ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096684

MULTI-RESOLUTION LOCATION-BASED TRAINING FOR MULTI-CHANNEL CONTINUOUS SPEECH SEPARATION

Hassan Taherian[1], and *DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA

taherian.1@osu.edu, dwang@cse.ohio-state.edu

ABSTRACT

The performance of automatic speech recognition (ASR) systems severely degrades when multi-talker speech overlap occurs. In meeting environments, speech separation is typically performed to improve the robustness of ASR systems. Recently, location-based training (LBT) was proposed as a new training criterion for multi-channel talker-independent speaker separation. Assuming fixed array geometry, LBT outperforms widely-used permutation-invariant training in fully overlapped utterances and matched reverberant conditions. This paper extends LBT to conversational multi-channel speaker separation. We introduce multi-resolution LBT to estimate the complex spectrograms from low to high time and frequency resolutions. With multi-resolution LBT, convolutional kernels are assigned consistently based on speaker locations in physical space. Evaluation results show that multi-resolution LBT consistently outperforms other competitive methods on the recorded LibriCSS corpus.

Index Terms— Continuous speech separation, complex spectral mapping, location-based training.

1. INTRODUCTION

Multi-talker speaker separation methods based on deep neural networks (DNNs) have achieved impressive progress in recent years [1]. Existing separation methods are mainly concerned with fully overlapped utterances in short segments. For practical applications, however, a separation model is required to process long audio recordings with occasional speech overlaps. Recently, continuous speech separation (CSS) is proposed to handle overlapped speech in conversational or meeting environments [2]. CSS usually divides a long audio stream into short segments. Each segment is independently processed by a separation model with permutation-invariant training (PIT) criterion. Separated signals are then concatenated to create individual speaker streams.

This research was supported in part by two National Science Foundation grants (ECCS-1808932 and ECCS-2125074), the Ohio Supercomputer Center, and the Pittsburgh Supercomputer Center (NSF ACI-1928147).

Many studies have been proposed to improve different aspects of the CSS framework [3, 4, 5, 6, 7, 8, 9, 10, 11]. We introduced a modulation factor based on segment overlap ratio to dynamically adjust the separation loss [3]. In [4], a recurrent selective attention network is used to separate one speaker at a time. The work in [5] and [6] proposed new training criteria that generalizes PIT to capture long speech contexts. Li et al. [7] proposed a dual-path separation model to leverage inter-segment information from a memory embedding pool. Wang et al. adopted a multi-stage approach by combining a multi-channel separation model with masking-based beamforming and a post-filtering network to achieve strong separation performance [8]. Other works employed unsupervised multi-channel source separation for CSS. Saijo et al. proposed a spatial loss that uses the estimated direction of arrival to impose spatial constraints on the demixing matrix [9]. Boeddeker et al. [10] introduced an initialization scheme for a beamformer based on a complex Angular Central Gaussian Mixture model.

In our previous study, we introduced a new training criterion, named location-based training (LBT), to assign DNN outputs according to speaker locations in physical space [12]. We showed that LBT performs better than PIT for fully overlapped utterances in simulated and matched reverberant conditions. This paper extends LBT to conversational multi-channel speaker separation and unmatched reverberant conditions. We train a multichannel input single-channel output (MISO) system which directly estimates the real and imaginary spectrograms of speakers from multi-channel mixture. We also propose a multi-resolution loss for convolutional neural networks by estimating the complex spectrograms from low to high time and frequency resolutions. With the multi-resolution loss, convolutional kernels are consistently paired with speakers according to their locations. We demonstrate that multi-resolution LBT generalizes well to real conversational recordings, despite using simulated room impulse responses (RIRs) for training only. Experiments on the LibriCSS corpus [2] show that the proposed separation model produces superior separation performance compared to other competitive methods.

