2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) | 979-8-3503-2372-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/WASPAA58266.2023.10248096

EXPLORING THE INTEGRATION OF SPEECH SEPARATION AND RECOGNITION WITH SELF-SUPERVISED LEARNING REPRESENTATION

Yoshiki Masuyama,[1⋆] *Xuankai Chang,*[2⋆] *Wangyou Zhang,*[3] *Samuele Cornell,*[4]
Zhong-Qiu Wang,[2] *Nobutaka Ono,*[1] *Yanmin Qian,*[3] *Shinji Watanabe,*[2]

[1]Tokyo Metropolitan University, Japan [2]Carnegie Mellon University, USA
[3]Shanghai Jiao Tong University, China [4]Università Politecnica delle Marche, Italy

ABSTRACT

Neural speech separation has made remarkable progress and its integration with automatic speech recognition (ASR) is an important direction towards realizing multi-speaker ASR. This work provides an insightful investigation of speech separation in reverberant and noisy-reverberant scenarios as an ASR front-end. In detail, we explore multi-channel separation methods, mask-based beamforming and complex spectral mapping, as well as the best features to use in the ASR back-end model. We employ the recent self-supervised learning representation (SSLR) as a feature and improve the recognition performance from the case with filterbank features. To further improve multi-speaker recognition performance, we present a carefully designed training strategy for integrating speech separation and recognition with SSLR. The proposed integration using TF-GridNet-based complex spectral mapping and WavLM-based SSLR achieves a 2.5% word error rate in reverberant WHAMR! test set, significantly outperforming an existing mask-based MVDR beamforming and filterbank integration (28.9%).

***Index Terms*—** speech separation, speech recognition, self-supervised learning, joint training, beamforming

1. INTRODUCTION

Speech separation and enhancement (SSE) is a crucial front-end for various applications such as speaker diarization, automatic speech recognition (ASR), and spoken language understanding [1–3]. The speech separation field has been revolutionized by the invention of deep clustering [4] and permutation invariant training (PIT) [5], which allow us to train deep neural networks (DNNs) for speech separation in a supervised manner. Previous speech separation methods based on time-frequency (T-F) masking [4–7] used a DNN to estimate the T-F mask for each speaker from the short-time Fourier transform (STFT) of the observed mixture. Meanwhile, time-domain methods [8–10] have demonstrated promising results by directly processing time-domain signals in an end-to-end (E2E) manner. Recently, fully complex STFT-domain methods have been proven to be extremely effective [11–13]. In particular, TF-GridNet [13] has achieved state-of-the-art (SotA) performance on several SSE benchmarks [4, 6, 14], including both monaural and multi-channel cases. Despite these impressive recent improvements in separation performance, it is still unclear how and when they can also lead to better ASR performance.

Most conventional SSE models are trained to minimize signal-level differences between the separated and target speech [8, 9].

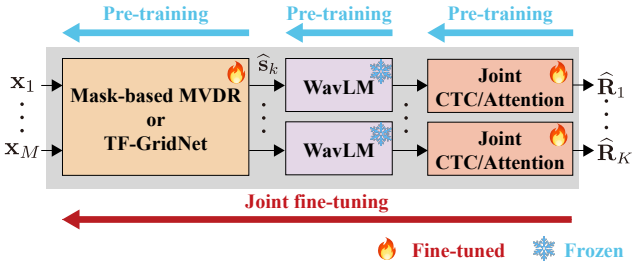

Figure 1: Overview of our E2E integration. We pre-train speech separation, SSLR, and ASR models separately, and fine-tune the speech separation and ASR models jointly while freezing WavLM.

