Real-time Streaming Wave-U-Net with Temporal Convolutions for Multichannel Speech Enhancement

Vasiliy Kuzmin, Fyodor Kravchenko, Artem Sokolov, Jie Geng

1 Huawei Russia Research Institute
2 Huawei Nanjing Research Institute, China
3 HSE University, Nizhniy Novgorod, Russia

Abstract

In this paper we describe our work that we have done to participate in Task1 of ConferencingSpeech2021 challenge. This task set a goal to develop the solution for multi-channel speech enhancement in a real-time manner. We propose a novel system for streaming speech enhancement. We employ Wave-U-Net architecture with temporal convolutions in encoder and decoder. We incorporate self-attention in decoder to apply attention mask retrieved from skip-connection on features from down-blocks. We explore history cache mechanisms that work like hidden states in recurrent networks and implemented them in proposal solution. It helps us to run inference with chunks length 40 ms and Real-Time Factor 0.4 with the same precision.

Index Terms: Multi-channel speech enhancement, speech enhancement, wave-u-net, self-attention, deep learning

1. Introduction

Speech enhancement (SE) tries to separate out individual audio sources from an input mixture of clean audio and noise. Multi-channel approaches, generally, do it better than single-channel ones considering that they could process spatial information taken from time differences between signal reach separate items of microphone arrays (MA). Conventional approaches like multi-channel Wiener filter [1] or beamforming [2] are progressively displaced by deep learning (DL) based techniques [3] or by mixture with them [4]. Deep learning approaches with U-Net [5] based architectures was very successfully on various computer vision tasks connected to image and video processing where we need to transform one image to another or extract topological information that we need. Furthermore, this architecture was transferred to speech related tasks close to SE as voice separation [6] and speech enhancement itself [7]. This architecture works well as with signal waveform as with its Short Time Fourier Transform (STFT) [8].

The version adopted for processing time domain signal called Wave-U-Net is also actively exploited nowadays for the same tasks [9]. To obtain real-time streaming speech enhancement the network should ignore the future frames and provide the Real Time Factor (RTF) < 1. Recurrent networks like unidirectional Long-Short Time Memory (LSTM) and Temporal Convolution Networks (TCN) could satisfy the first restrictions as they are based on casual convolutions which have zero look ahead [10]. Moreover, convolutions are very fast and allow to build real-time systems. LSTM mechanisms provide hidden states retrieving and sequences could be handled step-by-step by passing previous states for next iteration. But for convolution this trick is not supported from-the-box. In [12] the authors showed that history cache with hidden states could be saved and reused. Thus, it is also possible to fulfill inference with small pieces of stream, but it requires to "manually" fed on hidden states from previous iteration with the input. This paper proposes a neural network architecture based on Wave-U-Net with temporal convolutions (TC Wave-U-Net) which uses cache with hidden states for the inference.

To make prediction more precise various self-attention mechanisms are incorporated in Wave-U-Net systems [13]. We also do it in the paper as casual convolution gives in to convolution with future context and extra elements are required to yield acceptable precise.

This method was submitted to the ConferencingSpeech 2021 challenge [1] in INTERSPEECH 2021 conference. Summarily, our primary contributions are the follows:

- We developed the novel model for the multi-channel speech enhancement task that can be inferences real-time manner, have zero look ahead and based on successful U-Net architecture. This model have only 8.31 millions parameters and can be run on any device.

- We applied historical context cache which allowed us to decrease receptive field during inference as well as decrease total number of float points operation.

The rest of the paper is organised as follows. First, in the following section, we briefly review the background of the problem. In Section 3 we describe our Wave-U-Net based Temporal Convolution network with attentions. The details of of our experiments and training procedure are then presented in Section 4. We present the results and comparison with the baseline system in Section 5. Finally, in Section 6 we provide some conjectures as why and how to configure historical cache, further research in this direction followed by conclusion.

2. Background

During participating ConferencingSpeech 2021 challenge we aimed to solve Task1 which is Multi-channel speech enhancement with single microphone array. This task expects the noisy audio processing from single linear array with non-uniform distributed microphones. Real time factor considering less than 1.0 while running on device as well as participant should use frame of length 40 ms.

