ASoBO: Attentive Beamformer Selection for Distant Speaker Diarization in Meetings

Théo Mariotte, Anthony Larcher, Silvio Montrésor, Jean-Hugh Thomas

ASoBO: Attentive Beamformer Selection for Distant Speaker Diarization in Meetings

Théo Mariotte[1,2], *Anthony Larcher*[2], *Silvio Montrésor*[1], *Jean-Hugh Thomas*[1]

[1]LIUM, Institut Claude Chappe, Le Mans Université, France
[2]LAUM IA-GS UMR CNRS 6613, Le Mans Université, France

theo.mariotte@univ-lemans.fr

Abstract

Speaker Diarization (SD) aims at grouping speech segments that belong to the same speaker. This task is required in many speech-processing applications, such as rich meeting transcription. In this context, distant microphone arrays usually capture the audio signal. Beamforming, i.e., spatial filtering, is a common practice to process multi-microphone audio data. However, it often requires an explicit localization of the active source to steer the filter. This paper proposes a self-attention-based algorithm to select the output of a bank of fixed spatial filters. This method serves as a feature extractor for joint Voice Activity (VAD) and Overlapped Speech Detection (OSD). The speaker diarization is then inferred from the detected segments. The approach shows convincing distant VAD, OSD, and SD performance, e.g. 14.5% DER on the AISHELL-4 dataset. The analysis of the self-attention weights demonstrates their explainability, as they correlate with the speaker's angular locations.

Index Terms: speaker diarization, distant speech, multi-microphone, explainable AI

1. Introduction

Speaker diarization (SD) is an automatic speech processing task that answers the question *Who spoke and when?* in an audio stream. It is of major interest for rich meeting transcriptions where the speaker activity is required [1]. Two categories of systems appear in the literature [2]: end-to-end neural diarization (EEND) [3,4] and pipeline systems [5,6]. The former infers speaker activities from the raw audio signal and usually requires large synthetic datasets to be trained. The latter comprises sub-blocks that (1) detect speaker-homogeneous segments, and (2) cluster these segments to group them by speakers. The segmentation step can be divided into two tasks: Voice Activity Detection (VAD) to detect speech segments [7, 8], and Overlapped Speech Detection (OSD) to identify segments where several speakers are simultaneously active [9, 10]. Speaker change detection (SCD) [11] can also be performed to detect boundaries between speakers in speech segments. However, this task is out of the scope of this paper.

SD in meetings is a challenging task due to spontaneous speech, an unknown number of speakers, and difficult acoustic conditions [12, 13]. This scenario remains challenging, as shown by the recently organized challenges [1, 14]. A common approach is to record meetings with a multi-microphone device [1,15,16] such as uniform circular arrays (UCA) [17,18]. Microphone arrays have been widely studied in the literature [17, 18]. Specifically, beamforming extracts a signal steered in a given direction, e.g. by weighting and combining channels in the Fourier domain [18]. Both signal- [13, 17–19] and neural-based [20–22] approaches have been investigated. While neural beamformers increase the number of trainable parameters, signal-based approaches often require estimating the source direction of arrival (DoA).

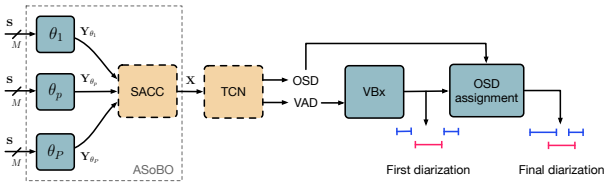

Figure 1: *High-level architecture diagram of the proposed ASoBO and the speaker diarization pipeline. The θ_p blocks represent the fixed spatial filters. Only dash line blocks are trained.*

In this work, we introduce the Attentive Selection of Beamfomer Outputs (ASoBO) as a multi-microphone front-end for VAD+OSD. A set of signal-based beamformers is steered in fixed, separated angular directions. The outputs of the beamformers are weighted and combined with a Self-Attention Channel Combinator (SACC) [23]. This results in a single-channel enhanced representation of the multi-microphone input signal. ASoBO prevents the DoA estimation step while limiting the number of trainable parameters. A first SD is obtained by first applying the VBx system [6] to the VAD output. The final SD is inferred by detecting and assigning OSD segments [24, 25]. The proposed ASoBO shows convincing SD performance on two multi-microphone datasets recorded in the meeting scenario. Furthermore, we show that the speaker's angular direction can be inferred in an unsupervised way from the self-attention weights. The code is available at `https://git-lium.univ-lemans.fr/speaker/sidiar/`.