This could lead to mismatches with respect to the subsequent ASR task. To address this issue, several attempts [15–22] have been made by integrating SSE and ASR models with joint optimization. For robust ASR, a neural beamformer and a joint connectionist temporal classification (CTC)/attention-based encoder-decoder were integrated and optimized with the ASR objectives [18]. This integration was extended to multi-speaker settings including MIMO-Speech [20]. It aims to directly improve the performance of multi-speaker ASR while preserving the modularity of the entire system, as opposed to a fully E2E black-box approach [23–25]. The intermediate separated speech achieves a good separation quality [20], although any signal-level criteria are not used for training.

Self-supervised learning (SSL) models such as Wav2Vec 2.0 [26], HuBERT [27], and WavLM [28] have shown considerable potential in a wide range of speech processing tasks [29, 30]. Recently, IRIS [31] demonstrated impressive results with an E2E model that integrates monaural speech enhancement, WavLM, and ASR models. MultiIRIS [32] expanded IRIS to perform multi-channel speech enhancement and demonstrated the effectiveness of the joint training under noisy and reverberant conditions.

Building upon MultiIRIS, this paper investigates MIMO-IRIS: an E2E integration of speech separation, SSLR extraction, and ASR for multi-channel multi-speaker overlapping scenarios. We explore the combination of SSLR-based ASR models [33] with TF-GridNet [13] as well as well-established beamforming techniques as illustrated in Fig. 1. We perform an extensive experimental validation on the spatialized WSJ0-2mix [6] and WHAMR! [14] datasets, assessing both separation and ASR performance. Interestingly, our experiments show that the correlation between speech separation and ASR performance is not precisely positive. The separation performance after fine-tuning degraded the separation performance while the word error rate (WER) decreases. This is especially true for TF-

⋆Equal contribution.

979-8-3503-2372-6/23/$31.00 ©2023 IEEE

GridNet-based complex spectral mapping, while mask-based beamforming [34, 35] results in less degradation. Despite this, our best MIMO-IRIS model after joint training achieves SotA ASR performance on the WHAMR! dataset with a WER of 2.5%, comparable to SotA results on clean single-speaker WSJ evaluation sets [33].

2. END-TO-END MULTI-CHANNEL MULTI-SPEAKER ASR WITH SPEECH SEPARATION AND SSLR

Given an L-sample, M-channel mixture signal $\mathbf{X} = (\mathbf{x}_m)_{m=1}^M \in \mathbb{R}^{M\times L}$ consisting of K speakers and noises $\mathbf{N} = (\mathbf{n}_m)_{m=1}^M$, we formulate the mixing process as follows:

$$\mathbf{x}_m = \sum_{k=1}^{K} \mathbf{s}_{k,m} + \mathbf{n}_m, \tag{1}$$

where $\mathbf{s}_{k,m} \in \mathbb{R}^L$ is the source image of speaker k at microphone m. The transcription sequence for speaker k is denoted as $\mathbf{R}_k$. This section describes each part of the proposed E2E system, depicted in Fig. 1, including speech separation, SSLR extraction, and ASR.

2.1. Speech Separation

The goal of speech separation is to estimate each speaker's signal $\hat{\mathbf{s}}_{k,r}$ at a reference microphone $r \in \{1, \dots, M\}$ from the mixture $\mathbf{X}$, which can be written as:

$$\{\widehat{\mathbf{s}}_1, \dots, \widehat{\mathbf{s}}_K\} = \texttt{SS}(\mathbf{X}). \tag{2}$$

Depending on the number of input microphones, the task can be divided into monaural and multi-channel speech separation.

2.1.1. Monaural speech separation

While our main focus is on multi-channel speech separation, we briefly explain monaural speech separation as TF-GridNet was originally proposed for the monaural case. In monaural speech separation, masking and mapping are two popular approaches [7]. Both can be performed in the complex T-F domain or in the time domain.

