Multi-Channel Conversational Speaker Separation via Neural Diarization

Hassan Taherian and DeLiang Wang, *Fellow, IEEE*

***Abstract*—When dealing with overlapped speech, the performance of automatic speech recognition (ASR) systems substantially degrades as they are designed for single-talker speech. To enhance ASR performance in conversational or meeting environments, continuous speaker separation (CSS) is commonly employed. However, CSS requires a short separation window to avoid many speakers inside the window and sequential grouping of discontinuous speech segments. To address these limitations, we introduce a new multi-channel framework called "speaker separation via neural diarization" (SSND) for meeting environments. Our approach utilizes an end-to-end diarization system to identify the speech activity of each individual speaker. By leveraging estimated speaker boundaries, we generate a sequence of embeddings, which in turn facilitate the assignment of speakers to the outputs of a multi-talker separation model. SSND addresses the permutation ambiguity issue of talker-independent speaker separation during the diarization phase through location-based training, rather than during the separation process. This unique approach allows multiple non-overlapped speakers to be assigned to the same output stream, making it possible to efficiently process long segments—a task impossible with CSS. Additionally, SSND is naturally suitable for speaker-attributed ASR. We evaluate our proposed diarization and separation methods on the open LibriCSS dataset, advancing state-of-the-art diarization and ASR results by a large margin.**

***Index Terms*—Multi-channel speaker diarization, conversational speaker separation, location-based training, multi-speaker speech recognition.**

I. Introduction

TALKER-INDEPENDENT speaker separation systems are increasingly tailored to address more realistic scenarios [1]. One such environment is conversational or meeting settings. Conversational speech is characterized by its extended duration, an arbitrary number of participating speakers, and varying degrees of speech overlap. To address the challenges posed by conversational speech, the notion of continuous speaker separation (CSS) has been introduced [2]. CSS is designed to process long audio recordings and manage overlapped speech involving an arbitrary number of speakers. In the CSS approach, an audio recording is broken into shorter, partially overlapping segments, typically ranging from 2–3 seconds. Each segment should contain at most two speakers. By doing so, the CSS task is simplified to a two-talker concurrent speaker separation task for each segment. During the separation process, each segment is treated independently, resulting in two estimated speech signals. In cases where segments contain no overlapped speech, the processing reduces to speech enhancement, and the enhanced signal is mapped to one of the two streams, while the other generates a zero signal. As CSS produces two speech estimates for each segment, there is a requirement to group the estimates of the current segment with those in the previous segment. Sequential grouping, also referred to as "stitching", is essential for handling same-talker speech that spans multiple segments. This is commonly achieved by comparing the separation results in the overlapped regions between consecutive segments.

Since its inception, the CSS framework has been the subject of numerous studies aiming to enhance various components [3], [4], [5], [6], [7], [8]. In [3], a modulation factor based on the segment overlap ratio is introduced to dynamically adjust a separation loss. Chen et al. [4] proposed an early exit mechanism in Transformer layers for multi-channel speaker separation, speeding up inference by exiting when successive layer outputs are similar. Li et al. [5] proposed a dual-path separation model that leverages inter-segment information through a memory embedding pool. Wang et al. [6] employed a multi-stage strategy, combining a multi-input single-output (MISO) separation model with deep learning based beamforming followed by a post-filtering network. Extending upon this approach, the study in [8] integrates multi-input multi-output (MIMO) separation, incorporating a multi-resolution loss [7].

Despite the significant progress, the CSS framework faces several challenges. The first challenge is its limited segment size, stemming from the requirement that each segment must contain no more than two speakers [2]. A shorter segment length creates a bigger difficulty to group the separated utterances of the same talker over a period of time. When processing single-talker segments, a separation model occasionally fails to isolate the speaker in one stream with the other stream silent. Consequently, a speaker is erroneously split into two streams, adversely impacting downstream speech applications like automatic speech recognition (ASR) as they process each stream as originating from a distinct speaker [7]. To capture longer

Manuscript received 15 November 2023; revised 14 March 2024; accepted 15 April 2024. Date of publication 25 April 2024; date of current version 1 May 2024. This work was supported in part by the National Science Foundation under Grant ECCS-2125074, in part by Research Contract from Meta Reality Labs, in part by Ohio Supercomputer Center, and in part by Pittsburgh Supercomputer Center under Grant NSF ACI-1928147. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiao-Lei Zhang. *(Corresponding author: Hassan Taherian.)*

Hassan Taherian is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: taherian.1@osu.edu).

DeLiang Wang is with the Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2024.3393726

speech utterances, two recent studies introduced new training criteria based on permutation-invariant training (PIT) [9], [10], [11]. In [10], the authors proposed "Graph-PIT", an extension of PIT that can separate a varying number of speakers from a two-talker separation model. This model employs graph coloring to optimally assign multiple speakers to two streams. Zhang et al. [11] introduced Group-PIT, which organizes a long reference signal into utterance groups, employing the PIT criterion for these groups instead of individual utterances.

The second CSS challenge lies in the stitching process. Typically, the overlap between neighboring segments is set to 50% of the segment length to provide an adequate context for aligning adjacent segments. However, this approach introduces computational inefficiency since each segment is processed twice. Moreover, spectral distance based stitching is prone to errors, resulting in the misalignment of adjacent segments.

The third challenge arises when dealing with long recordings, e.g. a one-hour meeting involving many participants. While stitching can group a continuous single-talker signal in consecutive frames, it does not address the challenge of grouping discontinuous signals of the same talker, e.g., how to group the utterance of a talker in the first 5 minutes of a one-hour meeting and the utterance of the same talker in the last five minutes of the meeting. The grouping of discontinuous utterances of the same talker is crucial for subsequent tasks such as speaker-attributed ASR.

In this paper, we present a new framework, termed "speaker separation via neural diarization" (SSND), for multi-channel conversational speaker separation. Unlike the traditional approach to speaker diarization comprising the stages of speech activity detection, speaker embedding extraction, and clustering, SSND demarcates the speech activities of individual speakers by employing a deep neural network (DNN) for end-to-end diarization. Leveraging the estimated utterance boundaries from neural diarization, we generate a sequence of speaker embeddings. These embeddings, in turn, facilitate the assignment of speakers to two output streams of the separation model. The SSND approach tackles the permutation ambiguity issue of talker-independent separation during the diarization phase, rather than during separation. This distinction permits non-overlapped speakers to be assigned to the same output stream, enabling the processing long recordings missing from standard CSS. Furthermore, there is no stitching in SSND, and hence duplicate processing of segments is eliminated, resulting in computational efficiency. Another advantage of SSND lies in the inherent integration of speaker separation and diarization, enabling sequential grouping of the discontinuous utterances of the same talker.

