Multi-microphone Automatic Speech Segmentation in Meetings Based on Circular Harmonics Features

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

Multi-microphone Automatic Speech Segmentation in Meetings Based on Circular Harmonics Features

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

[1]LAUM UMR CNRS 6613 IAGS, Le Mans Université, France
[2]LIUM, Le Mans Université, France

theo.mariotte@univ-lemans.fr, anthony.larcher@univ-lemans.fr

Abstract

Speaker diarization is the task of answering *Who spoke and when?* in an audio stream. Pipeline systems rely on speech segmentation to extract speakers' segments and achieve robust speaker diarization. This paper proposes a common framework to solve three segmentation tasks in the distant speech scenario: Voice Activity Detection (VAD), Overlapped Speech Detection (OSD), and Speaker Change Detection (SCD). In the literature, a few studies investigate the multi-microphone distant speech scenario. In this work, we propose a new set of spatial features based on direction-of-arrival estimations in the circular harmonic domain (CH-DOA). These spatial features are extracted from multi-microphone audio data and combined with standard acoustic features. Experiments on the AMI meeting corpus show that CH-DOA can improve the segmentation while being robust in case of deactivated microphones.

Index Terms: speech segmentation, multi-microphone, speaker diarization

1. Introduction

Speaker diarization is the task of answering *Who spoke and when?* in an audio stream [1, 2]. Many speaker diarization approaches are based on pipeline architectures [2, 3, 4]. Those approaches rely on a speech segmentation step that extracts speaker-homogeneous segments. Speaker clustering is then performed by extracting and grouping speaker embeddings via clustering algorithms [2]. This paper focuses on automatic speech segmentation, which can be divided into three sub-tasks: Voice Activity Detection (VAD), Overlapped Speech Detection (OSD), and Speaker Change Detection (SCD).

VAD detects speech segments in the audio signal. It is the first step in most speaker diarization pipelines [2]. Finally, since overlapping speech is one of the major sources of errors in speaker diarization pipelines [5, 6], OSD is required. It consists in detecting speech segments in which multiple speakers are simultaneously active. SCD is also required to detect speaker turns in the audio signal, i.e., when the currently active speaker is changing.

Early studies on VAD [7, 8], OSD [9, 10, 11] and SCD [12, 13] are based on the statistical modeling of acoustic features. The latter is originally solved by comparing the statistics of two adjacent segments.

Statistical models have then been replaced by neural networks due to their strong modeling capacities. VAD [5, 14, 15], OSD [5, 16, 17], and SCD [18, 19, 20] can be solved by modeling a sequence of acoustic features and performing a frame-level binary classification. SCD is also tackled as the regression of functions in which maxima are located at turn locations [20].

Most VAD, OSD, and SCD studies are conducted on single-channel data. In the meeting context, recording signals with a distant device offers practical benefits since it does not require participants to carry an individual microphone. Microphone arrays are commonly used as distant devices to capture additional spatial information. Few studies have been conducted on multi-microphone speech segmentation [17, 21, 22, 23]. In particular, Cornell *et al.* [17] explore the use of Interaural Phase Difference (IPD) spatial features for joint VAD and OSD and report a noticeable performance gain. Hu *et al.* [21] investigate the use of Time Difference of Arrival (TDOA) features to detect speaker changes. Although the authors show diarization performance gain, experiments were only conducted on simulated data. To the best of our knowledge, no other work has been reported on the use of spatial features for distant SCD.

Several spatial features have been investigated in various multi-microphone speech processing tasks [24, 25, 26, 27]. In particular, SongGong *et al.* [28] propose a speaker localization method based on circular harmonics (CH). Although these features require the use of a circular array, they depend little on the number of available microphones. Hence, circular harmonics are an interesting framework for feature extraction to rely less on the array configuration.