In masking-based approaches, a DNN is trained to estimate a mask for each speaker, and the mask is point-wisely applied to the encoded representation of the mixture $\mathbf{X}$:

$$\mathbf{Z} = \texttt{SSEnc}(\mathbf{X}), \tag{3}$$

$$\{\widehat{\mathbf{G}}_1, \dots, \widehat{\mathbf{G}}_K\} = \texttt{MaskEstimationNet}(\mathbf{Z}), \tag{4}$$

$$\widehat{\mathbf{S}}_k = \widehat{\mathbf{G}}_k \odot \mathbf{Z}, \tag{5}$$

$$\widehat{\mathbf{s}}_k = \texttt{SSDec}(\widehat{\mathbf{S}}_k), \tag{6}$$

where $\widehat{\mathbf{G}}_k$ denotes the estimated mask for speaker k, and $\odot$ denotes the Hadamard product. In T-F masking, `SSEnc` and `SSDec` are STFT and inverse STFT, respectively. Meanwhile, they are usually trainable one-dimensional convolutional layers and deconvolutional layers in the time-domain methods.

In mapping-based approaches, a DNN is trained to directly predict the encoded representation of each speaker. In detail, (4) and (5) are replaced by

$$\{\widehat{\mathbf{S}}_1, \dots, \widehat{\mathbf{S}}_K\} = \texttt{MappingNet}(\mathbf{Z}). \tag{7}$$

The mapping-based approaches in the T-F domain, or complex spectral mapping, have gained increasing attention due to the appearance of powerful DNN architecture called TF-GridNet [13]. TF-GridNet predicts the real and imaginary components of each speaker from those of the observed mixture. It has outperformed the best time-domain masking-based methods [10]. Furthermore, it has been successfully adapted to multi-channel speech separation.

2.1.2. Multi-channel speech separation

Multi-channel speech separation takes advantage of spatial information afforded by multiple microphones and has been used in robust ASR [2, 34, 35]. For the purpose of robust ASR, two popular approaches have been developed multi-channel separation: using DNN estimates to derive a conventional beamformer and using DNN to directly estimate each speaker's signal.

In the first approach, the minimum variance distortionless response (MVDR) beamformer has been widely used due to its distortionless property and generalization capability [34–37]. It incurs few processing artifacts by using the constrained time-invariant linear filters and is a preferable front-end of ASR backends [20, 21]. Neural mask-based beamforming estimates a T-F mask for each speaker $\widehat{\mathbf{G}}_k$ and computes a spatial covariance matrix as follows:

$$\widehat{\mathbf{V}}_k[f] = \frac{1}{\sum_t \widehat{G}_k[t,f]} \sum_{t=1}^{T} \widehat{G}_k[t,f]\mathbf{z}[t,f]\mathbf{z}[t,f]^{\mathsf{H}}, \tag{8}$$

where $\mathbf{z}[t,f] = [Z_1[t,f], \dots, Z_M[t,f]]^{\mathsf{T}}$, $Z_m[t,f]$ is the STFT coefficient of $\mathbf{x}_m$, $(\cdot)^{\mathsf{T}}$ denotes the transpose, and $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose. An MVDR beamformer $\widehat{\mathbf{w}}_k[f]$ is given by

$$\hat{\mathbf{w}}_k[f] = \frac{\widehat{\mathbf{V}}_{\setminus k}^{-1}[f]\widehat{\mathbf{V}}_k[f]}{\mathrm{trace}\Big(\widehat{\mathbf{V}}_{\setminus k}^{-1}[f]\widehat{\mathbf{V}}_k[f]\Big)}\mathbf{u}, \tag{9}$$

where $\widehat{\mathbf{V}}_{\setminus k}[f]$ denotes the sum of the spatial covariance matrices of the noise and all the speakers except for speaker k, and $\mathbf{u} \in \mathbb{R}^M$ is a one-hot vector indicating the reference microphone. The beamforming output is computed as:

$$\widehat{S}_k[t,f] = \widehat{\mathbf{w}}_k^{\mathsf{H}}[f]\mathbf{z}[t,f], \tag{10}$$

and converted to the time domain via inverse STFT as in (6).

