ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413900

TIME-DOMAIN LOSS MODULATION BASED ON OVERLAP RATIO FOR MONAURAL CONVERSATIONAL SPEAKER 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

Existing speaker separation methods deliver excellent performance on fully overlapped signal mixtures. To apply these methods in daily conversations that include occasional concurrent speakers, recent studies incorporate both overlapped and non-overlapped segments in the training data. However, such training data can degrade the separation performance due to triviality of non-overlapped segments where the model reflects the input to the output. We propose a new loss function for speaker separation based on permutation invariant training that dynamically reweighs losses using the segment overlap ratio. The new loss function emphasizes overlapped regions while deemphasizing the segments with single speakers. We demonstrate the effectiveness of the proposed loss function on an automatic speech recognition (ASR) task. Experiments on the recently introduced LibriCSS corpus show that our proposed single-channel method produces consistent improvements compared to baseline methods.

Index Terms— Monaural speech separation, automatic speech recognition, overlapped speech, speaker separation

1. INTRODUCTION

Speaker separation is one of the fundamental tasks in signal processing and it aims to separate several concurrent speakers. Speaker separation needs to be performed in order to improve the robustness of automatic speech recognition (ASR) and speaker diarization as these systems usually assume no overlap between speakers' utterances. Speaker separation can be categorized as talker-dependent, where speakers remain the same during training and testing, or as talker-independent, where test speakers can be different from training ones [1]. In this study, we address talker-independent monaural speaker separation consists of segments with two concurrent speakers, and single-talker segments.

Recent speaker separation methods have achieved impressive separation performance [2, 3, 4, 5]. However, these studies are mainly concerned with fully overlapped utterances. Without extension, such methods do not perform well in a conversational (or meeting) environment, in which overlapped speech accounts for just a small proportion of the entire conversation [6]. This is partly caused by the mismatch in overlap proportions between training and test conditions. A simple way to adapt is to perform overlap detection and speaker separation in tandem, i.e., only separating the signal during detected overlap intervals. The problem with this approach is that it requires an accurate overlap detector since a false alarm would result in erroneous separation. Another straightforward approach is to include both overlapped and non-overlapped utterances in the training data [7, 8]. However, with this training scheme, the gradients can be dominated by trivial predictions in the non-overlapped regions where the model would be expected to map the input to itself as the output. With a larger proportion in the training data, the non-overlapped segments could overwhelm the training and lead to degenerate models.

In this work, we propose a novel loss function that adaptively rescales the time-domain loss with utterance-level permutation invariant training (uPIT) [9] based on segment overlap ratio. Our loss function is designed to downplay the easy segments (non-overlapped regions), thus emphasizing the hard segments (overlapped regions). We find this simple loss function to be highly effective. We also develop a Dense-UNet model [3] that is influenced by the overlap information with Feature-wise Linear Modulation (FiLM) layers [10]. To process continuous audio streams, we employ speaker embeddings to align the separated outputs of segments. We find improvements with the both proposed model and loss function in terms of lower word error rate (WER) in the LibriCSS dataset [11]. Note that although our focus in this paper is single-channel separation, our approach can be easily extended to multi-channel processing using masking based beamforming [1, 12].

2. OVERLAP RATIO MODULATION

The task of conversational speaker separation is to separate independent speech sources x, where speech overlap occurs in the conversation. In this work, we assume that there are 2

Authorized licensed use limited to: The Ohio State University. Downloaded on June 11,2021 at 16:22:23 UTC from IEEE Xplore. Restrictions apply.

concurrent speakers in the overlapped regions. For the segments that contain no speech overlap, the separation model emits the input to one of its outputs while the other output channel produce zero or negligible noise.

Recent studies employ time-domain signal-to-noise ratio (SNR) loss for training the separation model and report substantial improvements in the separation performance [3, 4, 13]. This loss function is combined with uPIT [9] to address the source permutation problem [1]:

$$\mathcal{L}_{SNR} = \min_{\theta_n \in S} \frac{1}{2} \sum_{n=1}^{2} 10 \log_{10} \frac{\|x_n - \hat{x}_{\theta_n}\|^2}{\|x_n\|^2} \quad (1)$$

where S and $\hat{x}$ denote the permutation space and the estimated source, respectively. For the conversational speaker separation task, we train the separation model with a combined dataset that contains both overlapped and overlapped-free audio segments. However, using Eq. (1) as the training function causes numerical instabilities since one of the ground-truth signals is zero in the non-overlapped regions. Following [4], we address this issue by replacing the denominator by a constant that corresponds to the average norm of the training data.

