MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Teacher-Student Deep Clustering For Low-Delay Channel Speech Separation

Aihara, Ryo; Hanazawa, Toshiyuki; Okato, Yohei; Wichern, Gordon; Le Roux, Jonathan

TR2019-003 March 29, 2019

Abstract

The recently-proposed deep clustering algorithm introduced significant advances in monaural speaker-independent multi-speaker speech separation. Deep clustering operates on magnitude spectrograms using bidirectional recurrent networks and K-means clustering, both of which require offline operation, i.e., algorithm latency is longer than utterance length. This paper evaluates architectures for reduced latency deep clustering by combining: (1) block processing to efficiently propagate the memory encoded by the recurrent network, and (2) teacher-student learning, where low-latency models learn from an offline teacher. Compared to our best performing offline model, we only lose 0.3 dB SDR at a latency of 1.2 seconds and 0.7 dB SDR at a latency of 0.6 seconds on the publicly available wsj0-2mix dataset. Moreover, by providing a detailed analysis of the failure cases for our low-latency speech separation models, we show that the cause of this performance gap is related to frame-level permutation errors, where the network fails to accurately track speaker identity throughout an utterance.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

TEACHER-STUDENT DEEP CLUSTERING FOR LOW-DELAY SINGLE CHANNEL SPEECH SEPARATION

Ryo Aihara[1]*, Toshiyuki Hanazawa*[1]*, Yohei Okato*[1]*, Gordon Wichern*[2]*, Jonathan Le Roux*[2]

[1]Information Technology R&D Center, Mitsubishi Electric Corporation, Japan
[2]Mitsubishi Electric Research Laboratories (MERL), USA

ABSTRACT

The recently-proposed deep clustering algorithm introduced significant advances in monaural speaker-independent multi-speaker speech separation. Deep clustering operates on magnitude spectrograms using bidirectional recurrent networks and K-means clustering, both of which require offline operation, i.e., algorithm latency is longer than utterance length. This paper evaluates architectures for reduced latency deep clustering by combining: (1) block processing to efficiently propagate the memory encoded by the recurrent network, and (2) teacher-student learning, where low-latency models learn from an offline teacher. Compared to our best performing offline model, we only lose 0.3 dB SDR at a latency of 1.2 seconds and 0.7 dB SDR at a latency of 0.6 seconds on the publicly available wsj0-2mix dataset. Moreover, by providing a detailed analysis of the failure cases for our low-latency speech separation models, we show that the cause of this performance gap is related to frame-level permutation errors, where the network fails to accurately track speaker identity throughout an utterance.

Index Terms— cocktail party problem, speaker-independent speech separation, deep clustering, low-latency, chimera network

1. INTRODUCTION

The "cocktail party problem," i.e., speaker-independent multi-talker speech separation has long been known as challenging task in the speech processing community [1]. Before the deep clustering algorithm [2] was proposed, most speech separation approaches focused on multiple microphone scenarios [3, 4], speaker-dependent models [5, 6] or tasks with limited vocabulary and grammar [7]. The main hurdle for single-channel speech separation is the "permutation problem" where the correspondence between the outputs of an algorithm and the true sources is an arbitrary permutation.

In order to solve the permutation problem, deep clustering avoids estimating the target class directly, instead computing a high-dimensional embedding for each time-frequency (T-F) bin. A deep neural network learns the embeddings such that they are close to each other when they belong to the same speaker and far apart otherwise. At test time, the speaker assignment of each T-F bin can be determined by clustering the embeddings with an algorithm such as k-means. In [8], a more direct optimization is proposed where embedding estimation and clustering are combined using end-to-end training. A key property of this approach is that the same networks can be used with any number of sources.

Mask inference learning (MI) is an alternative approach, which estimates the target class for a fixed number of sources directly. This approach was first used in [2] as a baseline method, using a combination of long short term memory (LSTM) recurrent neural networks (RNNs). The segment-level permutation-free objective is used in [9]. In [10], bi-directional LSTMs (BLSTMs) and segment-level permutation-free training were adopted, where it was shown to perform nearly as well as deep clustering.