In this paper, we tackle VAD, OSD, and SCD tasks with the same architecture. We propose the use of CH to extract spatial features for multi-microphone speech segmentation. This choice is motivated by the common use of circular microphone arrays to capture distant speech in meetings [29, 30]. The proposed spatial features consist of direction-of-arrival estimation in the CH framework (CH-DOA). Spatial features are combined with commonly used acoustic features to solve segmentation tasks. As far as authors are aware, this is the first investigation on the use of CH features for distant speech segmentation. Furthermore, we report the impact of IPD spatial features for SCD since no work has been found considering their use for this task. We demonstrate that adding spatial information drastically improves the detection of speaker turns. Finally, we present encouraging results of CH-DOA-based OSD and SCD systems under mismatched array conditions. The code will be available soon in a large-scale diarization toolkit[1].

The paper is organized as follows. Sect. 2 presents VAD, OSD, and SCD tasks. Sect. 3 presents the CH-DOA feature extraction. Sect. 4 introduces the speech segmentation model along with the dataset and the experimental protocol before presenting results in Sect. 5 and conclusions in Section 6.

2. Segmentation Tasks

This section describes the segmentation tasks considered. The labeling procedure for each task is presented in figure 1.

[1]https://git-lium.univ-lemans.fr/speaker

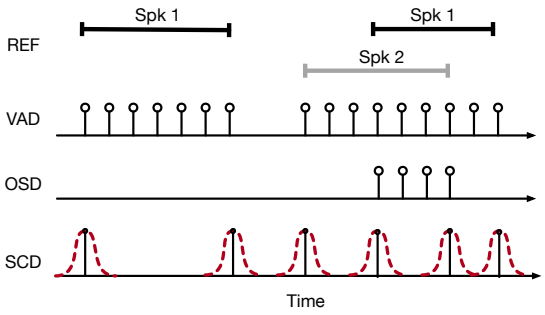

Figure 1: *VAD, OSD, and SCD labels along with the reference diarization. VAD and OSD are formulated as a binary frame classification task while SCD is solved as the regression of Gaussian-like functions.*

2.1. Voice Activity and Overlapped Speech Detection

VAD is formulated as a framewise binary sequence classification task [3]. Let $\boldsymbol{X} = [\boldsymbol{X}_0, \dots, \boldsymbol{X}_t, \dots, \boldsymbol{X}_{T-1}]$ be a sequence of features extracted from the audio signal with T being the length of the sequence. The VAD task aims at predicting the sequence $\boldsymbol{P} = [\boldsymbol{P}_0, \dots, \boldsymbol{P}_t, \dots, \boldsymbol{P}_{T-1}]$, where $\boldsymbol{P}_t = \{p(N_{spk} = 0|\boldsymbol{X}_t), p(N_{spk} \geq 1|\boldsymbol{X}_t)\}$ is the pseudo-probability of the t-th frame to belong to each class. N_{spk} denotes the number of active speakers in the current frame.

OSD formulation is similar to VAD. However, positive labels correspond to frames containing strictly more than one active speaker: $\boldsymbol{P}_t = \{p(N_{spk} \leq 1|\boldsymbol{X}_t), p(N_{spk} \geq 2|\boldsymbol{X}_t)\}$.

2.2. Speaker Change Detection

In this paper, SCD is formulated as a regression task [20]. This approach consists in estimating functions in which maxima are located at speaker change points. Speaker changes are represented as Gaussian-like functions[2] following the binary label encoding of [31]. In this work, the speech-to-non-speech transition is considered as a speaker turn (Fig. 1).

3. Circular Harmonics Features

Circular harmonics (CH) are a set of elementary 2-d functions. They are similar to cylindrical harmonics or spherical harmonics in the 3-d domain [32]. An acoustic signal can be represented as a weighted sum of CH components. This section presents this formulation and introduces the CH-DOA features used for speech segmentation.

