MULTIMODAL SPEAKER DIARIZATION OF REAL-WORLD MEETINGS USING D-VECTORS WITH SPATIAL FEATURES

Wonjune Kang[1], Brandon C. Roy[2,3], and Wesley Chow[2,3]

[1]Massachusetts Institute of Technology, USA [2]MIT Media Lab, USA [3]Cortico, USA

ABSTRACT

Deep neural network based audio embeddings (d-vectors) have demonstrated superior performance in audio-only speaker diarization compared to traditional acoustic features such as mel-frequency cepstral coefficients (MFCCs) and i-vectors. However, there has been little work on multimodal diarization systems that combine d-vectors with additional sources of information. In this paper, we present a novel approach to multimodal speaker diarization that combines d-vectors with spatial information derived from performing beamforming given a multi-channel microphone array. Our system performs spectral clustering on a combination of speaker embeddings and spatial features that are computed using the Steered-Response Power Phase Transform (SRP-PHAT) algorithm. We evaluate our system on the AMI Meeting Corpus and an internal dataset of real-world conversations. By using both acoustic and spatial features for diarization, we achieve significant improvements over a d-vector only baseline and show potential to achieve comparable results with other state-of-the-art multimodal diarization systems.

***Index Terms*—** Speaker diarization, d-vector, beamforming, sound source localization, spectral clustering

1. INTRODUCTION

Speaker diarization is the process of partitioning an audio stream into speaker segments and labeling them with the speakers' identities [1]. Informally, it can be summarized as the problem of determining "who spoke when". With recent advances in speech processing technologies such as automatic speech recognition and speaker identification, diarization has emerged as an important task in order to give the outputs of these systems greater meaning and context.

Speaker diarization has traditionally been addressed as a single-modality problem, in which a single channel of audio is used to extract feature embeddings such as mel-frequency cepstral coefficients (MFCCs) [2] or i-vectors [3, 4] from segmented speech, after which these embeddings are clustered using algorithms such as *k*-means, agglomerative hierarchical clustering, or spectral clustering. The speaker boundaries from clustering can then be further refined through a resegmentation step such as Variational Bayes [5, 6]. In recent years, deep neural network audio embeddings for speaker recognition (d-vectors, x-vectors) [7, 8] have been successfully used in diarization systems, often showing significantly improved performance over previously popular i-vector based approaches [9, 10, 11].

However, there are many settings in which additional modalities can be leveraged to improve speaker diarization performance. Notably, in the conference meeting domain, specialized hardware such as microphone arrays and cameras often allow for the collection of multiple audio channels and video. This information can be used for techniques such as sound source localization (SSL) or audio-visual correspondence. Previous works have sought to utilize spatial information by using acoustic beamforming to compute inter-channel delay features, which were combined with MFCCs at the weighted log-likelihood level [12, 13]. Other approaches have utilized visual information, such as by clustering MFCCs fused with video features [14] or combining visual analysis of motion and mouth movements with SSL [15]. However, none of these works used the currently state-of-the-art deep learning approaches for acoustic speaker modeling. Recently, [16] proposed a multimodal diarization system that uses a deep audio-visual synchronization network to enroll speaker models, then uses a combination of audio-visual correlation, deep speaker embeddings, and SSL to predict the speaker. However, despite achieving high diarization performance, this method requires significantly more hardware and processing for the video data, which may not be feasible for many types of meeting rooms in practice.

In this paper, we propose a novel approach to speaker diarization when a multi-channel microphone array is available by supplementing d-vectors with acoustic beamforming. Our key contribution is the use of steered-response powers from the Steered-Response Power Phase Transform (SRP-PHAT) algorithm [17] as spatial features that can be incorporated into the clustering step. In our approach, we build off of the methodology introduced in [10], which combines long short-term memory (LSTM) based d-vectors with spectral clustering; however, the spatial features can more generally be combined with other types of speaker modeling and clustering. We present two ways of utilizing the spatial information: an "early fusion" method, in which the spatial features are directly stacked with the speaker embedding before clustering, and a "late fusion" method, in which each set of features is combined by a weighted sum at the similarity matrix level. We evaluate our approach on the publicly available AMI Meeting Corpus [18] and an internal dataset of real-world conversations, and find that it achieves significantly improved performance compared to a d-vector only baseline. In addition, it demonstrates the potential to yield competitive results with state-of-the-art audio-visual approaches that also use sound source localization, despite not utilizing video information.