In the framework of deep learning, multitask learning is known as a powerful approach and its advantage is confirmed in graphical modeling [5], spectral clustering [11], and computational auditory scene analysis (CASA) [12]. Deep clustering and MI were combined in [13] in the sense of multitask learning and this architecture is referred to as a chimera network. In this approach, the deep clustering loss functions as a regularizer for mask inference achieving superior performance, and avoiding the need to run a clustering algorithm at inference time.

Unfortunately, the best performing deep clustering, mask inference, and chimera models are based on bidirectional recurrent networks (e.g., BLSTM), which require running a forward and backward pass over an entire utterance before separation results are obtained. This high-latency operation is unacceptable in many applications, e.g., as a front-end for speech recognition systems, however, simply replacing an offline BLSTM with a forward-only LSTM leads to unacceptable performance degradation. A trade-off between latency and separation performance can be achieved with the latency-controlled BLSTM (LC-BLSTM) [14,15], a block-based BLSTM where the input is cut into overlapping blocks and the latency is reduced to the block-size. While LC-BLSTM demonstrated promising results in a speech enhancement (i.e., separating speech from non-stationary background noise) task, its performance in multi-speaker separation is yet to be investigated.

Another, possible approach to closing the performance gap between BLSTM networks and low-latency LSTM/LC-BLSTM networks is to use teacher-student learning (also known as distillation [16]). In our case, a low-latency student network is trained to match the hidden layer outputs of a pre-trained offline BLSTM teacher. In this study, we propose several low-latency chimera network variations, and our contributions can be summarized as follows: (1) evaluating LC-BLSTM architectures for speech separation, (2) investigating teacher-student learning for speech separation, and (3) a detailed analysis of the errors made by low-latency speech separation systems. Through experiments with the publicly available wsj0-2mix dataset in Section 5 we show that an LC-BLSTM with teacher-student learning can achieve performance comparable to that of an offline BLSTM with most of the remaining performance gap due to frame-level permutation errors where the network fails to accurately maintain speaker identity over the course of an utterance.

2. DEEP CLUSTERING FOR SPEECH SEPARATION

We use as our basic architecture for single-channel speech separation the chimera++ network introduced in [13], illustrated in Fig. 1(a). The chimera++ architecture, adapted from [17], uses a shared se-

