INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Mixture Encoder for Joint Speech Separation and Recognition

Simon Berger[1,2], *Peter Vieting*[1], *Christoph Boeddeker*[3], *Ralf Schlüter*[1,2], *Reinhold Haeb-Umbach*[3]

[1]Machine Learning and Human Language Technology,
Computer Science Department, RWTH Aachen University, Germany
[2]AppTek GmbH, Germany
[3]Paderborn University, Germany

{sberger,vieting,schlueter}@cs.rwth-aachen.de, {boeddeker,haeb}@nt.upb.de

Abstract

Multi-speaker automatic speech recognition (ASR) is crucial for many real-world applications, but it requires dedicated modeling techniques. Existing approaches can be divided into modular and end-to-end methods. Modular approaches separate speakers and recognize each of them with a single-speaker ASR system. End-to-end models process overlapped speech directly in a single, powerful neural network. This work proposes a middle-ground approach that leverages explicit speech separation similarly to the modular approach but also incorporates mixture speech information directly into the ASR module in order to mitigate the propagation of errors made by the speech separator. We also explore a way to exchange cross-speaker context information through a layer that combines information of the individual speakers. Our system is optimized through separate and joint training stages and achieves a relative improvement of 7% in word error rate over a purely modular setup on the SMS-WSJ task.

Index Terms: speech separation, speech recognition, meeting transcription

1. Introduction

Multi-speaker automatic speech recognition (ASR) describes the task of automatically transcribing speech in a scenario where multiple speakers speak at the same time and it is highly relevant for many applications. From a classical cocktail party scenario to meetings in everyday life, natural conversations typically contain overlapped speech. Transcribing this type of data requires dedicated techniques as standard single-speaker ASR methods do not perform well in this context [1].

Existing works for multi-speaker ASR can be divided into two main lines of research. First, there are modular approaches [1, 2] which deploy a speech separation frontend and an ASR backend. The frontend separates the mixture signal into multiple signals that each contain only a single speaker. These can subsequently be recognized by a standard single-speaker ASR system. On the other hand, there are end-to-end approaches [3, 4, 5] that process the overlapped speech directly using one large neural network (NN) without explicit separation.

Modular approaches have the advantage of a clearly interpretable division between the components. This allows to listen to the separated speakers and enables an easier analysis of a system. At the same time, the ASR module relies on an imperfect separation produced by the separator. Since hard decisions are already enforced on this intermediate level, errors are propagated and cannot easily be corrected by the acoustic model (AM).

In contrast, end-to-end models avoid this by having only one global decision. However, they require more powerful NNs and do not utilize the strengths of existing speaker separation methods. Furthermore, they have to rely on permutation invariant training (PIT) [6] or use heuristic assignment policies for the correct mapping of label posteriors to transcriptions in ASR training. This can lead to high computational costs and is stated to be less reliable than solving the permutation problem on signal level [2].

In this work, we propose an approach that can be viewed as a middle course between these different methods. We leverage explicit speech separation by incorporating a classical frontend that is pre-trained in a permutation invariant way. In addition, the proposed mixture encoder processes the mixture speech signal and encodes information that is fed into the AM in addition to the separated signals. This information can be used to mitigate the error propagation effect. Since we train on simulated (artificially overlapped) data, the label ambiguity problem can be avoided by solving the permutation using the output of the pre-trained speech separator and the reference single speaker signals during AM training. In addition to the separate training of separator and AM, we optimize the whole system in a final joint training stage. We show improvements in ASR accuracy using this proposed approach over a purely modular setup.