3.1. Circular harmonics framework

Let us consider a uniform circular microphone array (UCA) composed of M microphones with a radius r. The $X_m(f,t)$ signal captured by the m-th microphone can be expressed in the short-time Fourier transform (STFT) domain as a weighted sum of circular harmonics [33]:

$$X_m(f,t) = \sum_{n=-\infty}^{\infty} C_n(f,t) e^{jn\phi}, \tag{1}$$

[2]The variance σ^2 of the Gaussian functions used as labels is randomized during training. It follows a uniform distribution $\mathcal{U}$ such as $\sigma^2 \sim \mathcal{U}_{[2,7]}$.

where f denotes a frequency bin and t the frame index. In equation (1), $j = \sqrt{-1}$, ϕ is the DOA of the sound source, $e^{jn\phi}$ is the n-th order CH and $C_n(f,t)$ the associated coefficient. By using a circular microphone array, the sound field is sampled at some discrete locations. CH coefficients are estimated as follows [28, 33]:

$$\tilde{C}_n(f,t) = \frac{1}{M} \sum_{m=1}^{M} X_m(f,t) e^{-jn\psi_m}, \tag{2}$$

where $\tilde{C}_n(f,t)$ is the estimated CH coefficient and $\psi_m = (m-1)\frac{2\pi}{M}$ denotes the angle of the m-th microphone.

3.2. CH-DOA feature extraction

Spatial filtering, i.e. beamforming, can be performed in the CH domain [33]. This is also known as *modal beamforming* and can be expressed as follows [28]:

$$B_n(f,t) = \sum_{-N}^{N} \frac{\tilde{C}_n(f,t)}{j^n J_n(kr)} e^{jn\theta}, \tag{3}$$

with $B_n(k,t)$ being the n-th order beamformed signal and $k = 2\pi f/c$ the wave number with c the speed of sound. $J_n(kr)$ is the n-th order Bessel function of the first kind and θ indicates the steering direction.

The Pseudo-Intensity Vector (PIV) uses only zero- and first-order beamformers. The zero-order beam is obtained from equation (3) by setting $N = 0$:

$$B_0(f,t) = \frac{\tilde{C}_0(f,t)}{J_0(kr)}. \tag{4}$$

For $N = 1$, two orthogonal beams can be defined oriented towards the $\theta_x = 0$ and $\theta_y = \pi/2$ respectively. The beam $B_{1x}(f,t)$ (respectively B_{1y} with θ_y) is expressed following:

$$B_{1x}(f,t) = \sum_{-1}^{1} \frac{\tilde{C}_n(f,t)}{j^n J_n(kr)} e^{jn\theta_x}. \tag{5}$$

Then, the PIV components I_x and I_y can be calculated as:

$$\begin{bmatrix} I_x(f,t) \\ I_y(f,t) \end{bmatrix} = \frac{1}{2}\Re\left\{ B_0^*(f,t) \begin{bmatrix} B_{1x}(f,t) \\ B_{1y}(f,t) \end{bmatrix} \right\} \tag{6}$$

where $\Re$ denotes the real part and * the complex conjugate.

The PIV is supposed to be oriented in the propagation direction of the impinging acoustic wave. Thus, the angular direction of the PIV in the frame of reference of the microphone corresponds to the source DOA [28]:

$$\hat{\phi}(f,t) = \arctan\left\{ \frac{I_y(f,t)}{I_x(f,t)} \right\}. \tag{7}$$

The estimated DOA $\hat{\phi}$ is used as a spatial feature for speech segmentation and is denoted as CH-DOA. The following section presents how these features are integrated into our segmentation systems. The CH-DOA features offer a similar computational complexity as IPD/CSIPD since it only relies on the multi-microphone STFT without any additional loop.

4. Experimental protocol

This section presents the experimental protocol to evaluate the impact of CH-DOA spatial features on multi-microphone speech segmentation.