For embedding extraction, we utilize EEND with an encoder-decoder-based attractor calculation module (EEND-EDA) [12], but extend EEND-EDA to multi-channel scenarios with a different training criterion that can handle a larger number of speakers. Specifically, we propose to use location-based training (LBT) [13] to resolve permutation ambiguity in speaker diarization. We show that that LBT significantly outperforms the PIT criterion for diarization of many speakers. Our SSND framework achieves state-of-the-art diarization and ASR results, surpassing all existing CSS based methods on the open LibriCSS dataset [2].

The rest of the paper is organized as follows. In Section II, we describe our proposed multi-channel diarization model, and the SSND framework. We present the experimental setup in Section III, and evaluation and comparison results in Section IV. Concluding remarks are provided in Section V.

II. Algorithm Description

A. Multi-Channel Diarization

Speaker diarization traditionally employs a clustering-based approach [14], [15]. This approach revolves around grouping speaker embeddings, such as x-vectors [16], into clusters. Such a method usually involves three distinct stages. First, a speech activity detection model is utilized to identify speech intervals. Then, speaker embeddings are extracted, and finally a clustering technique such as spectral clustering (SC) is applied. However, this method has its limitations. Its stages are independent and cannot be trained jointly to minimize diarization errors. Plus, clustering methods struggle with speaker overlaps, as they assume a single speaker within a segment.

End-to-end neural diarization (EEND) has been introduced to streamline the diarization process by using a single neural network model [17]. Unlike clustering-based diarization, the EEND method can handle speaker overlaps. Furthermore, by incorporating an encoder-decoder-based attractor calculation (EDA) [12], EEND can handle an unknown number of speakers. For a sequence of frame-wise x_t, where $t = [1, \ldots, T]$, EEND encodes these into a series of embeddings, $e_t \in \mathbb{R}^E$, via a DNN-based encoder. From these frame-wise embeddings, the EDA module derives a number of attractors a_c for $c = [1, \ldots, C]$ speakers. Subsequently, speech activity probabilities, $p_t = [p_1, p_2, \ldots, p_C]$, are derived by taking the dot product of the frame-wise embeddings and the speaker-wise attractors, followed by applying a sigmoidal function. Finally, the speech activities of different speakers are estimated by using a decision threshold τ.

EEND is formulated for monaural recordings. In this study, we extend EEND-EDA to multi-channel recordings by integrating spatial features. Fig. 1 illustrates our proposed multi-channel EEND-EDA (MC-EEND) model. Our model utilizes both spectral and spatial features [18]. For spectral features, we utilize log-Mel filterbanks, while for spatial features, we employ the inter-channel phase difference (IPD) between the reference microphone and the other microphones. For each pair of microphones, the cosine and sine of the IPD are concatenated. These IPD features are then processed through a series of convolutional blocks, each of which is composed of a convolutional layer, a PReLU activation function, and a group normalization layer. Subsequently, the processed IPD features are concatenated with the log-Mel features. The combined features are fed to the EEND encoder, which comprises several Transformer layers without positional encodings. To derive speaker attractors using EDA, we shuffle the order of the embeddings.

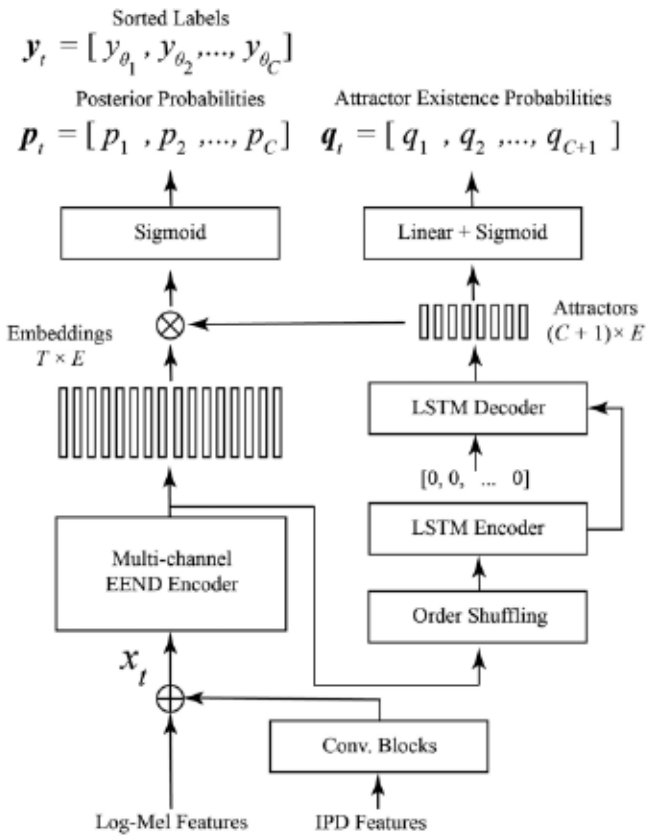

Fig. 1. Schematic diagram of the proposed MC-EEND with location-based training.

For EEND to be effective in real-world applications, it must be trained in a talker-independent manner to accommodate untrained speakers. Much like talker-independent speaker separation, the difficulty lies in aligning diarization output layers with the respective speaker labels. Without proper output-speaker assignment, EEND training would not converge due to conflicting gradients. This is known as the permutation ambiguity problem [9], [19]. Previous studies employ PIT to tackle this problem by analyzing the losses across all possible output-speaker pairings [12], [17]. However, unlike speaker separation where the number of concurrent speakers can be reasonably limited to two or three, diarization may involve many speakers. Using PIT in such cases becomes problematic as it has factorial or polynomial training complexity [20], [21], making it inefficient for a large number of speakers.