2. BASELINE AUDIO-ONLY SYSTEM

2.1. D-vector model

Our baseline audio-only system is based on the work in [10], which adapted the text-independent speaker verification d-vector model from [8] for speaker diarization. The speaker embedding network is trained using generalized end-to-end (GE2E) loss [8], and consists of three LSTM layers of 768 nodes each and a projection layer of 256 nodes. In this system, audio signals are first transformed into frames of width 25ms and step 10ms, and log-mel-filterbank energies of dimension 40 are extracted from each frame. These frames are used to build sliding windows of a fixed length, which are fed

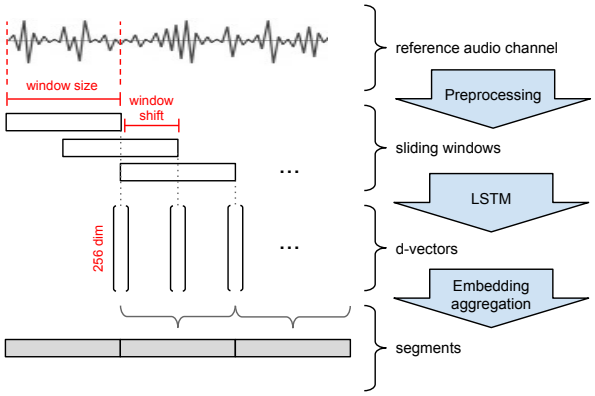

Fig. 1. The baseline d-vector system pipeline. Individual d-vectors are aggregated to form a single speaker embedding per segment.

into the LSTM as inputs. The last-frame output of the LSTM is used as the d-vector representation for a given sliding window. For each speech segment that is to be labeled, the d-vectors corresponding to the sliding windows ending in that segment are ℓ_2-normalized and averaged to form a single embedding for that segment. The key elements of the system pipeline are shown in Figure 1.

Our model was trained from scratch on the development sets of the VoxCeleb [19] and VoxCeleb2 [20] datasets, which amount to approximately 7K speakers and 1.2M utterances. We note that this differs from the model used in [10], which was trained on a collection of voice search data with around 18K speakers and 36M utterances. Based on empirical testing, we also used slightly different parameters from [10] for our experiments. Specifically, our speaker embedding network is trained on windows of fixed size 360ms, compared to 1600ms [10] or a uniform distribution within [240ms, 1600ms] [21]. Instead of sliding windows of size 240ms and 120ms overlap, we use 360ms (matching the model training) and 180ms, respectively. In addition, we use a maximal segment-length limit of 1000ms instead of 400ms.

2.2. Spectral clustering

We use the spectral clustering algorithm proposed in [10]. We note that this algorithm is different from "standard" spectral clustering, as it performs eigendecomposition directly on the similarity matrix and not on a Laplacian matrix. However, we will continue to refer to it as spectral clustering for consistency. We largely follow the same steps as described in [10] to construct the similarity matrix A, including refinement operations: Gaussian blur, row-wise soft thresholding, symmetrization, diffusion, and row-wise max normalization.

Following eigendecomposition of A, we estimate the number of speakers $\widehat{K}$ by using the heuristic introduced in [22]. It was seen experimentally that the n sorted eigenvalues of A ($\lambda_1 > \lambda_2 > ... > \lambda_n$) exhibit exponential decay, and that the number of speakers in a conversation consistently corresponds to the point at which the gradient of these eigenvalues exceeds a threshold θ. Therefore, to determine the number of clusters, we fit a smooth exponential $e^{-\alpha k}$ to the eigenvalues, where $k = 1, ..., n$. We then take $\widehat{K}$ to be the smallest value of k for which the derivative of the exponential exceeds θ:

$$\widehat{K} = \arg \min_{k \in \{1,...,n\}} \left[-\alpha e^{-\alpha k} \geq \theta \right] \tag{1}$$

3. MULTIMODAL DIARIZATION SYSTEM

Our complete diarization system is as follows. First, all audio channels are processed by a speech enhancement system, which is based on the one described in [23]; it is an LSTM-based speech denoising model trained to predict clean log-power spectra features given noisy log-power spectra features with acoustic context. We use WebRTC [24] to perform voice activity detection (VAD) on a single reference channel of the denoised audio and determine sections of speech, which are then divided into shorter non-overlapping segments that determine the temporal resolution of the diarization. For each segment, our system computes a speaker embedding from the reference channel and a vector of steered-response powers on the surrounding space. The two sets of features are then fused to construct a similarity matrix A, which is used to perform spectral clustering on all segments.