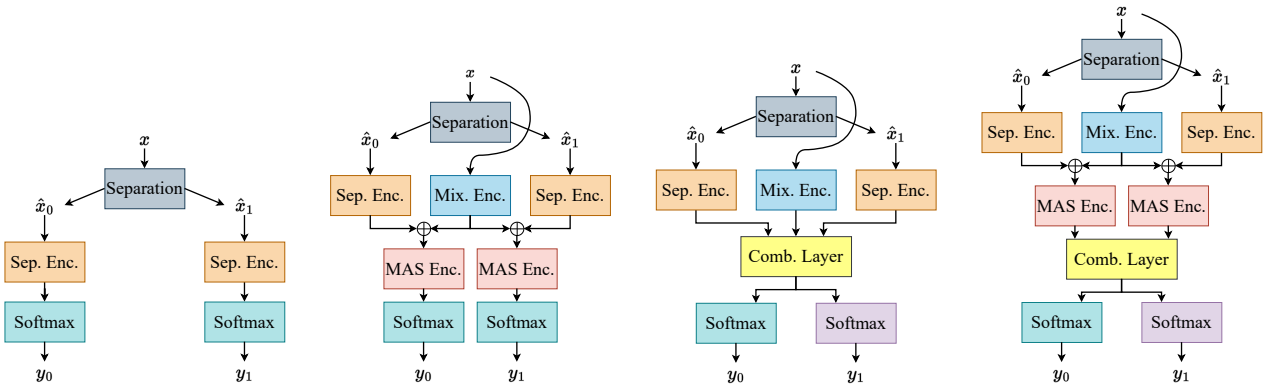
2. Related work

There has been significant research that focuses on a speech separation frontend which is used together with an ASR backend in a modular way for multi-speaker ASR [7, 8]. Several studies have shown that a joint fine-tuning of the speech separation- and ASR-module improves the performance of such modular systems [2, 9, 10, 11, 12].

End-to-end multi-speaker ASR systems that do not require explicit speech separation have also been investigated [3, 5, 13]. These approaches usually either employ a mixture encoder followed by two speaker-dependent encoders [3, 5, 13, 14] or they include an explicit mask-based speech separation module, the output of which is applied to the mixture encoder [15, 16, 17]. Such end-to-end approaches train the entire network using an ASR criterion and utilize assignment strategies such as PIT [18], serialized output training (SOT) [4], deterministic assignment training (DAT) [5] or heuristic error assignment training (HEAT) [16] to solve the permutation problem.

In the field of diarization on audio with overlapping speech, the target speaker-voice activity detection (TS-VAD) model [19] has shown an impressive performance in the CHiME-6 challenge [20]. For each speaker embedding vector at the input, it predicts the temporal activity of the respective speaker. The key for its success is a final combination layer that encodes all speaker information simultaneously while the preceding layers only see one speaker embedding vector each. In this work,

10.21437/Interspeech.2023-1815



(a) *Modular structure.* (b) *Structure with mixture encoder and MAS encoders.* (c) *Structure with combine layer instead of MAS encoders.* (d) *Structure with combine layer in addition to MAS encoders.*

Figure 1: *Full structure of multi-speaker model variants. Same coloring indicates shared parameters.*

we adapt this idea for multi-speaker ASR.

To the best of our knowledge, this is the first work that utilizes an explicit pre-trained speech separation network in a modular way while also leveraging the mixed audio directly through a mixture encoder inside the ASR module and allows the flow of cross-speaker context information through a combination layer.

3. Methods

3.1. Speech Separator

The input to our system is a mixture audio x which consists of multiple overlapping noisy and reverberant single speaker signals. The speech separation module produces estimates of the single speaker audios $\hat{x}_s$ whereby we only consider the two-speaker scenario, i.e. $s \in \{0, 1\}$.

To achieve this, we use a mask-based NN source separator in the short time Fourier transform (STFT) domain. This means, a NN is trained to predict one mask for each speaker and then the masks are multiplied with the observation to obtain estimates for each speaker's signal. After an inverse STFT, we use a signal-to-distortion ratio (SDR)-based loss [21] in PIT [6] to train the NN. To be more specific, a soft upper bound [22] for the SDR is used, to reduce the contribution of easy examples to the gradient-based optimization.

In recent years, new architectures have been proposed showing superior performance on reverberation-free data [23, Table 1]. However, for reverberated data a simple architecture like multiple bi-directional long-short term memory (BLSTM) layers [6] combined with a time domain loss [21] still exhibits similar performance, as it is shown in [24].

3.2. Acoustic Model

This section discusses the different composition possibilities for a network leveraging both the individual speaker audios $\hat{x}_0$ and $\hat{x}_1$ produced by a speech separation module, as well as incorporating the mixed audio x directly to produce output labels y_0 and y_1 for the two speakers.