Moreover, it is shown in [4] that training with the combined dataset degrades the separation performance compared to a model that is trained only with fully overlapped dataset. We speculate that the gradient update is influenced by easy estimation of the non-overlapped segments which makes the separation training ineffective. The similar problem arises in the field of object detection where the model should detect scarce objects from countless of easily classified background examples. This problem is typically addressed via adjusting the loss scales based on example difficulty and avoiding major gradient updates on trivial predictions [14, 15].

Inspired by the study in object detection [15], we propose a modulating factor that regulates the loss function scale based on the overlap ratio of the training data:

$$\mathcal{L}_{ORM} = (\sqrt{1+p} - \beta)\mathcal{L}_{SNR} \quad (2)$$

where $0 \leq p \leq 1$ is the ratio of the overlapped frames over the total number of the frame sequence and $0 < \beta < 1$ is the offset hyperparameter. The new loss function which we call it overlap ratio modulation is simple and can be easily derived from simulated training data. Intuitively, the modulating factor reduces the loss contribution from non-overlapped regions and instead emphasize on the segments that are fully overlapped. For example, if $p = 1$ i.e., the entire frame sequence is overlapped, we increase the magnitude of the loss function to increase the contribution of fully overlapped segments. By the same token, when $p = 0$, meaning no frame is overlapped in the sequence, the loss function is suppressed, and thus we down-weight the less informative segments when updating the model.

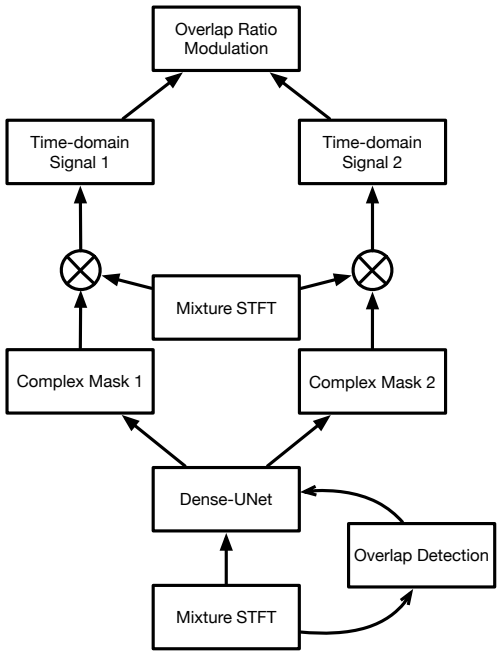

Fig. 1: Diagram of the separation model with FiLM layers. We use TCN for frame-level overlap detection.

3. EXPERIMENTAL SETUP

3.1. Separation Model

We employ the Dense-UNet model proposed by [3] as our baseline separation model. An input mixture is represented as a stack of real and imaginary short-time Fourier transform (STFT). The model outputs two complex ratio masks cRM_n which are multiplied with the input mixture to estimate the reconstructed sources in the complex domain [16]:

$$\hat{X}_n(t,f) = \text{cRM}_n(t,f) \otimes Y(t,f) \quad (3)$$

where $\hat{X}_n(t,f)$ and $Y(t,f)$ respectively represent the STFT value of the estimated source n, and input mixture at time t and frequency f. Symbol $\otimes$ denotes point-wise complex multiplication. In the end, inverse STFT is used to resynthesize the waveforms. The Dense-UNet model comprises 4 downsampling layers and 4 upsampling layers interleaved with 9 densely-connected convolutional neural network (CNN) blocks. In each block, layers are connected to every other layer in a feed-forward fashion:

$$z_l = F(conv([z_{l-1}, z_{l-2}, \cdots, z_0])) \quad (4)$$

where z_0 and z_l are the input feature maps and the output of the l^{th} layer, respectively. $conv$ denotes the convolutional layer with 64 channels, a kernel size of 3×3 and a stride of 1×1. $[...]$ denotes the concatenation operation and F represents the exponential linear unit activation followed layer normalization. There are 5 dense layers in each block.