In this study, we introduce LBT [13] for diarization. We employ the spatial locations of speakers to determine output-speaker assignments. Given a polar coordinate system with the microphone array's center as the origin, we define the diarization loss function as:

$$\mathcal{L}_{\text{EEND}} = \frac{1}{TC} \sum_{t=1}^{T} H(\boldsymbol{y}_t, \boldsymbol{p}_t) \tag{1}$$

where $\boldsymbol{y}_t = [y_{\theta_1}, y_{\theta_2}, \ldots, y_{\theta_C}]$ represents the binary speaker label vector for frame t. This vector is arranged in ascending order according to speaker azimuths relative to the microphone array [13]. The binary cross entropy function, $H(.,.)$, is defined as:

$$H(\boldsymbol{y}_t, \boldsymbol{p}_t) = -\sum_{c=1}^{C} y_{c,t} \log p_{c,t} + (1 - y_{c,t}) \log(1 - p_{c,t}). \tag{2}$$

LBT, or azimuth-based training specifically, has a linear computational complexity [13], making it efficient to train end-to-end diarization systems with many speakers. In the EDA module, attractors are trained using attractor existence probabilities, denoted as q_t (see Fig. 1). The attractor probabilities are calculated by using a fully connected layer, followed by sigmoidal activation. The EDA module's loss function is described as:

$$\mathcal{L}_{\text{EDA}} = \frac{1}{C+1} H(\boldsymbol{l}, \boldsymbol{q}) \tag{3}$$

where $\boldsymbol{l} \in \mathbb{R}^{C+1}$ is a binary vector with its first C elements set to 1 and its final element set to 0. The total loss for MC-EEND is expressed as:

$$\mathcal{L}_{\text{D}} = \mathcal{L}_{\text{EEND}} + \mathcal{L}_{\text{EDA}}. \tag{4}$$

B. Speaker Separation Via Neural Diarization (SSND)

Our SSND framework uses the MC-EEND diarization model. However, the SSND approach can be coupled with any diarization model. Within this framework, we assume that the number of concurrent speakers at each frame does not exceed two, as more than two speakers rarely talk simultaneously in real-world conversations [2], [22]. To represent each speaker, we extract speaker embeddings from the MC-EEND encoder. Specifically, for each speaker, we average the embedding vectors over all the frames when the speaker is active with all others silent, which may be discontinuous:

$$\hat{e}_c = \frac{1}{T_c} \sum_t e_t \quad \text{where} \quad \begin{cases} \hat{y}_{c,t} = 1, \\ \hat{y}_{c',t} = 0, \text{ for } c' \neq c. \end{cases} \tag{5}$$

Here, $\hat{y}_c$ represents the estimated speech activities derived from MC-EEND and T_c is the number of frames where only speaker c is active. Fig. 2(a) illustrates the embedding extraction process.

Based on extracted speaker embeddings, we create two sequences in the following manner. Using the order of diarized speech activities, each speaker-active interval is assigned to one of two embedding sequences. For such an interval, the extracted embedding of the corresponding speaker is assigned to every frame of the interval as illustrated on the left side of Fig. 2(b). We initially assign the first interval to the first embedding sequence. At the onset of a current interval, if both sequences are silent, we check whether the underlying speaker of the current interval is the same as that of the previously ended interval. If yes, the current interval is assigned to the same sequence as the previously ended sequence; if not, the interval is assigned to the other sequence. If only one sequence is silent, the current interval is assigned to this sequence. This procedure is illustrated in Fig. 2(b). We note that this assignment of speaker-active intervals to two embedding sequences guarantees no overlap between the intervals assigned to each sequence regardless of the number of speakers, as long as no more than 2 talkers speak simultaneously.

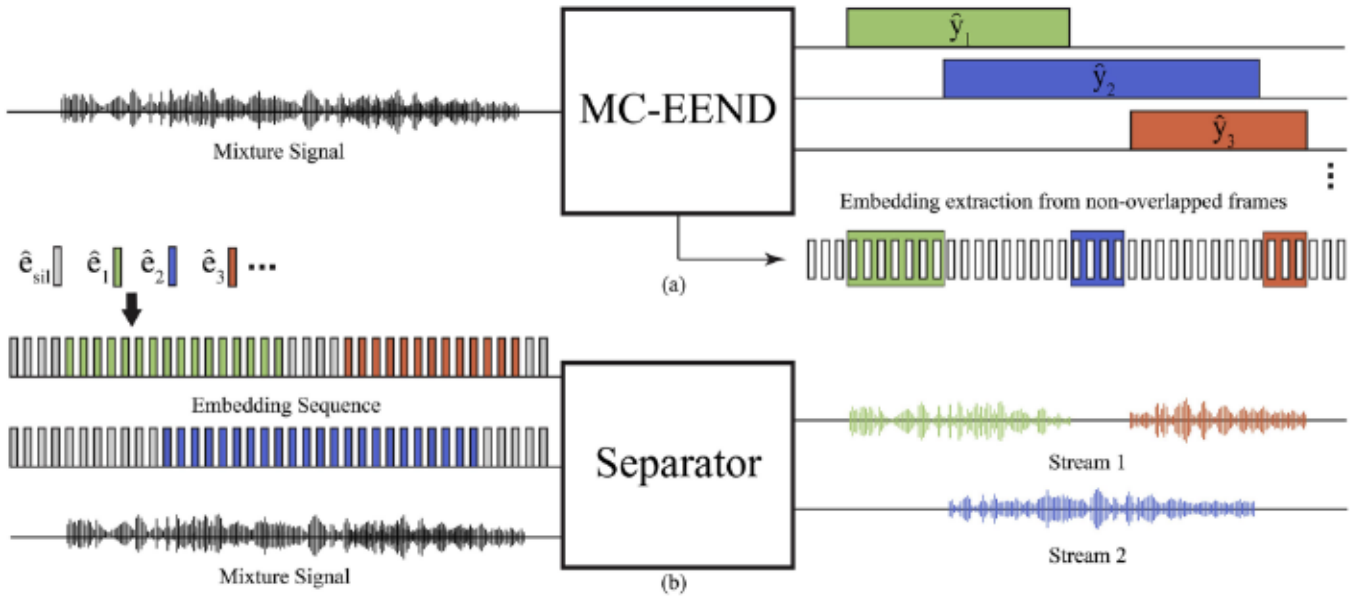

Fig. 2. Proposed SSND framework: (a) embedding extraction with MC-EEND based on estimated non-overlapped frames, and (b) constructing embedding sequences and feeding to speaker separation along with mixture signal.

For silent frames, we use a zero embedding vector $\hat{e}_{sil}$. The resulting two embedding sequences are then fed to a separation network along with the multi-channel mixture signal, as shown in Fig. 2(b).

We employ two architectures for the separator network (see Fig. 2(b)), both of which operate in the short-time Fourier transform (STFT) domain. The first architecture, TF-GridNet [23], processes time-frequency units in a grid-like manner. TF-GridNet consists of several blocks, each with three main components. The first two components utilize bi-directional long short-term memory (BLSTM) to process full-band spectral features within each frame and the temporal information within each frequency. The last component incorporates a self-attention module, to process full-band information across frames to capture long-range contexts. The second architecture, SpatialNet [24], processes spectral and temporal information similarly to TF-GridNet but employs only narrow-band and cross-band modules. The narrow-band module employs multi-head self-attention, while the cross-band module uses convolutional layers.