2. Segmentation for speaker diarization

The proposed speaker diarization (SD) system is pipeline-based. This section describes the VAD+OSD formulation and its use for SD. The overall architecture is presented in figure 1.

2.1. Feature extraction

Let $\mathbf{X} = [\mathbf{X}_1, \dots \mathbf{X}_t, \dots \mathbf{X}_T] \in \mathbb{R}^{F \times T}$ be a sequence of feature vectors where F is the number of features, T the number of time frames and t the time frame index. This sequence is extracted from the raw audio signal $\mathbf{s} \in \mathbb{R}^{M \times L}$, with M being the number of microphones and L being the number of samples. Feature extraction can be defined as a function $g : \mathbb{R}^{M \times L} \rightarrow \mathbb{R}^{F \times T}$, which maps the raw input signal to the sequence of feature vectors. In this paper, we propose a new design for the g function that uses a set of beamformers, i.e. spatial filter banks, followed by a self-attention model to combine the filter bank output channels. The method is presented in Section 3.

2.2. Frame classification

The sequence of feature vectors serves as input for a VAD+OSD system. Let $\mathbf{y} = [y_1, \dots y_t, \dots y_T] \in \mathbb{R}^T$ be a sequence of reference binary labels aligned with the sequence $\mathbf{X}$. VAD+OSD is solved by optimizing the parameters $\hat{\boldsymbol{\theta}}$ of the model $f : \mathbf{X}, \boldsymbol{\theta} \rightarrow \hat{\mathbf{y}}$ which maps the feature sequence to a sequence of predicted labels $\hat{\mathbf{y}} = [\hat{\mathbf{y}}_1, \dots \hat{\mathbf{y}}_t, \dots \hat{\mathbf{y}}_T] \in \mathbb{R}^{C \times T}$, with C being the number of classes. An element of $\hat{\mathbf{y}}_t$ contains the pseudo-probability for the frame $\mathbf{X}_t$ to belong to each class. The system is trained to predict $C = 3$ classes. The first class corresponds to the non-speech scenario with $N_{spk} = 0$ active speaker. The second and third classes correspond to $N_{spk} = 1$ and, $N_{spk} \geq 2$ respectively. Therefore, VAD can be solved by combining the two last outputs, i.e. $N_{spk} > 0$. OSD is inferred from the $N_{spk} \geq 2$ output.

2.3. Integrating segmentation for speaker diarization

VAD and OSD predictions are used to solve speaker diarization (SD) as illustrated in figure 1. The speech segments detected with VAD (all speech segments, overlap included) are used to extract speaker embeddings. The embeddings are then clustered, and the segmentation is refined using the VBx approach [6]. Once the first SD, OSD can be used to assign a second speaker to overlapping speech regions by an additional post-processing step. In this work, we use the approach from [24] that assigns the closest speaker in time to the overlapping segment. This approach has shown on-par performance as more complex approaches such as VB-based methods [25].

3. Self-Attentive beamformer selection

This section describes the proposed feature extraction algorithm. An abstract view of the method is depicted in figure 1.

3.1. Super-directive beamforming

Super-directive beamforming is a commonly used algorithm for spatial filtering [12]. The narrowband weights of such a filter can be expressed as follows:

$$\mathbf{w}_p^H(f) = \frac{\mathbf{v}_p^H(f)\boldsymbol{\Sigma}_N^{-1}(f)}{\mathbf{v}_p^H(f)\boldsymbol{\Sigma}_N^{-1}(f)\mathbf{v}_p(f)}, \tag{1}$$

where f is the frequency, $\mathbf{v}_p(f) \in \mathbb{C}^{M \times 1}$ a steering vector and $\boldsymbol{\Sigma}_N(f) \in \mathbb{R}^{M \times M}$ the noise covariance matrix. In this work, we use the standard isotropic noise assumption usually considered in super-directive beamforming [12]. When considering a UCA, the m-th element of the steering vector $v_{p,m}$, oriented towards the θ_p angular direction, can be expressed as [18] : $v_{p,m}(f) = \exp\left(j2\pi f r c^{-1} \cos(\theta_p - \psi_m)\right)$, with m being the index of the microphone with angle ψ_m, c the speed of sound and r the radius of the UCA.

