INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Overlapped Speech Detection based on Spectral and Spatial Feature Fusion

Weiguang Chen[1], *Van Tung Pham*[2], *Eng Siong Chng*[2], *Xionghu Zhong*[1]

[1]College of Computer Science and Electronic Engineering, Hunan University, China
[2]School of Computer Engineering, Nanyang Techonological University, Singapore
cwg@hnu.edu.cn, vtpham@ntu.edu.sg, ASESChng@ntu.edu.sg, xzhong@hnu.edu.cn

Abstract

Overlapped speech is widely present in conversations and can cause significant performance degradation on speech processing such as diarization, enhancement, and recognition. Detection of overlapped speech, in particular when the speakers are in the far-field, is a challenging task as the overlapped part is usually short, and heavy reverberation and noise may present in the conversation scenario. Existing solutions overwhelmingly rely on spectral features extracted from single microphone signal to perform the detection. In this paper, we propose a novel detection approach which is able to use a microphone array and fuse the spatial and spectral features extracted from multi-channel array signal. Two categories of spatial features, directional statistics which are projected to spherical location grids and generalized cross-correlation function based on phase transform (GCC-PHAT), are considered to model the speaker's spatial characteristic. Such spatial features are then fused with the spectral features to detect the overlapped speech by using a Gated Multimodal Unit (GMU). The performance of the proposed approach is studied under AMI and CHiME-6 corpora. Experimental results show that the proposed feature fusion approach achieves better performance than methods using spectral features only.

Index Terms: overlapped speech detection, spatial features, spectral features, microphone array, directional statistics

1. Introduction

In a real conversation, different speakers can speak at the same time over a short period frequently. As such, overlapped speech may present in the recorded signal and can cause significant performance degradation or even system failure on speech processing such as speaker localization [1] and diarization [2, 3], speaker identification [4], and automatic speech recognition (ASR) [5, 6]. Different from Voice Activity Detection (VAD) which only detects the presence of speech signal, overlapped speech detection (OSD) needs to differentiate the signal from single speaker or multiple speakers. In [7], a subset of the AMI meeting corpus is studied and showed that the length of overlapped speech segments is mostly between 0.5s and 1.5s, which is short and makes OSD more challenging than VAD. In addition, open datasets are rarely specially designed for OSD task and accurately labelled datasets are not always available [8].

In recent years, various approaches have been developed for OSD to improve the speaker diarization and ASR performance. Usually, speech features such as spectral and signal envelope are extracted from the single channel speech signal. A deep neural network (DNN) is then applied to train the detection system. In [9], a combination of Mel-frequency cepstral coefficients (MFCC) and speech envelope is employed, and a DNN based OSD method is developed. In [10], a convolutional neural network (CNN) based model is introduced to jointly solve OSD and speaker counting problems. The authors also demonstrate that supervised methods are able to perform better than human subjective recognition on these problems. In [2], a long short-term memory (LSTM) based model is proposed to predict the presence of overlapped speech by using the spectral features; force alignment is also applied to correct the labels of overlapped speech by using a pre-trained ASR model. A Temporal Convolutional Network (TCN) architecture is designed for OSD and speaker counting in [11], in which computational complexity is also studied and it is shown that TCN is more computationally efficient than LSTM and Convolutional Recurrent Neural Networks (CRNN) [12]. In [13], a CNN based module is introduced to extract features directly from the mixed signals to perform end-to-end OSD and speaker counting. It is worth mentioning that the results of these works are obtained from single channel signal and mostly artificially mixed speech signals.

