TIME-DOMAIN SPEECH EXTRACTION WITH SPATIAL INFORMATION AND MULTI SPEAKER CONDITIONING MECHANISM

Jisi Zhang[1], *Cătălin Zoriă*[2], *Rama Doddipatla*[2] *and Jon Barker*[1]

[1]University of Sheffield, Department of Computer Science, Sheffield, UK
[2]Toshiba Cambridge Research Laboratory, Cambridge, UK

ABSTRACT

In this paper, we present a novel multi-channel speech extraction system to simultaneously extract multiple clean individual sources from a mixture in noisy and reverberant environments. The proposed method is built on an improved multi-channel time-domain speech separation network which employs speaker embeddings to identify and extract multiple targets without label permutation ambiguity. To efficiently inform the speaker information to the extraction model, we propose a new speaker conditioning mechanism by designing an additional speaker branch for receiving external speaker embeddings. Experiments on 2-channel WHAMR! data show that the proposed system improves by 9% relative the source separation performance over a strong multi-channel baseline, and it increases the speech recognition accuracy by more than 16% relative over the same baseline.

Index Terms— multi-channel source separation, multi-speaker extraction, noise, reverberation

1. INTRODUCTION

Speech separation aims to segregate individual speakers from a mixture signal, and it can be used in many applications, such as speaker diarization, speaker verification or multi-talker speech recognition. Deep learning has allowed an unprecedented separation accuracy compared with the traditional signal processing based methods, however, there are still challenges to address. For instance, in blind source separation, the order of the output speakers is arbitrary and unknown in advance, which forms a speaker label permutation problem during training. Clustering based methods [1] or, more recently, Permutation Invariant Training (PIT) technique [2] have been proposed to alleviate this issue. Although the PIT forces the frames belonging to the same speaker to be aligned with the same output stream, frames inside one utterance can still flip between different sources, leading to a poor separation performance. Alternatively, the initial PIT-based separation model can be further trained with a fixed label training strategy [3], or a long term dependency can be imposed to the output streams by adding an additional speaker identity loss [4, 5]. Another issue in blind source separation is that the speaker order of the separated signals during inference is also unknown, and needs to be identified by a speaker recognition system.

An alternative solution to the label permutation problem is to perform target speaker extraction [6–8]. In this case, the separation model is biased with information about the identity of the target speaker to extract from the mixture. Typically, a speech extraction system consists of two networks, one to generate speaker embeddings, and another one to perform speech extraction. The speaker embedding network outputs a speaker representation from an enrollment signal uttered by the target. The speaker embedding network can be either jointly trained with the speech extraction model to minimise the enhancement loss or trained on a different task, i.e., a speaker recognition task, to access larger speaker variations [9]. The target speaker embedding is usually inserted into the middle-stage features of the extraction network by using multiplication [7] or concatenation operations [8, 10], however, the shared middle-features in the extraction model may not be optimal for both tasks of speaker conditioning and speech reconstruction.

Most of the existing speech extraction models enhance only one target speaker each time and ignore speech from other speakers. When multiple speakers are of interest, the extraction model has to be applied several times, which is inconvenient and requires more computational resources. Therefore, a system capable of simultaneously extracting multiple speakers from a mixture is of practical importance. Recently, a speaker-conditional chain model (SCCM) has been proposed that firstly infers speaker identities, then uses the corresponding speaker embeddings to extract all sources [11]. However, SCCM is still trained with the PIT criterion, and the output order of separated signals is arbitrary. Lastly, when multiple microphones are available, the spatial information has been shown to improve the performance of both separation and extraction [7, 12] systems in clean and reveberant environments. So far, the spatial information has not been tested with a multi-speaker extraction system, nor it has been evaluated in noisy and reverberant environments.