In the second approach, a DNN directly estimates the encoded representation of each speaker by replacing the input of (7) to the concatenation of the encoded representation of microphone m. Compared to the output of linear beamformers, the output of the second approach tends to have fewer non-target signals but more distortion on the target speech. Although earlier studies suggested that linear beamformers would be preferable for robust ASR [20, 21], modern ASR back-ends and separation front-ends have become much more powerful nowadays. Hence, we expect that modern back-ends could handle speech distortion in separated signals, and modern front-ends can produce much less distortion in separated signals. We will compare their performance in our experiments, where TF-GridNet [13] and the joint CTC/attention-based encoder-decoder [38] are used for speech separation and ASR, respectively.

2.2. SSLR Extraction and E2E-ASR

We extract SSLR from each separated signal $\widehat{\mathbf{s}}_k$ in (2) and pass it to E2E-ASR in the same way as in previous studies [31, 32]:

$$\widehat{\mathbf{R}}_k = \texttt{ASR}(\texttt{SSLR}(\widehat{\mathbf{s}}_k;\, \theta^{\text{ssl}});\, \theta^{\text{asr}}), \tag{11}$$

where θ^{ssl} and θ^{asr} represent the parameters of the SSLR extractor $\texttt{SSLR}(\cdot)$ and ASR model $\texttt{ASR}(\cdot)$, respectively. Specifically, WavLM [28] is used to extract robust SSLR by applying the weighted sum of all transformer encoder embeddings. The weights are optimized with the following ASR model. E2E-ASR is based on the joint CTC/attention-based encoder-decoder framework [38].

2.3. MIMO-IRIS: Integration of Separation, SSLR and ASR

To recognize multi-speaker speech, one can directly send the outputs of the speech separation model to a pre-trained ASR model. This solution is, however, not optimal because ASR models are typically trained with single-speaker speech, while the separated speech usually contain residual interference. Following IRIS [31] and MultiIRIS [32], we integrate the speech separation model, SSLR extractor, and E2E-ASR model into a single model as shown in Fig. 1. The speech separation model can generate multiple streams, one for each speaker, and the ASR model is shared among all separated streams along with the SSLR extractor. During the training, PIT is applied to the CTC loss in the ASR model to determine the optimal permutation. The following attention-based decoder uses this permutation to select the corresponding reference transcript for each input stream in the teacher-forcing training. Our E2E model can be extended from (11) as:

$$\{\hat{\mathbf{R}}_1, \ldots, \hat{\mathbf{R}}_K\} = \texttt{ASR}(\texttt{SSLR}(\texttt{SS}(\mathbf{X}; \theta^{\text{ss}}); \theta^{\text{ssl}}); \theta^{\text{asr}}), \quad (12)$$

where θ^{ss} represents the parameters of the speech separation model, as discussed in Section 2.1. The loss function of the ASR task is the same as in MIMO-Speech [20]. We omit the details here.

The E2E model could be trained from scratch with multi-task learning, including speech separation and ASR objectives. Such training, however, requires intensive computation. In addition, previous studies on the integration of speech enhancement, SSLR extraction, and E2E ASR reported that the integrated model resulted in sub-optimal performance when trained from scratch [31, 32]. We thus propose a two-stage approach. First, the speech separation model is pre-trained on commonly-used speech separation datasets, e.g., spatialized WSJ0-2mix [4, 7] and WHAMR! [14]. Second, the ASR model is pre-trained on monaural clean speech datasets, e.g., the WSJ corpus. Finally, the entire integrated model is fine-tuned with the ASR objective, as shown in Fig. 1. Following previous studies, we freeze the WavLM, which is pre-trained on a large amount of external data. This strategy is efficient and requires only a few optimization epochs to achieve excellent performance in speech enhancement [31, 32].

3. EXPERIMENTS

We validate the effectiveness of our integration on two-speaker mixtures under anechoic/reverberant and clean/noisy conditions. Our experiments were conducted using the ESPnet-SE++ toolkit[1] [3].

