Integrating Audio, Visual, and Semantic Information for Enhanced Multimodal Speaker Diarization

Luyao Cheng[1], Hui Wang[1], Siqi Zheng[1], Yafeng Chen[1], Rongjie Huang[2],
Qinglin Zhang[1], Qian Chen[1], Xihao Li[3],

[1]Alibaba Group
[2]Zhejiang University
[3]University of North Carolina at Chapel Hill
shuli.cly@alibaba-inc.com

Abstract

Speaker diarization, the process of segmenting an audio stream or transcribed speech content into homogenous partitions based on speaker identity, plays a crucial role in the interpretation and analysis of human speech. Most existing speaker diarization systems rely exclusively on unimodal acoustic information, making the task particularly challenging due to the innate ambiguities of audio signals. Recent studies have made tremendous efforts towards audio-visual or audio-semantic modeling to enhance performance. However, even the incorporation of up to two modalities often falls short in addressing the complexities of spontaneous and unstructured conversations. To exploit more meaningful dialogue patterns, we propose a novel multimodal approach that jointly utilizes audio, visual, and semantic cues to enhance speaker diarization. Our method elegantly formulates the multimodal modeling as a constrained optimization problem. First, we build insights into the visual connections among active speakers and the semantic interactions within spoken content, thereby establishing abundant pairwise constraints. Then we introduce a joint pairwise constraint propagation algorithm to cluster speakers based on these visual and semantic constraints. This integration effectively leverages the complementary strengths of different modalities, refining the affinity estimation between individual speaker embeddings. Extensive experiments conducted on multiple multimodal datasets demonstrate that our approach consistently outperforms state-of-the-art speaker diarization methods.

1 Introduction

In the fields of human-computer interaction (HCI) and human-robot interaction (HRI), addressing multi-party dialogue problem is frequently essential (Ouchi and Tsuboi 2016; Gu et al. 2021). In a conversation situation where multiple speakers are involved, one crucial challenge that must be tackled along with automatic speech recognition (ASR) and natural language processing (NLP) is to correctly recognize and assign temporal segments of speech or transcribed texts to corresponding speakers. In this procedure, human individuals rely on a combination of listening, observing, and understanding to easily pinpoint who is speaking at any given moment. This spontaneous process of deciphering speech underscores a growing desire for machines capable of executing the same task with high accuracy. In the field of speech signal processing, this task is referred to

as Speaker Diarization aiming at determining "who spoke when" within an audio stream (Wang et al. 2018; Zhang et al. 2019; Anguera et al. 2012; Anguera, Wooters, and Hernando 2007; Tranter and Reynolds 2006; Zheng et al. 2021).

Traditional unimodal speaker diarization approaches only rely on acoustic cues to differentiate between speakers (Sell and Garcia-Romero 2014; Park et al. 2020; Landini et al. 2022; Du et al. 2022). These methods often suffer from low-quality acoustic environments, which are typically characterized by the presence of background noise, reverberation, and overlapping speech from multiple speakers (Reynolds and Torres-Carrasquillo 2005; Park et al. 2022). In last decade, much research attempted to investigate joint modeling of multiple modality information from a variety of view points, enhancing the generalization and reliability of speaker diarization. (Chung, Lee, and Han 2019; Xu et al. 2022; Chung et al. 2020; Gebru et al. 2017) exploit the audio-visual association, based on the assumption that a speech signal is typically synchronized with the facial attributes and lip motion of active speakers. (Flemotomos and Narayanan 2022; Park and Georgiou 2018; Zuluaga-Gomez et al. 2022) employed both lexical and acoustic features to identify roles in a specific two-speaker scenario, where the speakers assume distinct roles and are expected to follow different linguistic patterns. (Cheng et al. 2023a,c; Flemotomos, Georgiou, and Narayanan 2020) proposed to develop supplementary language sub-tasks to detect semantic change points within dialogues, aiming to provide guidance for the audio-only diarization process.

While there are some known attempts to jointly model audio-visual or audio-textual information for speaker diarization, there is an obvious absence in addressing the comprehensive utilization of all three modalities, audio, visual, and textual information, in a unified framework. As summarized in Table 1, each modality offers distinct and complementary strengths. Audio signals, being the primary source of speech content, provide direct access to vocal characteristics such as pitch, timbre, and speaking rate, which are essential for speaker identification. Visual cues can assist in distinguishing speakers by capturing unique facial features and tracking lip movements over time in low-quality acoustic environment. Textual data, transcribed from ASR modules, provide rich contextual and semantic content, which can reveal clear linguistic patterns and identify speaker-turns

Table 1: Advantages and disadvantages of different modalities for speaker diarization.

Modality	Advantages	Disadvantages
Audio	Conveys direct voice characteristics. Tracks speaker activity seamlessly.	Vulnerable to environmental interference. Fails in handling simultaneous speech.
Vision	Offers distinctive visual cues. Robust to complicated acoustic conditions.	Cannot handle off-screen voices. Sensitive to camera quality, angle, and distance.
Text	Identifies speaker-turn with semantic breaks. Provides semantic context.	Sensitive to transcription errors. Often presents ambiguity regarding speaker identity.

based on semantic breaks. Concurrently, the inherent limitations of each individual modality also constrain the efficacy of unimodal speaker diarization. We believe that the complementary natures of all three modalities, when carefully combined, can potentially yield a performance leap beyond the sum of their individual contributions. To successfully integrate them under one unified framework, we introduce a clustering method based on constrained optimization. By carefully constructing visual and semantic constraints, we effectively incorporate multimodal information in the joint constraint propagation.