Authorized licensed use limited to: The Ohio State University. Downloaded on June 11,2021 at 16:22:23 UTC from IEEE Xplore. Restrictions apply.

We also extend the baseline Dense-UNet model by incorporating the FiLM layers [10] in each dense block. The idea of using the FiLM layers is to influence the separation model via a feature-wise affine transformation based on conditioning information. Specifically, we replace the dense layers in Eq. (4) with:

$$z_l = F(h_l(c) * conv([z_{l-1}, z_{l-2}, \cdots, z_0]) + h'_l(c)) \quad (5)$$

where $h_l(c)$ and $h'_l(c)$ are affine transformations of the condition vector c. We use frame-level overlap sequence predictions as the condition vector. A Temporal convolutional network (TCN) [17] is utilized as a frame-level binary classifier for overlap detection. A diagram of the proposed model is illustrated in Fig. 1.

The TCN consists of 6 of dilated convolutional blocks with an exponentially increasing dilation factor and a final classification layer with 2 units and softmax activation. Each dilated convolutional block has a 1-D CNN layer with 256 channels, batch normalization and ReLU activation. We concatenate the real and imaginary of mixture's STFT as the input features. The TCN is trained with cross-entropy loss. The TCN outputs of T sequence frames are concatenated and used as the condition vector $c \in \mathbb{R}^{2T}$ in the separation model. The separation model and overlap detection are jointly trained.

3.2. Dataset

We test our separation models with the LibriCSS corpus [11]. The recordings consists 10 hours of audio from Librispeech development set that are retransmitted with loudspeakers to capture real room reverberation. The dataset has 10 sessions, each of which is divided into 6 mini-sessions that have different overlap ratios from 0% to 40%. The overlap ratio is defined as the total overlapped region length over the total speech length. The 0% overlap ratio contains two scenarios, one has a short pause (0.1-0.5s) between utterances while the other has a long silence (2.9-3.0s). The LibriCSS corpus is recorded with seven-channel circular microphone array and use the first microphone for monaural speaker separation.

The separation accuracy is evaluated by the default ASR system provided with the LibriCSS corpus [11]. The evaluation protocol of LibriCSS contains two evaluation configuration: *utterance-wise evaluation* where the ground-truth utterance boundaries are provided and each utterance is processed with the speaker separator independently and *continuous evaluation* where the boundaries are unknown and recordings are processed in long segments that contains 8-10 utterances.

We train our model with `train-clean-{100,360}` subset of LibriSpeech dataset. We created 240 hours of conversational speech where the probability of utterances are fully overlapped, partially overlapped or not overlapped is 45%, 45% and 10%, respectively. Each conversation is convolved with a room impulse response (RIR) with reverberation time (T60) between 0.2 and 0.7 seconds using the simulation procedure described in [18]. Afterwards, stationary ambient noise is added to the mixture signal at a random SNR from 5 to 25 dB. In this study, we use STFT with a frame length of 32ms and a frame shift of 8 ms.

3.3. Stream Processing

We should address how to process a conversation that spans several hours with the separation model. In the work by [7, 19], the audio stream is processed in short sliding windows. However, since PIT training is order agnostic, the outputs of adjacent windows should be aligned to prevent the speaker signal from being swapped to different output. For the output alignment, the optimal permutation is selected based on the mean squared error between the estimated spectrograms calculated over the shared frames of two adjacent windows [7]. After the alignment, the output streams are generated by using current window frames that are not shared with the previous window.