3.1. Datasets

We evaluated our systems on the spatialized WSJ0-2mix [6] and WHAMR! [14] datasets, both of which support anechoic and reverberant two-speaker mixture simulations. The training, validation, and test sets of both datasets contain 20,000, 5,000, and 3,000 mixtures, respectively. Room impulse responses were simulated and convolved with dry source signals from WSJ0-2mix [4]. The signal-to-distortion ratio (SDR) [39] with respect to the input mixture is 0.07 dB in spatialized WSJ0-2mix. WHAMR! [14] is one of the most challenging datasets for speech separation, as it contains two-channel real-recorded environmental noise. For WHAMR!, the SDR with respect to the input mixture is -4.61 dB. To leverage the pre-trained WavLM [28], which was trained on 16 kHz, we used the 16 kHz version of both datasets in our experiments. We combined both anechoic and reverberant conditions of the training and validation sets to form the new training and validation sets, respectively.

3.2. Training Configurations

The ASR model ($\texttt{ASR}(\cdot)$ in (11) and (12)) consists of a Conformer-based encoder of 12 layers and a Transformer-based decoder of 6 layers by following a previous study [32]. The encoder and decoder have 2,048 hidden units and 4 attention heads. We reduced the dimensions of the speaker-wise SSLR from 1,024 to 80 by a fully-connected layer before feeding it to the ASR model. The ASR model and the learnable weight for the WavLM embeddings were pre-trained on the clean WSJ corpus. We used the Adam optimizer with a warm-up and the peak learning rate of 1.0×10^{-3}. During inference, we also used a Transformer-based character-level language model. On the clean single-speaker WSJ evaluation set, the ASR model achieved a WER of 1.3%.

As the speech separation model ($\texttt{SS}(\cdot)$ in (2) and (12)), our mask-based MVDR beamformer employed a 3-layer bidirectional long short-term memory of 512 units with a projection layer to estimate the T-F masks as in [20, 40]. STFT was implemented with the Hann window of 512 samples with a 128-sample shift. The mask estimation network was optimized with the convolutive transfer function invariant signal-to-distortion ratio (CI-SDR) loss [41] on beamforming outputs. Meanwhile, TF-GridNet consists of 6 blocks, where the TF-unit embedding dimension was 48. To reduce the computation, we increased the window shift size to 256 samples in STFT. TF-GridNet was optimized with a sum of the L_1 loss on the waveform and on the STFT magnitude[2] following [42], where the weight for the waveform loss was 0.99. Both mask estimation network and TF-GridNet were pre-trained with the Adam optimizer. Then, the joint fine-tuning of the speech separation and ASR models was performed using the stochastic gradient descent method with a learning rate of 1.0×10^{-3} and momentum of 0.9. We used the *max* condition of the spatialized WSJ0-2mix and WHAMR! datasets, mixtures of the non-trimmed utterances, in the joint fine-tuning.

3.3. Results on Clean Multi-channel Speech Separation

Table 1 presents the results on the spatialized WSJ0-2mix dataset. First, we show the results of the cascaded monaural TF-GridNet and ASR performance, an SDR of 19.4 dB and a WER of 4.8%. It outperformed an existing cascaded system with a time-domain masking-based method [43]. We then show the results in multi-channel cases, where the speech separation models were fine-tuned with the ASR objective. The TF-GridNet model consistently outperformed the MVDR beamformer not only in terms of SDRs but also in terms of WERs. This result demonstrates that the unconstrained complex spectral mapping is advantageous as an ASR

[1] Our source codes and configurations will be available through ESPnet: https://github.com/espnet/espnet.

[2] In our preliminary experiments, we also used the loss presented in [42] to train the mask-based beamformer. This resulted in worse WERs on the validation sets than using the CI-SDR loss [41].

Table 1: Separation and WER results on single-channel WSJ0-2mix and spatialized WSJ0-2mix.