Specifically, we first utilize voice activity detection(VAD) to obtain active speech segments and extract speaker embeddings, similar to the prevalent audio-only speaker diarization systems. Then we introduce additional visual components such as face recognition and lip movement detection to obtain visual connections among speakers, establishing pairwise constraints on visually active speakers. Subsequently, text-based dialogue detection and speaker-turn detection models are used to construct an understanding of semantic structures of the conversations. Finally, a joint pairwise constraint propagation algorithm is introduced to estimate a refined affinity matrix of speaker embeddings and facilitate speakers clustering process. In this framework, multimodal modeling is explicitly formulated as a constrained optimization problem.

To comprehensively evaluate the effectiveness of our method, we conducted experiments using multiple multimodal datasets. These included a video dataset that we collected and manually annotated, as well as several popular open-source datasets. The results indicate a significant performance improvement and a robust generalization capability of our method.

To the best of our knowledge, this paper is the first effort to systematically integrate three modalities - visual, audio, and semantic information to improve the performance of speaker diarization. This study represents contribution to the field of multimodal speaker diarization, enhancing the existing literature with richer modality information. Through this augmented approach, we aim to catalyze subsequent advancements and broaden the scope for future research in the domain.

The main contributions of this paper are as follows:

- We propose a novel framework to incorporate audio, visual and semantic information into speaker diarization, leveraging the complementary strengths of three modalities.

- We introduce a joint pairwise constraint propagation method into the speaker clustering process, effectively enhancing the understanding of conversational structure from distinct modality perspectives.

- We contribute a video evaluation set that contains speaker identity labels, their corresponding speech activity timestamps, and speech content. This addresses the absence of comparable public benchmark for multimodal speaker diarization, providing a much-needed standard for performance assessment.

2 Related Work

2.1 Audio-only Speaker diarization

Audio-only speaker diarization has been studied extensively (Anguera et al. 2012; Park et al. 2022). A typical speaker diarization systems employ a multi-stage framework (Ajmera, Lathoud, and McCowan 2004; Sell and Garcia-Romero 2014; Landini et al. 2022; Park et al. 2020; Zheng and Suo 2022), including voice activity detection (VAD) (Gelly and Gauvain 2018), speech segmentation (Xia et al. 2022), acoustic embedding extraction (Sell and Garcia-Romero 2014; Snyder et al. 2018; Yu et al. 2021; Zheng, Lei, and Suo 2020; Chen et al. 2023) and clustering (Von Luxburg 2007; Landini et al. 2022). Recently, end-to-end neural diarization (EEND) where individual submodules in traditional systems can be replaced by one neural network has received more attention with promising results (Fujita et al. 2019a,b; Horiguchi et al. 2020). Due to lack of available large-scale data, EEND is usually trained on simulated dataset and suffers from generation issue in real-word applications. The transformer-based architecture of encoders and decoders causes computational inefficiencies when processing long audio sequences. In general, it tends to be used combining with clustering-based diarization in most methods (Kinoshita, Delcroix, and Tawara 2021b,a).

2.2 Audio-visual Speaker diarization

Facial activities and lip motion are highly related to speech (Yehia, Rubin, and Vatikiotis-Bateson 1998). Visual information contains a strong clue for the identification of speakers and the location of speaker changes (Yoshioka et al. 2019), which can be used to significantly improve the accuracy of speaker diarization. Most methods leverage the audio and visual cues for diarization using synchronization between talking faces and voice tracks (Ahmad et al. 2019; Chung, Chung, and Kang 2019; Xu et al. 2022). Another
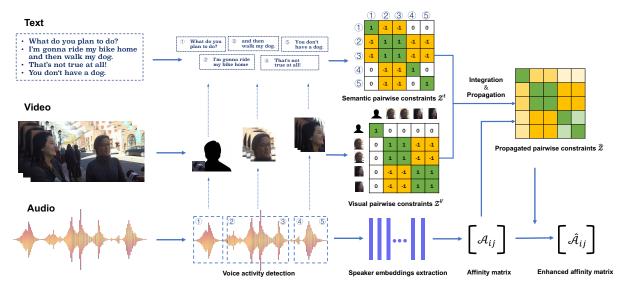
Figure 1: An overview of our proposed multimodal speaker diarization system. It incorporates additional visual and textual processing modules that independently extract visual and semantic constraints. By integrating and propagating knowledge derived from these different insights, comprehensive multimodal pairwise constraints are generated, serving as a robust guidance for enhancing the audio-based diarization.

work (Chung, Lee, and Han 2019) adopted a synchronisation network to get self-enrolled speaker profiles and reformulate the task into a supervised classification problem. Recently, an interesting and promising direction is to use separate neural networks to process data streams of two modalities and directly output speech probabilities for all speakers simultaneously (kui He, Du, and Lee 2022), similar to audio-only EEND frameworks.