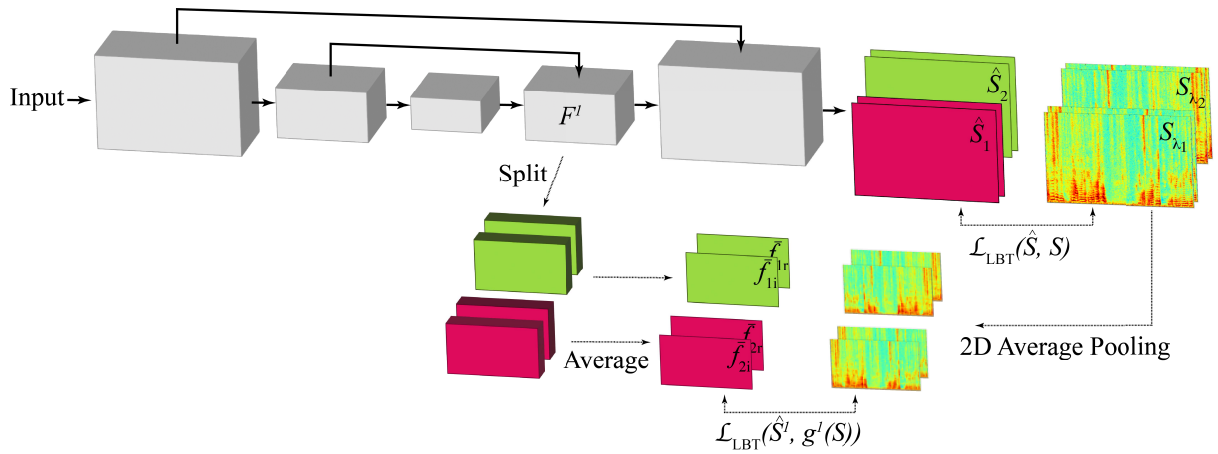

Fig. 1: Schematic diagram of multi-resolution location-based training for a MC-CSM network with $K_e = 3$ encoder layers and $K_d = 2$ decoder layers. A strided 2×2 depthwise convolutional layer is used after each dense block for downsampling and upsampling in encoder and decoder, respectively. Each cube represents an output feature map. Skip connections link encoders to the corresponding decoders. Feature maps in the decoder middle layers are divided into four groups and assigned to speakers according to their locations.

2. SYSTEM DESCRIPTION

2.1. Location-based training for CSS

The objective of CSS is to generate multiple non-overlapped streams from a multi-channel signal that contains occasional speech overlaps [2]. The CSS framework can handle any number of speakers as long as there are at most N active speakers in a segment at the same time. In this study, we assume $N = 2$ as more than two concurrent speakers rarely happens for a short length segment. Most prior works in CSS employ utterance-wise PIT for training the separation model [2, 8, 13]:

$$\mathcal{L}_{\text{PIT}}(\hat{S}, S) = \min_{\phi \in \Phi} \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n, S_{\phi(n)}), \tag{1}$$

where $\hat{S}$ and S are estimated and clean speech signals in the short-time Fourier transform (STFT) domain, respectively. $\mathcal{L}$ denotes a loss function, symbol Φ the set of all permutations and ϕ refers to one permutation. For the segments with no speech overlap, the PIT-based separation model puts the input to one of its outputs while the other output channel produces a zero signal.

Recently LBT was proposed to resolve the permutation ambiguity problem in multi-channel talker-independent speaker separation [12]. LBT leverages distinct spatial locations of multiple speakers in physical space and produces superior separation performance compared to PIT. Moreover, LBT has a linear training complexity to the number of speakers which computationally much more efficient than PIT with a factorial complexity. In this study, we extend the LBT criterion to CSS:

$$\mathcal{L}_{\text{LBT}}(\hat{S}, S) = \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n, S_{\lambda_n}), \tag{2}$$

where $\lambda_n \in \{1, 2\}$ is speaker index sorted in ascending order based on speaker azimuths or distances relative to the microphone array. Different from PIT, Eq. (2) organizes DNN outputs on the basis of speaker azimuths or distances. For non-overlap segments, the input is mapped to the first output and a zero signal is assigned to the second output.

2.2. Multi-channel complex spectral mapping

Our separation model is based on multi-channel complex spectral mapping (MC-CSM) which achieves better separation performance than masking-based beamforming [14, 15]. We employ the Dense-UNet architecture proposed in [16] for MC-CSM. The Dense-UNet input is the real and imaginary components of the mixture STFT at all microphones as well as the spectral magnitude of the first microphone. The Dense-UNet contains an encoder and decoder with $K_e = 5$ and $K_d = 4$ layers of densely-connected convolutional blocks, respectively. Each dense block has 5 convolutional layers with 76 channels, a kernel size of 3×3, and a stride of 1×1. After the last dense block, we use a 1×1 convolutional layer with 4 channels to produce estimates of the real and imaginary components for each speaker. MC-CSM contains around 6.9 million parameters. For more details about the Dense-UNet architecture, we refer the reader to [16]. The model is trained with the loss function proposed by [17]:

$$\begin{aligned} \mathcal{L}(\hat{S}, S) = &\left\| \hat{S}^{(r)} - S^{(r)} \right\|_1 + \left\| \hat{S}^{(i)} - S^{(i)} \right\|_1 \\ &+ \left\| |\hat{S}| - |S| \right\|_1, \end{aligned} \tag{3}$$

where superscripts r and i denote real and imaginary parts, $|\,.\,|$ computes magnitude and $\|\,.\,\|_1$ computes ℓ_1 norm.