Figure 1a shows a purely modular AM structure that recognizes the separated audio relying only on the separator output. This baseline approach is modeled by the following equations:

$$z_s^{Sep} = SepEnc(\hat{x}_s) \tag{1}$$

$$p(y_s \mid \hat{x}_s) = \text{softmax}(W \cdot z_s^{Sep} + b). \tag{2}$$

Here, $s \in \{0, 1\}$ represents the speaker index, *SepEnc* refers to the separation encoder and W and b are weight and bias of a single linear layer that projects to the dimension of the label space.

If we choose to incorporate x directly, various different possibilities of encoder compositions arise. Figure 1b shows a structure that includes the separation encoders that operate on $\hat{x}_0$ and $\hat{x}_1$, as well as a mixture encoder that receives x as its input. The outputs of the separation encoders and the mixture encoder are added element-wise before being fed into a mixture-aware speaker (MAS) encoder *MASEnc* to produce a common encoding that captures $\hat{x}_s$ and x. The equations for this approach are as follows:

$$z_s^{MAS} = MASEnc\left(SepEnc(\hat{x}_s) + MixEnc(x)\right) \tag{3}$$

$$p(y_s \mid \hat{x}_s, x) = \text{softmax}(W \cdot z_s^{MAS} + b). \tag{4}$$

To investigate the individual contributions, we also explore composition options where certain encoders are disabled, i.e. replaced by an identity function (see Table 1). To handle dimensional mismatches, we use concatenation instead of element-wise addition for $SepEnc(\hat{x}_s)$ and $MixEnc(x)$ if either of these encoders is replaced by identity.

We investigate one more configuration option, which is the addition of a combination layer (Figures 1c and 1d) that processes encodings of all three signals $\hat{x}_0$, $\hat{x}_1$ and x jointly. This option is inspired by [19] where such a combination layer is employed for the TS-VAD model in diarization of overlapping speech. The motivation for this combination layer is to enable an "information exchange" between the two halves of the model and constitute a form of soft context between the speakers. Further variants that exploit cross-speaker context are a topic for future research. In one variant we omit the MAS encoder and directly concatenate ($\|$) the outputs of the three preceding encoders (Figure 1c) in order to form the input of the combination layer:

$$z^{Comb} = Comb\left(z_0^{Sep} \| MixEnc(x) \| z_1^{Sep}\right) \tag{5}$$

$$p(y_s \mid \hat{x}_0, x, \hat{x}_1) = \text{softmax}(W_s \cdot z^{Comb} + b_s). \tag{6}$$

Here, $Comb$ represents the combination layer and W_s, b_s are the parameters of a speaker-specific linear layer. Two separate sets of parameters for the linear layers are required since they share the same input z^{Comb} which jointly encodes information for both speakers.

Our final variant (Figure 1d) is similar to the previous one but we keep the MAS encoders and concatenate their outputs to form the input of the combination layer:

$$z^{Comb} = Comb\left(z_0^{MAS} \| z_1^{MAS}\right). \quad (7)$$

3.3. Training Strategy

Our training process is divided into three phases. In the first phase, we pre-train the speech separation module independently using an SDR-loss (see [24]). During the second phase, we freeze the separator and incorporate our acoustic model, which is optimized using a frame-wise cross entropy (CE) loss. To obtain the targets needed for this loss function, we align the clean, un-mixed utterances with a Gaussian mixture model (GMM). To ensure that the alignments are compatible with the mixed utterances, we pad them with silence-labels analogous to the zero-padding for the waveforms before mixing. In the third phase, we unfreeze the separator and jointly optimize it together with the AM using the same frame-wise CE loss. Similar to previous works such as [2, 10, 11, 12], this joint optimization approach can be expected to result in improved performance.

4. Experimental Setup

4.1. Dataset

For the experiments, we used the SMS-WSJ [25] dataset. It uses the WSJ0 and WSJ1 utterances [26], down-sampled to $8\,\mathrm{kHz}$, and reverberates them with simulated room impulse responses (RIRs) [27] with a sound decay time T_{60} in the range of $0.2\,\mathrm{s}$ to $0.5\,\mathrm{s}$. To create a mixture, two reverberated WSJ utterances are added, where the shorter utterance is padded to the same length as the longer utterance. Further, Gaussian noise with an average signal to noise ratio (SNR) of $25\,\mathrm{dB}$ is added to the mixture. While this dataset contains multiple microphones, we only use the first microphone in this work.