4.1. Dataset

Experiments are conducted on the AMI meeting corpus [29]. This dataset is about 100h of speech acquired during realistic meetings. The majority of participants are non-native English speakers and were asked to conduct a design project. Speech can be either spontaneous or scripted depending on the session. Meetings have been recorded using various devices. Experiments are conducted on the AMI *Array 1* data, which is an 8-microphone circular array placed in the center of the table. Training, development, and evaluation partitions follow the protocol proposed in [4]. Labels for VAD, OSD, and SCD are extracted from the manual annotation of the segments. Speech signals are sampled at 16kHz.

4.2. Segmentation architecture

The segmentation architecture – figure 2 – is composed of the following modules. The acoustic feature module extracts a representation $\boldsymbol{A} \in \mathbb{R}^{F_a \times T}$ from the multi-microphone signal with F_a being the acoustic feature size. The spatial feature module extracts a representation $\boldsymbol{S} \in \mathbb{R}^{F_s \times T}$ from the same signal with F_s being the spatial feature size. Both kinds of features are concatenated on the first dimension to produce a F-long feature vector. The feature sequence is fed to the sequence modeling network which outputs the prediction $\boldsymbol{P} \in \mathbb{R}^{C \times T}$ with C being the output size.

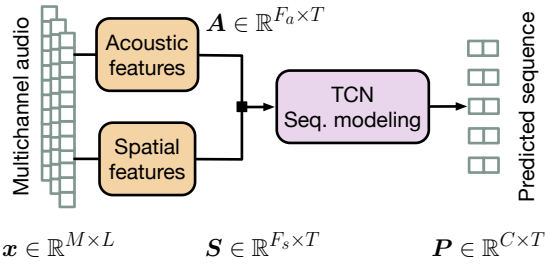

Figure 2: *Segmentation architecture used for VAD, OSD, and SCD. Acoustic and spatial features are extracted from the multi-microphone signal.* ■ *denotes the concatenation operation. The model has two outputs ($C = 2$) for classification and one output ($C = 1$) for regression.*

4.2.1. Acoustic features

Acoustic features are extracted from the signal captured by the first channel of the microphone array. Mel Frequency Cepstral Coefficients (MFCC) and Log-Mel spectrogram are used as acoustic features following [17]. Both types of features are extracted on a 25ms sliding window with a 10ms shift. The mel spectrogram is obtained using 80 mel-scale filters before conversion to the log scale. Time-frequency masking is applied as data augmentation during training. This results in a vector of $F_a = 80$ features.

20 MFCC are computed from a mel-spectrogram extracted using 40 triangular filters since it leads to the best performance. The first MFCC coefficient (energy) is removed and both Δ and $\Delta\Delta$ are computed. This results in a vector of $F_a = 59$ features.

4.2.2. Spatial features

The proposed CH-DOA features are computed for each frequency bin of the STFT which results in a vector of $F_s = 257$ features. To ensure alignment with the acoustic features, spatial features are extracted on 25ms sliding windows with a 10ms shift. Interaural Phase Difference (IPD) and cosine and sine IPD (CSIPD) are considered as baseline spatial features following [17]. In this work, we consider 4 microphone pairs with the microphones in opposition. This results in a vector of $F_s = 1028$ and $F_s = 2056$ features respectively.

4.2.3. Sequence modeling

Sequence modeling is performed using the Temporal Convolutional Network[3]. It consists of 1-d convolutional layers with exponentially increasing dilatation to learn a large temporal context. Our architecture is composed of 3 stacked TCN blocks of 5 convolutional layers. A residual connection is added after each TCN block. Before feeding the sequence of feature vectors to the TCN blocks, a 1-d bottleneck convolution compresses the feature sequence from F-dimensional vectors to 128 dimensions. Prediction is obtained by adding a 1-d convolution followed by a softmax activation function for the classification tasks or a linear activation in the case of regression (SCD).