Similar to the embedding sequences, we create two output streams for separated speech signals as illustrated in Fig. 2(b). Clean speech signals are assigned to a stream based on the corresponding speaker embedding assignment. For silent frames, we use a zero signal. We train the separator model with an ℓ_1 norm loss on the real and imaginary components of the estimated signal and target signal at the reference microphone [25], with an additional magnitude loss term [26]:

$$\mathcal{L}_{\text{sep}}\left(\hat{S}, S\right) = \frac{1}{2}\sum_{n=1}^{2}\mathcal{L}\left(\hat{S}_n, S_n\right) \tag{6}$$

$$\mathcal{L}\left(\hat{S}, S\right) = \left\|\hat{S}^{(r)} - S^{(r)}\right\|_1 + \left\|\hat{S}^{(i)} - S^{(i)}\right\|_1 + \left\||\hat{S}| - |S|\right\|_1, \tag{7}$$

where $\hat{S}$ and S denote the estimated and clean speech signals in the STFT domain. Superscripts r and i denote real and imaginary parts, $|.|$ computes magnitude, and $\|.\|_1$ indicates ℓ_1 norm.

In the proposed SSND framework, the separator network resolves the permutation ambiguity problem by using the embedding sequences for output-speaker assignment. Unlike CSS based on PIT where the number of outputs is equal to the number of speakers, this approach allows multiple non-overlapped speakers to share the same output stream. As a result, SSND can handle long recordings containing numerous speakers. It is worth noting that our approach differs from research on personalized speech enhancement [27] and speaker separation [28] that use embeddings for speaker extraction. In these studies, speaker embedding does not change over a stream. In contrast, speaker embedding in our approach may change corresponding to a speaker change (e.g., over the first stream of Fig. 2(b)).

III. Experimental Setup

A. Datasets

We assess the proposed approach for both diarization and multi-speaker ASR tasks using the LibriCSS corpus [2]. This corpus is structured into ten one-hour sessions. Each session is further subdivided into six ten-minute mini-sessions, which are characterized by varying speech overlap levels. These levels include 0S (no overlap with short pauses ranging from 0.1–0.5 seconds between utterances), 0 L (no overlap with long pauses lasting between 2.9–3.0 seconds), and then speaker overlap at 10%, 20%, 30%, and 40%. Every recording in the dataset was sampled at 16 kHz. The recordings for LibriCSS were drawn from the LibriSpeech development set. To capture real room acoustics, utterances were replayed through loudspeakers, and recorded by a circular microphone array with six microphones positioned in a circle with a radius of 4.25 cm and an additional

central microphone. We designate the center microphone as the reference microphone. For speaker diarization and speech recognition, we adopt session-wise evaluation by utilizing every mini-session for assessment, and all 10 sessions are used for evaluation.

To generate training data for diarization, we simulate meeting-style conversations based on a recipe in LibriCSS[1]. Each session includes eight speakers, with two utterances for each speaker. These clean speech signals originate from the Librispeech training set [29]. The speaker overlap ratio is chosen randomly from 0 to 45%. Furthermore, silences varying from 0.5 to 3.0 seconds are inserted between neighboring utterances with a 0.26 probability. On average, the sessions generated for diarization training have a duration of 3 minutes. We employ simulated room impulse responses (RIRs) with reverberation time (T60) chosen in the range of 0.2 and 0.6 seconds. Speaker azimuths are randomly generated and maintain a minimum separation of 5 degrees. We adjust the sound levels between the speakers in the range of -3.5 to 3.5 dB. We also add simulated diffuse noise with the signal-to-noise ratio (SNR) in the range of 10 to 30 dB, where all reverberant speech utterances are considered as the signal in the SNR calculation.

To generate training data for speaker separation, we use the same recipe but opt for shorter sessions. Each session in our data generation has 1–2 utterances per speaker, chosen at random with each utterance under 10 seconds. Overlap ratios are randomly selected between 0.4 to 0.5. Silence intervals, randomly varying from 0.5 to 1.0 seconds, are inserted with a probability of 0.05. To generate target speech signals, clean speech is convolved with the direct-path (anechoic) signal at the reference microphone. To train the separation model, each mixture is divided into 5-second segments. Of these, 11% contain a single speaker, and the remaining 89% feature two. Prior to mixing the training utterances, we use our diarization model to extract an embedding vector for each speaker from the reverberant utterances of the speaker. This is done to ensure that embedding vectors accurately represent the corresponding speakers, as using mixture utterances may not produce adequate single-talker frames to extract quality speaker embeddings for training purposes.

B. Diarization Training and Evaluation Metrics

For the EEND encoder, we employ eight stacked Transformer blocks, each equipped with 16 attention heads without positional encodings. These encoder blocks generate $E = 256$ dimensional frame-wise embeddings. The window size for STFT is set to 25 ms with a window shift of 10 ms, and the square root of the Hann window is used as the analysis window. A 512-point discrete Fourier transform is employed, resulting in the extraction of 257-dimensional complex spectra. For spectral features, we extract a 23-dimensional log-Mel filterbank, which is subsequently concatenated with those of the seven preceding and seven succeeding frames. As for spatial features, we process IPD features through seven convolutional blocks, having channel configurations of (8, 8, 16, 16, 32, 32, 32). These blocks use kernel sizes of (3, 5) and strides of (1, 2) along the time and frequency axes, respectively, except for the initial block which employs a kernel size of (15, 1). These log-Mel features are subsequently concatenated with the spatial features. All input features are normalized to have zero mean and unit variance.

We use the Adam optimizer for training the diarization model with the learning rate of 0.001, which is coupled with the Noam scheduler [30] including 125 K warm-up steps. We set the number of speakers to 8 for both training and testing. During diarization training, we use segments of 4-minute duration to ensure that all speakers are active within each segment. Sessions shorter than 4 minutes are zero padded at the end. To process long audio segments, we reduce the number of frames in each batch via subsampling with a factor of 5, resulting in a 50-ms frame shift for training. For speech activity decisions, we use a threshold of $\tau = 0.5$. To avoid generating exceedingly short segments, a 31-frame median filter is applied.

Diarization performance is evaluated using the NIST diarization error rate (DER) [31], which calculates the combined durations of missed speech, false alarm, and speaker confusion errors, divided by the total duration of speech. We use a 0-second collar tolerance at utterance boundaries.

C. Separation Training and Evaluation Metrics

Our separation models employ the following DNN architectures:

- A TF-GridNet featuring 4 blocks, a 192-unit BLSTM, a kernel size of 4, a stride of 1, and $D = 48$ channels.
- A large TF-GridNet with 6 blocks, a 256-unit BLSTM, and $D = 64$ channels.
- A SpatialNet with 12 blocks, $D = 192$ channels, narrow-band hidden dimensions of 384, and cross-band hidden dimensions of 16.