3.1. Acoustic beamforming

We use the Steered-Response Power Phase Transform (SRP-PHAT) algorithm for acoustic beamforming. SRP-PHAT can be interpreted as a grid-search procedure that attempts to maximize the steered-response power $P(\mathbf{x})$ from a set of candidate source locations using a steered delay-and-sum beamformer. It can be summarized by the following expressions:

$$P(\mathbf{x}) \triangleq \sum_{n \in N} \left| \sum_{m=1}^{M} s_m(n - \tau_m(\mathbf{x})) \right|^2 \tag{2}$$

$$\widehat{\mathbf{x}}_s = \arg \max_{\mathbf{x} \in \mathcal{G}} P(\mathbf{x}) \tag{3}$$

where N denotes all sample indices within the beamforming window, $s_m(n)$ is the discrete-time output signal from the m-th microphone, and $\tau_m(\mathbf{x})$ is the time-lag due to the propagation from a source located at $\mathbf{x}$ to the m-th microphone. $\widehat{\mathbf{x}}_s$ is the estimated spatial location of the true sound source $\mathbf{x}_s$, and $\mathcal{G}$ is the set of candidate source locations to be considered.

This formulation provides the single most likely direction of arrival. However, it has a disadvantage in that it requires a large number of computations. To accurately obtain the location of the source, the spatial resolution of $\mathcal{G}$ must be relatively high, which requires that $P(\mathbf{x})$ be computed for a large number of points; if the spatial resolution is reduced, $\widehat{\mathbf{x}}_s$ may differ significantly from $\mathbf{x}_s$. In addition, we found during our experiments that even with a high-resolution grid, it is difficult to incorporate scalar directional values with d-vector embeddings in a meaningful and significant way.

To this end, we use the values of $P(\mathbf{x})$ slightly differently. Instead of selecting the single global maximum from a high-resolution grid, we flatten all steered-response powers computed from a coarser grid into a vector of these values for the entire space surrounding the microphone array. Concretely, we compute the following vector:

$$\mathbf{s}_t = [P_t(\mathbf{x}_1), P_t(\mathbf{x}_2), ..., P_t(\mathbf{x}_n)]^\top, \ \forall \mathbf{x}_i \in \mathcal{G} \tag{4}$$

where t denotes the time of the beamforming and n determines the spatial resolution of $\mathcal{G}$. In our experiments, $\mathbf{s}$ is computed with a beam window length of 600ms and shift of 150ms, with $n = 90$ (i.e. $P(\mathbf{x})$ is computed every 4 degrees in the 360 degree space). These spatial vectors are then aggregated using the same logic as for the d-vectors. For each speech segment, all $\mathbf{s}$ corresponding to beamforming windows that end in that segment are ℓ_2-normalized and averaged to form a single vector for that segment. Figure 2 shows the main elements of our pipeline for computing spatial features.

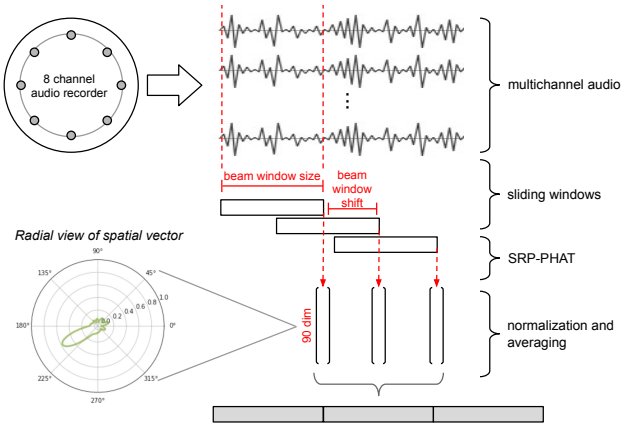

Fig. 2. A diagram of our pipeline for computing spatial features. Each vector can be seen as a flattening of the steered response powers computed on the radial space surrounding the microphone array.