4.2. Speech Separation Model

For the separation model, we use a re-implementation of the PIT-BLSTM from [24], with 3 BLSTM layers followed by 2 feed-forward layers. With the baseline ASR model from SMS-WSJ [25], this separator achieves a word error rate (WER) of $32.83\,\%$ on eval'92, which can be compared to the $35.70\,\%$ from [24]. For the pre-training, we use dynamic mixing [28]. This means, we sample new speaker combinations on demand and reverberate them with the precomputed RIRs.

4.3. Acoustic Model Training

To train a full ASR model, we first employ a training phase where the parameters of the speech separation module are frozen and only the AM is trained. The output of the separator is in the waveform domain and we extract 40-dimensional Gammatone features for both the overlapped and separated audios at the beginning of the separation and mixture encoders. We use a hybrid hidden Markov model (HMM) based acoustic model. The model utilizes BLSTM encoders, which have varying layer counts for each encoder variant. The model's output alphabet is comprised of 9001 CART labels. The BLSTM layer width is 400 for all encoder types and 800 for the combination layer. During training, we add an auxiliary softmax output directly on the separation encoders, which optimizes a CE loss with the same targets as the final output layers and a scale of 0.3. This is done to aid in better convergence. To solve the problem of assigning the correct target sequences to the two softmax outputs, we use the STFT magnitude of the reference reverberant signals and of the separated signals and choose the permutation that minimizes the mean squared error (MSE) between them. We train the model for 20 epochs using a Newbob learning rate (LR) schedule with an initial LR of $4e^{-4}$ and employ $L2$ weight decay of factor $1e^{-2}$, 10% dropout and 0.1 gradient noise as regularization techniques. During this phase, we also utilize SpecAugment [29], a data augmentation technique that randomly masks time and frequency blocks of the input features. The training is performed using the RETURNN toolkit [30].

4.4. Joint Training

After individually pre-training the speech separation module and acoustic model, we optimize them jointly in a third training phase. We unfreeze the parameters of the speech separator and continue training with the frame-wise CE loss for another 20 epochs, using a Newbob LR schedule and an initial LR of $3e^{-5}$. We found that SpecAugment was not helpful in this training phase and therefore disable it.

4.5. Recognition

Recognition is carried out using the RASR toolkit [31] and utilized a 3-gram language model (LM). To recognize the overlapped speech, one speaker was recognized at a time, i.e. two recognition runs are performed with only one of the softmax outputs being enabled in each run. After recognition, the hypothesized transcriptions are assigned and scored against the reference transcriptions in the permutation that minimized the WER.

5. Results

Table 1: *Performance of different AMs structures. Shown are the number of BLSTM layers of individual components, total number of model parameters (including speech separator) and WERs.*

Num encoder layers				# Par.	WER [%]	
Sep	Mix	MAS	Comb	[Mil.]	dev'93	eval'92
6	-	-	-	51	20.4	14.6
8				80	20.8	14.9
-	6	4		67	21.7	15.1
6	-			74	19.6	14.1
	4	-		64	19.9	14.1
			1	114	19.5	13.8
		2	-	80	19.6	14.0
		1	1	120	**19.1**	**13.6**

Table 1 shows the results of our experiments with different encoder sizes and compositions. Our modular baseline system with 6-layer separation encoders achieves a WER of 20.4% on dev'93 and 14.6% on eval'92. We observe that purely increasing the size of the modular system does not yield any improvement but in fact leads to a small degradation in performance,

with a WER of 20.8% on dev'93 and 14.9% on eval'92. This can be attributed to the increased difficulty in training the larger model and the higher risk of overfitting.

Removing the separation encoders clearly underperforms, with a WER of 21.7% on dev'93 and 15.1% on eval'92. One possible explanation for this is that the cleaner audio produced by the speech separator is not leveraged enough in this model structure. However, in all other cases, the inclusion of mixed audio through a mixture encoder and/or MAS encoder led to an improvement over the baseline. Without a combination layer, including both mixture and MAS encoder does not lead to a significant improvement over having only one of them active, so the increased model size in this variant is not justified by the performance and only including one of them is sufficient. Furthermore, adding a combination layer requires a large increase in the number of model parameters due to the concatenated inputs, but it results in the best performance, reducing the WER to 19.1% on dev'93 and 13.6% on eval'92. This represents a relative improvement of about 7% over the baseline.