The STFT parameters for the TF-GridNet models are set to a window length of 32 ms and shift of 10 ms. For the SpatialNet model, the window length and shift are set to 32 ms and 16 ms, respectively. Both setups extract 257-dimensional complex spectra for the 16 kHz sampling rate. We modify the TF-GridNet and SpatialNet models to incorporate speaker embedding sequences with a mixture signal. Specifically, we divide the input encoder layer, a two-dimensional convolution for TF-GridNet and a one-dimensional convolution for SpatialNet, each with D channels, into two separate encoders. The two encoders, with $D - 2$ and 2 channels, process the mixture signal and the embedding sequences, respectively. The outputs of these two encoders are then stacked and fed into the subsequent blocks of each network.

Before training, the sample variance of each mixture segment is normalized to 1.0, and the corresponding scaling factor is applied to the clean target sources. We employ the Adam optimizer, with the ℓ_2 norm of gradients capped at 1.0. The learning rate is initialized to 0.001, and halved if no improvement in validation loss is observed over three epochs. All models use mixed precision to expedite training.

[1]Available online at: https://github.com/jsalt2020-asrdiar/jsalt2020_simulate

TABLE I
DER RESULTS (IN %) OF COMPARISON DIARIZATION SYSTEMS ON LIBRICSS

Separation Method	Diarization Method	Overlap Ratio						All
		0S	0L	10%	20%	30%	40%	
-	X-vector + SC [34]	9.29	10.25	14.04	18.76	23.82	27.43	18.19
Mask-based MVDR [2]	X-vector + SC	11.49	13.42	11.63	14.22	16.98	16.18	14.18
MIMO-BF-MISO [37]	X-vector + SC	9.33	10.4	8.97	9.52	11.66	9.54	9.9
MISO-BF-MISO [6]	DOA-based [38]	11.95	10.69	11.25	12.22	13.04	14.31	12.36
-	RPN [34]	4.5	9.1	8.3	6.7	11.6	14.2	9.5
-	TS-VAD [34]	6.0	**4.6**	6.6	7.3	10.3	9.5	7.6
-	TS-SEP [39]	-	-	-	-	-	-	6.49
-	SC-EEND (PIT)	23.2	23.48	29.55	27.1	34.94	39.22	30.36
-	MC-EEND (PIT)	7.56	6.57	5.09	7.18	8.44	12.17	8.05
-	MC-EEND (LBT)	**4.94**	6.12	**3.36**	**4.09**	**4.88**	**5.05**	**4.68**

For the multi-speaker ASR evaluation, two pretrained ASR models from ESPnet are employed [32]. The first ASR model is an end-to-end Transformer-based system [33], [34]. This model is equipped with 12 self-attention blocks in the encoder and 6 in the decoder, and trained on the 960 h Librispeech corpus. The second ASR model[2] is a conformer-based system, leveraging self-supervised learning (SSL) features derived from WavLM [35]. This ASR model achieves a WER of 1.9% on the clean test set of LibriSpeech. We refer to the first ASR model as E2E and the second as E2E-SSL.

We report multi-speaker ASR performance using concatenated minimum-permutation word error rate (cpWER) [36]. This metric evaluates speaker-attributed ASR and is computed by sequentially joining all the utterances of each separated and target speaker. Following this joining, all speaker pairs are scored. The permutation yielding the lowest WER is then selected.

We also conduct the continuous-input evaluation of LibriCSS [2]. For this evaluation, each 10-minute mini-session recording, is pre-segmented into segments spanning 60 to 120 seconds. Each of these segments includes 8 to 10 utterances. The objective here is to accurately recognize all the utterances within a segment in a speaker-agnostic manner. While the ASR backend evaluates both streams individually, it combines the decoding results to determine the final WER. For this evaluation, we employ the default ASR backend from the LibriCSS dataset for consistent comparisons with other algorithms.

IV. EVALUATION RESULTS AND COMPARISONS

A. Diarization Results

Table I presents the DER results for the proposed MC-EEND models and comparison baselines on LibriCSS. To provide a comprehensive perspective on these results, we compare our MC-EEND models with a number of other diarization methods, all of which have been evaluated on the same LibriCSS corpus. The x-vector+SC diarization method [34] achieves 18.19% DER. This diarization method can be combined with CSS-based separation for DER reduction. Here, clustering is performed for both separated streams concurrently, as speaker segments can be assigned to either separated stream. Raj et al. [34] reported that using mask-based minimum variance distortionless response (MVDR) beamforming [2] prior to diarization improves DER results.

We have conducted experiments using a more powerful CSS-based separation model. Specifically, we employ the MIMO-BF-MISO model [37] for separation. This model is based on MIMO complex spectral mapping through a TF-GridNet architecture. Additionally, it includes a beamformer and an enhancement model for post-filtering. As part of post-processing, the separation model performs speaker localization to reduce speaker splitting errors. Coupling this separation model and x-vector+SC cuts DER by half, resulting in a 8.29% absolute DER reduction.

The next baseline employs a diarization model that relies on direction of arrival (DOA) estimation [38]. This model utilizes a CSS-based MISO-BF-MISO system [6] to perform speaker separation, estimate the DOA for each separated speaker, and then group the separation results across segments according to the DOA estimates. Region proposal network (RPN) is supervised method that integrates both segmentation and embedding extraction steps into one neural network, optimizing them jointly [40]. After obtaining the embeddings, they are clustered using K-means clustering based on the oracle number of speakers.

Another notable diarization baseline is target-speaker voice activity detection (TS-VAD), a two-stage method [41]. Initially, diarization estimates are obtained using a clustering-based method. Subsequently, a DNN is employed to refine these initial diarization estimates. Specifically, the DNN model takes

[2]Available online at: https://huggingface.co/espnet/simpleoier_librispeech_asr_train_asr_conformer7_wavlm_large_raw_en_bpe5000_sp

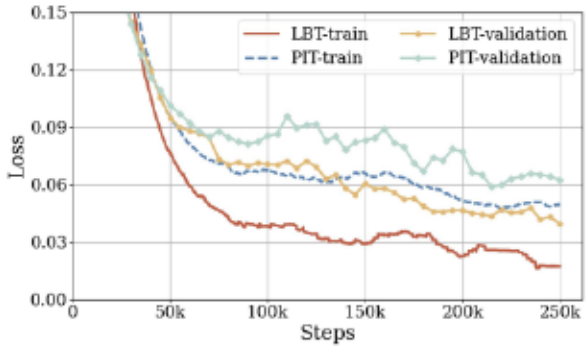

Fig. 3. Comparison of diarization loss curves (L_{EEND}) in training and validation with LBT and PIT criteria.