2.3 Audio-textual Speaker diarization

Some previous works (Zuluaga-Gomez et al. 2022; Flemotomos and Narayanan 2022; Park and Georgiou 2018; Paturi, Srinivasan, and Li 2023) utilized semantic information derived from transcription to estimate the role profiles and detect speaker change point, demonstrating improvement in specific role-playing conversations, such as job interviews and doctor-patient medical consultations. Other works (Kanda et al. 2021; Xia et al. 2022; Khare et al. 2022) enhanced ASR models to capture speaker identity through joint training of paired audio and textual data, which typically require substantial annotated multi-speaker speech data. More recent works (Park et al. 2023; Wang et al. 2024; Cheng et al. 2023a) employed large language models as post-processing to correct word speaker-related boundaries according to local semantic context.

2.4 Pairwise Constrained Clustering

As mentioned in Section 2.1, speaker diarization systems often necessitate the introduction of an unsupervised speaker clustering algorithm, owing to the need to handle an indeterminate number of speakers. When introducing multimodal information, due to the inability to directly compare cross-modal information for similarity, it becomes essential to incorporate other modal information as weakly supervised sig-

nals into the speaker clustering; this process is known as constrained clustering (Bibi, Alqahtani, and Ghanem 2023).

Pairwise constrained clustering is one such classic methodology. Pairwise constraints (must-link and cannot-link) are well-studied and they provide the capability to define any ground truth set partitions (Davidson and Ravi 2007). Since weakly supervised signals often do not encompass all target samples for clustering, various pairwise constraint propagation methods (Lu and Peng 2011) have been proposed to increase the number of pairwise constraints from a limited number of initial ones. Initially confined to data mining datasets (Hoi, Jin, and Lyu 2007), the application of pairwise constrained clustering has expanded into multimodal areas such as vision and text (Yang et al. 2014; Yan et al. 2006). Advancing with theoretical progress, pairwise constraint propagation algorithms have increasingly integrated complex optimization techniques, including Non-negative Matrix Factorization (NMF)(Fu 2015), Inexact Augmented Lagrange Multiplier (IALM)(Liu et al. 2019), and deep learning outcomes (Zhang et al. 2021a,b).

3 Methods

The proposed framework's overview is depicted in Figure 1. We will introduce joint pairwise constraint propagation in Sec. 3.1, visual constraints construction in Sec. 3.2, and semantic constraints construction in Sec. 3.3, separately.

3.1 Joint Pairwise Constraint Propagation with multimodal Information

Considering that the audio contains comprehensive speaker-related information over time, we employ audio-based models, specifically a VAD model and a speaker embedding extractor, to obtain a sequence of acoustic speaker embeddings

$E = \{e_1, e_2, ..., e_N | e_i \in \mathbb{R}^D\}$, by applying sliding windows to the audio data. Subsequently, we compute the affinity matrix $\mathcal{A} = \{\mathcal{A}_{ij}\}_{N \times N}$, where $\mathcal{A}_{ij} = g(e_i, e_j)$ and $g(\cdot)$ represents the similarity measurement. It should be noted that this affinity matrix may contain errors resulting from acoustic environmental interference.

Assuming we have access to speaker-related cues from additional sources of information, we can derive various types of constraint pairs: must-link $\mathcal{M}$ and cannot-link $\mathcal{C}$, defined as:

$$\mathcal{M}^k = \{(e_i, e_j) | l(e_i) = l(e_j)\},$$
$$\mathcal{C}^k = \{(e_i, e_j) | l(e_i) \neq l(e_j)\}, \quad (1)$$

where $l(\cdot)$ denotes the speaker label associated with an acoustic speaker embedding, and k is the index of sources type. For different modality information, the criteria for establishing $\mathcal{M}$ and $\mathcal{C}$ is different, which will be described in Sec. 3.2 and Sec. 3.3 according to specific situation. Then each constraint is initially encoded into a matrix $\mathcal{Z}^k$:

$$\mathcal{Z}^k_{ij} = \begin{cases} +1 & \text{if } (e_i, e_j) \in \mathcal{M}^k, \\ -1 & \text{if } (e_i, e_j) \in \mathcal{C}^k, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

A series of constraint matrix $\mathcal{Z}^k$ are integrated into a final constraint matrix $\mathcal{Z}$. During the integration process, some scenarios are relatively straightforward. For instance, if an embedding pair (e_i, e_j) belongs to $\bigcap_k \mathcal{M}^k$, then (e_i, e_j) is considered as a must-link constraint pair. Conversely, if (e_i, e_j) resides in $\bigcap_k \mathcal{C}^k$, it is a cannot-link constraint pair due to agreement between all modalities. However, there are evidently more complex scenarios, where the constraint matrices conflict with one another, such as $(e_i, e_j) \in (\mathcal{M}^1 \cap \mathcal{C}^2)$ or $(e_i, e_j) \in (\mathcal{M}^2 \cap \mathcal{C}^1)$. To address these issues, we introduce acoustic information as the arbiter in the final determination. To summarize, we compute the integrated constraint scores following the given formula:

$$\mathcal{Z}' = \sum_k \alpha_k \mathcal{Z}^k + \beta \mathcal{A} - \theta \quad (3)$$

where α_k, β represent the weight hyper-parameters for different modalities, and θ is the bias. Then, $\mathcal{Z}'$ is converted into a binarized constraint matrix $\mathcal{Z}$ according to a threshold δ.

$$\mathcal{Z}_{ij} = \begin{cases} +1 & \text{if } \mathcal{Z}'_{ij} > \delta, \\ -1 & \text{if } \mathcal{Z}'_{ij} < -\delta, \\ 0 & \text{else.} \end{cases} \quad (4)$$