Overall, our experiments demonstrate that incorporating mixed audio through a mixture or MAS encoder, along with a combination layer to enable the sharing of cross-speaker context information, can significantly improve the performance of our ASR system.

6. Discussion

As demonstrated by our results in Table 1, our novel approach to acoustic modeling in multi-speaker ASR results in clear improvements over the baseline method, which cannot be replicated by simply making the baseline network larger. Because of the parameter sharing and the way that mixture encoder and separation encoders are combined, the structure with mixture encoder and/or MAS encoder (Figure 1b) can easily be applied to an arbitrary amount of speakers. However, the extension of the combination layer to an arbitrary speaker count is not trivial and requires further investigation. Our performance increase using the combination layer suggests that more exploration of cross-speaker context may be a valuable topic for future research.

Our proposed encoder configuration could also be applied to other AM architectures such as the recurrent neural network transducer (RNN-T), which opens up further research opportunities.

Table 2: *Comparison of results reported in the literature for the clean WSJ corpus.*

Model	WER [%]	
	dev'93	eval'92
LF-MMI [32]	-	2.9
ours	5.4	3.0

Finally, our methods are compared to results reported in the literature. Table 2 shows a comparison on the clean WSJ corpus. Here, we train and evaluate an acoustic model with the same configuration as the AM used as our baseline with one encoder followed by a softmax output (see Figure 1a). This model shows similar performance to the state of the art presented in [32].

Table 3 shows results on the SMS-WSJ corpus. For these results, it is worth noting that the ASR performance of modular multi-speaker systems is highly dependent on the quality of the speech separation module, making direct comparisons of the AM with literature results difficult. Nevertheless, it is clear that the proposed model with mixture encoding outperforms other recent approaches [33, 24] and even multi-channel techniques like [11] despite using only a single channel. Only the TF-GridNet separator presented in [8] outperforms all other reported approaches by a large margin even using the baseline Kaldi ASR backend. This supports the effectiveness of their separator. In contrast, we use a rather standard speech separator and focus on improving acoustic modeling for ASR. Future work could include a combination of the separator from [8] with our ASR backend.

Table 3: *Comparison of results reported in the literature for SMS-WSJ corpus. #Ch denotes the number of microphone channels used.*

Model	#Ch	WER [%]	
		dev'93	eval'92
Joint Sep. & ASR [11]	6	-	15.4
2-spkr ASR [11]	1	-	35.0
U-Net TCN [33]		-	28.7
SepFormer [24]		-	26.5
TF-GridNet [8]		-	7.9
ours		19.1	13.6

7. Conclusions

In this research, we propose a middle-ground approach for multi-speaker ASR that combines the benefits of explicit speech separation with the direct incorporation of mixture speech information into a hybrid HMM-based ASR module. Our approach also includes a layer that combines encoder information of individual speakers to exchange cross-speaker context information. The system is optimized through separate and joint training stages.

Our experiments show that the proposed approach outperforms the conventional modular baseline method by around 7% relative improvement in WER. Our proposed method of incorporating mixture audio can easily be applied to a higher speaker count and the general encoder structure can also be used for other AM architectures, such as RNN-T, which opens up opportunities for future research.

The inclusion of a combination layer which merges the encoder data of individual speakers to share cross-speaker context information, drawing inspiration from work in speaker diarization, also clearly improves the ASR accuracy. This finding indicates that more research on cross-speaker context may prove valuable in the future. Overall, our approach is a promising solution for multi-speaker ASR that utilizes explicit speech separation, creates fertile opportunities for further extensions and investigations, and has the potential to enhance performance in real-world applications.

8. Acknowledgements

This research was partially supported by the Deutsche Forschungsgemeinschaft (DFG) under project no. 448568305.

9. References

[1] T. Menne, I. Sklyar, R. Schlüter, and H. Ney, "Analysis of deep clustering as preprocessing for automatic speech recognition of sparsely overlapping speech," in *Proc. Interspeech*. Graz, Austria: ISCA, 2019.