Although significant improvement of OSD performance has been achieved by using different speech features and various DNN based methods are proposed, OSD of far-field speech signal remains a challenging task as heavy reverberation and noise may present in the received signal. In [2, 11], OSD using microphone array recordings is studied. However, the authors simply average the predictions across different channels rather than using the spatial characteristics captured by the microphone array. In this work, we propose an array signal based detection approach by fusing the spatial features and the spectral features extracted from the multi-channel signal. Two categories of spatial features obtained from GCC-PHAT and directional statistics are considered to characterize the spatial information of different speakers. The GCC-PHAT coefficients of all microphone pairs are computed and concatenated to construct the first type of spatial features. The second one is obtained by projecting directional statistics into spherical location grids using a learnable kernel matrix. A GMU module [14] is then used to fuse the extracted spatial and spectral features, followed by a bi-directional LSTMs (BLSTM) to perform the OSD task. The performance of the proposed approach is studied under AMI [15] and CHiME-6 [16] datasets. Experimental results show that compared to using spectral features only, much better performance can be achieved by the proposed fusion approach. In addition, the proposed directional statistics is superior to the GCC-PHAT based spatial features.

2. Overlapped speech detection Method

The task of OSD is to identify speech segments where two or more speakers are active simultaneously. It can be formulated as a binary classification problem, i.e., given a sequence of feature vectors and corresponding labels of overlapped segments, we can train a model such that the overlapped speech can be identified based on the probability of the model output. Assume that an M-channel microphone array

http://dx.doi.org/10.21437/Interspeech.2021-2138

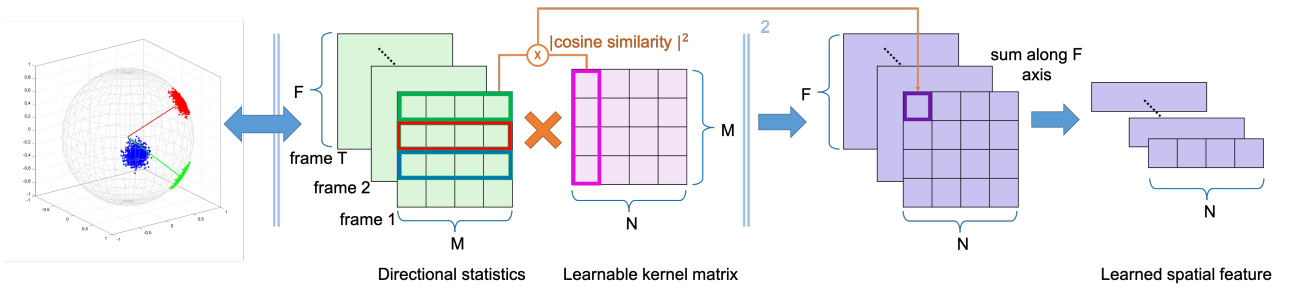

Figure 1: *Spatial features learned from the directional statistics. The directional statistics are distributed on the unit hypersphere and each cluster corresponds to a source location. Speaker spatial characteristics are obtained by projecting directional statistics to N spherical grids. F is the number of frequency bins and M denotes the number of microphones.*

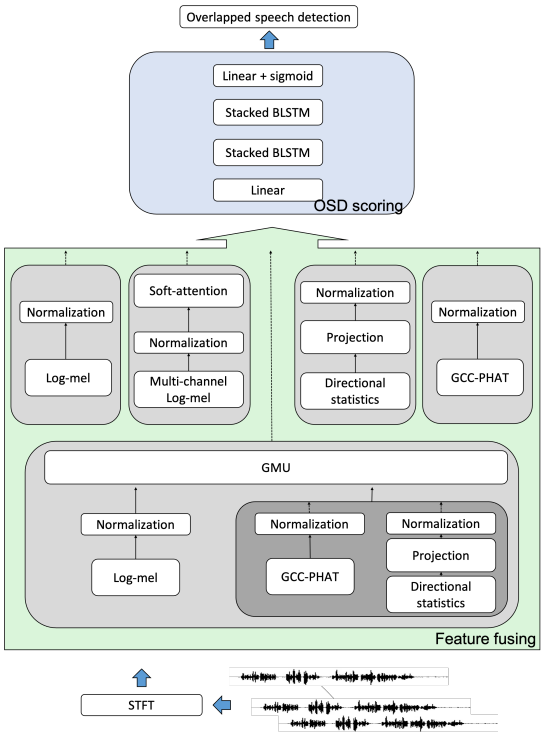

Figure 2: *Diagram of the model architecture for OSD. The upper part with light blue color is the scoring component used to detect overlapped speech segments. The bottom part with light green color denotes the feature fusing component and the dotted lines represents various optional input features.*