TABLE II
DER (IN %) FOR DIFFERENT FRAME SHIFTS AND DIARIZATION THRESHOLDS. MI, FA AND CF REFER TO MISSED SPEECH, FALSE ALARM, AND CONFUSION ERRORS, RESPECTIVELY

Frame Shift	Threshold	DER	MI	FA	CF
30 ms	$\tau = 0.5$	**4.68**	1.86	**2.16**	0.66
40 ms	$\tau = 0.5$	4.97	1.51	2.87	0.59
50 ms	$\tau = 0.5$	5.72	1.50	3.66	0.56
30 ms	$\tau = 0.3$	7.98	0.84	6.48	0.66
40 ms	$\tau = 0.3$	7.41	0.69	6.28	0.44
50 ms	$\tau = 0.3$	7.62	**0.53**	6.98	**0.11**

acoustic features along with representative embeddings for each speaker as inputs and generates frame-level activities for each speaker. To address the permutation ambiguity problem, TS-VAD arranges the DNN outputs based on the order of input embeddings, and assumes the knowledge of the total number of speakers in a meeting. TS-VAD achieves strong diarization results in the CHiME-6 challenge [41].

An extension to TS-VAD is target-speaker separation (TS-SEP), which combines speaker diarization and separation into a unified process [39]. More specifically, TS-SEP extends the final output layer of TS-VAD to generate time-frequency masks for individual speakers. To achieve robust diarization and separation, TS-SEP incorporates several additional techniques. First, it employs weighted prediction error (WPE) for speech dereverberation [42] before mask estimation. Second, TS-SEP applies mask-based MVDR beamforming, and the resulting masks are further refined through guided source separation (GSS) [43]. This baseline achieves a 6.49% DER.

Compared to all the aforementioned baselines, our proposed MC-EEND trained with LBT achieves a 4.68% DER, surpassing all other diarization methods. It is worth stressing that MC-EEND attains these results without employing any speech separation techniques. Furthermore, MC-EEND is a single-stage system, making it easier to train than multi-stage diarization methods that depend on other modules. The results also demonstrate that MC-EEND generalizes well to real conversational recordings, despite using only simulated RIRs for training.

For reference, we have also trained an MC-EEND diarization model using the PIT criterion. To accommodate training for 8 speakers, we employ a PIT criterion that leverages the Hungarian algorithm [21] with an $O(C^3)$ computational complexity. From Table I we observe that the MC-EEND model with LBT significantly outperforms that with PIT. Fig. 3 displays the diarization loss curves for LBT and PIT, and it is evident from the curves that the diarization loss for LBT is considerably lower. This suggests that the PIT criterion may be suboptimal when dealing with a larger number of speakers.

To investigate this further, we have trained a single-channel EEND (SC-EEND), which is based on the original EEND-EDA [12], using the PIT criterion. Table I indicates that SC-EEND with the PIT criterion yields poor separation performance. This is consistent with the findings from other studies [12], [44] reporting the poor performance of PIT-based EEND models for many speakers.

B. Diarization Tuning for SSND

We extract speaker embeddings for SSND from our MC-EEND model trained with LBT. Prior to the extraction of embeddings, we fine-tune our MC-EEND model to achieve optimal cpWER results. Of the three diarization errors—missed speech, confusion, and false alarm—the first two are particularly detrimental for ASR. Missed speech errors lead to deletion errors, while confusion errors result in a deletion error for one speaker and an insertion error for the other speaker. In contrast, false alarm errors do not contribute to cpWER because ASR systems do not generate recognition hypotheses for silent frames. Additionally, we have empirically observed that overly strict segment boundaries contribute to deletion errors at the beginning and ending of an utterance. Therefore, tolerating false alarm errors can be a strategic choice to improve ASR performance.

To enhance SSND performance, one can tune the diarization threshold (τ) as well as frame shift. Table II provides diarization errors using different frame shifts and threshold values. From the table the best DER is obtained at the smallest shift (30 ms) with $\tau = 0.5$. With this setting, missed speech and false alarm errors appear balanced. When we lower the threshold to 0.3, there is a marked reduction in missed speech errors, but at the expense of increased false alarm errors.

In terms of combined missed speech and confusion errors, the best setting is a 50 ms frame shift and $\tau = 0.3$. Consequently, we adopt this MC-EEND setting for subsequent separation and ASR experiments.

C. Segment Size and Shift Analysis for SSND

In this section, we investigate the influence of segment size and shift on the performance of SSND. This analysis utilizes the TF-GridNet and SpatialNet architectures, and an E2E ASR method with oracle utterance boundaries. The cpWER results for various segment sizes and shifts processed through SSND are shown in Fig. 4.

The default CSS segment size and shift (2.4/1.2 seconds) offer limited contextual information and lead to subpar cpWER performance, particularly with TF-GridNet. It is important to

TABLE III
cpWER Results (In %) for Different Separation and Diarization Methods

Separation Method	Diarization Method	ASR	Overlap Ratio						Avg
			0S	0L	10%	20%	30%	40%	
Unprocessed	Oracle	E2E	5.23	5.19	11.43	19.31	28.46	38.32	19.72
MIMO-BF-MISO [37]	Oracle	E2E	3.45	3.87	3.15	4.46	5.35	5.76	4.44
SSND (TF-GridNet Large)	Oracle	E2E	4.51	4.01	3.74	4.28	5.13	5.39	4.58
SSND (SpatialNet)	Oracle	E2E	4.04	3.97	3.37	3.54	4.51	4.66	4.04
SSND (TF-GridNet Large)	Oracle	E2E-SSL	2.51	2.34	2.46	2.55	2.91	2.93	2.64
SSND (SpatialNet)	Oracle	E2E-SSL	2.28	2.38	2.36	2.27	2.68	2.47	2.42
Unprocessed	X-vector + SC	E2E	13.95	12.2	20.12	29.64	35.06	41.81	27.01
Mask-based MVDR [2]	X-vector + SC	E2E	8.78	13.07	10.51	15.37	17.54	17.63	14.13
MIMO-BF-MISO [37]	X-vector + SC	E2E	7.05	8.05	7.31	8.48	10.81	10.03	8.76
SSND (TF-GridNet Large)	MC-EEND ($\tau = 0.5$)	E2E	5.11	4.88	5.55	6.89	7.98	9.25	6.84
SSND (TF-GridNet Large)	MC-EEND ($\tau = 0.3$)	E2E	5.97	3.74	4.69	5.32	6.18	7.28	5.69
SSND (SpatialNet)	MC-EEND ($\tau = 0.3$)	E2E	5.56	3.52	3.98	4.76	5.58	6.55	5.13
SSND (TF-GridNet Large)	MC-EEND ($\tau = 0.3$)	E2E-SSL	3.77	2.02	2.56	3.29	3.78	4.11	3.36
SSND (SpatialNet)	MC-EEND ($\tau = 0.3$)	E2E-SSL	3.67	1.97	2.46	3.06	3.52	4.07	3.22