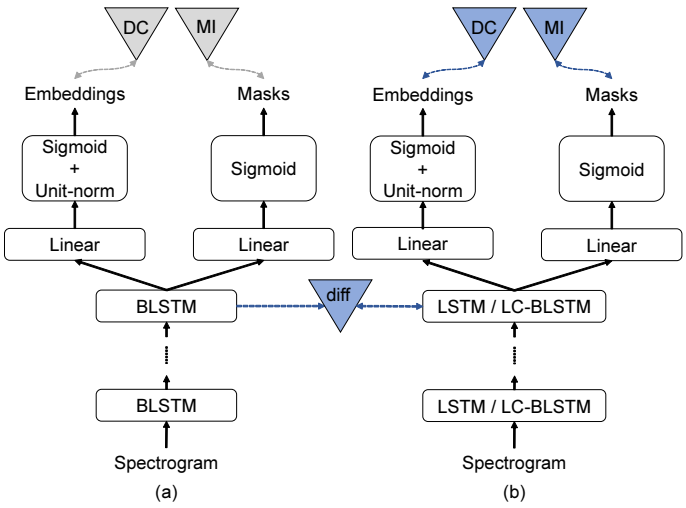

Fig. 1. (a) Teacher chimera network, (b) Student chimera network

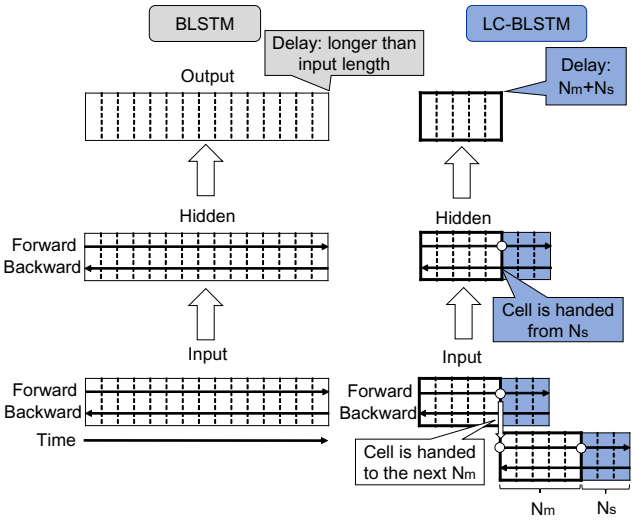

Fig. 2. Illustrations of BLSTM and LC-BLSTM

quence of BLSTM layers followed by a deep clustering head and a mask-inference (MI) head. The network is trained in a multi-task fashion, where the loss function is defined as a weighted sum of a deep clustering loss $\mathcal{L}_{\text{DC}}$ and an MI loss $\mathcal{L}_{\text{MI}}$ as follows:

$$\mathcal{L}_{\text{chi}} = \alpha\mathcal{L}_{\text{DC}} + (1-\alpha)\mathcal{L}_{\text{MI}}, \quad (1)$$

where α denotes a weight.

The deep clustering head computes a unit-length embedding vector $\mathbf{v}_i \in \mathbb{R}^{1\times D}$ for the i-th time-frequency (T-F) element $X_i = X_{t,f}$ corresponding to T-F indices t and f. Similarly, $y_i \in \mathbb{R}^{1\times C}$ is a one-hot label vector representing which source in a mixture dominates the i-th T-F unit. Vertically stacking the$N$ T-F bins, we form the embedding matrix $\mathbf{V} \in \mathbb{R}^{N\times D}$ and the label matrix $\mathbf{Y} \in \mathbb{R}^{N\times C}$. In this case, $\mathbf{YY}^\top$ is considered as a binary affinity matrix that represents the cluster assignments in a permutation-independent way. Following [13], we use the whitened k-means variant of the deep clustering objective, which gave best performance for speech separation:

$$\mathcal{L}_{\text{DC,W}} = \|V(V^\top V)^{-\frac{1}{2}} - Y(Y^\top Y)^{-1}Y^\top V(V^\top V)^{-\frac{1}{2}}\|_{\text{F}}^2, \quad (2)$$

where $||\cdot||_F^2$ denotes the Frobenius norm. We also follow [13] in discounting the influence of low-energy T-F bins by using magnitude ratio weights applied to the embedding and label matrices.

The MI head, which directly estimates masks to be applied to the mixture, uses a logistic sigmoid activation as in [13] and is trained as in [10, 13] using a truncated version of the phase-sensitive spectrum approximation (PSA) loss [18], referred to as truncated PSA (tPSA), and defined using the L_1 norm:

$$\mathcal{L}_{\text{MI,tPSA},L_1} = \min_{\pi\in\mathcal{P}} \sum_c \Big\| \hat{M}_c \circ |X| - \text{T}_0^{|X|}\left(|S_{\pi(c)}| \circ \cos(\theta_X - \theta_{\pi(c)})\right) \Big\|_1, \quad (3)$$

where $\mathcal{P}$ is the set of permutations on $\{1,\ldots,C\}$, $|X|$ and θ_X the mixture magnitude and phase, $\hat{M}_c$ the c-th estimated mask, $|S_c|$ and θ_c the magnitude and phase of the c-th reference source, and $\text{T}_a^b(x) = \min(\max(x,a),b)$ denotes truncation to the range $[a,b]$.

3. LATENCY-CONTROLLED BLSTM

BLSTMs are not practical for low-latency applications. Indeed, as illustrated on the left-hand side of Fig. 2, the forward LSTM operates from the first frame to the last frame of the input and the backward LSTM operates from the last frame to the first frame of the input. The output of each direction is concatenated and used as an input of the next layer. Therefore, one needs to wait until the BLSTM sees a whole utterance.