In this paper, we reformulate our previous multi-channel speech separation design in [12] as a multi-talker speech extraction system. The proposed system uses embeddings from all speakers in the mixture to simultaneously extract all sources, and does not require PIT to solve the label permutation problem. There are three main contributions in this work. Firstly, we improve our previous multi-channel system in [12] by swapping the Temporal fully-Convolutional Network (TCN) blocks with U-Convolutional blocks, which yielded promising results for a recent single-channel speech separation model [13]. Secondly, the previous modified system is reformulated to perform multi-speaker extraction, and, lastly, a novel speaker conditioning mechanism is proposed that exploits the speaker embeddings more effectively. The evaluation is performed with multi-channel noisy and reverberant 2-speaker mixtures. We show that combining the updated multi-channel structure and the proposed speaker conditioning mechanism leads to a significant improvement in terms of both the separation metric and speech recognition accuracy.

The rest of paper is organised as follows. In section 2, we introduce the proposed multi-channel speech extraction approach. Section 3 presents implementation details and the experiment setup. Results and analysis are presented in Section 4. Finally, the paper is concluded in Section 5.

2. MULTI-CHANNEL END-TO-END EXTRACTION

Recently, neural network based multi-channel speech separation approaches have achieved state-of-the-art performance by directly processing time-domain speech signals [12, 14]. These systems incorporate a spectral encoder, a spatial encoder, a separator, and a decoder. In [12], spatial features are input to the separator only. In this work, we simplify the previous framework by combining the spatial and spectral features as depicted in Figure 1. We found the proposed approach is beneficial for the speech extraction task. The spectral encoder and spatial encoder independently generate N-dimensional single-channel representations and S-dimensional multi-channel representations, respectively. The spectral encoder is a 1-D convolutional layer, and the spatial encoder is a 2-D convolutional layer. The encoded single-channel spectral features and two-channel spatial features are concatenated together to form multi-channel representations with a dimension of $(N + S)$, which are accessed by both the separation module and the decoder. The separator will estimate linear weights for combining the multi-channel representations to generate separated representations for each source. Finally, the decoder (1-D convolutional layer) reconstructs the estimated signals by inverting the separated representations back to time-domain signals.

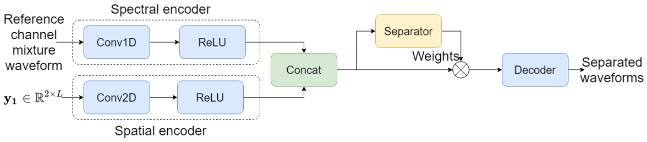

Fig. 1. Updated multi-channel model structure

Compared with our previous work [12], we also upgrade the separator by replacing the original TCN [15] blocks with U-Convolutional blocks (U-ConvBlock), which have proven to be more effective in modelling sequential signals in the single-channel speech separation task [13]. Furthermore, a system built on U-ConvBlock requires fewer parameters and floating point operations compared with the systems built on TCN or recurrent neural network architectures [16]. The U-ConvBlock (Figure 2) extracts information from multiple resolutions using Q successive temporal downsampling and Q upsampling operations similar to a U-Net structure [17]. The channel dimension of the input to each U-ConvBlock is expanded from C to C_U before downsampling, and is contracted to the original dimension after upsampling. The updated separation module is shown in Figure 3 and consists of a instance normalisation layer, a bottleneck layer, B stacked U-ConvBlocks and a 1-D convolutional layer with a non-linear activation function. We choose to use an instance normalisation layer [18] rather than global layer normalisation for the first layer-normalisation, as the latter would normalise over the channel dimension which is inappropriate given the heterogeneous nature of the concatenated features.

2.1. Proposed speech extraction structure

Building on the modified system described above, in this section we introduce a novel multi-channel speech extraction system which simultaneously tracks multiple sources in the mixture. In general, the system uses embeddings from multiple speakers as input, which are used to condition single-source outputs with a consistent speaker order. Common strategies for supplying speaker information to the extraction model are to modulate the speaker features on middle-level features inside the separation model [6, 19] or concatenate the speaker features with the mixture speech representations [8]. However, it is not trivial to find a single optimal layer at which to insert the speaker features. For instance, the shared middle-features in the extraction model may not be optimal for both speaker conditioning and speech reconstruction.