is employed, the received signal can be formed as $\boldsymbol{x}(t,f) = [x_1(t,f), x_2(t,f), ..., x_M(t,f)]^T$ in short-time Fourier transform (STFT) domain, with t and f denoting the time frame and frequency bin respectively. Log-mel filter banks are usually computed from the STFT spectrogram and used as spectral features which is also an input of the DNN. Following [11], 80 log-mel filter banks are extracted for each channel at time t, and corresponding coefficients are written as a vector $\boldsymbol{x}_m(t)$. In the next section, the spatial features and the fusion architecture will be introduced.

2.1. Spatial features

The phase difference of signals between two different channels can be employed to characterize the location information of the speaker, and consequently identify the overlapped speech.

GCC-PHAT based spatial features. The GCC-PHAT function between any two microphones can be computed as [17]

$$g_{ij}(t,\tau) = \sum_f \mathcal{R}\left(\frac{x_i(t,f)x_j(t,f)^*}{|x_i(t,f)x_j(t,f)^*|}e^{j2\pi f\tau}\right), \quad (1)$$

where i and j denote the i-th and j-th microphone respectively, and $\mathcal{R}(\cdot)$ is the real part of a complex number, and τ is the time-delay, and $(\cdot)^*$ and $|\cdot|$ are the conjugation and amplitude of a complex number respectively. Following [18], 51 correlation coefficients around τ=0 are employed to represent the spatial features in this work. However, GCC-PHAT features are sensitive to noise and reverberation [19] and it suffers from the resolution problem when the speakers are closely located [20].

Spatial features extracted from directional statistics. In [19, 21], the normalized spectral vector, referred to as directional statistics, is demonstrated to be robust in noise and reverberant environments for blind source separation and diarization. The directional statistics can be defined as [19]

$$\boldsymbol{y}(t,f) = \frac{\boldsymbol{x}(t,f)}{\|\boldsymbol{x}(t,f)\|}. \quad (2)$$

where $\|\cdot\|$ denotes the Euclidean norm. Assume that the speaker location hypersphere is divided into N grids where N is a predefined number, and let $\boldsymbol{w}_y(n) \in \mathcal{C}^{M\times 1}$ be a complex vector which is able to map the directional statistics $\boldsymbol{y}(t,f)$ to the n-th location grid, as shown in Fig. 1. We construct a learnable complex kernel matrix $W_y = [\boldsymbol{w}_y(1), \boldsymbol{w}_y(2), \cdots, \boldsymbol{w}_y(n), \cdots, \boldsymbol{w}_y(N)] \in \mathcal{C}^{M\times N}$ to calculate the cosine similarity between $\boldsymbol{w}_y(n)$ and $\boldsymbol{y}(t,f)$. The similarity matrix $S(t) = [\boldsymbol{a}_1(t), \boldsymbol{a}_2(t), \cdots, \boldsymbol{a}_N(t)] \in \mathcal{R}^{F\times N}$ is formulated as

$$a_n(t,f) = \frac{\boldsymbol{w}_y^T(n)\boldsymbol{y}(t,f)}{\|\boldsymbol{w}_y(n)\|\,\|\boldsymbol{y}(t,f)\|}, \quad (3)$$

$$\boldsymbol{a}_n(t) = [a_n(t,1), a_n(t,2), \cdots, a_n(t,F)]^T, \quad (4)$$

where F is the number of frequency bins. After taking the power of the absolute value of the similarity, the final spatial feature vectors $\boldsymbol{z}(t) = [\bar{a}_1(t), \bar{a}_2(t), \cdots, \bar{a}_N(t)]^T$ are obtained by summing the matrix along frequency indices, given by

$$\bar{a}_n(t) = \sum_f |a_n(t,f)|^2. \quad (5)$$

As such, the directional statistics are projected to the spatial grids to represent the spatial features of the multi-channel signal. Fig. 1 gives an illustration of the spatial features extraction process. Different from existing methods in which directional statistics are usually modeled by complex Gaussian mixture models [22] or Watson mixture models [23] in a bin-wise manner, and an alignment step is usually needed to solve the permutation ambiguities, the learnable kernel matrix developed in our approach categorizes all frequency bins which share the same spatial information into a cluster, and hence avoids the permutation problem. In addition, as the spatial information over different frequency bins are summarized together in each frame, the computational complexity can be reduced.