3.2. Multimodal fusion

The aforementioned processes convert multi-channel audio inputs of arbitrary length into two sets of fixed-length vectors: speaker embeddings and spatial steered-response power vectors. These features are then combined to construct a similarity matrix for spectral clustering. Here, we present two ways of doing so.

3.2.1. Early fusion

In our early fusion method, the speaker embeddings and spatial vectors are individually ℓ_2-normalized, and then directly combined by stacking them together. This results in a set of vectors whose dimensions are the sum of the dimensions of the two sets of features. These concatenated vectors are used to construct a similarity matrix A, where A_{ij} is defined as the cosine similarity between the ith and jth segment vectors if $i \neq j$, and the diagonal elements are set to the maximum value of each row: $A_{ii} = \max_{j \neq i} A_{ij}$.

3.2.2. Late fusion

In our late fusion method, each set of features is used to construct a separate similarity matrix (denoted below by A_d and A_s for the d-vector and spatial features, respectively) by the same methodology as described above, using cosine similarity as the affinity measure. Then, the two matrices are combined by a weighted sum to yield the final similarity matrix that will be used for spectral clustering:

$$A = \alpha A_d + (1 - \alpha) A_s \tag{5}$$

3.2.3. Link between early and late fusion

For clarity, we derive a link between the early and late fusion methods. In early fusion, the similarity matrix A has entries $A_{ij} = \cos(x_i, x_j)$ where $x_i = (d_i, s_i)$, the concatenation of the i-th ℓ_2-normalized d-vector (d_i) and spatial vector (s_i). As a result, $\|x_i\| = \sqrt{2}$, $\forall i$ and we have:

$$A_{ij} = \frac{1}{2} x_i^\top x_j \tag{6}$$

The inner product of x may be expanded into the inner products of its constituents, yielding

$$A_{ij} = \frac{1}{2}[d_i^\top d_j + s_i^\top s_j] \tag{7}$$

This is equivalent to the corresponding entry of the similarity matrix for late fusion when $\alpha = \frac{1}{2}$, and we see that early fusion is a special case of late fusion (when not taking any matrix refinement operations into account). Thus, late fusion involves an additional hyperparameter that provides an opportunity to further optimize performance.

4. EXPERIMENTS

4.1. Datasets

4.1.1. AMI corpus

The AMI Meeting Corpus [18] consists of 100 hours of meeting recordings and contains a range of signals, including close-talking and far-field microphones, individual and room-view video cameras, and outputs from a slide projector and an electronic whiteboard. The audio is recorded from an 8-element uniformly spaced circular microphone array with a radius of 10cm, sampled at 16kHz.

All meetings are recorded in English in three locations with different acoustic properties, and include mostly non-native speakers. It has been previously used to evaluate many audio-only and audio-visual speaker diarization systems. We perform evaluations on meetings in the ES (Edinburgh) and IS (IDIAP) categories, which contain approximately 30 and 17 hours of data, respectively. Of the IS meeting files, `IS1002a`, `IS1003b`, `IS1005d`, and `IS1007d` were not used because of missing data.

4.1.2. Internal conversation dataset

Our internal conversation dataset was collected as part of the Local Voices Network (LVN) project [25] developed by Cortico, and consists of conversations amongst 3 to 6 people that are between 60 to 90 minutes in length. Each conversation focuses on topics of concern to the partipants' local community, and are held in natural settings in which no particular instructions are given to the participants regarding the recording of the audio. The lengths of speaker turns vary widely, from short bursts of frequent changes to long stretches of several minutes with just one speaker.

The audio is recorded using an 8-element uniformly spaced circular microphone array with a radius of 8.55cm, sampled at 16 kHz. The recording device is portable and conversations may be held in a wide variety of settings. As a result, the acoustic environments of the conversations tend to vary widely, with differing sources and amounts of background noise. Our evaluation dataset consists of 8 files containing approximately 10 hours of audio, carefully annotated by hand. In the case of overlapped speech, only the identity of the loudest speaker was annotated.