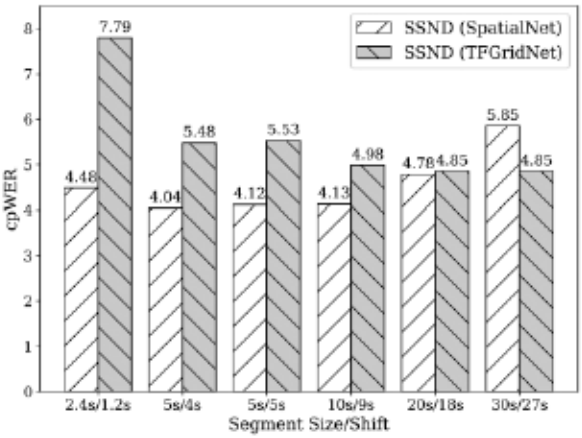

Fig. 4. Effect of segment size and segment shift on cpWER in SSND.

note that our approach uses no stitching, meaning these cpWERs are not due to potential stitching errors. For SSND using the TF-GridNet model, increasing the segment size up to 30 seconds (i.e. 12 times the default CSS segment size) results in improved performance. This finding is significant as it suggests the model ability to process longer segments without compromising performance. Interestingly, in scenarios with no overlap between adjacent segments, e.g., segment size/shift of 5/5 seconds, cpWER remains good. This observation implies that one can reduce the computational cost for SSND with little performance degradation. In the case of SSND with SpatialNet, increasing the segment size improves ASR performance up to 10 seconds. For longer segments (20 seconds and 30 seconds), performance starts to decline. This trend may be attributed to the lack of a full-band self-attention module and recurrent layers in SpatialNet to capture long-range contextual information. Based on the results in Fig. 4, we adopt a segment size/shift of 30/27 seconds for TF-GridNet models and 5/4 seconds for SpatialNet in the subsequent experiments.

D. Speaker-Attributed ASR Results

In this section, we present the speaker-attributed ASR results on the LibriCSS dataset. Table III displays the cpWER results using various separation and diarization methods. For comparisons, we establish a strong baseline using a CSS method based on MIMO-BF-MISO built on TF-GridNet [37]. To obtain cpWER results with oracle utterance boundaries for the MIMO-BF-MISO system, the cross-correlation function between the separated audio streams and the reference utterances is used to determine channel assignments accurately.

A notable observation is that the MIMO-BF-MISO method reduces cpWER by a large margin—from 19.72% to 4.44%—over the unprocessed mixtures when coupled with E2E ASR using oracle utterance boundaries. Furthermore, SSND with TF-GridNet Large produces ASR results on par with MIMO-BF-MISO, while SSND with SpatialNet surpasses the performance of both models. Using E2E-SSL ASR, we achieve excellent recognition results for the TF-GridNet and SpatialNet SSND models, with 2.64% and 2.42% cpWER respectively. These results are close to the clean transcription level of 1.9%. Interestingly, we observe that the difference in performance between TF-GridNet and SpatialNet becomes marginal when E2E-SSL is used for ASR.

Using estimated speaker boundaries provided by diarization, we notice a significant increase in cpWER for the MIMO-BF-MISO system coupled with clustering-based diarization, to an 8.76% cpWER. However, this performance remains substantially better than the mask-based MVDR result at 14.13% cpWER.

For SSND using the TF-GridNet Large model, there is a small increase in cpWER to 5.69% over oracle speaker boundaries. The impact of diarization tuning becomes evident when an MC-EEND diarization model is employed with $\tau = 0.5$, degrading cpWER to 6.84%. SSND with SpatialNet surpasses the performance of TF-GridNet Large, reaching a 5.13% cpWER.

TABLE IV
PERFORMANCE COMPARISONS OF SPEAKER-ATTRIBUTED ASR SYSTEMS ON LIBRICSS DATASET

Ref.	Separation	Diarization	ASR	cpWER
[38]	CSS	DOA-based	TDNN-F [34]	12.98
[34]	CSS	X-vector + SC	E2E	12.7
[45]	–	–	SA-ASR	11.6
[46]	Speakerbeam	TS-VAD	E2E	18.8
[46]	GSS	TS-VAD	E2E	11.2
[39]	TS-SEP		E2E	6.42
[39]	TS-SEP		E2E-SSL	5.36
Ours	SSND (SpatialNet)		E2E	5.13
Ours	SSND (SpatialNet)		E2E-SSL	**3.22**

TABLE V
WER RESULTS (IN %) OF COMPARISON SYSTEMS FOR CONTINUOUS-INPUT EVALUATION ON LIBRICSS USING DEFAULT ASR BACKEND

	Overlap Ratio						Avg
	0S	0L	10%	20%	30%	40%	
Unprocessed	15.4	11.5	21.7	27	34.3	40.5	25.06
CSS							
MVDR [2]	11.9	9.7	13.4	15.1	19.7	22.0	15.29
Conformer [47]	11.0	8.7	12.6	13.5	17.6	19.6	13.83
MISO-BF-MISO [6]	7.7	7.5	7.4	8.4	9.7	11.3	8.66
TF-GridNet	9.0	10.8	10	10.4	12.0	12.9	10.85
MIMO-BF-MISO [37]	**6.8**	6.8	**6.7**	6.9	8.4	9.0	7.43
SSND							
TF-GridNet	7.9	7.4	7.8	7.8	9.8	10.3	8.50
SpatialNet	7.2	**6.5**	**6.7**	**6.6**	**8.3**	**8.6**	**7.33**

Moreover, using E2E-SSL for ASR yields a marked improvement in cpWER for both TF-GridNet Large and SpatialNet models, achieving cpWERs of 3.36% and 3.22% respectively.

We compare the speaker-attributed performance of the proposed SSND model with other representative algorithms in Table IV. The CSS-based MISO-BF-MISO system coupled with DOA-based diarization and a hybrid ASR model [34] reports a 12.98% cpWER [38]. Another system outlined in [34] employs mask-based MVDR, x-vector+SC diarization, and E2E ASR, and achieves a 12.7% cpWER [34]. A single-channel E2E speaker-attributed ASR system [45] derives diarization estimates from the internal state of the recognizer, and achieves a 11.6% cpWER. The system in [46] uses TS-VAD for diarization and reports the cpWER scores of 18.8% and 11.2% for single- and multi-channel setups utilizing a speakerbeam and GSS, respectively. Finally, the TS-SEP system [39] reaches the cpWER values of 6.42% and 5.36% with E2E and E2E-SSL ASR, respectively. The cpWER results in Table IV demonstrate that our proposed SSND model surpasses all previous results, achieving the remarkable cpWER scores of 5.13% and 3.22% for E2E and E2E-SSL, respectively. Our results establish a new state-of-the-art benchmark for speaker-attributed ASR on the LibriCSS dataset.