2.2. Model architecture

After obtaining the spatial and spectral features, our model for OSD mainly consists of two steps: feature fusing and OSD scoring, as illustrated in Fig. 2. For the first step, a normalization is applied to different features so that the features have normal distributions. A Gated Multimodal Unit (GMU) [14] is then used to fuse the features. Consequently, the embedding sequence generated by GMU are input into a BLSTM model for scoring. It is worth mentioning that the BLSTM model used here is the same as that in [3], which includes two stacked BLSTM recurrent layers, two feed-forward layers and a final classification layer with sigmoid activation. The output of BLSTM model is a probability sequence which indicates the classification of overlapped speech at the frame level.

GMU feature fusion. Feature fusion can be considered from two perspectives: 1) data-level early fusion; or 2) late fusion at decision-level. In [24], these two fusion methods are employed for speaker diarization. For the early fusion, the spatial and spectral features are directly stacked before the clustering, and for the late fusion, the features are combined by a weighted sum at the similarity matrix level. In this work, we follow the early fusion approach. Assume that normalized spatial features are $\bar{\boldsymbol{z}}(t)$, and the normalized spectral features from the m-th channel are $\bar{\boldsymbol{x}}_m(t)$. Instead of concatenating the features directly, a GMU module is employed to map spatial features $\bar{\boldsymbol{z}}(t)$ and spectral features $\bar{\boldsymbol{x}}_m(t)$ into an embedding space such that the spatial embedding $\boldsymbol{h}_z(t)$ and spectral embedding $\boldsymbol{h}_x(t)$ have the same size. Different importance is then assigned to each type of feature using a sigmoid gate neuron σ. The GMU module is detailed as follows

$$\boldsymbol{h}_z(t) = \tanh(W_z \bar{\boldsymbol{z}}(t)), \tag{6}$$

$$\boldsymbol{h}_x(t) = \tanh(W_x \bar{\boldsymbol{x}}_m(t)), \tag{7}$$

$$\boldsymbol{e}(t) = \sigma(W_e[\bar{\boldsymbol{z}}(t), \bar{\boldsymbol{x}}_m(t)]), \tag{8}$$

$$\boldsymbol{h}(t) = \boldsymbol{e}(t) \circ \boldsymbol{h}_z(t) + (1 - \boldsymbol{e}(t)) \circ \boldsymbol{h}_x(t), \tag{9}$$

where $\{W_z, W_x, W_e\}$ are the parameters to be learned and "$\circ$" is the Hadamard product. Note that different from the soft-attention mechanism which assigns single attention weight to the whole feature vector, the GMU model computes different importance for each element in the spatial embedding $\boldsymbol{h}_z(t)$ and spectral embedding $\boldsymbol{h}_x(t)$. As such, a feature can be weighted more when it is more important to the OSD detection. For example, in a quiet meeting room, spatial features may contribute more than spectral features for OSD, more importance is hence given to the spatial features by using GMU, and vice versa.

We also compare our fusion method with the multi-channel spectral features based OSD method in [11], in which the prediction probabilities were simply averaged across different channels. To make a fair comparison, a soft-attention mechanism [25, 26] is implemented to fuse multi-channel spectral features. We calculate the attention weight $e_m(t)$ for the m-th channel so that the output can be computed as a weighted sum of multi-channel features, given by

$$e_m(t) = \text{softmax}(V \times \tanh(A\bar{\boldsymbol{x}}_m(t) + b)), \tag{10}$$

$$\hat{\boldsymbol{x}}(t) = \sum_{m=1}^{M} e_m(t)\bar{\boldsymbol{x}}_m(t), \tag{11}$$

where $\{A, V, b\}$ denotes the shared parameters over all M channels and $\hat{\boldsymbol{x}}(t)$ is the input of the BLSTM model.