Let $\mathbf{S} \in \mathbb{C}^{M \times F \times T}$ be the short-time Fourier Transform (STFT) of the multi-microphone input signal $\mathbf{s}$. The output of the p-th filter – steered in the θ_p direction – at frequency f, is obtained following

$$\mathbf{Y}_p(t, f) = \mathbf{w}_p^H(f)\mathbf{S}(t, f). \tag{2}$$

The output $\mathbf{Y}_p \in \mathbb{C}^{T \times F}$ is a single-channel signal steered towards the θ_p direction. One can build a set of spatial filters $\mathcal{W} = \{\mathbf{W}_p\}_{p=1}^P$ steered in P unique angular directions, where $\mathbf{W}_p = [\mathbf{w}_p(f_i)], i = 1, \dots, F$ is the broadband filter coefficients. Filtering $\mathbf{S}$ by all the filters from $\mathcal{W}$ results in a new multichannel signal $\mathbf{Y} \in \mathbb{C}^{T \times P \times F}$, where the p-th channel corresponds to the beamformed version of $\mathbf{S}$ in the θ_p direction.

3.2. Beamformer selection

Once the input signal is filtered by the spatial filter bank $\mathcal{W}$, a second step selects the optimal directions at the frame level. This filter selection is performed using the Self-Attention Channel Combinator (SACC) module, which is efficient for audio channel selection [23, 26]. Contrary to [26], the SACC is applied to beamforming outputs instead of the multi-microphone signal directly. We first only keep the magnitude of the beamforming output $\mathbf{Y}$. Then, $|\mathbf{Y}|^2$ is projected by three linear layers to the query and key $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{T \times P \times D}$, and the value $\mathbf{V} \in \mathbb{R}^{T \times P \times 1}$, with D being the output dimension of the linear transformation. The attention weights $\mathbf{w}_{SA} \in \mathbb{R}^{T \times P}$ are computed as follows:

$$\mathbf{w}_{SA} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}}\right)\mathbf{V}, \tag{3}$$

with $\cdot^T$ the transpose operator applied to each frame of $\mathbf{K}$. The spectrogram resulting from the attentive selection $\bar{\mathbf{Y}}$ is calculated by first weighting $\mathbf{Y}$ with $\mathbf{w}_{SA}$. A sum is then applied on the channel dimension P. For a given frame t, this operation can be expressed as:

$$\bar{\mathbf{Y}}_t = \sum_{p=1}^{P} \mathrm{softmax}\left(\mathbf{w}_{SA,t}\right) \odot \mathbf{Y}_t, \tag{4}$$

with $\odot$ being the element-wise product on the channel dimension, i.e. one weight at t is applied to all the frequencies of $\mathbf{Y}$. The $\mathrm{softmax}$ activation function is applied to the channel dimension such that $w_{SA,t,p} \in [0, 1]$. The final feature sequence $\mathbf{X}$ is obtained by converting $\bar{\mathbf{Y}}$ to the Mel scale with 64 triangular filters [23].

4. Experimental protocol

4.1. Datasets

The experiments are conducted on two datasets featuring distant multi-microphone speech with known array geometry: AMI [15] and AISHELL-4 [16]. The AMI corpus is about 100h of meetings in English with up to 5 participants recorded using different devices. In this work, we use the audio recorded by the 8-microphone, 10cm-radius UCA placed in the center of the table during the sessions. The data partition follows the protocol proposed in [6] since it guarantees no speaker overlap between the subsets. The AISHELL-4 dataset provides 120 hours of conference recordings with 4 to 8 participants. Audio is recorded with an 8-microphone, 5cm-radius UCA usually placed in the center of the table. Meetings are in Mandarin and were recorded in various acoustic environments. Both datasets are sampled at 16kHz.