E. Speaker-Agnostic ASR Results

This section assesses our SSND model using the continuous-input evaluation of the LibriCSS dataset. Additionally, we compare with other works using the default ASR backend. It is important to note that these results do not incorporate diarization since the segment boundaries (with several utterance from different speakers) are provided for this evaluation, so the corresponding ASR is referred to as speaker-agnostic. The system in [47] estimates real-valued time-frequency masks using a conformer architecture.

Table V presents the continuous-input evaluation results of the proposed SSND and comparison methods. The results show that our SSND based on TF-GridNet significantly outperforms the corresponding CSS-based system. This can be attributed to the utilization of a larger context—30 seconds as opposed to the 2.4 seconds used in CSS. Using SpatialNet, our SSND results show a further improvement over TF-GridNet, achieving 7.33% WER on average and surpassing the performance level of the best-performing CSS system using MIMO-BF-MISO [37]. It should be noted that, unlike MIMO-BF-MISO, the SSND system does not perform additional processing such as beamforming and localization. On the other hand, SSND makes use of speaker embedding sequences as additional inputs.

V. CONCLUDING REMARKS

In this paper, we have proposed a new multi-channel diarization model, MC-EEND, which produces state-of-the-art diarization performance on the LibriCSS dataset. We find that PIT falls short when diarizing many speakers. With multi-channel recordings, we demonstrate that the LBT criterion effectively resolves the permutation ambiguity problem in talker-independent diarization.

Furthermore, we have introduced the SSND framework, a novel approach that seamlessly integrates speaker diarization with speaker separation, making it well-suited for speaker-attributed ASR. Our SSND framework achieves state-of-the-art performance for speaker-attributed ASR, as well as speaker-agnostic ASR (standard CSS), on the LibriCSS dataset. Unlike CSS, the SSND framework is capable of processing long segments regardless the number of participating speakers. SSND avoids stitching needed in CSS and ensures that consecutive segments are sequentially organized.

Future research will extend MC-EEND to causal and real-time implementation and moving speakers, and connect to speaker localization and tracking. Additional research is also needed to deal with many speakers in a single-channel setup.

ACKNOWLEDGMENT

The authors would like to thank Dr. Zhong-Qiu Wang, Dr. Ashutosh Pandey, Dr. Daniel Wong and Dr. Buye Xu for helpful discussions.

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, no. 10, pp. 1702–1726, Oct. 2018.

[2] Z. Chen et al., "Continuous speech separation: Dataset and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7284–7288.

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

[4] S. Chen et al., "Don't shoot butterfly with rifles: Multi-channel continuous speech separation with early exit transformer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6139–6143.

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

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

[7] H. Taherian and D. L. Wang, "Multi-resolution location-based training for multi-channel continuous speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[8] H. Taherian, A. Pandey, D. Wong, B. Xu, and D. L. Wang, "Multi-input multi-output complex spectral mapping for speaker separation," in *Proc. Interspeech*, 2023, pp. 1070–1074.

[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, no. 10, pp. 1901–1913, Oct. 2017.

[10] T. v. Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "Segment-less continuous speech separation of meetings: Training and evaluation criteria," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 576–589, 2023.

[11] W. Zhang et al., "Separating long-form speech with group-wise permutation invariant training," in *Proc. Interspeech*, 2022, pp. 5383–5387.

[12] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. García, "Encoder-decoder based attractors for end-to-end neural diarization," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 1493–1507, 2022.

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

[14] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, Sep. 2006.

[15] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Comput. Speech Lang.*, vol. 72, 2022, Art. no. 101317.

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5329–5333.

[17] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *Proc. Interspeech*, 2019, pp. 4300–4304.

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

[19] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 31–35.

[20] H. Tachibana, "Towards listening to 10 people simultaneously: An efficient permutation invariant training of audio source separation using sinkhorn's algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 491–495.

[21] S. Dovrat, E. Nachmani, and L. Wolf, "Many-speakers single channel speech separation with optimal permutation training," in *Proc. Interspeech*, 2021, pp. 3890–3894.

[22] Ö. Çetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, 2006, pp. 293–296.

[23] 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," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3221–3236, 2023.

[24] C. Quan and X. Li, "SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 1310–1323, 2024.

[25] 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.

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

[27] H. Taherian, S. E. Eskimez, T. Yoshioka, H. Wang, Z. Chen, and X. Huang, "One model to enhance them all: Array geometry agnostic multi-channel personalized speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 271–275.

[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] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.

[30] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[31] O. Sadjadi, C. Greenberg, E. Singer, L. Mason, and D. Reynolds, "NIST 2021 speaker recognition evaluation plan," Tech. Rep., 2021, [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=932697

[32] S. Watanabe et al., "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[33] S. Karita et al., "A comparative study on transformer vs RNN in speech applications," in *Proc. Autom. Speech Recognit. Understanding*, 2019, pp. 449–456.

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

[35] S. Chen et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[36] S. Watanabe et al., "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. Int. Workshop Speech Process. Everyday Environ.*, 2020, pp. 1–7.

[37] H. Taherian, A. Pandey, D. Wong, B. Xu, and D. L. Wang, "Leveraging sound localization to improve continuous speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 1–5.

[38] Z.-Q. Wang and D. Wang, "Localization based sequential grouping for continuous speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 281–285.

[39] C. Boeddeker, A. S. Subramanian, G. Wichern, R. Haeb-Umbach, and J. Le Roux, "TS-SEP: Joint diarization and separation conditioned on estimated speaker embeddings," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 1185–1197, 2024.

[40] Z. Huang et al., "Speaker diarization with region proposal network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6514–6518.

[41] I. Medennikov et al., "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. Interspeech*, 2020, pp. 274–278.

[42] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[43] C. Boeddecker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *Proc. Int. Workshop Speech Process. Everyday Environ.*, 2018, pp. 35–40.

[44] T. Cord-Landwehr, C. Boeddeker, C. Zorilă, R. Doddipatla, and R. Haeb-Umbach, "Frame-wise and overlap-robust speaker embeddings for meeting diarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.

[45] N. Kanda et al., "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 8082–8086.

[46] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, and T. Nakatani, "Speaker activity driven neural speech extraction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6099–6103.

[47] S. Chen et al., "Continuous speech separation with conformer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 5749–5753.