	SDR [dB]	PESQ	STOI	WER (%)
Monaural				
Time-domain⋆ [43]	13.8	-	-	22.9
TF-GridNet⋆	19.40	3.41	0.976	4.8
Anechoic eight-channel				
MVDR (**proposed**)	12.83	3.86	0.987	2.1
- w/o fine-tuning	14.53	3.90	0.989	7.8
TF-GridNet (**proposed**)	15.28	3.14	0.983	**1.7**
- w/o fine-tuning	**26.43**	**4.09**	**0.995**	3.2
- w/o WavLM				6.3
Reverberant eight-channel				
MVDR (**proposed**)	4.56	2.76	0.859	3.6
- w/o fine-tuning	5.11	2.76	0.864	30.5
TF-GridNet (**proposed**)	12.32	3.17	0.956	**1.8**
- w/o fine-tuning	**18.81**	**3.89**	**0.983**	2.4
- w/o WavLM				28.2

⋆ The monaural models were not jointly fine-tuned.

front-end when using modern speech separation models. Furthermore, even the monaural TF-GridNet is more effective than the MVDR beamformer without joint fine-tuning.

To clarify the effectiveness of WavLM as a robust SSLR extractor, we evaluated the ASR model using filterbank features without joint fine-tuning. According to the bottom row of Table 1, its WER was degraded to 28.2% from 2.4% with WavLM in the reverberant condition. This result confirms the importance of the robust SSLR even with the powerful complex spectral mapping. In the weighted sum for extracting SSLR, the weight concentrated on the last layer, around 0.83, similar to previous studies [31, 32].

As an interesting finding, joint fine-tuning further reduced the WERs in both anechoic and reverberant conditions while degrading the separation performance. This degradation was less severe for the MVDR beamforming as its output is constrained to be distortion-less. Meanwhile, TF-GridNet-based unconstrained complex spectral mapping faced severe performance degradation, despite the better WER. In the anechoic case, the multi-channel TF-GridNet can achieve an SDR of 26.43 dB and a WER of 3.2% without fine-tuning. However, the separation performance dropped to 15.28 dB after joint fine-tuning. In detail, we observed buzzy artifacts in the intermediate separated signals[3].

3.4. Results on Noisy Multi-channel Speech Separation

In this section, we present our experimental results of the WHAMR! dataset, which are summarized in Table 2. In the top panel, we report the performance of monaural TF-GridNet on both noisy anechoic and reverberant conditions. As with the results on the spatialized WSJ0-2mix, the monaural TF-GridNet outperformed the mask-based MVDR beamformer integrated with weighted prediction error dereverberation [44]. The difference is even more significant due to the limitation of the number of microphones and noisy/reverberant characteristics of the data.

The best model overall is again the multi-channel TF-GridNet, which reached the best signal-level metrics before fine-tuning. After joint fine-tuning, the SDR decreased significantly, but the WER

[3]Examples of spectrograms and audio signals are available online: `https://yoshikimas.github.io/mimo-iris`.

Table 2: Separation and WER results on WHAMR!.

	Noisy/Anechoic		Noisy/Reverberant	
	SDR [dB]	WER (%)	SDR [dB]	WER (%)
Monaural				
TF-GridNet⋆	9.27	14.5	9.07	18.3
Two-channel				
MIMO-Speech [40]	-	-	-2.27	28.9
Time-domain [45]	-	-	-	20.9
MVDR (**proposed**)	-1.42	42.2	-1.30	44.4
TF-GridNet (**proposed**)	9.11	**2.3**	7.84	**2.5**
- ASR-only fine-tuning	**13.12**	4.4	**11.05**	6.5
- w/o fine-tuning		6.5		10.5

⋆ The monaural TF-GridNet was not jointly fine-tuned.

improved by over 400% relative factor in the noisy/reverberant condition. The performance is outstanding with WERs of 2.3% and 2.5% in anechoic and reverberant conditions, respectively, which are close to the performance achieved on the clean WSJ dataset. We also fine-tuned the ASR model while freezing the separation model, and its results are in the second bottom row of Table 2. While it outperformed the model without fine-tuning, its WER did not reach that of the joint fine-tuning model. This result confirms the advantage of the joint fine-tuning of both front-end and back-end. We emphasize that the ASR performance without fine-tuning still outperformed the previous MIMO-Speech [44] and the cascade combination of the time-domain speech separation and ASR models [45].