In order to cope with such issues, latency-controlled BLSTM (LC-BLSTM) networks were proposed for automatic speech recognition in [14, 19]. We here consider their application to speech separation, as an alternative to the low-latency approximations to BLSTM considered in [15] for speech enhancement. The LC-BLSTM architecture is illustrated in the right-hand side of Fig. 2. In LC-BLSTM, the input utterance is cut into non-overlapping blocks of fixed length N_m called the main block. Each main block has a sub block of fixed length N_s, which is appended to the right context. The forward LSTM now operates from the first frame of a main block to the last frame of its sub block. The memory cell of the last frame of the main block is handed over to the next main block. The backward LSTM operates from the last frame of the sub block to the first frame of the main block, and its memory cell is always initialized with 0. The outputs of sub blocks are propagated to the last LC-BLSTM layer but are not propagated to the linear layer in Fig. 1. Also, the gradients from the sub blocks are not back-propagated. If N_s is set to 0, LC-BLSTM is equivalent to a block BLSTM where each block operates independently.

4. TEACHER-STUDENT DEEP CLUSTERING

In order to improve the performance of low-latency models involving stacks of either LSTM or LC-BLSTM layers in place of the BLSTM stack, we consider applying teacher-student learning to deep clustering based speech separation. The procedure is illustrated in Fig. 1. A BLSTM-chimera network as presented in Sec. 2 is used as the teacher. As the student, we use either a stack of LSTM layers, which enables frame-wise operation, or a stack of LC-BLSTM layers, which enables block-wise operation. The teacher network is first trained using Eq. (1). The student network is then optimized under the following equation:

$$\mathcal{L}_{\text{stu}} = \alpha\mathcal{L}_{\text{DC}} + (1-\alpha)\mathcal{L}_{\text{MI}} + \beta\mathcal{L}_{\text{diff}}, \quad (4)$$

where $\mathcal{L}_{\text{diff}}$ denotes a distance between the weights of the final hidden layer of the teacher and student networks, and β a weight for

that distance. We consider here two variants for the teacher-student distance:

$$\mathcal{L}_{\text{diff},L_p} = ||\mathbf{h}_N^t - \mathbf{h}_N^s||_p^p, \quad p \in \{1, 2\}, \tag{5}$$

where $\mathbf{h}_N^t$ and $\mathbf{h}_N^s$ denote the output of the final layer of the teacher and the student recurrent network, respectively. In the above equations, $\mathbf{h}_N^t$ and $\mathbf{h}_N^s$ should have the same number of units. If the number of units is different, a projection layer to expand or contract that number can be used.

5. EXPERIMENTS

5.1. Experimental Conditions

We evaluate the proposed algorithms on the publicly available wsj0-2mix corpus [2], which is widely used in speaker-independent speech separation works. It contains 20,000, 5,000, and 3,000 two-speaker mixtures in its 30 h training, 10 h validation, and 5 h test sets, respectively. The speakers in the validation set are seen during training, while the speakers in the test set are completely unseen. The SNR of each mixture is randomly chosen between 0 dB and 10 dB. All data were downsampled to 8 kHz before processing to reduce computational and memory costs. For the spectrogram analysis, the window length is 32 ms and the hop size is 8 ms. The square root Hann window is employed as the analysis window and the synthesis window is designed accordingly to achieve perfect reconstruction after overlap-add. A 256-point DFT is performed to extract 129-dimensional log magnitude input features.

All the speech separation systems evaluated in this paper are based on the chimera network architecture. Although the model is trained in a multi-task fashion, we only use the MI loss during validation for model selection. At run time, we use the output from the MI branch as the masks for separation. We use four recurrent layers and apply a dropout of 0.3 to the output of each layer except the last one [13, 20]. The networks are trained from scratch on 400-frame segments, and each segment is trained independently. The run-time separation is always performed on the entire utterance. In the case of training LC-BLSTM networks, the block-wise operation is conducted within each segment. The Adam algorithm [21] is used for optimization of the networks. All systems are implemented using the Chainer deep learning toolkit [22]. We enhanced the magnitude only, and the noisy phase is used directly for time-domain re-synthesis.