3. Experimental evaluation

The performance of the proposed method is studied under AMI dataset [15] and CHiME-6 dataset [16], which are developed for the performance evaluation of distant ASR. The AMI dataset is a meeting corpus consisting of 100 hours of recordings. Two circular microphone arrays with 10cm radius are placed on the table located at the center of the meeting room. Ground truth overlapped speech labels were obtained by manually annotating the speech signal from a close-talk microphone. In [2, 11], label force alignment is applied to correct annotation errors by using a pre-trained ASR model. However, it is worth mentioning that both force alignment and manual annotation have errors [8]. In our experiments on AMI, we only use the data of the 1-th microphone array for training and testing as the data from the 2-th array is incomplete. The recordings in CHiME-6 corpus are made by using six 4-channel linear microphone arrays. Labels after force alignment are used as ground truth. Compared to AMI dataset, CHiME-6 is more noisy and challenging.

The signals are first transformed to the STFT domain using a 1024 sample (corresponding to 64 ms at 16kHz sampling rate) Hanning window with a hop size of 512 samples. We extract 80 log-mel filter bank features as spectral features which can also be replaced by MFCC. The dimension of GCC-PHAT features is 51×28 in AMI corpus and 51×6 in CHiME-6 dataset. Directional statistics are projected onto a 64 grid space. The input sequence length for training is set to 600 frames. A single layer of 128 BLSTM cells is used in the first stacked BLSTM module and 5 layers are used in the second BLSTM module. Adam optimizer with a learning rate of 0.001 and binary cross-entropy loss are employed. As the same as in [11], the average precision (AP) is employed to evaluate the OSD performance.

Analysis of different features. In this experiment, the discriminative property of the spatial and spectral features are studied. We have plotted the last layer of BLSTM outputs using t-Distributed Stochastic Neighbor Embedding (t-SNE) [27], as shown in Fig. 3. The t-SNE is able to present lower dimensional embeddings of data points while preserving global inter-cluster data structure. Therefore, it can be used to identify the clusters in the data and investigate the discrimination of different features in OSD task. The better the plotted embedding data points can be clustered, the more the features are separable, and vice versa. It can be observed from Fig. 3 that for the embeddings from log-mel or GCC-PHAT features, the data points representing the overlapped speech segments spread over the picture, but the data points from directional statistics are more concentrated and separable. Hence, the directional statistics are more separable than the log-mel and GCC-PHAT features. More accurate OSD result can thus be obtained by using directional statistics. It can also be observed that the fusion embeddings from

Table 1: *Average Precision (%) on the development and evaluation sets of AMI and CHiME-6 cropora based on BLSTM model.*

Feature	Method	AMI		CHiME-6	
		dev	eval	dev	eval
Log-mel	BLSTM	70.75	58.59	53.87	45.04
GCC-PHAT		69.23	58.43	34.68	33.11
Directional Statistics		77.60	69.42	54.29	42.50
Multi-channel Log-mel	BLSTM + Attention	72.68	61.17	60.00	44.03
Log-mel + GCC-PHAT	BLSTM + GMU	76.89	68.55	60.20	53.14
Log-mel + Directional Statistics		**80.66**	**74.75**	**60.76**	**53.97**

Table 2: *Average Precision (%) on the development and evaluation sets of AMI and CHiME-6 cropora based on TCN model.*

Feature	Method	AMI		CHiME-6	
		dev	eval	dev	eval
Log-mel	TCN	71.68	60.72	63.11	59.36
GCC-PHAT		74.88	65.47	38.12	39.71
Directional Statistics		79.79	72.95	55.88	51.03
Multi-channel Log-mel	TCN + Attention	76.33	62.74	62.84	55.99
Log-mel + GCC-PHAT	TCN + GMU	78.11	71.15	**63.43**	**59.65**
Log-mel + Directional Statistics		**81.53**	**74.96**	63.29	58.52