In this study we propose a novel approach for selecting optimal permutation based on speaker embeddings. Instead of using sliding windows, we divide the input audio stream into independent segments and process each segment with the separation model. Then, a unit-length speaker embedding is extracted for the outputs that contain speech. For the silent outputs we assign a zero embedding vector. The optimal permutation for two adjacent segments is selected based on cosine similarity:

$$\underset{\theta_n \in S}{\operatorname{argmax}} \sum_{n=1}^{2} \langle e_n^{t-1}, e_{\theta_n}^{t} \rangle \quad (6)$$

where e_n^t denotes the speaker embedding for n^{th} output of segment t. The advantage of this method is twofold. First, the processing is based on independent segments, hence the output streams are more consistent than previous method which requires more frame concatenation with sliding windows. Second, the extracted speaker embeddings can be used in other applications such as speaker diarization and speaker-dependent ASR. The speaker embeddings are based on d-vectors. Following [20], we train d-vectors with 3 layers of long short-term memory (LSTM) with 256 units and generalized end-to-end loss. VoxCeleb corpus [21, 22] and the training part of LibriSpeech dataset are used for training d-vectors.

4. EVALUATION RESULTS

We first analyze the effect of segment length on processing the audio stream. We train 3 separation models with different segment lengths. Table 1 shows the WER results for continuous evaluation. We observe that the model trained with 1.6s

Authorized licensed use limited to: The Ohio State University. Downloaded on June 11,2021 at 16:22:23 UTC from IEEE Xplore. Restrictions apply.

Table 1: Continuous evaluation results (%WER) for separation model with different segment lengths. The separation model is based on Dense-UNet with FiLM layers. 0S and 0L denote 0% overlap with short and long pause, respectively.

	Overlap ratio (%)					
	0S	0L	10	20	30	40
50 frames (0.4s)	13.8	12.4	17.4	22.5	27.8	31.7
200 frames (1.6s)	12.3	11.5	16.0	21.2	27.2	30.5
400 frames (3.2s)	15.2	14.6	24.4	31.6	40.7	44.4

Table 2: Utterance-wise evaluation results (%WER). 'ORM' and 'UNet + FiLM' refer to overlap ratio modulation and Dense-UNet with FiLM layers, respectively.

		Overlap ratio (%)					
	ORM	0S	0L	10	20	30	40
No separation	—	11.8	11.7	18.8	27.2	35.6	43.3
Baseline [11]	—	12.7	12.1	17.6	23.2	30.5	35.6
UNet	—	9.8	9.0	13.6	19.2	25.1	29.8
UNet	✓	9.4	8.9	13.3	18.4	**23.3**	28.2
UNet + FiLM	—	**9.3**	8.9	13.6	20.1	25.5	30.4
UNet + FiLM	✓	9.5	**8.9**	**12.6**	**18.2**	23.4	**27.3**
Conformer [23]	—	12.9	12.2	15.1	20.1	24.3	27.6

segments outperforms the one trained with longer segment. This may be due to the greater variety in number of speakers with longer segments. Using very short segments (0.4s) also degrades the performance which can be attributed to reduced d-vector accuracy with short segments. For the rest of the experiments, we set the segment length to 1.6s.

The separation models performance for utterance-wise evaluation is presented in Table 2. We set $\beta = 0.2$ for training with overlap ratio modulation. The baseline model uses bidirectional LSTM to estimate real-valued masks [11]. It can be seen that the basic Dense-UNet yields significantly better WER scores in all scenarios compared to baseline method [11], especially in non-overlapped conditions. When the Dense-UNet is trained with a simple extension of overlap ratio modulation, we achieve consistent improvements for all scenarios. Note that the error reduction for overlapped conditions is greater compared to non-overlapped conditions. This indicates that emphasizing on the overlapped regions during training improves the separation performance without distorting the non-overlapped signals.

With regard to the comparison between the two speaker separation models, we do not observe meaningful improvements in overlapped conditions when the FiLM layers are included in Dense-UNet. By contrast, further WER reduction is achieved when the Dense-UNet with FiLM layers is combined with overlap ratio modulation. We present the performance of separation models for continuous evaluation in Table 3. One can observe similar trends to utterance-wise evaluation, when overlap ratio modulation and FiLM layers are incorporated in the training.

Table 3: Continuous evaluation results (%WER).