4. CONCLUSION

In this paper, we investigated the integration of speech separation, SSLR extraction, and ASR with well-established beamforming techniques as well as the latest SotA techniques including TF-GridNet. Our experiments were perfromed under anechoic/reverberant and clean/noisy conditions using the spatialized WSJ0-2mix and WHAMR! datasets. In detail, we explored how both separation performance and WER are affected by joint fine-tuning. Our experimental results show that the purely DNN-based speech separation method, TF-GridNet-based complex spectral mapping, can considerably outperform the mask-based MVDR beamforming preferred as an ASR front-end. Joint fine-tuning degraded the separation performance while significantly improving the WER, which is inconsistent with the tendency reported in a speech enhancement paper [32]. Our future work should focus on how this degradation can be prevented, e.g. by using continual learning strategies. Overall our best system, based on multi-channel TF-GridNet, WavLM, and E2E ASR, was able to reach performance on par with the one achieved on clean, single-speaker WSJ [33].

5. ACKNOWLEDGEMENTS

Y. Masuyama was partially supported by JSPS KAKENHI Grant Numbers JP21J21371 and JST CREST Grant Number JPMJCR19A3. X. Chang, Z.-Q. Wang, and W. Zhang used the Bridges2 system at PSC and Delta system at NCSA through allocation CIS210014 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program. S. Cornell was partially supported by Marche Region within the funded project "Miracle" POR MARCHE FESR 2014-2020.

6. REFERENCES

[1] D. Raj and *et al.*, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *Proc. SLT*, 2021, pp. 897–904.

[2] B. Li *et al.*, "Acoustic modeling for google home," *Proc. Interspeech*, pp. 399–403, 2017.

[3] Y.-J. Lu *et al.*, "ESPnet-SE++: Speech enhancement for robust speech recognition, translation, and understanding," in *Proc. Interspeech*, 2022, pp. 5458–5462.

[4] J. R. Hershey *et al.*, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[5] D. Yu *et al.*, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017, pp. 241–245.

[6] Z. Q. Wang *et al.*, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. ICASSP*, 2018, pp. 1–5.

[7] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

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

[9] Y. Luo *et al.*, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020, pp. 46–50.

[10] C. Subakan *et al.*, "Attention is all you need in speech separation," in *Proc. ICASSP*, 2021, pp. 21–25.

[11] L. Yang *et al.*, "TFPSNet: Time-frequency domain path scanning network for speech separation," in *Proc. ICASSP*, 2022, pp. 6842–6846.

[12] K. Tan *et al.*, "Neural spectrospatial filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 605–621, 2022.

[13] Z. Q. Wang *et al.*, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *arXiv:2211.12433*, 2022.

[14] M. Maciejewski *et al.*, "WHAMR!: Noisy and reverberant single-channel speech separation," in *Proc. ICASSP*, 2020, pp. 696–700.

[15] M. L. Seltzer *et al.*, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Trans. Speech, Audio process.*, vol. 12, no. 5, pp. 489–498, 2004.

[16] B. Li *et al.*, "Neural network adaptive beamforming for robust multichannel speech recognition," *Proc. Interspeech*, pp. 1976–1980, 2016.

[17] J. Heymann *et al.*, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *Proc. ICASSP*, 2017, pp. 5325–5329.