5.2. Experimental Results

Table 1 compares the offline BLSTM network, with several LSTM network variations in terms of scale-invariant SDR [23]. The LSTM-chimera with 1,200 units performs best and is used in subsequent experiments. For Teacher-Student learning, a BLSTM with 1,200 units (600 units in each direction) is used as the teacher and a 1,200 unit LSTM is the student. From Table 1 we see a performance gap between BLSTM and LSTM of more than 2 dB, and no improvement from Teacher-Student learning.

The SDR results for LC-BLSTM-chimera are shown in Table 2 where (N_m, N_s) denote the number of frames in the main-block and sub-block, respectively. In the case of 150 frames latency (1.2 s), LC-BLSTM with $(N_m, N_s) = (100, 50)$ obtains the best performance, and we note the effectiveness of overlapping blocks since $(N_m, N_s) = (100, 50)$ performs better than $(150, 0)$. In the case of 75 frames latency (0.6 s), $(N_m, N_s) = (50, 25)$ performed better than $(75, 0)$ by 0.3 dB.

Table 1. SDR [dB] results for teacher-student LSTMs

Method	#units	TS-cost	β	SDR
BLSTM	1,200	Teacher	0	11.0
LSTM	1,000	-	0	8.7
LSTM	1,200	-	0	8.7
LSTM	1,400	-	0	8.6
TS-LSTM	1,200	L1	0.001	8.6
TS-LSTM	1,200	L1	0.01	8.7
TS-LSTM	1,200	L1	0.1	8.6
TS-LSTM	1,200	L2	0.001	8.7
TS-LSTM	1,200	L2	0.01	**8.7**
TS-LSTM	1,200	L2	0.1	8.4

Table 2. SDR [dB] results for LC-BLSTMs

Method	(N_m, N_s)	SDR
BLSTM	-	11.0
LC-BLSTM	(150, 0)	10.2
LC-BLSTM	(100,50)	**10.5**
LC-BLSTM	(50,100)	10.5
LC-BLSTM	(75, 0)	9.9
LC-BLSTM	(50,25)	**10.2**
LC-BLSTM	(25,50)	10.1

Table 3 compares several configurations of LC-BLSTM-chimera with teacher-student learning (TS-LC-LSTM) where a BLSTM with 1,200 units is used as the teacher and LC-BLSTM with $(N_m, N_s) = (100, 50)$ or $(50, 25)$ is used as the student. Unlike, the LSTM results of Table 1 we observe some SDR improvements from Teacher-Student learning in the LC-BLSTM cases. Specifically, we note maximum improvements of 0.2 dB for 150 frames (1.2 s) of latency, and 0.1 dB improvement for the 75 frame (0.6 s) case.

5.3. Failure Analysis

The recent study in [24] demonstrated that when separating speech from noise, the drop in performance between offline and zero-latency models was quite small, and performance equal to the offline model was achieved with only 0.2 s of latency. However, as shown in Section 5.2, even with 1.2 s of latency we still suffer a 0.3 dB drop in SDR. Similarly, TasNet [25], another recent speech separation algorithm, reported a performance drop of 3.8 dB between causal and non-causal models on the wsj0-2mix dataset. In this section we investigate why future context is so critical for speech separation.

Figs. 3 to 5 display scatter-plots for the BLSTM, LSTM, and TS-LC-BLSTM, respectively, where color indicates density. We note, that the shape of the scatter-plots for SDR improvement values above

Table 3. SDR [dB] results for teacher-student LC-BLSTMs

Network Architecture	TS-cost	β	(N_m, N_s) (100, 50)	(50, 25)
BLSTM	Teacher	-	11.0	
LC-BLSTM	-	0	10.5	10.2
TS-LC-BLSTM	L1	0.001	10.6	10.3
TS-LC-BLSTM	L1	0.01	10.7	10.3
TS-LC-BLSTM	L1	0.1	10.5	10.1
TS-LC-BLSTM	L2	0.001	10.7	10.3
TS-LC-BLSTM	L2	0.01	**10.7**	**10.3**
TS-LC-BLSTM	L2	0.1	10.6	10.3