The constraint matrix $\mathcal{Z}$ may be sparse. Constraint information is confined to discrete and localized regions. It is essential to deploy a constraint propagation algorithm to efficiently broadcast the constraint information in $\mathcal{Z}$ on a larger scale. Specifically, we employ E2CP (Lu and Peng 2011) algorithm to obtain propagated constraints $\hat{\mathcal{Z}}$:

$$\hat{\mathcal{Z}} = (1 - \lambda)^2 (\mathbf{I} - \lambda \mathbf{L}_e)^{-1} \mathcal{Z} (\mathbf{I} - \lambda \mathbf{L}_e)^{-1}, \quad (5)$$

where $\mathbf{L}_e = \mathbf{D}_e^{-1/2} \mathcal{A} \mathbf{D}_e^{-1/2}$ is the normalized Laplacian matrix, and $\mathbf{D}_e$ is the degree matrix of $\mathcal{A}$ and $\mathbf{I}$ is a identity matrix. The parameter $\lambda \in [0, 1]$ modulates the impact degree of the constraints. The refined affinity matrix

$\hat{\mathcal{A}} \in \mathbb{R}^{N \times N}$ is then updated to incorporate the influences of the propagated constraints $\hat{\mathcal{Z}}$:

$$\hat{\mathcal{A}}_{ij} = \begin{cases} 1 - (1 - \hat{\mathcal{Z}}_{ij})(1 - \mathcal{A}_{ij}) & \text{if } \hat{\mathcal{Z}}_{ij} \geq 0, \\ (1 + \hat{\mathcal{Z}}_{ij}) \mathcal{A}_{ij} & \text{if } \hat{\mathcal{Z}}_{ij} < 0. \end{cases} \quad (6)$$

Upon calculating the affinity matrix $\hat{\mathcal{A}}$, it is then fed into the clustering process to derive the ultimate speaker diarization results. It is worth noting that there is no limit to the number of constraint types k. We can extract diverse constraint matrices related to different modal data. These constraint matrices can be perceived as prior knowledge, directing the clustering attention toward a particular view of the data. In this paper, we fix k at 2, thereby extracting two distinct constraint types: visual constraint $\mathcal{Z}^v$ and semantic constraint $\mathcal{Z}^t$.

3.2 Visual constraints construction

The speaker-related visual constraints is constructed through the following steps, similar to (Chung et al. 2020; Xu et al. 2022).

Face tracking. The initial phase involves detecting and tracking faces within the video content over time. A CNN-based face detector (Liu, Huang, and Wang 2018) is used to continuously locate faces across frames, thereby creating a consistent track for each face present, using a position-based tracker. Only those face tracks that correspond with audible speech segments, as identified by VAD, are retained for further processing.

Active speaker detection. This step takes the cropped face video and corresponding audio as input and decides whether the tracked faces correspond to active speakers at any given moment. To achieve this, a two-stream network (Tao et al. 2021), consisting of temporal encoders and an attention-based decoder, is used to incorporate audio-visual synchrony and determine target speaker activity in the audio stream. In order to ensure the effectiveness of the subsequent process, we established a threshold to filter out video frames with relatively low confidence levels.

Face clustering. A face recognition CNN (Huang et al. 2020) is utilized to extract embeddings for every face track. The embeddings are extracted at uniform intervals, such as every 200 ms, in each face track. These are then clustered using Agglomerative Hierarchical Clustering (AHC).

By integrating these steps, constraints based on visual information are obtained. Faces clustered to the same speaker are considered as must-link constraints, while those clustered to different speakers are cannot-link constraints. Each face is aligned with respective acoustic embeddings e_i along the time axis. If an acoustic embedding corresponds to multiple faces, we will select the speaker associated with the majority of those faces. As a result, the obtained visual constraint $\mathcal{Z}^v$ is consistent with the shape of the acoustic affinity $\mathcal{A}$.

3.3 Semantic constraints construction

To extract speaker-related information from the transcriptions, we constructed two Spoken Language Processing
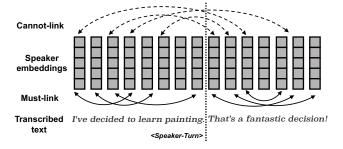
Figure 2: Semantic constraint construction based on dialogue detection and speaker-turn detection. Text segments judged as non-dialogue indicate that the associated embeddings are related through must-link constraints, depicted by solid connections below. Conversely, a detected transition point dictates that embeddings spanning this point should be connected with cannot-link constraints, as represented by dashed connections above.

(SLP) tasks: (1) **Dialogue Detection** discriminates between multi-speaker dialogues and monologues, conceptualized as a binary classification challenge. (2) **Speaker-Turn Detection** assesses each sentence in a sequence to estimate speaker change, functioning as a sequence labeling problem that identifies semantically significant speaker role transitions. Semantic constraints can be formulated based on the outputs of these two tasks. Specifically, must-link $\mathcal{M}^t$ is formed between two embeddings if they are sourced from the same non-dialogue segment. Conversely, cannot-link $\mathcal{C}^t$ is established between embeddings separated by a detected speaker-turn boundary, as illustrated in Figure 2.