[18] T. Ochiai *et al.*, "Multichannel end-to-end speech recognition," in *Proc. ICML*, 2017, pp. 2632–2641.

[19] W. Minhua *et al.*, "Frequency domain multi-channel acoustic modeling for distant speech recognition," in *Proc. ICASSP*, 2019, pp. 6640–6644.

[20] X. Chang *et al.*, "MIMO-Speech: End-to-end multi-channel multi-speaker speech recognition," in *Proc. ASRU*, Dec. 2019, pp. 237–244.

[21] W. Zhang *et al.*, "Closing the gap between time-domain multi-channel speech enhancement on real and simulation conditions," in *Proc. WASPAA*, 2021, pp. 146–150.

[22] T. von Neumann *et al.*, "Multi-talker ASR for an unknown number of sources: Joint training of source counting, separation and ASR," *Proc. Interspeech*, pp. 3097–3101, 2020.

[23] H. Seki *et al.*, "An end-to-end language-tracking speech recognizer for mixed-language speech," in *Proc. ICASSP*, 2018, pp. 4919–4923.

[24] N. Kanda *et al.*, "Serialized output training for end-to-end overlapped speech recognition," in *Proc. Interspeech*, 2020, pp. 2797–2801.

[25] I. Sklyar *et al.*, "Streaming multi-speaker ASR with RNN-T," in *Proc. ICASSP*, 2021, pp. 6903–6907.

[26] A. Baevski *et al.*, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020.

[27] W. N. Hsu *et al.*, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[28] S. Chen *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[29] S. W. Yang and *et al.*, "SUPERB: Speech processing universal performance benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[30] H. S. Tsai and *et al.*, "SUPERB-SG: Enhanced speech processing universal performance benchmark for semantic and generative capabilities," in *Proc. ACL*, 2022, pp. 8479–8492.

[31] X. Chang *et al.*, "End-to-End Integration of Speech Recognition, Speech Enhancement, and Self-Supervised Learning Representation," in *Proc. Interspeech 2022*, 2022, pp. 3819–3823.

[32] Y. Masuyama *et al.*, "End-to-end integration of speech recognition, dereverberation, beamforming, and self-supervised learning representation," in *Proc. SLT*, 2023, pp. 260–265.

[33] X. Chang *et al.*, "An exploration of self-supervised pretrained representations for end-to-end speech recognition," in *Proc. ASRU*, 2021, pp. 228–235.

[34] J. Heymann *et al.*, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 196–200.

[35] H. Erdogan *et al.*, "Improved MVDR beamforming using single-channel mask prediction networks," *Proc. Interspeech*, pp. 1981–1985, 2016.

[36] S. Gannot *et al.*, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 692–730, 2017.

[37] T. Yoshioka *et al.*, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. ICASSP*, 2018, pp. 5739–5743.

[38] S. Kim *et al.*, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. ICASSP*, 2017, pp. 4835–4839.

[39] E. Vincent *et al.*, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[40] W. Zhang *et al.*, "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.

[41] C. Boeddeker and *et al.*, "Convolutive transfer function invariant SDR training criteria for multi-channel reverberant speech separation," in *Proc. ICASSP*, 2021, pp. 8428–8432.

[42] Y. J. Lu and *et al.*, "Towards low-distortion multi-channel speech enhancement: The ESPNET-SE submission to the L3DAS22 challenge," in *Proc. ICASSP*, 2022, pp. 9201–9205.

[43] T. von Neumann and *et al.*, "End-to-end training of time domain audio separation and recognition," in *Proc. ICASSP*, 2020, pp. 7004–7008.

[44] W. Zhang *et al.*, "End-to-end far-field speech recognition with unified dereverberation and beamforming," in *Proc. Interspeech*, 2020, pp. 324–328.

[45] J. Zhang *et al.*, "Time-domain speech extraction with spatial information and multi speaker conditioning mechanism," in *Proc. ICASSP*, 2021, pp. 6084–6088.