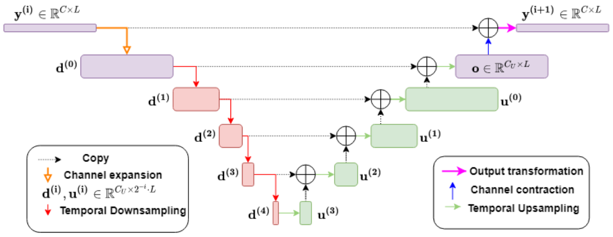

Fig. 2. U-Conv block structure

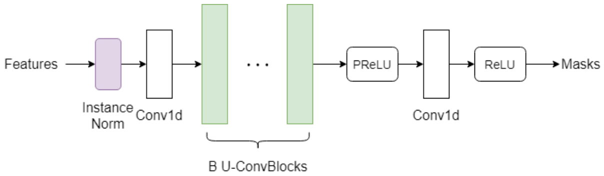

Fig. 3. Improved separator with U-Conv blocks

To address this issue, we propose a new 'speaker stack' for processing the input speaker representations to coordinate with the main separation stack, as shown in Figure 4. The speaker stack takes the encoded multi-channel features and generates two high-level sequential features, which are suitable to receive speaker information from externally computed speaker embeddings. The output of the speaker branch containing speaker information is encouraged to learn similar characteristics as the original multi-channel features and can be concatenated together as input to the separation stack. Note that the encoder is shared for both the speaker stack and the separation stack. The speaker stack, illustrated in Figure 5, first em-

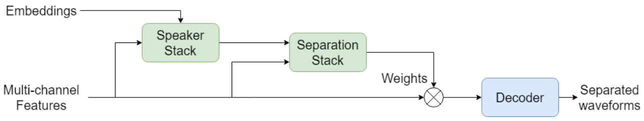

Fig. 4. Proposed multi-channel speech extractor with dedicated speaker stack

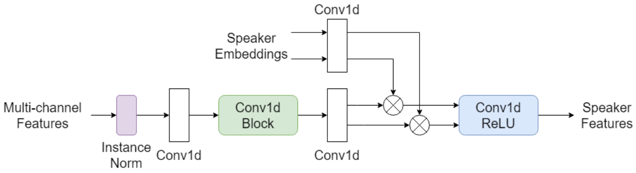

Fig. 5. Internal structure of proposed speaker stack

ploys an instance normalisation, a bottleneck 1-D CNN and a single TCN block to receive multi-channel features. Then, the output of the TCN block will be factorised by an adaptation layer into multiple features for modulation with multiple speaker embeddings, which are transformed with a 1×1 convolutional layer to the same feature dimension. The modulated signals from each speaker embedding are concatenated together and processed with a 1-D convolutional layer and a ReLU non-linear activation function to form E-dimensional speaker information features, which have the same time length as the multi-channel features.

The speaker stack and the separation stack are jointly trained to directly optimise the scale-invariant signal-to-noise ratio (SI-SNR) metric [20],

$$\begin{aligned} \text{SI-SNR} &= 10\log_{10}\frac{||\mathrm{s}_{target}||^2}{||\mathrm{e}_{noise}||^2} \\ \mathrm{s}_{target} &= \frac{\langle \hat{s}, s\rangle s}{||s||^2}, \quad \mathrm{e}_{noise} = \hat{s} - s_{target} \end{aligned} \tag{1}$$

where $\hat{s}$ and s denote the estimated and clean source, respectively, and $||s||^2 = \langle s, s\rangle$ denotes the signal power. In contrast with PIT, we condition the decoded signals on the speaker representations and keep the output speaker order consistent with the order of input speaker embeddings.

3. EXPERIMENT SETUP

3.1. Data simulation

The evaluation is performed on the WHAMR! dataset [21], which consists of simulated noisy and reverberant 2-speaker mixtures. WHAMR! is based on Wall Street Journal (WSJ) data, mixed with noise recorded in various urban environments [22], and artificial room impulse responses generated by using pyroomacoustics [23] to approximate domestic and classroom environments. There are 20k sentences from 101 speakers for training, and 3k sentences from 18 speakers for testing. The speakers in the test set do not appear during training of the speaker recognition model nor they appear during training of the speaker extraction system. All data are binaural (2-channels) and have 8 kHz sampling rate.