		Overlap ratio (%)					
	ORM	0S	0L	10	20	30	40
No separation	—	15.4	11.5	21.7	27	34.3	40.5
Baseline [11]	—	17.6	16.3	20.9	26.1	32.6	36.1
UNet	—	12.7	11.3	16.3	21.5	27.2	30.3
UNet	✓	**11.6**	11.5	15.5	**20.1**	25.6	30.6
UNet + FiLM	—	12.3	11.5	16.0	21.2	27.2	30.5
UNet + FiLM	✓	12.1	**11.1**	**15.3**	20.4	**25.0**	**28.8**
Conformer [23]	—	13.3	11.7	16.3	20.7	25.6	29.3

We also compare our separation model to the model introduced in [23]. This model is based on state-of-the-art conformer architecture with 58.72M parameters. Our Dense-UNet with FiLM layers and the LSTM for d-vector estimation have 12.26M and 1.42M parameters, respectively. With fewer parameters, our best model outperforms the conformer in all scenarios for both utterance-wise and continuous evaluation. We should mention that a very recently posted paper [24] reports state-of-the-art results for the LibriCSS evaluation. This study uses complex spectral mapping to train the separation model. In addition, a speech enhancement network is used on top of the separation model to further reduce WER. Our focus in this study is on the separation model, and we can expect further improvement by introducing speech enhancement in future work (see [25]).

5. CONCLUDING REMARKS

This paper introduces a modulation factor based on segment overlap ratio to dynamically adjust the speaker separation loss. Consistent with the object detection task, our experimental results demonstrate that a simple modification of the time-domain loss improves the separation performance without distorting the signals on non-overlapped regions. Future research will explore different forms of overlap ratio modulation and incorporate a speech enhancement module.

6. ACKNOWLEDGMENTS

This research was supported in part by a National Science Foundation grant (ECCS-1808932) and the Ohio Supercomputer Center.

Authorized licensed use limited to: The Ohio State University. Downloaded on June 11,2021 at 16:22:23 UTC from IEEE Xplore. Restrictions apply.

7. 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] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, "FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *Int. Conf. on Multimedia Modeling*, 2020, pp. 653–665.

[3] 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.

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

[5] 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. Acoust., Speech Signal Process.*, 2020, pp. 46–50.

[6] D. R. Rutter, *Communicating by telephone*. Pergmon Press, 1987.

[7] T. Yoshioka, I. Abramovski, C. Aksoylar, Z. Chen, M. David, D. Dimitriadis, Y. Gong, I. Gurvich, X. Huang, Y. Huang *et al.*, "Advances in online audio-visual meeting transcription," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding,*, 2019, pp. 276–283.

[8] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5739–5743.

[9] 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 Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 1901–1913, 2017.

[10] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *AAAI Conference on Artificial Intelligence*, 2018.

[11] 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. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7284–7288.

[12] H. Taherian, Z.-Q. Wang, J. Chang, and D. L. Wang, "Robust speaker recognition based on single-channel and multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1293–1302, 2020.

[13] 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, pp. 1256–1266, 2019.

[14] S. Ryou, S.-G. Jeong, and P. Perona, "Anchor loss: Modulating loss scale based on prediction difficulty," in *Proc. IEEE Int. Conf. Computer Vision*, 2019, pp. 5992–6001.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Computer Vision*, 2017, pp. 2980–2988.

[16] 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.

[17] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv:1803.01271*, 2018.

[18] Z.-Q. Wang and D. L. Wang, "Integrating spectral and spatial features for multi-channel speaker separation," in *Proc. Interspeech*, vol. 2018, 2018, pp. 2718–2722.

[19] T. Yoshioka, H. Erdogan, Z. Chen, X. Xiao, and F. Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," in *Proc. Interspeech*, 2018, pp. 3038—-3042.

[20] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4879–4883.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[23] S. Chen, Y. Wu, Z. Chen, J. Li, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," *arXiv:2008.05773*, 2020.

[24] Z.-Q. Wang, P. Wang, and D. L. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speaker separation," *arXiv:2010.01703*, 2020.

[25] Y. Liu, M. Delfarah, and D. L. Wang, "Deep CASA for talker-independent monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6354–6358.

Authorized licensed use limited to: The Ohio State University. Downloaded on June 11,2021 at 16:22:23 UTC from IEEE Xplore. Restrictions apply.