4.3. Training and evaluation procedures

VAD, OSD, and SCD systems are trained on the AMI train subset. The classification models – VAD and OSD – are optimized using cross-entropy loss. SCD models are optimized with the Mean Squared Error training objective. Model weights are updated using the ADAM optimizer with a learning rate set to $lr = 0.001$. Each system is trained on 2s audio segments randomly sampled from the training set. The batch size is set to 64. Overlaps augmentation [16] is applied to 50% of the training segments. Models are trained on Nvidia RTX6000 GPU cards.

Models are evaluated on the AMI evaluation set. Inference is performed on 2s segments with a 0.5s shift. The VAD, OSD, and SCD detection thresholds are tuned on the development set. VAD is evaluated in terms of False Alarm rate (FA) and missed detection (Miss). OSD is evaluated using F1-score and average precision (AP) [22]. Finally, SCD is evaluated using purity (P) and coverage (C) as suggested in [18, 19]. The F1-score, i.e. the harmonic mean, of purity and coverage is also reported and is denoted as Segmentation Error (SER). We also report the 95% confidence interval calculated on the file-level performance for each metric.

5. Experimental study

This section presents the experimental results obtained with CH-DOA features on VAD, OSD, and SCD. The performance on each task is presented in table 1.

5.1. VAD performance

Considering only MFCC shows robust VAD performance with 3.5% Miss and 3.0% FA. Adding IPD or CSIPD slightly degrades the performance as shown by the imbalanced Miss and FA. The proposed CH-DOA shows similar performance as the MFCC with 3.4% FA and 3.1% Miss. VAD performance is similar when considering Log-Mel features. Using Log-Mel features offers 3.2% FA and 3.5% Miss. Again, IPD and CSIPD fail at improving the VAD performance. CH-DOA features show a similar performance as the Log-Mel with 3.2% FA and 3.3% Miss but without improvement. Since the performance is the same between MFCC and CH-DOA, it seems the model

[3] `https://github.com/popcornell/OSDC`

Table 1: *Performance of VAD, OSD, and SCD systems on the AMI meeting corpus with each type of features. The number of parameters is given for the VAD and OSD systems. The number of parameters of the regression model–SCD–is slightly lower. Bold values indicate the best-performing model for each type of acoustic feature. MFCC and Log-Mel correspond to single-channel models.*

		VAD		OSD		SCD		
	# param.	$\text{FA}_{\%\downarrow}$	$\text{Miss}_{\%\downarrow}$	$\text{F1-score}_{\%\uparrow}$	$\text{AP}_{\%\uparrow}$	$P_{\%\uparrow}$	$C_{\%\uparrow}$	$SER_{\%\uparrow}$
MFCC	0.26M	**3.5**±0.5	**3.0**±1.3	64.5±5.6	65.1±7.2	82.2±2.2	79.2±2.6	80.7±0.9
+ IPD	0.33M	4.2±0.8	**3.0**±1.2	60.7±5.5	62.5±7.5	74.8±2.8	76.4±2.9	75.6±0.7
+ CSIPD	0.40M	**3.0**±0.3	4.1±1.6	**71.7**±4.5	**75.9**±5.5	81.9±1.3	**85.9**±2.7	**83.9**±1.2
+ CH-DOA	0.28M	**3.4**±0.4	**3.1**±1.4	**69.3**±4.6	**73.0**±5.8	**84.6**±1.6	**84.3**±3.4	**84.4**±1.4
Log-Mel	0.27M	**3.2**±0.4	**3.5**±1.4	66.1±5.9	68.9±7.9	83.9±1.7	80.4±2.3	82.1±1.2
+ IPD	0.33M	**3.0**±0.4	4.3±1.3	65.2±5.9	65.9±7.2	79.5±3.0	76.8±4.2	78.1±1.3
+ CSIPD	0.40M	3.7±0.4	**3.1**±1.3	**73.4**±5.3	**75,6**±6.1	85.5±1.8	**83.9**±3.9	**84.7**±1.6
+ CH-DOA	0.28M	**3.2**±0.4	**3.3**±1.3	67.3±5.5	68.3±7.1	**87.2**±1.4	**82.5**±3.4	**84.8**±1.5

does not use spatial information for VAD. Other information fusion schemes could be investigated instead of feature concatenation.