A multi-channel speech enhancement problem could be described the following way:

Let take $C$ - channel signal on the $t$ time step as $Y_t = [y_1, ..., y_C]$
We assume that \( y^c_t \) is given as the following mathematical expression:
\[
y^c_t = x^c_t + n^c_t = h^c \otimes x^c_t + n^c_t
\]
where \( x^c_t \) the clean signal recorded by the \( C \)-th microphone, \( n^c_t \) is the additive noise signal and \( h^c \) is room impulse response (RIR). The convolution of RIR and dry signal \( x^c_t \) called reverberant signal \( x^r_t \). The aim of Task 1 is to estimate signal \( x^{orig}_t \), where \( orig \in \{1, \ldots, C\} \), by removing additive noise and by dereverberation for whole utterance \( t = 1, \ldots, T \). The \( x^{orig}_t \) represents the original mono channel audio from multi-channel mixture \( y^c_t \).

3. Architecture: Wave-U-Net with temporal convolutions

In this section, we explain the TC Wave-U-Net architecture with attentions and history cache we implemented for streaming inference.

3.1. Structure of TC Wave-U-Net

We follow the Wave-U-Net structure for our multi-channel speech enhancement solution. Our framework consists of an encoder, middle part (bottleneck) and decoder. We use history padding to avoid the non zero look ahead. Schematically system is depicted on Figure 1. We note that our model takes C-channel encoder, middle part (bottleneck) and decoder. We use history context length which taken by the similar way with [13] to find out relevant features came as product \( P \) and \( A \). Afterward, we concatenate Up Block output with self-attention output and result goes to TC block.

The papers [13], [15], [16] inspired us to integrate a self-attention to our architecture. We fed on self-attention an up block and skip connection tensors from the same hierarchical level. Afterwards, we concatenate Up Block output with self-attention output and result goes to TC block.

Self-attention mechanism is intended to capture the global dependencies. It is successfully exploit, for instance, in such fields of deep learning as speech recognition [17] and machine translation [18]. In our architecture we employed attention by the similar way with [13] to find out relevant features came as product \( P \) and \( A \). Attention uses \( k, q \) and \( v \) as 1D convolution operations for the representation of input to embedding space. The products \( k \) and \( q \) are then summarised and, after parametric linearity, as product \( P \) follows to attention mask calculation \( A(P) \) which is a convolution with kernel size 1. The output of attention is a term-wise product of \( W_i = A(P)D_i \). In our experiments, attention block helps the network to converge deeper, while without it, a model converges faster, however it gets stuck in some space.

3.2. Streaming Inference

For streaming inference we employed idea from [12]. Paper represents how to extend the history context length which taken by a network into account if the network was being trained on fixed-size segments of data. The authors applied the mechanisms of hidden states reusing on the Transformer architecture and achieved considerable gain in the length of context remembered.
We conducted various experiments with our streaming TC Wave-U-Net model to challenge its speech enhancement quality and demonstrate efficiency of cache mechanism.

4. Experiments

We conducted various experiments with our streaming TC Wave-U-Net model to challenge its speech enhancement quality and demonstrate efficiency of cache mechanism.

4.1. Datasets

The for training clean data we used audio speech datasets of Chinese language AISHELL-1, AISHELL-3, sets Librispeech (train-clean-360) and VCTK for the English language. The organizers were provided the lists permitted audio files from that datasets with loudness larger than 15dB that could be used for training procedure. The total duration of clean training speech was around 550 hours. The noise set was composed from two parts. Part 1 is selected from public noise datasets MUSAN and Audioset with total duration about 120 hours. The Part 2 was a real meeting room noise recordings. The total amount of provided clips was 98 items. Also we got simulated Room Impulse Responses (RIR) for the reverberation affect obtaining. Provided framework allowed to augment data on-the-fly during training. It considerably slow-down the training procedure. We generated and saved data locally.

4.2. Evaluation Items

Thus, the following denoising systems were taken to evaluate the efficacy of our proposed approach:

- **Baseline.** LSTM-based solution, that was provided by organizers itself. Basically the model process audio in time and frequency domains. The baseline has 8 channels raw audio input. It calculates inter-channel phase difference between pairs of microphone channels, complex STFT for input signal and pass he concatenated tensor through 3 LSTMs. Each recurrent layer has 512 hidden states and, finally, they followed by a projection layer. Resulted mask is applied on channel 1 of original signal and inverse STFT returns the output of baseline solution.

- **Wave-U-Net.** The implementation which we took from open GitHub repository[1] It exploits 12 convolution layers in encoder and decoder. We just made changes in code to establish the ability of C-channel raw waveform processing instead of original mono channel input.

- **TC Wave-U-Net.** Our proposed model with time domain input, casual convolutions in encoder decoder and bottleneck and self-attentions in decoder.