2.3. Multi-resolution location-based training

The key idea of multi-resolution LBT is to estimate the complex spectrograms using dedicated convolutional kernels in the decoder by progressively increasing the time and frequency resolutions. Fig. 1 illustrates multi-resolution LBT for a Dense-UNet with $K_e = 3$ encoder layers and $K_d = 2$ decoder layers. For every middle layer k in the decoder, we split the channel dimension of the output feature maps F^k into four groups:

$$f^k_{1r}, f^k_{1i}, f^k_{2r}, f^k_{2i} = \text{Split}(F^k) \tag{4}$$

where f^k_{nr} and f^k_{ni} are the real and imaginary feature maps for speaker $n \in \{1, 2\}$, respectively. Then, the feature maps within each group are averaged across the channel dimension. A low-resolution estimate of the speech signals in the STFT domain is created at layer k for each speaker:

$$\hat{S}^k_n = \bar{f}^k_{nr} + i\bar{f}^k_{ni} \tag{5}$$

where $\bar{f}$ is the averaged feature maps and i denotes the imaginary unit. The LBT loss between lower resolution estimates and clean signals is calculated for every middle layer and added to the final loss:

$$\mathcal{L}_{\text{LBT}}(\hat{S}, S) + \sum_{k}^{K_d-1} \mathcal{L}_{\text{LBT}}(\hat{S}^k, g^{K_d-k}(S)) \tag{6}$$

where $g^x(.)$ is a 2D average pooling function with a kernel size of 2×2 and a stride of 2×2, applied recursively for x iterations. Intuitively, we assign a group of convolutional kernels to learn the real and imaginary spectrograms of each speaker with different time and frequency resolutions. The kernel assignment is consistent at every decoder layer based on sorted speaker azimuth angles or distances.

It is important to note that the approach we are proposing is distinct from multi-scale loss [18], which uses a shared decoder for a PIT-based separation model and enforces the production of a waveform estimate after each layer. For the multi-resolution loss to be effective, it is crucial to ensure that convolution kernels are consistently assigned. While we acknowledge that combining the multi-resolution loss with the PIT criterion is possible, it may not perform optimally as the PIT criterion allows for label permutation and the order of convolution kernels can differ from one mixture to another.

3. EXPERIMENTAL SETUP

3.1. Datasets

We evaluate the proposed separation models with the LibriCSS corpus [2] which is a conversational speech recognition task with speech overlaps. This dataset contains 10 hours of recordings from the Librispeech development set that are retransmitted with loudspeakers to capture real room reverberation. The loudspeakers are positioned around a table and their distance to the array ranges from 33 to 409 cm. The recordings are divided into 6 sessions with different overlap ratios: 0S (no overlap with a 0.1-0.5s pause between utterances), 0L (no overlap with a 2.9-3.0s pause between utterances), 10%, 20%, 30% and 40% overlaps. The LibriCSS corpus is recorded using a 7-channel circular microphone array with 4.25 cm radius.

We follow the setup described in [8] for simulating training and validation data using the image method [19, 20]. We created 192K two-speaker mixtures with different overlap ratios using the LibriSpeech dataset. Simulating the LibriCSS recording device, each mixture is convolved with 7-channel microphone array RIRs. For generating RIRs, we simulate rectangular rooms with random length, width and height dimensions in the range of [5×5×3, 10×10×4] meters, with the microphone array placed in the center of the room. The speaker positions are uniformly sampled from 360 candidate azimuth angles in the range of -180° to 180° with a 1° resolution. The speaker-array distances are randomly selected such that the two speakers are at least 5 cm far apart. The reverberation time (T60) is randomly sampled between 0.2 and 0.6 s. Stationary ambient noise is also added to the mixture signal at a random SNR from 10 to 30 dB.

The separation models are evaluated using the default ASR system provided with the LibriCSS corpus [2]. The LibriCSS contains two evaluation scenarios: utterance-wise and continuous evaluations. In the utterance-wise scenario, word error rate (WER) is calculated for both separated signals of pre-segmented mixtures, and the one with lower WER is considered. The continuous scenario evaluates long mixture segments that contain 8-10 utterances. The decoding results from both streams are combined for evaluation.