3.2. Speech extraction network

The multi-channel separation network in [12] trained with PIT has been set as the baseline for comparison. The hyper-parameters of the baseline model are the same as those for the best model in the original paper, chosen as follows, $N = 256$, $S = 36$, $R = 3$, $X = 7$, $L = 20$, and the batch size $M = 3$. For the U-ConvBlock based separation module, the hyper-parameters are set as SuDoRM-RF 1.0x in [13] namely, $L = 21$, $B = 16$, $Q = 4$, $C = 256$, $C_U = 512$, and the training batch size $M = 4$. Each utterance is split into multiple segments with a fixed length of 4 seconds. The dimension of speaker features, E, in the speaker stack is set to 128. The ADAM optimizer [24] is used for training with a learning rate of $1e - 3$, which will be halved if the loss of validation set is not reduced in 3 consecutive epochs. All models are trained with 100 epochs. The input for all the models is the reveberant mixture with noise and the targets are the clean individual sources.

3.3. Speaker recognition network

We retrained the time-domain speaker recognition model SincNet [25] for speaker embedding generation. Employing the same configuration as in the original paper, SincNet is trained on the clean training set of WSJ0 (101 speakers), using speech segments of 200 ms with 10 ms overlap. The output of the last hidden layer of final SincNet model represents one frame-level speaker embedding for each 200 ms segment, and an utterance-level embedding is derived by averaging all the frame predictions.

Randomly selecting a single enrollment utterance for generating the speaker embedding leads to poor extraction performance. Therefore, to increase the robustness, we follow an averaging strategy to obtain one global embedding for each speaker [26]. Specifically, each global speaker embedding is obtained by averaging several embeddings generated from multiple randomly selected utterances belonging to the same speaker. During training, one global speaker embedding is generated by averaging all the utterance-level embeddings from the training utterances belonging to the corresponding speaker. During evaluation, 3 utterances are randomly selected for each speaker, and the utterance-level embeddings from the selected utterances are averaged to form one global embedding. Experiments showed that increasing the number of utterances beyond 3 does not improve performance.

3.4. Acoustic model

To evaluate the speech recognition performance, two acoustic models have been trained using the WSJ corpus. One model (AM1) was trained on roughly 80 hrs of clean WSJ-SI284 data plus the WHAMR! single-speaker noisy reverberant speech, and the other one (AM2) was trained on the data used for AM1 plus the separated signals from the WHAMR! mixture in the training set processed by the proposed model. The audio data is downsampled to 8 kHz to match the sampling rate of data used for separation experiments. The acoustic model topology is a 12-layered Factorised TDNN [27], where each layer has 1024 units. The input to the acoustic model is 40-dimensional MFCCs and a 100-dimensional i-Vector. A 3-gram language model is used during recognition. The acoustic model is implemented with the Kaldi speech recognition toolkit [28]. With our set-up, the ASR results obtained with AM1 on the standard clean WSJ Dev93 and Eval92 are 7.2% and 5.0% WER, respectively.

4. RESULTS AND ANALYSIS

4.1. Improved Multi-channel separation network

Table 1 reports the separation performance for the improved multi-channel separation network with various configurations. The first observation is that the dimension of the spatial features does not have to be fixed to a small value (typically 36) as mentioned in the previous work. The results show that when the dimension increases, more useful spatial information is extracted and the model benefits more from the multi-channel signals. Replacing the TCN blocks with the stacked U-ConvBlocks provides a larger receptive field due to successive downsampling operations, and the latter model yields 0.5 dB SI-SNR improvement. The configuration depicted in the last row of Table 1 is used for the rest of the experiments.