There is an inherent transitivity associated with must-link constraints. That is, if $(e_i, e_j) \in \mathcal{M}^t$ and $(e_j, e_k) \in \mathcal{M}^t$, then it can be inferred that $(e_i, e_k) \in \mathcal{M}^t$. Unfortunately, such a property does not extend to cannot-link constraints. If $(e_i, e_j) \in \mathcal{C}^t$ and $(e_j, e_k) \in \mathcal{C}^t$, we cannot ascertain the relationship between e_i and e_k, as in a real meeting scenario. Following a dialogue between speakers A and B, either speaker C may begin speaking, or speaker A may continue. Hence, semantic constraints derived solely from aforementioned methods can only determine the relationship between embeddings adjacent to a speaker-turn, and cannot influence embeddings separated by extended temporal intervals.

4 Experiments

4.1 Datasets

To evaluate our methods, we have constructed a diverse evaluation video dataset sourced from in-the-wild scenarios. This dataset includes dialogues featuring 2 to 10 English-speaking participants from interviews, talk shows, meetings, press conferences, round-table discussions, and TV shows, totaling approximately 6.3 hours of content. Individual videos range from 7 to 29 minutes and have been meticulously annotated with speaker identities, speech activity timestamps, and content. We intend to release this dataset publicly to advance multimodal research.

In addition to our evaluation dataset, we conducted supplementary experiments using three public multimodal datasets: AVA-AVD (Xu et al. 2022), AIShell-4 (Fu et al. 2021), and Alimeeting (Yu et al. 2022). The AVA-AVD dataset, which focuses on the multimodal analysis of audio-visual diarization, includes over six languages and provides rich scenarios and face annotations; however, it lacks ground truth transcripts. In contrast, AIShell-4 and Alimeeting are both Mandarin datasets that offer speaker-labeled transcripts, making them particularly suited for various speech tasks. The combination of these varied datasets further validates the effectiveness and applicability of our methods across different domains.

4.2 Implementation Details

In our system, the audio-based diarization modules follow the pipeline outlined in (Cheng et al. 2023a). Our speaker embedding extractor is an adaptation of CAM++ (Wang et al. 2023)[1], which has been trained on VoxCeleb dataset (Nagrani et al. 2020). For the visual componets, we employ a series of pretrained models for different tasks: RFB-Net (Liu, Huang, and Wang 2018)[2] for face detection, TalkNet (Tao et al. 2021)[3] for active speaker detection, and CurricularFace model (Huang et al. 2020)[4] for extracting face embeddings. To transcribe audio into text, we utilize the ASR model, Paraformer (Gao et al. 2022), which has been trained with the aid of the FunASR (Gao et al. 2023) toolkits[5]. These off-the-shelf pretrained models, crucial to our system, are all accessed through open-source platforms.

For semantic tasks, we pretrained muliple models with open-source meeting datasets for different scenarios. Specifically, we employed AMI (Carletta et al. 2005), ICSI (Janin et al. 2003) and CHiME-6 (Watanabe et al. 2020) to generate English semantic models, and used Alimeeting and AIShell-4 training datasets to obtain Mandarin semantic models. In our experiments, a sliding window strategy was employed, featuring a window size of 96 words and a shift of 16 words, to construct training sets for dialogue detection and speaker-turn detection training from transcripts with speaker annotations within these datasets. All that training was conducted using a pre-trained BERT model (Devlin et al. 2019). Subsequently, we employed the methods described in Section 3.3 to construct the semantic constraints.

The VBx approach (Landini et al. 2022) represents a canonical method for speaker diarization, where the original paper employs speaker embeddings based on the x-vector model. We substituted this with the more robust CAM++ model. Furthermore, given that the post-processing

---

[1]The pretrained CAM++ came from https://github.com/modelscope/3D-Speaker

[2]The pretrained RFB-Net came from https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB

[3]The pretrained TalkNet came from https://github.com/TaoRuijie/TalkNet-ASD

[4]The pretrained CurricularFace model came from https://modelscope.cn/models/iic/cv_ir101_facerecognition_cfglint

[5]The ASR and forced-alignment models came from https://github.com/modelscope/FunASR

Table 2: The results of speaker diarization with multimodal constraints

Cluster Algorithm	Constraints	Cluster Metrics			Speaker Metrics(%)		
		ARI↑	NMI↑	DER↓	JER↓	TextDER↓	CpWER↓
VBx	No Constraints	0.903	0.893	10.31	29.28	4.23	18.03
SC	No Constraints	0.918	0.899	9.37	27.21	3.25	17.04
E2CP + SC	Semantic Constraints	0.924	0.904	9.12	25.98	3.02	16.86
E2CP + SC	Visual Constraints	0.924	0.905	9.13	26.02	3.02	16.83
E2CP + SC	Semantic + Visual Constraints	**0.925**	**0.908**	**9.01**	**22.57**	**2.89**	**16.36**

Table 3: Constraints derived from various modalities. We separately evaluated the accuracy and coverage of these constraints.

Constraints	Accuracy(%)			Coverage(%)		
	Must-Link	Cannot-Link	Total	Must-Link	Cannot-Link	Total
Semantic Constraints	99.75	84.80	99.40	1.23	0.08	0.49
Visual Constraints	99.07	97.87	99.32	22.81	21.78	22.53
Semantic + Visual Constraints	99.11	97.83	99.34	23.65	21.84	22.87

step of the E2CP method integrates spectral clustering (SC) (Von Luxburg 2007), we also explored the performance of a method utilizing solely speaker embeddings and SC. The aforementioned two audio-only methods will serve as the baselines for this study.