For both scenarios, the input mixture is processed in 2.4 s segments with a segment shift of 1.2 s. The processed segments are concatenated using the stitching algorithm proposed by [2]. All signals are sampled at 16 kHz. We use STFT with a frame length of 32 ms and a frame shift of 8 ms. The sample variance of the multi-channel input mixture is normalized. Global mean and variance normalization is also applied to STFT features.

3.2. Speaker counting

For the segments without speaker overlaps, the separation model is expected to produce a zero signal in one of its outputs. However, it is observed that the model sometimes emits an intelligible residual signal for non-overlapped segments, leading to more decoding errors. Following [8], we utilize a frame-level speaker counting (SC) to find non-overlapped frames and remove the signal residual by combining the two separated outputs. For each frame, SC produces the probabilities for three classes: zero, one or two speakers. The SC

Table 1: WER results (in %) of comparison systems for utterance-wise and continuous evaluation on LibriCSS. '$MISO_1$+SC' and 'MC-CSM+SC' refer to $MISO_1$ and MC-CSM with a speaker counting network, respectively.

			Utterance-wise						**Continuous**					
	Criterion	Multi-resolution	0S	0L	10%	20%	30%	40%	0S	0L	10%	20%	30%	40%
Unprocessed	–	–	11.8	11.7	18.8	27.2	35.6	43.3	15.4	11.5	21.7	27.0	34.3	40.5
BLSTM [2]	PIT	–	8.3	8.4	11.6	16.0	18.4	21.6	11.9	9.7	13.4	15.1	19.7	22.0
Conformer [13]	PIT	–	7.2	7.5	9.6	11.3	13.7	15.1	11.0	8.7	12.6	13.5	17.6	19.6
$MISO_1$ [8]	PIT	–	7.7	7.5	7.9	9.6	11.3	13.0	10.7	10.5	10.9	11.5	13.8	15.3
$MISO_1$+SC [8]	PIT	–	–	–	–	–	–	–	7.9	8.5	8.5	10.5	12.3	14.3
MC-CSM	PIT	–	7.0	7.5	7.9	9.3	11.3	13.0	10.0	9.4	11.1	11.8	14.3	15.2
	Azimuth	–	8.6	8.8	9.2	10.3	11.3	12.4	13.5	14.5	13.6	13.4	14.7	15.4
	PIT	✓	7.6	7.6	7.8	9.4	11.0	12.8	11.4	12.5	11.3	13.0	15.1	16.3
	Azimuth	✓	7.3	7.9	7.6	**9.2**	**10.8**	**11.4**	9.2	9.8	9.1	10.5	12.2	**12.8**
	Distance	✓	8.7	10	10.5	12.9	14.6	17.8	12.2	12.8	12.4	14.8	17.1	19.0
MC-CSM+SC	PIT	–	**6.5**	**6.9**	7.4	9.3	11.7	13.6	8.0	**7.4**	8.8	10.7	12.6	13.8
	Azimuth	–	7.4	7.7	8.1	9.5	11.3	12.5	7.9	7.9	8.5	9.4	**11.6**	12.9
	PIT	✓	**6.5**	7.3	**7.3**	9.6	11.3	13.5	8.1	8.1	8.3	10.2	12.4	14.0
	Azimuth	✓	6.7	7.1	7.5	**9.2**	10.9	12.1	**7.7**	7.7	**8.0**	**9.3**	**11.6**	13.0
	Distance	✓	7.6	8.3	9.3	12.1	14.2	17.8	8.3	8.4	9.4	11.7	14.9	16.8

model is based on a modified MC-CSM architecture. Specifically, SC consists of the MC-CSM encoder followed by two bi-directional long short-term memory (BLSTM) layers and a softmax layer. The SC labels for training were generated using a pre-trained DNN based voice activity detector [21].

4. EVALUATION RESULTS

We trained MC-CSM using PIT and LBT criteria with and without the multi-resolution loss. For LBT, output-speaker assignments are ordered based on speaker azimuth angles. Table 1 presents the WER results of MC-CSM models for utterance-wise and continuous evaluations. We observe MC-CSM trained with azimuth criteria (Eq. (2)) yields comparable results to PIT in higher overlap ratios but it performs worse than PIT in lower overlap ratios. However, clear improvement is observed with azimuth-based training when SC is used to suppress the residual signal in non-overlapped segments. Combining the multi-resolution loss (Eq. (6)) and azimuth-based training shows excellent speech recognition performance and outperforms PIT with or without speaker counting. For example, MC-CSM based on multi-resolution LBT leads to a 2.4% absolute WER reduction compared to PIT on the 40% overlap condition in continuous evaluation. The results indicate that LBT models trained with simulated RIRs generalize well to real microphone array recordings. On the other hand, PIT-based separation model performance does not improve when the multi-resolution loss is used. This is because the PIT criterion is order-agnostic and features maps are not sorted consistently with the multi-resolution loss for different mixtures.