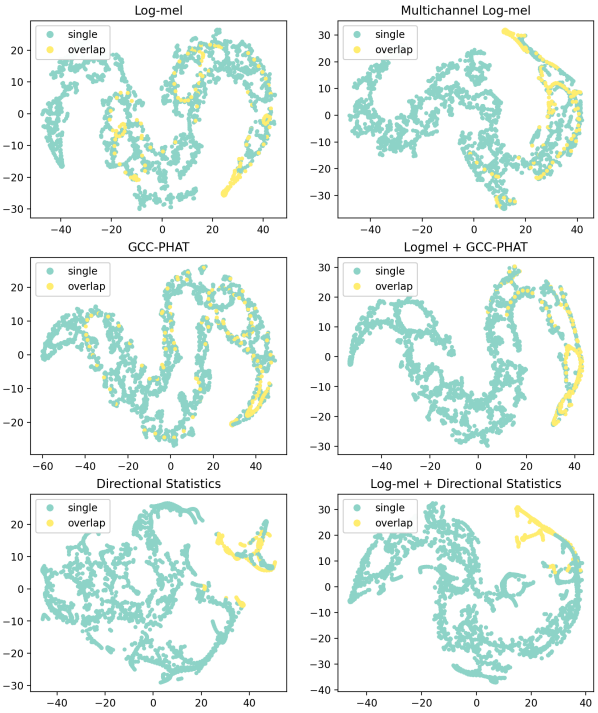

Figure 3: *t-SNE scatter plots of embeddings from the last BLSTM layer based on different input features extracted from a recording of AMI corpus not used in training. Intuitively, the more separable of the embedding plotting, the more discriminative of the features, and thus better OSD results can be obtained.*

the spatial and spectral features are more distinguishable than other features in both single and overlapped speech segments and in particular, embeddings from fusion of log-mel and directional statistics can be best clustered. In next experiment, we will demonstrate that better OSD performance can be achieved when the features are more separable.

Analysis of OSD results. The results of the OSD from both AMI and CHiME-6 datasets is given in Table 1. It shows that the proposed spatial and spectral features fusion approach outperforms existing methods which use spectral features only. In particular, the AP score of the proposed approach increases about 16% on AMI dataset. However, the performance improvement on CHiME-6 is relatively limited as the CHiME-6 dataset is very noisy; AP scores of all methods degrade significantly. Moreover, on AMI corpus, fusion method using proposed directional statistics performs better than that using GCC-PHAT as the directional statistics is more distinguishable than the GCC-PHAT features. This further validates our analysis showed in Fig. 3.

We also implement the method proposed in [11] in which the BLSTM model is replaced by TCN model. The results are given in Table 2. It can be observed that for all different feature types, TCN model is able to achieve higher scores than BLSTM method, and the proposed fusion method performs better than other methods on AMI dataset. It is worth mentioning that on CHiME-6 dataset, fusing the spatial features generated by GCC-PHAT can obtain slightly higher score than using directional statistics due to heavy noise on CHiME-6 dataset.

4. Conclusions

In this paper, a spatial and spectral features fusion method is proposed for OSD. Two different spatial features, GCC-PHAT function and projection of directional statistics are studied. A GMU module is then introduced to fuse the spatial and spectral features. Analysis of different combinations of various features shows that the fusion of the projection of directional statistics and spectral features can achieve more distinguishable embeddings, and thus more accurate OSD results can be obtained. Experimental results on AMI and CHiME-6 corpora also demonstrate that better performance can be achieved by fusion the spatial and spectral features.

5. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61971186.

6. References

[1] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[2] N. Sajjan, S. Ganesh, N. Sharma, S. Ganapathy, and N. Ryant, "Leveraging lstm models for overlap detection in multi-party meetings," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5249–5253.

[3] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7114–7118.

[4] V.-T. Tran and W.-H. Tsai, "Speaker identification in multi-talker overlapping speech using neural networks," *IEEE Access*, vol. 8, pp. 134 868–134 879, 2020.