4.3. Experimental Setup

We fed to network multi-channel audio recordings with additive random noise and convoluted with random RIRs. For training we capture random piece with 16384 samples from 16 kHz raw waveforms and used that data for input. Our network is constructed from 9 TC blocks in encoder and decoder, 1 casual convolution layer in bottleneck. Filter size for convolutions in encoder is 15 and in decoder accordingly, as mentioned in paper [19]. All channel sizes for encoder are 8, 24, 48, 72, 96, 120, 144, 168, 192, 216 and (240 in bottleneck). For decoder they are repeated in reverse order. Dilation parameter for each block is changed by the following way 1, 1, 2, 4, 5, 16, 32, 64. The receptive field of the encoder is 1807 samples or 112 ms.

4.4. Learning Target

For the learning objective we use weighted signal-to-distortion loss (wSDR) [20]. This is a time-domain loss function that could be defined by the following formula:

\[
L_{wSDR}(x, y, \tau) := \alpha L_{SDR}(y, \tau) + (1 - \alpha) L_{SDR}(z, \tau)
\]
where $L_{SDR}$ is a conventional signal-to-distortion (SDR) loss, $x$ is a mixture signal assumed as linear sum of clean speech signal $y$ and original noise $z$. Estimated noise $\hat{z}$ given as $\hat{z} = \frac{1}{\alpha} - y$. The $\alpha$ is an enhanced audio. Taken this into account the energy ratio $\alpha$ between dry speech $y$ and original noise $z$ is defined as $\alpha = \frac{\|y\|^2}{\|y\|^2 + \|z\|^2}$.

For the baseline system, we trained mean square error (MSE), signal-to-distortion and wSDR. Our experiments show that SDR performed better that MSE and wSDR was slightly better than SDR. In long training scenarios with SDR loss the model wasn’t able to continue to converge while with SDR loss even after 300 epoch the model steadily improved. We train all our models using 6\times Tesla V100.

For the final training we choose wSDR loss and Adam optimizer with starting learning rate $= 10^{-3}$. The decayed learning scheduler was applied with minimum value $10^{-8}$ after 250 epochs. Batch size is equal 1000.

The baseline was trained using default configuration parameters, SDR objective loss, Adam optimizer with learning rate $= 10^{-4}$ and with reduce on plateau scheduler. Chunks of 4 seconds was chooses by default as well. The dataset was identical to primary experiment.

### 4.5. Streaming Measurements Setup

We evaluated the performance of Baseline, Wave-U-Net and our TC Wave-U-Net in streaming mode. Wave-U-Net consist of simple convolution layers and does not satisfy the requirement of challenge for zero-look ahead. We changed code to conduct pseudo stream tensor processing. The chunk with 16384 samples (~1 sec) is fed to the model with shift 640 samples (40 ms) and 40 ms of predicted audio was taken from output. For the first times we provide chunk 40 ms, 80 ms, 120 ms etc. with zero-padded left part till it reached full size 16384. For TC Wave-U-Net with cache we pass chunk window with 1024 samples (64 ms), and move the window on 640 samples (40 ms).

### 5. Results

#### 5.1. Signal Quality

According to the conference rules, we push test dataset enhanced by our system to organizers that evaluate our resulting audio. Moreover, they provide mean opinion score (MOS) - measure of the human-judged overall quality of an audio. Each rater determines MOS, subjective speech MOS (S-MOS) and subjective noise MOS (N-MOS) for each cleaned recording. Next, confidence interval (CI) of MOS score is calculated. Each file is listened by more that 20 raters. In Table 1 we denoted the result of comparing of our enhanced audio with noised raw waveforms.

#### 5.2. Real Time Factor

In Table 2 we report Real Time Factor (RTF) (processing time divided by audio duration) in relative scale where lower values indicate faster processing and lower user-perceived latency.

We can see that the network with cache enabled outperforms the models working with whole receptive field. It is even better than other networks but it compensated by better precise. In the addition to mentioned above, in Table 2 we demonstrate the evaluation results of our model in streaming and non streaming models. Baseline and vanilla Wave-U-Net with pseudo streaming (see 5.5) are inferring close to the speed of TC Wave-U-Net with cache but with the lower quality than these models for non-streaming mode. PESQ for Baseline and Wave-U-Net with streaming 12% and 8% less, respectively, than non-streaming. Also Wave-U-Net in streaming mode don’t meet the requirement of challenge for zero-look ahead.

| Model                        | RTF | PESQ | STOI | E-STOI | SI-SNR |
|------------------------------|-----|------|------|--------|--------|
| BASELINE(NONSTREAMING)       | 0.7X| 1.76 | 8.68 |
| WAVE-U-Net(NONSTREAMING)     | 0.9X| 1.93 | 10.14|
| TC WAVE-U-Net(NONSTREAMING)  | 3.3X| 2.18 | 8.31 |
| TC WAVE-U-Net CACHE ENABLED(S) | X   | 2.19 | 8.31 |

### 6. Conclusions

This paper proposed our speech enhancement solution for Task1 on ConferencingChallenge2021. We provided the Wave-U-Net based network that outputs cleaned audio for passed raw waveform with reverberations and additive noise. Our system employed causal dilated convolutions for encoder, decoder and a bottleneck parts. It also involved self-attentions in decoder for better precise. We implemented historical cache and obtain fast streaming inference.