4.2. Implementation details

The ASoBO inputs complex STFT extracted on 25ms segments with 10ms shift. P spatial filters, steered in uniformly-spaced angular sectors between 0 and 2π are applied to the input STFT. We empirically found that $P = 4$ and $P = 8$ offer the best performance on AMI and AISHELL-4, respectively. The P-channel output signal is then processed by the SACC algorithm with a hidden size $D = 256$. The modeling of the ASoBO

feature sequence is performed with the same TCN-based architecture as [27]. It is composed of 3 TCN blocks with residual connections. Each block contains 5 1D convolutional layers with exponentially increasing dilation.

The speaker diarization is inferred with the VBx implementation proposed in [6]. This system uses a ResNet101 x-vector extractor followed by a VB-HMM clustering algorithm. We used the default AMI diarization setup from the available code[1]. The VAD segments, predicted by our systems, are used as an initial segmentation. X-vector clustering is initialized with Hierarchical Agglomerative Clustering (HAC) before performing VB clustering. Overlapped speech segments are assigned as a post-processing step using a heuristic approach [24].

4.3. Baselines

The ASoBO approach is compared to two baseline systems. As a lower-bound system, we consider the single-distant microphone (SDM) scenario. The segmentation is performed on the first microphone of the array. 20 Mel Frequency Cepstral Coefficients (MFCC) and its deltas are extracted from the audio signal on 25ms windows with 10ms shift. These features are fed directly to the segmentation system to obtain VAD and OSD predictions. As an upper-bound baseline, we consider the original implementation of the SACC architecture, which has shown strong OSD performance in the multichannel distant scenario [26]. The SACC is applied to the magnitude of the STFT calculated on 25ms windows with 10ms shift. The self-attention hidden dimension is set to $D = 256$. The speaker diarization performance is compared to [28] on both AMI and AISHELL-4 datasets. They report the performance obtained with both VBx and spectral clustering approaches.

4.4. Training and evaluation

The segmentation systems are trained on 200 epochs with batches of 64 segments. The training segment duration is fixed to 2s. The segmentation is solved as a multiclass classification task and is optimized in a supervised way using the cross-entropy loss. The weights of the models are optimized with the ADAM optimizer with the learning rate set to 0.001.

The evaluation is conducted on the evaluation set of both datasets. VAD and OSD predictions are inferred from a 2s sliding window with a 0.5s shift. VAD is evaluated regarding False Alarm (FA) and Missed Detection (MD) rates. The sum of both metrics, the Segmentation Error Rate (SER) is also reported [8]. OSD is evaluated in terms of Precision (P), Recall (R), and F1-score (F1). The speaker diarization is evaluated using the Diarization Error Rate (DER) [2]. We report the scores with ($\delta = 0.25$), and without ($\delta = 0$) forgiveness collar. Unless otherwise specified, values highlighted in bold indicate the best systems and the statistically equivalent ones ($p < 0.001$). We use the Wilcoxon signed-rank non-parametric test on the file-level scores [29].

5. Experimental study

This section presents the experimental results on both segmentation (VAD+OSD) and speaker diarization.

5.1. Segmentation performance

Table 1 presents the VAD and OSD performance on both AMI and AISHELL-4 datasets. On the AMI corpus, the SDM system reaches 6.57% SER on VAD and 65.4% F1-score on OSD. SACC outperforms the SDM model with 5.59% VAD SER and 68.4% OSD F1-score. The proposed ASoBO system shows mitigated performance with 6.53% SER on VAD but improves OSD compared to SDM with 67.2% F1-score. The VAD degradation can be explained by the high false alarm rate (4.16%). On the AISHELL-4 dataset, ASoBO reaches the best VAD performance with 4.39% SER compared to the SACC (4.57%) and the SDM (5.17%). The OSD scores on this dataset are low for each model. This can be explained by the low quality of the annotations on overlapping speech. The SDM shows a 31.3% F1-score and is largely outperformed by SACC (41.1%) and ASoBO (40.5%). In summary, the ASoBO system improves VAD+OSD concerning the SDM scenario. The segmentation performance on the AMI corpus is still mitigated compared to the original SACC. However, this approach also improves the segmentation on the AISHELL-4 dataset.