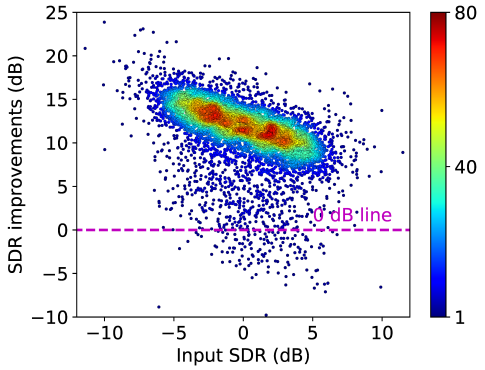

Fig. 3. SDR scatter-plot for BLSTM

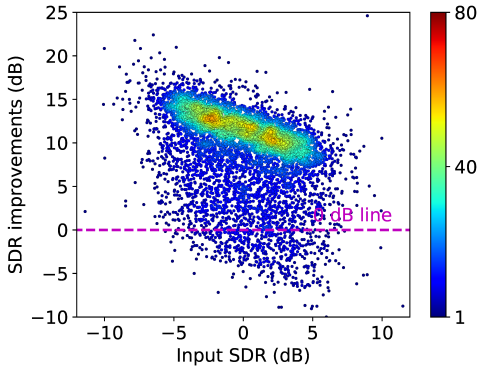

Fig. 4. SDR scatter-plot for LSTM

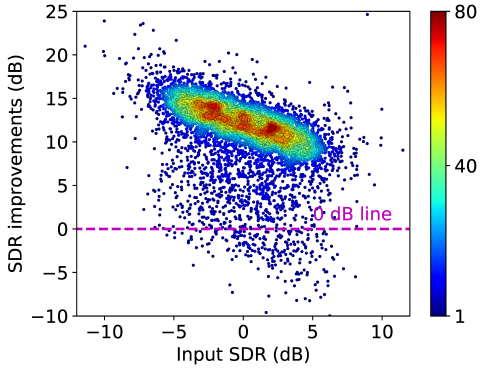

Fig. 5. SDR scatter-plot for TS-LC-BLSTM, with parameters $(N_m, N_s) = (100, 50)$, L2, $\beta = 0.01$

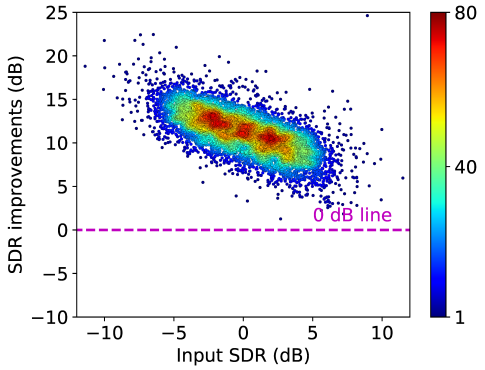

Fig. 6. SDR scatter-plot for LSTM with oracle frame permutation

Table 4. SDR [dB] results with and without oracle permutations

Architecture	(N_m, N_s)	Chimera mask	Oracle perm.
BLSTM	-	11.0	11.8
LC-BLSTM	(100,50)	10.5	11.7
TS-LC-BLSTM	(100,50)	10.7	11.9
LC-BLSTM	(50,25)	10.2	11.8
TS-LC-BLSTM	(50,25)	10.3	11.8
LSTM	-	8.7	11.2
TS-LSTM	-	8.7	11.3
IRM (Oracle)	-	-	12.7
IBM (Oracle)	-	-	13.5

10 dB are quite similar in all cases. However, the number of failure cases, defined here as SDR improvements below 0 dB are much larger for the LSTM case of Figure 4 than for the BLSTM and TS-LC-BLSTM cases. A major difference when separating overlapping speech signals as apposed to separating speech from noise is that during training we must attempt to solve the permutation problem, i.e., the correspondence between estimated and true sources. We typically solve this problem at the utterance-level, however as shown in [26] improvements can be observed when attempting to solve for the ideal permutation at the frame-level. To determine whether permutation errors are the cause of the failure cases shown in Figs. 3 to 5 we use the ground-truth separated signals to compute oracle permutations at the frame-level for the LSTM network output shown in Figure 6. We see that even when the LSTM has no future context, correctly solving the permutation problem eliminates all of the failure cases below 0 dB. Intuitively, what is happening in these failure cases is that the network is getting confused and switching between speakers in the middle of an utterance. While having no future context certainly leads to a greater number of failure cases (Figure 4), even the offline BLSTM makes some permutation errors (Figure 3).