5.2. OSD performance

Results show that adding IPD features (62.5% AP) degrades OSD with regard to MFCC features (65.1%). This degradation can be seen in both F1-score and AP metrics. Adding CSIPD features (75.9% AP) significantly improves OSD performance with a +10.8% absolute AP gain compared to MFCC. Then, the proposed CH-DOA feature (73.0% AP) reaches a similar performance as the CSIPD on both F1-score and AP with a little degradation. This model, however, has only 0.28M parameters to optimize against 0.40M for the CSIPD one.

Models trained with Log-Mel features behave similarly to the MFCC with a global performance gain, except with CH-DOA features. IPD features (65.9% AP) degrade the OSD performance with regard to Log-Mel (68.9% AP). Again, CSIPD (75.6% AP) offers the best OSD performance with a +6.7% AP and a +7.3% F1-score absolute improvements with respect to Log-Mel. In this configuration, CH-DOA features offer mitigated OSD performance (68.3% AP) without improving nor degrading the detection.

5.3. SCD performance

SCD models behave similarly to OSD models. When MFCC features are considered, IPD degrades the SCD performance (75.6% SER) with respect to MFCC (80.7% SER). Considering CSIPD features (83.9% SER) significantly improves SCD performance with a +3.2% absolute SER gain. CH-DOA (84.4% SER) reaches a similar performance, reaching a +3.7% absolute SER gain. This system also shows the best balance between purity and coverage.

Considering Log-Mel features with IPD degrades the detection by an absolute -4.0% SER while both CSIPD and CH-DOA improve SER by +2.6% and +2.7% respectively. The proposed CH-DOA features offer similar performance as CSIPD while reducing the number of trainable parameters. Moreover, this model is not constrained to the array configuration used in the training data as shown in the following section.

5.4. Robustness to the number of microphones

CH-DOA is based on zero- and first-order circular harmonics. Hence, the feature extraction is not supposed to rely on the number of available microphones in the UCA. This sub-section evaluates the two best-performing MFCC-based OSD and SCD models by desactivating 4 channels in the evaluation data. Performance on OSD and SCD is presented in table 2. Results on the OSD task show that CH-DOA features are more robust to a mismatch in the microphone number, reaching a 71.4% AP. This system remains better than MFCC (65.1% AP) with an absolute +6.3% AP improvement. CSIPD features are less robust to array mismatch with a 51.5% AP.

On SCD, the CH-DOA model shows the best performance on both purity (84.1%) and coverage (83.2%) while still improving single-channel MFCC (82.2%/79.2%). CSIPD features degrade the performance with $M = 4$, mostly on coverage (75.8%).

Table 2: *OSD and SCD performance on AMI array 1 evaluation data with $M = 4$ deactivated channels.*

	OSD		SCD	
$M = 4$	$\text{F1-score}_{\%\uparrow}$	$\text{AP}_{\%\uparrow}$	$P_{\%\uparrow}$	$C_{\%\uparrow}$
MFCC	64.5±5.6	65.1±7.2	82.2±2.2	79.2±2.6
+ CSIPD	55.4±6.7	51.5±7.9	81.1±1.7	75.8± 3.0
+ CH-DOA	**69,6**±5.3	**71.4**±6.3	**84.1**±1.7	**83.2**±3.4

6. Conclusions

This paper introduces a new set of spatial features based on direction-of-arrival (DOA) estimation in the circular harmonics (CH) domain. CH-DOA is investigated on three automatic speech segmentation tasks: Voice Activity Detection (VAD), Overlapped Speech Detection (OSD), and Speaker Change Detection (SCD). The proposed CH-DOA is compared with state-of-the-art spatial features and combined with commonly used acoustic features. Although limited to circular arrays, CH-DOA shows better segmentation performance than single-channel acoustic features, particularly on OSD and SCD. Furthermore, we demonstrate that adding spatial features significantly improves SCD and reach the best performance with CH-DOA (84.8% SER). Finally, CH-DOA shows encouraging robustness to array mismatch by still improving SCD and OSD under these conditions.