4.2. Evaluation protocols

We evaluated our system using Diarization Error Rate (DER), which consists of three components: missed detection, false alarm, and speaker confusion. We used the tool developed for the Rich Transcription 2009 evaluations by NIST (NIST RT-09) [26]. Following common convention in the literature, we exclude overlapped speech (multiple individuals speaking at the same time) from evaluation and tolerate errors less than 250ms in locating segment boundaries.

Table 1. Diarization Error Rate results on the ES and IS subsets of the AMI corpus and our internal dataset. Note that all experiments use the same VAD system, so the missed detection and false alarm rates are identical across different methods for each dataset. x-vector results are reported from Table 1 in [16].

Dataset	Method	Unknown Speaker Count				Oracle Speaker Count			
		Missed	FA	Confusion	DER	Missed	FA	Confusion	DER
AMI ES	d-vector only	7.95	3.44	18.63	30.02	7.95	3.44	17.03	28.42
	Early Fusion			8.76	20.15			8.86	20.25
	Late Fusion			8.06	19.45			10.44	21.83
	x-vector [16]	-	-	12.8	-	-	-	-	-
	Chung et al. [16]	-	-	2.8	-	-	-	-	-
AMI IS	d-vector only	8.25	1.72	17.16	27.13	8.25	1.72	16.87	26.84
	Early Fusion			11.47	21.44			11.51	21.48
	Late Fusion			12.13	22.1			13.81	23.78
	x-vector [16]	-	-	10.2	-	-	-	-	-
	Chung et al. [16]	-	-	4.9	-	-	-	-	-
Internal	d-vector only	7.31	1.84	19.02	28.17	7.31	1.84	13.79	22.94
	Early Fusion			13.98	23.13			14.66	23.81
	Late Fusion			14.54	23.69			7.78	16.93

4.3. Results

Table 1 summarizes the diarization performance results on the ES and IS subsets of the AMI corpus and our internal dataset, given both unknown and oracle speaker counts for clustering. Missed detection and false alarm rates are identical across the different methods for each dataset because we use the same VAD system; therefore, speaker confusion is the only metric that is affected by the different input features and fusion methods. We also include results for an audio-only x-vector [7] system (results reported from [16]) and the best audio-visual correspondence model from [16]. Since we use slightly different systems for VAD, we only list scores for speaker confusion in order to provide a fair comparison.

Despite being trained on the same data (VoxCeleb1 and VoxCeleb2), our d-vector only system does not perform as well as the x-vector model reported in [16] in terms of speaker confusion. This may be due to differences in the amount of training data needed to optimize performance given the model architectures. Although [10] reports comparable results between their d-vector model and x-vectors on the NIST SRE 2000 CALLHOME dataset, their model was trained on substantially more data than ours. It is unclear whether this difference accounts for the discrepancy in performance shown here, or whether x-vectors simply generalize better to speech recorded under different acoustic conditions.

After incorporating spatial information, our system achieves comparable or better performance than the x-vector model. On the ES subset of the AMI corpus, early and late fusion result in 53% and 57% relative improvement on speaker confusion, respectively; on the IS subset, they result in 33% and 29% relative improvement, respectively. Our multimodal system yet lags behind the best audio-visual model from [16]. However, the relative performance differences between the baseline audio-only models, along with the magnitude of the improvements from incorporating spatial information, suggest that our methods could potentially provide competitive results when combined with a better speaker embedding model, even without the use of video information.

The overall effectiveness of early fusion compared with late fusion is unclear, as neither method outperforms the other on all subsets of the evaluation data. However, given that early fusion can be seen as a special case of late fusion with a fixed value of α, there appears to be potential to further optimize late fusion by implementing tunable or adaptive α values for individual audio files.

On our internal dataset of conversations, early and late fusion result in 26% and 24% relative improvement, respectively. The slightly smaller relative improvement compared to the AMI corpus may have been due to the noisier and more inconsistent environments in which the audio was recorded. Interestingly, assigning the number of clusters based on the oracle speaker count did not seem to significantly improve performance, and even degraded it in many cases. However, providing the oracle speaker count resulted in the largest relative performance improvement on our internal dataset for our late fusion method (44%). These results suggest that, despite the many heuristics that have been suggested in the literature, finding the optimal number of clusters for spectral clustering is still a difficult task that can have a significant effect on DER.