Table 4 compares performance of the LSTM algorithms proposed in this paper, assuming access to oracle permutations. For reference the ideal ratio mask (IRM) and ideal binary mask (IBM) are also shown in Table 4. Notably, the SDR difference between the BLSTM and LSTM chimera output is 2.3 dB, but reduced to 0.6 dB with frame-level oracle permutations. We also note that the (50, 25) LC-BLSTM (0.6 s latency) is able to perform at a level consistent with the offline BLSTM, further confirming our hypothesis that correcting frame-level permutation errors is key for improving low-latency speech separation and should be a main focus of future works on the topic.

6. CONCLUSION

This paper proposed architectures for low-latency single-channel speech separation. By replacing the latency bottleneck BLSTM with the block-based LC-BLSTM, and using teacher-student learning, the SDR on the wsj0-2mix dataset was improved from 10.2 dB to 10.7 dB in the case of 1.2 s latency, and from 9.9 dB to 10.3 dB in the case of 0.6 s latency when compared with non-overlapping block BLSTM processing. Furthermore, we showed that the backward context in the BLSTM is important to solving the frame permutation problem. While our proposed techniques mitigate these permutation errors to a certain extent, future efforts will focus on directly tackling this issue. Other topics of future work include applying teacher-student learning to recent speech separation systems such as convolutional-TasNet [25] and evaluating low-latency approaches with the end-to-end phase reconstruction proposed in [27, 28].

7. REFERENCES

[1] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1990.

[2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[3] H. Erdogan, J. R. Hershey, S. Watanabe, and M. I. Mandel, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. ISCA Interspeech*, 2016.

[4] S. Makino, T. Lee, and H. Sawada, *Blind Speech Separation*. Springer, 2007.

[5] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer Speech Language*, vol. 24, no. 1, 2010.

[6] T. Virtanen, "Speech recognition using factorial hidden markov models for separation in the feature space," in *Proc. ISCA Interspeech*, 2006.

[7] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech & Language*, 2010.

[8] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. ISCA Interspeech*, 2016.

[9] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," in *arXiv preprint*, 2017.

[10] M. Kolbaek, 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, 2017.

[11] F. Bach and M. Jordan, "Learning spectral clustering, with application to speech separation," *The Journal of Machine Learning Research*, vol. 7, 2006.

[12] D. Wang and G. J. Brown, "Computational auditory scene analysis: Principles, algorithms, and applications," *Hoboken, NJ: Wiley-IEEE Press*, 2006.

[13] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[14] S. Xue and Z. Yan, "Improving latency-controlled BLSTM acoustic models for online speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[15] G. Wichern and A. Lukin, "Low-latency approximation of bidirectional recurrent networks for speech denoising," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS Deep Learning Workshop*, 2014.

[17] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[18] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[19] Y. Zhang, G. Chen, D. Yu, K. Yaco, S. Khudanpur, and J. Glass, "Highway long short-term memory RNNs for distant speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[20] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

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

[22] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: A next-generation open source framework for deep learning," in *Proc. Workshop on Machine Learning Systems in NIPS*, 2015.

[23] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or well done?" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[24] K. Wilson, M. Chinen, J. Thorpe, B. Patton, J. R. Hershey, R. A. Saurous, J. Skoglund, and R. F. Lyon, "Exploring tradeoffs in models for low-latency speech enhancement," in *Proc. IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[25] Y. Luo and N. Mesgarani, "TasNet: Surpassing ideal time-frequency masking for speech separation," in *arXiv preprint*, 2018.

[26] Y. Liu and D. Wang, "A CASA approach to deep learning based speaker-independent co-channel speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[27] G. Wichern and J. Le Roux, "Phase reconstruction with learned time-frequency representations for single-channel speech separation," in *Proc. IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[28] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff, and J. R. Hershey, "Phasebook and friends: Leveraging discrete representations for source separation," in *arXiv preprint*, 2018.