As introduced in Section 3.1, after obtaining multimodal pairwise constraints, our clustering process is divided into two submodules: constraint propagation and post-cluster. We employ E2CP as the core algorithm for constraint propagation. When only visual constraints are present, the parameter λ in E2CP is set to 0.8, while it is set to 0.95 when semantic constraints are incorporated. For the post-cluster phase, we adhered to the SC algorithm consistent with the baseline. Our method, inspired by the work presented in the paper (Park et al. 2020), incorporates refinement operations such as row-wise thresholding and symmetrization to enhance the performance of spectral clustering. In the row-wise thresholding of SC, the p-percentile parameter is set to 0.982. To account for the randomness of the clustering algorithm, our reported clustering metric is the average of 10 repetitions.

4.3 Evaluation Metrics

We report two popular clustering algorithm metrics: Normalized Mutual Information (NMI) (Strehl and Ghosh 2002) and Adjusted Rand Index (ARI) (Chac'on and Rastrojo 2020). To demonstrate the impact of the speaker diarization system, we will also report the Diarization Error Rate (DER) (Fiscus et al. 2006) with tolerance 0.25s and Jaccard Error Rate (JER) (Ryant et al. 2019). The DER is generally composed of three parts: DER = Missed Speech (MS) + False Alarms (FA) + Speaker Error (SPKE). As the transcribed text and forced-alignment module have been used in the pipeline, we directly report the Concatenated Minimum-permutation Word Error Rate (Watanabe et al. 2020). Additionally, we use the metric Text Diarization Error Rate (TextDER) (Gong, Wu, and Choi 2023), to evaluate the amount of text assigned to wrong speakers.

5 Results and Discussion

5.1 Results of Speaker Diarization

The results of speaker diarization in Table 2 show that adding either semantic or visual constraints individually in E2CP + SC system can lead to improvements in cluster and diarization metrics. With semantic constraints, the JER metric decreases from 27.21% to 25.98%, and the DER metric also shows a reduction from 9.37% to 9.12%, compared to the baseline SC's DER. The integration of visual constraints improves cluster precision, as reflected by an improved NMI of 0.905.

The combination of semantic and visual constraints in E2CP + SC achieves the best performance across all evaluated metrics. It demonstrates a DER of 9.01% and an NMI of 0.908, showcasing the synergistic effect of combining both types of constraints. TextDER and CpWER, which are from a textual perspective, also see significant improvement with the combined semantic and visual constraints. TextDER decreases from 3.25% to 2.89%, and CpWER from 17.04% to 16.36%. In our experiments, we keep the transcripts fixed, so the enhancement in CpWER is entirely attributable to the improvement of speaker diarization performance.

While relying solely on semantic constraints may appear to lag behind the use of visual constraints in clustering metrics, it has certain advantages in Speaker Metrics. Semantic constraints are often located near speaker change points, which can have a more substantial effect on the final diarization outcome, particularly for speaker identification within transcript text. On the other hand, visual information might suffer from asynchronicity during speaker change points, which can impede accurate speaker boundary determination. Therefore, the benefits of semantic information in adjudicating boundaries do not manifest in clustering metrics but confer a definitive advantage in speaker metrics.

5.2 Constraint Construction and Analysis

It is important to note that both visual and semantic decoding methods introduce some level of error, and the constraints constructed from these modalities may cover different sets

Table 4: The results of audio-visual speaker diarization experiments on AVA-AVD datasets.

	Modality	Methods	VAD	SPKE(%)↓	DER(%)↓
AVR-Net	Audio	VBx	Oracle	18.45	21.37
	Audio + Visual	AVA-AVD	Oracle	17.65	20.57
Ours	Audio	SC	Oracle	18.39	21.31
	Audio + Visual	E2CP + SC	Oracle	**17.40**	**20.32**

Table 5: The results of audio-text speaker diarization experiments on AIShell-4 and Alimeeting Datasets.

Dataset	Modality	Methods	CpWER(%)↓	TextDER(%)↓
AIShell-4	Audio	SC	17.31	5.97
	Audio + Semantic	Fusion	15.23	6.28
	Audio + Semantic	Ours	**14.95**	**4.98**
Alimeeting	Audio	SC	41.67	18.89
	Audio + Semantic	Fusion	36.15	14.50
	Audio + Semantic	Ours	**31.11**	**10.76**

of embedding pairs, so we need to examine the impact of these factors on the outcome. Table 3 presents the accuracy and coverage rates of constraints generated from both visual and semantic modalities.

It is evident that visual constraints outperform semantic constraints in terms of coverage. This difference can be attributed to the fact that the semantic tasks used in our model only evaluate relationships between embeddings within adjacent speaker turns, whereas visual constraints are assessed across embedding pairs with substantial temporal intervals. Furthermore, we have designed a method to effectively combine constraints from different modalities, and our findings indicate that semantic constraints provide a valuable supplement to visual constraints. In addition, we have conducted experiments in Appendix A to discuss the impact of constraints' quality and quantity on the results.

5.3 Constraint Propagation Parameter

As mentioned in Section 3.1, λ is a critical parameter during the constraint propagation process. By combining the analysis of Equations 5 and 6, it can be found that when λ tends towards 0, the final $\hat{\mathcal{Z}}$ will be closer to $\mathcal{Z}$, whereas when λ approaches 1, the resulting $\hat{\mathcal{A}}$ will be closer to $\mathcal{A}$.