Table 1: *VAD+OSD performance on the AISHELL-4 and AMI evaluation sets. # Param. represents the number of trainable parameters in millions.*

		VAD			OSD		
AMI	#Param.	FA	Miss	SER	P	R	F1
SDM	0.26M	4.33	2.24	6.57	73.8	68.8	65.4
SACC	0.40M	2.91	3.61	**5.59**	78.1	60.8	**68.4**
ASoBO	0.36M	4.16	2.15	6.53	70.8	69.3	67.2
AISHELL-4		FA	Miss	SER	P	R	F1
SDM	0.26M	3.69	1.48	5.17	20.4	67.1	31.3
SACC	0.40M	3.35	1.21	4.57	28.4	74.4	**41.1**
ASoBO	0.36M	2.29	2.10	**4.39**	28.7	69.0	**40.5**

5.2. Speaker diarization performance

Table 2 shows the speaker diarization performance on both AMI and AISHELL-4 evaluation sets. The VBx-based system from [28] reaches 25.1% and 18.0% DER on AMI and AISHELL-4 respectively. The spectral clustering-based model offers 23.7% and 16.1% on these datasets.

On the AMI corpus, the SACC offers the best speaker diarization performance with 23.1%. Note that the overlap assignment improves the diarization performance by a relative +11.5%. ASoBO offers close performance with 24.1% when OSD segments are assigned. This system is limited by the segmentation performance. It reaches 16.7% DER when a forgiveness collar is applied, which is 0.4% far from SACC. Both SACC and ASoBO improve or offer similar performance as the baseline. The SDM system reaches 25.0% DER and is largely outperformed by multichannel systems.

The trend is different on the AISHELL-4 dataset. First, the OSD segment assignment degrades the performance. This was expected based on the low OSD performance on this dataset and was also observed in [28]. On this dataset, both SDM and SACC offer close performance, with 16.7% and 16.4% DER. ASoBO reaches the best DER with 14.5%. It improves SACC by a relative +11.5%.

In summary, ASoBO is a good candidate for distant speaker diarization with pipeline systems. While the improvement in the AMI data is mitigated, the performance on the AISHELL-4 dataset is noticeable and very encouraging for such a system. Beyond performance, self-attentive selection of the beamformer makes ASoBO an explainable system as shown in section 6.

[1] https://github.com/BUTSpeechFIT/VBx

Table 2: *Diarization Error Rate (DER) with each segmentation system on AMI and AISHELL-4 evaluation sets.*

	AMI		AISHELL-4	
	$\delta = 0.25$	$\delta = 0$	$\delta = 0.25$	$\delta = 0$
VBx [28]	-	25.1	-	18.0
Spectral [28]	-	23.7	-	16.1
SDM	19.4	27.5	11.3	16.7
↪ w/ OSD	17.7	25.0	24.9	29.3
SACC	18.2	26.1	11.0	16.4
↪ w/ OSD	**16.3**	**23.1**	19.3	23.9
ASoBO	18.6	26.9	**9.2**	**14.5**
↪ w/ OSD	**16.7**	24.1	16.6	20.9

6. Weight explanation

The self-attention module is trained to select the appropriate filters for the segmentation task. The intuition is that the self-attention model selects the filter steered toward the active speakers. This section verifies this hypothesis by analyzing the combination weights on simulated data.

6.1. Data simulation

The AMI and AISHELL-4 datasets do not come with speaker position annotations. The analysis of self-attention weights is performed on a simulated dataset. The simulations are conducted using the LibriMix dataset [30] featuring single-channel speech mixtures. The original mixtures are spatialized with simulated Room Impulse Responses (RIRs). These are generated using the GpuRIR toolkit [31]. The simulation setup is similar to the AMI corpus, with an 8-microphone UCA with a 10cm radius. We simulate the 2- and 3-speaker mixtures, with a reverberation time T_{60} of 0.6 seconds. Two evaluation scenarios are considered: *easy* where the sources are always aligned with a beamformer, *i.e.* one of the P directions, and *hard* where the source position is randomly sampled around the p-th selected filter direction with $\pm 5^\circ$.