Table 1. Speech separation performance of improved multi-channel structure on WHAMR! test set

Model	S	SI-SNRi
Multi-TasNet (TCN)	36	12.1
Multi-TasNet (TCN)	64	12.2
Multi-TasNet (TCN)	128	12.4
Multi-TasNet (U-Conv)	128	12.9

4.2. Results of speech extraction system

Three subsets of experiments with different speaker information conditioning strategies are performed. The first experiment uses the multiplication strategy applied in SpeakerBeam [7], which modulates the speaker embedding on the middle-stage representations in the separation module, denoted as Multiply. The second experiment repeats and concatenates the speaker embeddings with the spectral and spatial representations before being fed into the separation module, denoted as Concat. Lastly, the third experiment uses the proposed conditioning mechanism, denoted as Split.

Table 2. Speech extraction performance with improved multi-channel structure on the WHAMR! test set

Model	PIT	SI-SNRi
Separation (Improved)	✓	12.9
Extraction (Concat)	✗	12.8
Extraction (Multiply)	✗	12.9
Extraction (Split)	✗	13.3
Extraction (Split)	✓	13.4

The results in Table 2 show that the extraction model cannot directly benefit from the speaker information through the multiplication or concatenation strategies. The reason for failure of direct multiplication is presumed to be that the shared middle-stage features are not optimal for both tasks of speaker conditioning and speech reconstruction. As for the concatenation, the multi-channel features and the speaker embedding are completely different signals and cannot be suitably processed by the convolutional layer, which assume time and frequency homogeneity. Conversely, the separation model with the proposed mechanism can benefit from the speaker information and outperforms the blind source separation system and other conditioning strategies. The proposed method uses a separated speaker branch to generate high-level features for speaker conditioning tasks to alleviate the shared feature problem. And the sequential speaker features from the speaker branch can have a similar signal characteristic to the multi-channel features, which is a suitable input to the convolutional layers.

It should be noted that the proposed speech extraction system can be evaluated without accessing reference clean speech to find the right permutation. When the system is evaluated with the PIT criterion to find the oracle permutation, there is only a small difference between the two results. This demonstrates that our system can successfully identify and track multiple speakers in noisy and reverberant acoustic conditions.

Table 3. Results on different and same gender mixtures

Model	#nchs	PIT	SI-SNRi	
			Diff.	Same
SuDo-RMRF [13]	1	✓	10.6	9.1
Multi-TasNet (TCN)	2	✓	12.4	12.4
Multi-TasNet (U-Conv)	2	✓	12.9	12.9
Extraction (Split)	2	✗	13.5	13.1
Extraction (Split)	2	✓	13.5	13.3

Table 3 reports the performance of various systems with different and same gender WHAMR! mixture speech. For blind source separation, a single-channel system can achieve better separation performance with different gender mixtures than same gender mixtures. With the spatial information, a multi-channel system improves performance in both conditions and reduces the gap between the two mixture conditions. With the additional speaker information, the performance in the different gender condition is further boosted. It can be also noticed that the same gender mixtures are more challenging, and more future work is needed to find better speaker representations in this case.

Table 4 compares the proposed approach with other competing systems evaluated on WHAMR!. The proposed speaker conditioning mechanism provides consistent separation performance gain in both single and multi-channel scenarios. With the additional information from multiple microphones and speaker enrollment, our system achieves the best performance.

Table 4. Comparative results of single and multi-channel speech separation/extraction on WHAMR! data

Model	#nchs	Building Unit	PIT	SI-SNRi
Conv-TasNet [29]	1	TCN	✓	9.3
SuDo-RMRF [13]	1	U-Conv	✓	9.9
Wavesplit [19]	1	TCN	✓	12.0
Nachmanis's [5]	1	RNN	✓	12.2
Multi-TasNet [12]	2	TCN	✓	12.1
Extraction (Split)	1	U-Conv	✗	11.1
Extraction (Split)	1	U-Conv	✓	11.1
Extraction (Split)	2	U-Conv	✗	13.3
Extraction (Split)	2	U-Conv	✓	13.4