Our evaluation showed that adding cache mechanism for the model with large receptive field not only can reduce it for the expected one, but also reduce floating-point calculations, thereby improving the inference speed. Even more we have shown that compared to pure pseudo streaming, our proposed method provides the same quality as non-streaming model or does it a bit better. Due to the low time constraints, we aren’t able to do the full research of the evaluation of calculation of the beginning of the cache, and have lefted it for the further research. Moreover, our final model is still training and we are waiting for final metrics.

Future directions of this work could include experiments with deeper versions of this architecture, new versions of TC blocks, incorporation of Channel Attention and using complex ratio masking for signal enhancement.

### 7. Acknowledgements

The work of Artem Sokolov is partially supported by RSF (Russian Science Foundation) grant 20-71-10010.
8. References

[1] M. C. Simmer KU, Bitzer J, “Post-filtering techniques. in micro-
phone arrays.” Springer, Berlin, Heidelberg, pp. 39–60, 2001.

[2] M. Brandstein, “Arrays, microphone. signal processing techniques
and applications.” Springer Science & Business Media, 2001.

[3] Wang, DeLiang, and Jitong Chen., “Supervised speech separation
based on deep learning: An overview.” IEEE/ACM Transactions
on Audio, Speech, and Language Processing, pp. 1702–1725, May
2018.

[4] Erdogan, Hakan, John R. Hershey, Shinji Watanabe, Michael I.
Mandel, and Jonathan Le Roux., “Improved mvdr beamforming
using single-channel mask prediction networks.” Interspeech, pp.
1981–1985, Sep. 2016.

[5] Ronneberger, O., Fischer, P. and Brox, T, “U-net: Convolu-
tional networks for biomedical image segmentation.” Interna-
tional Conference on Medical image computing and computer-assisted
intervention, pp. 234–241, Oct. 2015.

[6] Jansson, Andreas, Eric Humphrey, Nicola Montecchio, Rachel
Bittner, Aparna Kumar, and Tillman Weyde, “Singing voice sepa-
ration with deep u-net convolutional networks,” 18th Interna-
tional Society for Music Information Retrieval Conference, pp. 23–27,
2017.

[7] Ernst, O., Chazan, S.E., Gannot, S. and Goldberger, J., “Improved
speech enhancement with the wave-u-net.” 26th European Signal
Processing Conference (EUSIPCO), pp. 397–394, Sep. 2018.

[8] Soni, Meet H., Neel Shah, and Hemant A. Patil., “Time-frequency
masking-based speech enhancement using generative adversarial
network.” IEEE international conference on acoustics, speech and
signal processing (ICASSP), pp. 5039–5043, 2018.

[9] Macartney, C. and Weyde, T., “Speech dereverberation using fully
convolutional networks.” arXiv preprint arXiv:1811.11307, 2018.

[10] Hochreiter, Sepp, and Jürgen Schmidhuber., “Long short-term
memory.” Neural computation, pp. 1735–1780, Nov. 1997.

[11] Jansson, Andreas, Eric Humphrey, Nicola Montecchio, Rachel
Bittner, Aparna Kumar, and Tillman Weyde, “Channel-attention
dense u-net for multichannel speech enhancement.” Interna-
tional Conference on Acoustics, Speech and Signal Processing, pp. 836–
840, 2020.

[12] Ho, M.T., Lee, J., Lee, B.K., Yi, D.H. and Kang, H.G., “A cross-
channel attention-based wave-u-net for multi-channel speech en-
hancement.” Proceedings of Interspeech, pp. 4049–4053, 2020.

[13] Kriman, Samuel, Stanislav Beliaev, Boris Ginsburg, Jocelyn
Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li,
and Yang Zhang, “Quartet: Deep automatic speech recognition with 1d
time-channel separable convolu-
tions.” International Conference on Acoustics, Speech and Signal
Processing (ICASSP), pp. 6124–6128, 2020.

[14] Stoller, S. Ewert, and S. Dixon., “Wave-u-net: A multi-scale
neural network for end-to-end audio source separation.” arXiv
preprint arXiv:1806.03185, 2019.