We also trained MC-CSM with LBT by using speaker distances instead of azimuth angles. As shown in the table, distance-based training underperforms both azimuth-based training and PIT. The reason could be that the distance criterion is not discriminative enough in the LibriCSS dataset as loudspeakers are positioned symmetrically and their distances to the microphone array are close.

Finally, we compare our MC-CSM models with other competitive separation methods that are evaluated with the LibriCSS default ASR. For all methods, we list the best reported results, and leave unreported fields blank. The proposed systems in [2] and [13] estimate real-valued masks using BLSTM and conformer architectures with 21.80M and 58.72M parameters, respectively. The system in [8] performs complex spectral mapping using a UNet model (6.9M) with a temporal convolutional network. A dedicated speaker counting network was also utilized to further improve WER scores. Our LBT models trained with the multi-resolution loss produce much better WER scores compared to other systems. We should mention that WER scores can be further decreased by using a speech enhancement network on top of the separation model as done in [8]. Our focus in this study is on the separation model, and further improvements can be expected by introducing speech enhancement.

5. CONCLUDING REMARKS

In this study, we have extended LBT to conversational multi-channel speaker separation. We have proposed a multi-resolution loss to consistently assign convolutional kernels according to speaker locations. We have demonstrated LBT yields significantly better separation and ASR performance than other competitive methods on the LibriCSS corpus.

6. REFERENCES

[1] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1702–1726, 2018.

[2] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *Proc. ICASSP*, 2020, pp. 7284–7288.

[3] H. Taherian and D. L. Wang, "Time-domain loss modulation based on overlap ratio for monaural conversational speaker separation," in *Proc. ICASSP*, 2021, pp. 5744–5748.

[4] Y. Zhang, Z. Chen, J. Wu, T. Yoshioka, P. Wang, Z. Meng, and J. Li, "Continuous speech separation with recurrent selective attention network," in *Proc. ICASSP*, 2022, pp. 6017–6021.

[5] W. Zhang, Z. Chen, N. Kanda, S. Liu, J. Li, S. Emre Eskimez, T. Yoshioka, X. Xiao, Z. Meng, Y. Qian, and F. Wei, "Separating long-form speech with group-wise permutation invariant training," in *Proc. Interspeech*, 2022, pp. 5383–5387.

[6] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "Graph-PIT: Generalized permutation invariant training for continuous separation of arbitrary numbers of speakers," in *Proc. Interspeech*, 2021, pp. 3490–3494.

[7] C. Li, Z. Chen, and Y. Qian, "Dual-path modeling with memory embedding model for continuous speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1508–1520, 2022.

[8] Z.-Q. Wang, P. Wang, and D. L. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 2001–2014, 2021.

[9] K. Saijo and R. Scheibler, "Spatial loss for unsupervised multi-channel source separation," in *Proc. Interspeech*, 2022, pp. 241–245.

[10] C. Boeddeker, T. Cord-Landwehr, T. von Neumann, and R. Haeb-Umbach, "An initialization scheme for meeting separation with spatial mixture models," in *Proc. Interspeech*, 2022, pp. 271–275.

[11] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo, N. Kanda, J. Li, S. Wisdom, and J. R. Hershey, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *Proc. SLT Workshop*, 2021, pp. 897–904.

[12] H. Taherian, K. Tan, and D. L. Wang, "Multi-channel talker-independent speaker separation through location-based training," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2791–2800, 2022.

[13] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *Proc. ICASSP*, 2021, pp. 5749–5753.

[14] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 483–492, 2016.

[15] K. Tan, Z.-Q. Wang, and D. L. Wang, "Neural spectrospatial filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 605–621, 2022.

[16] Y. Liu and D. L. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 2092–2102, 2019.

[17] Z.-Q. Wang and D. L. Wang, "Deep learning based target cancellation for speech dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 941–950, 2020.

[18] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. ICML*, 2020, pp. 7164–7175.

[19] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[20] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP*, 2018, pp. 351–355.

[21] V. Manohar. ASpIRE SAD Model. [Online]. Available: https://kaldi-asr.org/models/m4