[5] T. Yoshioka, H. Erdogan, Z. Chen, X. Xiao, and F. Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," in *Proc. INTERSPEECH*, 2018, pp. 3038–3042.

[6] J. Yu, B. Wu, R. G. S.-X. Z. L. C. Yu, Y. X. Meng, D. Su, D. Yu, X. Liu, and H. Meng, "Audio-visual multi-channel recognition of overlapped speech," in *Proc. INTERSPEECH*, 2020, pp. 3496–3500.

[7] R. Vipperla, J. T. Geiger, S. Bozonnet, D. Wang, N. Evans, B. Schuller, and G. Rigoll, "Speech overlap detection and attribution using convolutive non-negative sparse coding," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4181–4184.

[8] M. Kunešová, M. Hrúz, Z. Zajíc, and V. Radová, "Detection of overlapping speech for the purposes of speaker diarization," in *Proc. International Conference on Speech and Computer*, 2019, pp. 247–257.

[9] V. Andrei, H. Cucu, and C. Burileanu, "Detecting overlapped speech on short timeframes using deep learning." in *Proc. INTERSPEECH*, 2017, pp. 1198–1202.

[10] V. Andrei, H. Cucu, and C. Burileanu, "Overlapped speech detection and competing speaker counting—humans versus deep learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 850–862, 2019.

[11] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Detecting and counting overlapping speakers in distant speech scenarios," in *Proc. INTERSPEECH*, 2020, pp. 3107–3111.

[12] F.-R. Stöter, S. Chakrabarty, B. Edler, and E. A. Habets, "Countnet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 268–282, 2018.

[13] W. Zhang, M. Sun, L. Wang, and Y. Qian, "End-to-end overlapped speech detection and speaker counting with raw waveform," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, 2019, pp. 660–666.

[14] J. E. A. Ovalle, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion." in *Proc. 5th International Conference on Learning Representations (Workshop)*, 2017.

[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 *Proc. International workshop on machine learning for multimodal interaction*, 2005, pp. 28–39.

[16] S. Watanabe, M. Mandel, J. Barker, and E. Vincent, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," *arXiv preprint arXiv:2004.09249*, 2020.

[17] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 2814–2818.

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

[19] M. Fakhry, N. Ito, S. Araki, and T. Nakatani, "Modeling audio directional statistics using a probabilistic spatial dictionary for speaker diarization in real meetings," in *Proc. IEEE International Workshop on Acoustic Signal Enhancement*, 2016, pp. 1–5.

[20] C. Boeddeker, T. Cord-Landwehr, J. Heitkaemper, C. Zorila, D. Hayakawa, M. Li, M. Liu, R. Doddipatla, and R. Haeb-Umbach, "Towards a speaker diarization system for the chime 2020 dinner party transcription," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments*, 2020, pp. 42–47.

[21] N. Ito, S. Araki, and T. Nakatani, "Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3238–3242.

[22] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. 24th European Signal Processing Conference*, 2016, pp. 1153–1157.

[23] N. Ito, S. Araki, and T. Nakatani, "Data-driven and physical model-based designs of probabilistic spatial dictionary for online meeting diarization and adaptive beamforming," in *Proc. 25th European Signal Processing Conference*, 2017, pp. 1165–1169.

[24] W. Kang, B. C. Roy, and W. Chow, "Multimodal speaker diarization of real-world meetings using d-vectors with spatial features," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6509–6513.

[25] H. Zhang, J. Zhang, and Y. Wang, "End-to-end models with auditory attention in multi-channel keyword spotting," *arXiv preprint arXiv:1811.00350*, 2018.

[26] X. Ji, M. Yu, J. Chen, J. Zheng, D. Su, and D. Yu, "Integration of multi-look beamformers for multi-channel keyword spotting," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7464–7468.

[27] P. G. Poličar, M. Stražar, and B. Zupan, "opentsne: a modular python library for t-sne dimensionality reduction and embedding," *BioRxiv*, 2019.