Table 5. Speech recognition results

System	#nchs	WER(%)	
		AM1	AM2
Mixture	-	79.1	77.0
Multi-TasNet [12]	2	37.7	-
Extraction (Split)	2	**31.6**	**20.9**
Noisy Oracle	-	19.8	20.0

Table 5 reports the ASR results. The proposed speech extraction model yields a significant WER reduction over the noisy reverberant mixture and outperforms the strong multi-channel separation baseline. The extraction system can introduce distortions to the separated signals (causing a mismatch problem between training and testing of the acoustic model), therefore, by decoding the data with AM2, the WER is further reduced by 34% relatively, which is close to the result obtained with oracle single-speaker noisy reverberant speech (last row in Table 5).

In future work, we plan to exploit other speaker recognition models for embedding generation, and to train these models with larger and more challenging datasets, such as VoxCeleb [30]. Moreover, we will investigate joint training of the speaker embedding and the proposed speech extraction networks, which is expected to benefit both tasks [10].

5. CONCLUSIONS

In this paper, we have presented a multi-channel speech extraction system with a novel speaker conditioning mechanism. By introducing an additional speaker branch for receiving external speaker features, this mechanism solves the problems caused by feature sharing from contradicting tasks and difference between multiple inputs, providing a more effective way to use the speaker information to improve separation performance. Informed by multiple speaker embeddings, the proposed system is able to simultaneously output corresponding sources from a noisy and reverberant mixture, without a label permutation ambiguity. Experiments on WHAMR! simulated 2-speaker mixtures have shown that the proposed multi speaker extraction approach outperforms a strong blind speech separation baseline based on PIT.

6. REFERENCES

[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2016, pp. 31–35.

[2] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[3] G.-P. Yang, S.-L. Wu, Y.-W. Mao, H.-y. Lee, and L.-s. Lee, "Interrupted and cascaded permutation invariant training for speech separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2020, pp. 6369–6373.

[4] L. Drude, T. von Neumann, and R. Haeb-Umbach, "Deep attractor networks for speaker re-identification and blind source separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2018, pp. 11–15.

[5] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," *arXiv preprint arXiv:2003.01531*, 2020.

[6] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černockỳ, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[7] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving speaker discrimination of target speech extraction with time-domain Speakerbeam," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2020.

[8] M. Ge, C. Xu, L. Wang, C. E. Siong, J. Dang, and H. Li, "SpEx+: A complete time domain speaker extraction network," *ArXiv*, vol. abs/2005.04686, 2020.

[9] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. Interspeech 2019*, 2019.

[10] X. Ji, M. Yu, C. Zhang, D. Su, T. Yu, X. Liu, and D. Yu, "Speaker-aware target speaker enhancement by jointly learning with speaker embedding extraction," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2020, pp. 7294–7298.

[11] J. Shi, J. Xu, Y. Fujita, S. Watanabe, and B. Xu, "Speaker-conditional chain model for speech separation and extraction," *arXiv preprint arXiv:2006.14149*, 2020.

[12] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2020, pp. 6389–6393.

[13] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm-rf: Efficient networks for universal audio source separation," *arXiv preprint arXiv:2007.06833*, 2020.

[14] R. Gu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2020, pp. 7319–7323.

[15] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Proc. European Conference on Computer Vision*, 2016, pp. 47–54.

[16] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2020, pp. 46–50.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[18] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[19] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *arXiv preprint arXiv:2002.08933*, 2020.

[20] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2019, pp. 626–630.

[21] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "WHAMR!: Noisy and reverberant single-channel speech separation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2020, pp. 696–700.

[22] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "WHAM!: Extending speech separation to noisy environments," in *Proc. Interspeech 2019*, 2019.

[23] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2018, pp. 351–355.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018.

[26] W. Li, P. Zhang, and Y. Yan, "Target Speaker Recovery and Recognition Network with Average x-Vector and Global Training," in *Proc. Interspeech 2019*, 2019, pp. 3233–3237.

[27] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," *Proc. Interspeech 2018*, pp. 3743–3747, 2018.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[29] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[30] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.