrogram and outputs single-channel complex mask.
- **MIMO-U-net**: It inputs multi-channel noisy spectrogram and outputs multi-channel complex mask, and a beamforming operation is applied based on the output mask. And this is our proposed system without post-filter.
- **MIMO-U-net+BF**: This is the proposed MIMO-U-net network followed by post-filter.

We evaluate these models on the 2000 testing clips via four objective measurements: PESQ[28], STOI[29], Extended-STOI[30] and Si-SNR[10]. The result is presented in Table 2. By utilizing multichannel information, models with IPD information or multichannel inputs outperform the SISO model. With multichannel input, MISO model achieves higher PESQ compared to SISO+IPD model, which means the network can automatically learn multichannel information better than IPD. By estimating multichannel masks and applying multiply and sum operations similar to beamforming, MIMO model can further improve the PESQ to 1.95, compared to that of 1.89 for MISO model. With post-filter, although the PESQ slightly decreases from 1.95 to 1.91, our internal subject listening test shows that MIMO-PF model achieves best listening experiences.

Secondly, we compare our system with the baseline on the development set provided by ConferencingSpeech 2021 challenge. The total number of parameters of the proposed neural network is ~1.97M, and the processing time of one frame(16ms) is on average ~4ms tested on an Intel Core i7 (2.6 GHz) CPU.

<table>
<thead>
<tr>
<th></th>
<th>PESQ</th>
<th>STOI</th>
<th>E-STOI</th>
<th>Si-SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td>1.278</td>
<td>0.728</td>
<td>0.587</td>
<td>1.893</td>
</tr>
<tr>
<td>SISO-U-net</td>
<td>1.841</td>
<td>0.844</td>
<td>0.740</td>
<td>7.475</td>
</tr>
<tr>
<td>SISO+IPD-U-net</td>
<td>1.855</td>
<td>0.847</td>
<td>0.746</td>
<td>7.501</td>
</tr>
<tr>
<td>MISO-U-net</td>
<td>1.890</td>
<td>0.852</td>
<td>0.758</td>
<td>7.959</td>
</tr>
<tr>
<td>MIMO-U-net+BF</td>
<td>1.950</td>
<td>0.861</td>
<td>0.764</td>
<td>8.008</td>
</tr>
<tr>
<td>MIMO-U-net+BF+PF (proposed)</td>
<td>1.919</td>
<td>0.857</td>
<td>0.759</td>
<td>7.935</td>
</tr>
</tbody>
</table>

Table 3: The objective scores on the data of MA No.1

<table>
<thead>
<tr>
<th></th>
<th>PESQ</th>
<th>STOI</th>
<th>E-STOI</th>
<th>Si-SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noisy</td>
<td>1.515</td>
<td>0.823</td>
<td>0.690</td>
<td>4.474</td>
</tr>
<tr>
<td>baseline</td>
<td>1.999</td>
<td>0.888</td>
<td>0.780</td>
<td>9.159</td>
</tr>
<tr>
<td>proposed</td>
<td>2.125</td>
<td>0.908</td>
<td>0.817</td>
<td>9.287</td>
</tr>
<tr>
<td>Task2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noisy</td>
<td>1.506</td>
<td>0.824</td>
<td>0.693</td>
<td>4.504</td>
</tr>
<tr>
<td>baseline</td>
<td>1.983</td>
<td>0.887</td>
<td>0.780</td>
<td>9.228</td>
</tr>
<tr>
<td>proposed</td>
<td>2.125</td>
<td>0.909</td>
<td>0.818</td>
<td>9.343</td>
</tr>
</tbody>
</table>

Table 4: MOS on evaluation test set of the challenge

<table>
<thead>
<tr>
<th></th>
<th>MOS</th>
<th>S-MOS</th>
<th>N-MOS</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noisy</td>
<td>2.56</td>
<td>2.93</td>
<td>3.03</td>
<td></td>
</tr>
<tr>
<td>proposed</td>
<td>4.02</td>
<td>3.87</td>
<td>3.87</td>
<td>0.02</td>
</tr>
<tr>
<td>improvement</td>
<td>1.46</td>
<td>0.94</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>Task2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noisy</td>
<td>2.51</td>
<td>2.88</td>
<td>2.99</td>
<td></td>
</tr>
<tr>
<td>proposed</td>
<td>4.14</td>
<td>3.93</td>
<td>3.92</td>
<td>0.02</td>
</tr>
<tr>
<td>improvement</td>
<td>1.63</td>
<td>1.05</td>
<td>0.93</td>
<td></td>
</tr>
</tbody>
</table>

5. Conclusions

In this paper, a causal U-net based neural beamformer is proposed for real-time multi-channel speech enhancement. With combining MIMO structure and a beamforming operation, the proposed system outperforms SISO and MISO U-net structure on PESQ, Si-SNR and STOI metrics. Besides, the proposed system has achieved better performance than the baseline system of ConferencingSpeech 2021 Challenge, and also significantly improves the MOS of noisy speech.
### 6. References