[2] T. von Neumann, K. Kinoshita, L. Drude, C. Boeddeker, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "End-to-end training of time domain audio separation and recognition," in *Proc. ICASSP*. Barcelona, Spain: IEEE, 2020, pp. 7004–7008.

[3] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," in *Proc. ACL*, Melbourne, Australia, 2018, pp. 2620–2630.

[4] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," *Proc. Interspeech*, pp. 2797–2801, 2020.

[5] I. Sklyar, A. Piunova, and Y. Liu, "Streaming multi-speaker ASR with RNN-T," in *Proc. ICASSP*. Toronto, Canada: IEEE, 2021, pp. 6903–6907.

[6] 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, no. 10, pp. 1901–1913, 2017.

[7] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*. San Francisco, USA: ISCA, 2016, pp. 545–549.

[8] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *arXiv preprint arXiv:2211.12433*, 2022.

[9] S. Settle, J. L. Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 4819–4823.

[10] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *Speech Communication*, vol. 104, pp. 1–11, 2018.

[11] W. Zhang, X. Chang, C. Boeddeker, T. Nakatani, S. Watanabe, and Y. Qian, "End-to-end dereverberation, beamforming, and speech recognition in a cocktail party," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 3173–3188, 2022.

[12] T. von Neumann, C. Boeddeker, L. Drude, K. Kinoshita, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "Multi-talker ASR for an unknown number of sources: Joint training of source counting, separation and ASR," in *Proc. Interspeech*. Shanghai, China: ISCA, 2020.

[13] I. Sklyar, A. Piunova, X. Zheng, and Y. Liu, "Multi-turn RNN-T for streaming recognition of multi-party speech," in *Proc. ICASSP*. Singapore: IEEE, 2022, pp. 8402–8406.

[14] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker ASR system without pretraining," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 6256–6260.

[15] A. Tripathi, H. Lu, and H. Sak, "End-to-end multi-talker overlapping speech recognition," in *Proc. ICASSP*. Barcelona, Spain: IEEE, 2020, pp. 6129–6133.

[16] L. Lu, N. Kanda, J. Li, and Y. Gong, "Streaming end-to-end multi-talker speech recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 803–807, 2021.

[17] ——, "Streaming multi-talker speech recognition with joint speaker identification," in *Proc. Interspeech*. Brno, Czechia: ISCA, 2021, pp. 1782–1786.

[18] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*. New Orleans, USA: IEEE, 2017, pp. 241–245.

[19] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny *et al.*, "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. Interspeech*. Shanghai, China: ISCA, 2020, pp. 274–278.

[20] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. S. Subramanian, J. Trmal, B. B. Yair, C. Boeddeker, Z. Ni, Y. Fujita, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020, pp. 1–7.

[21] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A dissecting approach," in *Proc. ICASSP*. Barcelona, Spain: IEEE, 2020, pp. 6359–6363.

[22] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3846–3857, 2020.

[23] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. ICASSP*. Toronto, Canada: IEEE, 2021, pp. 21–25.

[24] T. Cord-Landwehr, C. Boeddeker, T. Von Neumann, C. Zorilă, R. Doddipatla, and R. Haeb-Umbach, "Monaural source separation: From anechoic to reverberant environments," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2022, pp. 1–5.

[25] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv preprint arXiv:1910.13934*, 2019.

[26] D. B. Paul and J. Baker, "The design for the wall street journal-based CSR corpus," in *Speech and Natural Language: Proceedings of a Workshop*, Harriman, New York, 1992.

[27] 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.

[28] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2840–2849, 2021.

[29] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*. Graz, Austria: ISCA, 2019.

[30] A. Zeyer, T. Alkhouli, and H. Ney, "RETURNN as a generic flexible neural toolkit with application to translation and speech recognition," in *Proc. ACL*, Melbourne, Australia, 2018.

[31] D. Rybach, S. Hahn, P. Lehnen, D. Nolden, M. Sundermeyer, Z. Tüske, S. Wiesler, R. Schlüter, and H. Ney, "RASR - the RWTH aachen university open source speech recognition toolkit," in *IEEE ASRU*, 2011.

[32] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free MMI," in *Proc. Interspeech*. Hydarabad, India: ISCA, 2018.

[33] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2001–2014, 2021.