6.2. Speaker localization from combination weights

Identifying the steering directions of the system can be seen as a speaker localization task. Let $\mathbf{w}_{SA} \in \mathbb{R}^{T \times P}$ be a set of ASoBO combination weights predicted from a given audio segment, calculated with equation (3). The activations in $\mathbf{w}_{SA}$ indicate the selected angular directions as a function of time. The average steering direction is calculated by first averaging $\mathbf{w}_{SA}$ across time: $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{softmax}\left(\mathbf{w}_{SA}\right)_t$, where the $\mathrm{softmax}$ activation is applied on the channel dimension as in equation (4). Hard labels are obtained by applying a threshold $\tau \in [0, 1]$ to each element of $\bar{\mathbf{w}}$. The angular directions $\hat{\theta}_p$ can be compared to the ground truth directions θ_p using the precision (P), recall(R) and F1-score (F1) metrics.

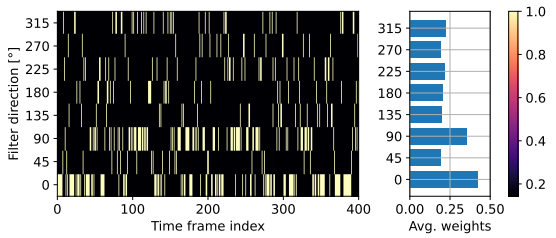

Figure 2: *(left) Combination weights for a spatialized utterance of Libri2Mix with speakers located at 0° and 90°. (right) Time-averaged weights from the same utterance.*

Table 3: *Speaker localization performance of ASoBO on spatialized Libri2Mix and Libri3mix development sets. L represents the number of simultaneously active sources.*

	$L = 2$			$L = 3$		
Scenario	P	R	F1	P	R	F1
Random	49.9	49.9	49.9	49,7	49,6	48,7
Easy	84.0	83.2	83.6	82.5	76.1	75.9
Hard	70.4	73.0	71.7	66.9	66.5	66.6

Figure 2 illustrates the combination weights obtained for two active sources at 0° and 90° respectively with the ASoBO system with $P = 8$ filters. The weight map (left) shows that the two directions of the active speaker are more often activated than the others. The averaged weights (right) confirm this behavior, with two peaks in the speakers' directions.

6.3. Localization performance

The spatial filter selection is evaluated as a speaker localization task. Each evaluation is conducted on a spatialized version of the LibriMix development set. Localization performance in the *easy* and *hard* scenarios, for 2- and 3-source mixtures, are presented in Table 3. This analysis is performed on the $P = 8$-filter ASoBO system trained on AISHELL-4. The *random* row corresponds to the random selection of the filters. For $L = 2$ sources, the system tends to select the closest direction to the speaker. This is shown by the 83.6% and 71.7% F1-score in the *easy* and *hard* scenarios, respectively. Note that the localization score is strongly degraded when the source is not aligned with the speaker. For $L = 3$, the system still localizes the sources accurately. The *easy* scenario shows a 75.9% F1-score. This score degrades–but remains higher than random selection–in the *hard* case, with 66.6%.

This study shows how the self-attention module can select the filter direction associated with the active speaker. The degradation in the *hard* scenario can explain the diarization performance limitation of our system on the AMI dataset. If the active speakers are not aligned with the filters, the self-attention might extract the features from the wrong spatial filter.

7. Conclusions

In this paper, we propose a multi-microphone segmentation algorithm for distant speaker diarization. This method consists of a set of beamformers steered in fixed directions whose outputs are selected with a self-attention module. The output of this system serves as a feature sequence for a joint Voice Activity (VAD) and Overlapped Speech Detection (OSD) system. This segmentation is used for speaker diarization with the VBx system. Experiments on the AMI and AISHELL-4 datasets have shown that the proposed approach improves the speaker diarization under distant conditions, with mitigated gain on AMI but significant improvement on AISHELL-4. The analysis of the weights of the self-attention model shows that we can perform pseudo-localization of the active speakers. This demonstrates how explainable such a system can be. In future work, we plan to evaluate the impact of mismatched array setup on the speaker diarization performance and to estimate the Direction of Arrival at inference time.

8. Ackowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101007666. This work was performed using HPC resources from GENCI–IDRIS (Grant 2022-AD011012565).

9. References

[1] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma *et al.*, "M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6167–6171.

[2] 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," *Computer Speech & Language*, vol. 72, p. 101317, 2022.

[3] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 296–303.

[4] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. Garcia, "Encoder-decoder based attractors for end-to-end neural diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1493–1507, 2022.

[5] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote.Audio: Neural Building Blocks for Speaker Diarization," in *ICASSP*, 2020, pp. 7124–7128.

[6] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks," *Computer Speech & Language*, vol. 71, p. 101254, 2022.

[7] G. Gelly and J.-L. Gauvain, "Optimization of rnn-based speech activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 646–656, 2017.

[8] M. Lavechin, M.-P. Gill, R. Bousbib, H. Bredin, and L. P. Garcia-Perera, "End-to-End Domain-Adversarial Voice Activity Detection," in *Proc. Interspeech 2020*, 2020, pp. 3685–3689.

[9] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-Aware Diarization: Resegmentation Using Neural End-to-End Overlapped Speech Detection," in *ICASSP*, 2020, pp. 7114–7118.

[10] M. Lebourdais, M. Tahon, A. LAURENT, and S. Meignier, "Overlapped speech and gender detection with WavLM pre-trained features," in *Proc. Interspeech 2022*, 2022, pp. 5010–5014.

[11] R. Yin, H. Bredin, and C. Barras, "Speaker change detection in broadcast tv using bidirectional long short-term memory networks," in *Interspeech 2017*. ISCA, 2017.

[12] M. Wölfel and J. McDonough, *Distant speech recognition*. John Wiley & Sons, 2009.

[13] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[14] A. Vinnikov, A. Ivry, A. Hurvitz, I. Abramovski, S. Koubi, I. Gurvich, S. Peer, X. Xiao, B. M. Elizalde, N. Kanda *et al.*, "Notsofar-1 challenge: New datasets, baseline, and tasks for distant meeting transcription," *arXiv preprint arXiv:2401.08887*, 2024.

[15] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers 2*. Springer, 2006, pp. 28–39.

[16] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, X. Xu, J. Du, and J. Chen, "AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario," in *Proc. Interspeech 2021*, 2021, pp. 3665–3669.

[17] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[18] J. Benesty, J. Chen, and I. Cohen, *Design of Circular Differential Microphone Arrays*, ser. Springer Topics in Signal Processing. Cham: Springer International Publishing, 2015, vol. 12.

[19] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[20] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

[21] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5325–5329.

[22] S. Cornell, M. Pariente, F. Grondin, and S. Squartini, "Learning filterbanks for end-to-end acoustic beamforming," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6507–6511.

[23] R. Gong, C. Quillen, D. Sharma, A. Goderre, J. Laínez, and L. Milanović, "Self-Attention Channel Combinator Frontend for End-to-End Multichannel Far-Field Speech Recognition," in *Interspeech*, 2021, pp. 3840–3844.

[24] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 2007, pp. 683–686.

[25] F. Landini, O. Glembek, P. Matějka, J. Rohdin, L. Burget, M. Diez, and A. Silnova, "Analysis of the but diarization system for voxconverse challenge," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5819–5823.

[26] T. Mariotte, A. Larcher, S. Montrésor, and J.-H. Thomas, "Channel-combination algorithms for robust distant voice activity and overlapped speech detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–14, 2024.

[27] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Overlapped speech detection and speaker counting using distant microphone arrays," *Computer Speech & Language*, vol. 72, p. 101306, 2022. [Online]. Available: https://doi.org/10.1016/j.csl.2021.101306

[28] D. Raj, D. Povey, and S. Khudanpur, "Gpu-accelerated guided source separation for meeting transcription," *arXiv preprint arXiv:2212.05271*, 2022.

[29] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine learning research*, vol. 7, pp. 1–30, 2006.

[30] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," 2020.

[31] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with GPU acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.