The use of information fusion schemes (e.g. cross attention) will be investigated in future work since spatial information seems less exploited on the VAD. The segmentation models remain to be evaluated in a full diarization pipeline.

7. References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 2, pp. 356–370, 2012.

[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] 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.

[4] 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.

[5] H. Bredin and A. Laurent, "End-To-End Speaker Segmentation for Overlap-Aware Resegmentation," in *Proc. Interspeech 2021*, 2021, pp. 3111–3115.

[6] P. García, J. Villalba, H. Bredin, J. Du, D. Castan, A. Cristia, L. Bullock, L. Guo, K. Okabe, P. S. Nidadavolu *et al.*, "Speaker detection in the wild: Lessons learned from jsalt 2019," *arXiv preprint arXiv:1912.00938*, 2019.

[7] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing letters*, vol. 6, no. 1, pp. 1–3, 1999.

[8] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Veselỳ, and P. Matějka, "Developing a speech activity detection system for the darpa rats program," in *Thirteenth annual conference of the international speech communication association*, 2012.

[9] K. Boakye, O. Vinyals, and G. Friedland, "Improved overlapped speech handling for speaker diarization," in *Interspeech*, 2011, pp. 941–944.

[10] D. Charlet, C. Barras, and J.-S. Lienard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates," in *ICASSP*, 2013, pp. 7707–7711.

[11] S. H. Yella and H. Bourlard, "Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1688–1700, 2014.

[12] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA speech recognition workshop*, vol. 1997, 1997.

[13] D. Liu and F. Kubala, "Fast speaker change detection for broadcast news transcription and indexing," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[14] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks." in *INTERSPEECH*. Lyon, France, 2013, pp. 728–731.

[15] 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.

[16] 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.

[17] 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

[18] H. Bredin, "Tristounet: triplet loss for speaker turn embedding," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 5430–5434.

[19] 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.

[20] M. Hrúz and Z. Zajíc, "Convolutional neural network for speaker change detection in telephone speaker diarization system," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4945–4949.

[21] M. Hu, D. Sharma, S. Doclo, M. Brookes, and P. A. Naylor, "Speaker change detection and speaker diarization using spatial information," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5743–5747.

[22] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Detecting and Counting Overlapping Speakers in Distant Speech Scenarios," in *Interspeech*, 2020, pp. 3107–3111.

[23] T. Mariotte, A. Larcher, S. Montrésor, and J.-H. Thomas, "Microphone Array Channel Combination Algorithms for Overlapped Speech Detection," in *Proc. Interspeech 2022*, 2022, pp. 4636–4640.

[24] J. Pak and J. W. Shin, "Sound localization based on phase difference enhancement using deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1335–1345, 2019.

[25] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.

[26] R. Gu, S.-X. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, and D. Yu, "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7319–7323.

[27] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "High-resolution speaker counting in reverberant rooms using crnn with ambisonics features," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 71–75.

[28] K. SongGong and H. Chen, "Robust Indoor Speaker Localization in the Circular Harmonic Domain," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 4, pp. 3413–3422, Apr. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9037178/

[29] 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.

[30] 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.

[31] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 74–79.

[32] E. G. Williams and J. A. Mann III, "Fourier acoustics: sound radiation and nearfield acoustical holography," 2000.

[33] A. M. Torres, J. Mateo, and M. Cobos, "Room Acoustics Analysis Using Circular Arrays: A Comparison Between Plane-Wave Decomposition and Modal Beamforming Approaches," *Circuits, Systems, and Signal Processing*, vol. 35, no. 5, pp. 1625–1642, May 2016.