Moreover, the parameter λ also signifies the level of confidence that the model places in the constraints matrix. By adjusting the λ value, the model can effectively handle different levels of error in the constraints, enabling the constrained propagation algorithm to adapt to models with varying performance. This adaptability is essential for effectively utilizing constraints in real-world scenarios.

For additional details and insights on the impact of different λ values on the algorithm's performance, please refer to the experiments illustrated in the Appendix B.

5.4 Compare with Audio-visual Diarization

In this section, we will demonstrate the effectiveness of our method on the audio-visual speaker diarization using the AVA-AVD (Xu et al. 2022) dataset. Due to the lack of annotated transcripts in the AVA-AVD dataset, in accordance with the original paper, only the SPKE and DER metrics are calculated.

The method proposed in (Xu et al. 2022) involves training an audio-visual relation network (AVR-Net) to simultaneously model audio and visual information. This approach requires a substantial amount of aligned audio-visual paired data for training. In contrast, our method utilizes a pretrained visual model to directly construct visual constraints that refine the affinity matrix from acoustic embeddings. Table 4 presents a comparison conducted on AVA-AVD, revealing the competitive performance of our method when utilizing

only visual constraints. This demonstrates strong generalizability of our approach.

5.5 Compare with Audio-textual Diarization

In this section, we will compare our method with existing audio-text speaker diarization approaches, specifically evaluating our performance against the system presented in (Cheng et al. 2023b) using the AIShell-4 and Alimeeting datasets. Since both datasets are in Mandarin, we utilize the open-sourced speaker embedding model in the same manner as their work for a fair comparison. To mitigate the influence of the ASR system, our experiments will focus on the results decoded from the ground-truth annotated text, as described in their work. Consistent with their findings, we will report the metrics CpWER and TextDER.

The method presented in (Cheng et al. 2023b) integrates acoustic and semantic information to address boundary issues in acoustic-only systems. In contrast, our approach uses semantic information to create pairwise constraints that directly impact spectral clustering. Table 5 presents the experimental results of our audio-text speaker diarization method, showing significant performance improvements on both the AIShell-4 and AliMeeting datasets. Specifically, the CpWER on the AIShell-4 dataset decreased from 15.23% to 14.95%, and on the AliMeeting dataset, it reduced from 36.15% to 31.11%. These results indicate that our method effectively utilizes semantic information compared to the approach in (Cheng et al. 2023b).

6 Conclusions

In this study, we propose a novel multimodal approach that jointly leverages audio, visual, and semantic information for enhanced speaker diarization. Additional visual and textual processing modules are incorporated to generate complementary visual and semantic constraints. A joint pairwise constraint propagation method is employed to integrate multimodal information into the speaker clustering process. Experimental results confirm the significant improvement in diarization performance.

We posit that our study represents a significant advancement in the field of multimodal speaker diarization. By incorporating an augmented array of modal information, we provide a framework that not only enriches the current understanding of speaker diarization processes but also catalyzes further innovation and exploration within this domain. Our work sets the stage for the development of more sophisticated systems that can accurately parse and attribute speaker identity, thereby broadening the horizon for future research endeavors.

References

Ahmad, R.; Zubair, S.; Alquhayz, H. A.; and Ditta, A. 2019. Multimodal Speaker Diarization Using a Pre-Trained Audio-Visual Synchronization Model. *Sensors (Basel, Switzerland)*, 19.

Ajmera, J.; Lathoud, G.; and McCowan, L. 2004. Clustering and segmenting speakers and their locations in meetings. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, I–605.

Anguera, X.; Bozonnet, S.; Evans, N.; Fredouille, C.; Friedland, G.; and Vinyals, O. 2012. Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2): 356–370.

Anguera, X.; Wooters, C.; and Hernando, J. 2007. Acoustic Beamforming for Speaker Diarization of Meetings. *IEEE Trans. Speech Audio Process.*, 15(7): 2011–2022.

Bibi, A.; Alqahtani, A.; and Ghanem, B. 2023. Constrained Clustering: General Pairwise and Cardinality Constraints. *IEEE Access*, 11: 5824–5836.

Carletta, J.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V.; Kraaij, W.; Kronenthal, M.; Lathoud, G.; Lincoln, M.; Lisowska, A.; McCowan, I.; Post, W.; Reidsma, D.; and Wellner, P. 2005. The AMI Meeting Corpus: A Pre-announcement. In Renals, S.; and Bengio, S., eds., *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, volume 3869 of *Lecture Notes in Computer Science*, 28–39. Springer.

Chac'on, J. E.; and Rastrojo, A. I. 2020. Minimum adjusted Rand index for two clusterings of a given size. *Advances in Data Analysis and Classification*, 17: 125–133.

Chen, Y.; Zheng, S.; Wang, H.; Cheng, L.; Chen, Q.; and Qi, J. 2023. An Enhanced Res2Net with Local and Global Feature Fusion for Speaker Verification. *CoRR*, abs/2305.12838.

Cheng, L.; Zheng, S.; Qinglin, Z.; Wang, H.; Chen, Y.; and Chen, Q. 2023a. Exploring Speaker-Related Information in Spoken Language Understanding for Better Speaker Diarization. In *Annual Meeting of the Association for Computational Linguistics*.

Cheng, L.; Zheng, S.; Zhang, Q.; Wang, H.; Chen, Y.; and Chen, Q. 2023b. Exploring Speaker-Related Information in Spoken Language Understanding for Better Speaker Diarization. In *Findings of the ACL 2023, Toronto, Canada, July 9-14, 2023*, 14068–14077.