5. CONCLUSION

In this work, we proposed a novel approach to multimodal speaker diarization by supplementing d-vector speaker embeddings with spatial features obtained using the SRP-PHAT algorithm. We presented two methods of combining these features before performing spectral clustering: early fusion, in which the two sets of features are stacked together before forming the similarity matrix, and late fusion, in which the features are combined by a weighted sum at the similarity matrix level. Our approach achieves significant improvements over a d-vector only baseline on the AMI Meeting Corpus and an internal dataset of conversations. In addition, it demonstrates the potential to yield competitive results with state-of-the-art audio-visual approaches that also use sound source localization, despite not using any video information. An interesting direction for future work is to consider adaptive weights for our late fusion method that could be tuned based on characteristics of individual audio files.

6. ACKNOWLEDGEMENTS

We would like to thank David van Dokkum for curating the dataset of LVN conversations, Sam Woolf for assistance with the microphone array hardware, and Doug Beeferman for the helpful discussion.

7. REFERENCES

[1] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol 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] Patrick Kenny, Douglas Reynolds, and Fabio Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[3] Stephen H Shum, Najim Dehak, Réda Dehak, and James R Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[4] Gregory Sell and Daniel Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 413–417.

[5] Gregory Sell and Daniel Garcia-Romero, "Diarization resegmentation in the factor analysis subspace," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4794–4798.

[6] Mireia Diez, Lukáš Burget, and Pavel Matejka, "Speaker diarization based on bayesian hmm with eigenvoice priors," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 147–154.

[7] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[8] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[9] Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4930–4934.

[10] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopez Moreno, "Speaker diarization with lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5239–5243.

[11] David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5796–5800.

[12] Jose M Pardo, Xavier Anguera, and Chuck Wooters, "Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences," in *Ninth International Conference on Spoken Language Processing*, 2006.

[13] Xavier Anguera, Chuck Wooters, and Javier 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] Nikolaos Sarafianos, Theodoros Giannakopoulos, and Sergios Petridis, "Audio-visual speaker diarization using fisher linear semi-discriminant analysis," *Multimedia Tools and Applications*, vol. 75, no. 1, pp. 115–130, 2016.

[15] P Cabañas-Molero, M Lucena, José Manuel Fuertes, Pedro Vera-Candeas, and Nicolás Ruiz-Reyes, "Multimodal speaker diarization for meetings using volume-evaluated srp-phat and video analysis," *Multimedia Tools and Applications*, vol. 77, no. 20, pp. 27685–27707, 2018.

[16] Joon Son Chung, Bong-Jin Lee, and Icksang Han, "Who Said That?: Audio-Visual Speaker Diarisation of Real-World Meetings," in *Interspeech*, 2019, pp. 371–375.

[17] Joseph Hector DiBiase, *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, Brown University Providence, RI, 2000.

[18] Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al., "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005, vol. 88, p. 100.

[19] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Interspeech*, 2017.

[20] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018.

[21] Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang, "Fully supervised speaker diarization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6301–6305.

[22] Huazhong Ning, Ming Liu, Hao Tang, and Thomas S Huang, "A spectral clustering approach to speaker diarization," in *Ninth International Conference on Spoken Language Processing*, 2006.

[23] Lei Sun, Jun Du, Chao Jiang, Xueyang Zhang, Shan He, Bing Yin, and Chin-Hui Lee, "Speaker diarization with enhancing speech for the first dihard challenge.," in *Interspeech*, 2018, pp. 2793–2797.

[24] Alan B Johnston and Daniel C Burnett, *WebRTC: APIs and RTCWEB protocols of the HTML5 real-time web*, Digital Codex LLC, 2012.

[25] Cortico, "Local Voices Network," `http://lvn.org`, 2019.

[26] Gerald Friedland, Adam Janin, David Imseng, Xavier Anguera, Luke Gottlieb, Marijn Huijbregts, Mary Tai Knox, and Oriol Vinyals, "The icsi rt-09 speaker diarization system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 371–381, 2011.