Cheng, L.; Zheng, S.; Zhang, Q.; Wang, H.; Chen, Y.; Chen, Q.; and Zhang, S. 2023c. Improving Speaker Diarization using Semantic Information: Joint Pairwise Constraints Propagation. *CoRR*, abs/2309.10456.

Chung, J. S.; Huh, J.; Nagrani, A.; Afouras, T.; and Zisserman, A. 2020. Spot the conversation: speaker diarisation in the wild. In *Interspeech*.

Chung, J. S.; Lee, B.-J.; and Han, I. 2019. Who said that?: Audio-visual speaker diarisation of real-world meetings. In *Interspeech*.

Chung, S.; Chung, J. S.; and Kang, H. 2019. Perfect Match: Improved Cross-modal Embeddings for Audio-visual Synchronisation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, 3965–3969. IEEE.

Davidson, I.; and Ravi, S. S. 2007. Intractability and clustering with constraints. In Ghahramani, Z., ed., *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, 201–208. ACM.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

Du, Z.; Zhang, S.; Zheng, S.; and Yan, Z. 2022. Speaker Overlap-aware Neural Diarization for Multi-party Meeting Analysis. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 7458–7469. Association for Computational Linguistics.

Fiscus, J. G.; Ajot, J.; Michel, M.; and Garofolo, J. S. 2006. The Rich Transcription 2006 Spring Meeting Recognition Evaluation. In Renals, S.; Bengio, S.; and Fiscus, J. G., eds., *Machine Learning for Multimodal Interaction, Third International Workshop, MLMI 2006, Bethesda, MD, USA, May 1-4, 2006, Revised Selected Papers*, volume 4299 of *Lecture Notes in Computer Science*, 309–322. Springer.

Flemotomos, N.; Georgiou, P.; and Narayanan, S. 2020. Linguistically Aided Speaker Diarization Using Speaker Role Information. In *Proceedings of Odyssey*, 117–124.

Flemotomos, N.; and Narayanan, S. 2022. Multimodal clustering with role induced constraints for speaker diarization.

Fu, Y.; Cheng, L.; Lv, S.; Jv, Y.; Kong, Y.; Chen, Z.; Hu, Y.; Xie, L.; Wu, J.; Bu, H.; Xu, X.; Du, J.; and Chen, J. 2021. AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, 3665–3669. ISCA.

Fu, Z. 2015. Pairwise constraint propagation via low-rank matrix recovery. *Comput. Vis. Media*, 1(3): 211–220.

Fujita, Y.; Kanda, N.; Horiguchi, S.; Nagamatsu, K.; and Watanabe, S. 2019a. End-to-End Neural Speaker Diarization with Permutation-free Objectives. In *Interspeech*, 4300–4304.

Fujita, Y.; Kanda, N.; Horiguchi, S.; Xue, Y.; Nagamatsu, K.; and Watanabe, S. 2019b. End-to-End Neural Speaker Diarization with Self-Attention. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 296–303.

Gao, Z.; Li, Z.; Wang, J.; Luo, H.; Shi, X.; Chen, M.; Li, Y.; Zuo, L.; Du, Z.; Xiao, Z.; and Zhang, S. 2023. FunASR: A Fundamental End-to-End Speech Recognition Toolkit. *ArXiv*, abs/2305.11013.

Gao, Z.; Zhang, S.; Mcloughlin, I.; and Yan, Z. 2022. Paraformer: Fast and Accurate Parallel Transformer for Non-autoregressive End-to-End Speech Recognition. In *Interspeech*.

Gebru, I. D.; Ba, S.; Li, X.; and Horaud, R. 2017. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5): 1086–1099.

Gelly, G.; and Gauvain, J.-L. 2018. Optimization of RNN-Based Speech Activity Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3): 646–656.

Gong, C.; Wu, P.; and Choi, J. D. 2023. Aligning Speakers: Evaluating and Visualizing Text-based Speaker Diarization Using Efficient Multiple Sequence Alignment. *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, 778–783.

Gu, J.; Tao, C.; Ling, Z.; Xu, C.; Geng, X.; and Jiang, D. 2021. MPC-BERT: A Pre-Trained Language Model for Multi-Party Conversation Understanding. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 3682–3692. Association for Computational Linguistics.

Hoi, S. C. H.; Jin, R.; and Lyu, M. R. 2007. Learning nonparametric kernel matrices from pairwise constraints. In Ghahramani, Z., ed., *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, 361–368. ACM.

Horiguchi, S.; Fujita, Y.; Watanabe, S.; Xue, Y.; and Nagamatsu, K. 2020. End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors. In *Interspeech*.

Huang, Y.; Wang, Y.; Tai, Y.; Liu, X.; Shen, P.; Li, S.; Li, J.; and Huang, F. 2020. CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 5900–5909. Computer Vision Foundation / IEEE.

Janin, A. L.; Baron, D.; Edwards, J.; Ellis, D. P. W.; Gelbart, D.; Morgan, N.; Peskin, B.; Pfau, T.; Shriberg, E.; Stolcke, A.; and Wooters, C. 2003. The ICSI Meeting Corpus. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, 1: I–I.

Kanda, N.; Xiao, X.; Gaur, Y.; Wang, X.; Meng, Z.; Chen, Z.; and Yoshioka, T. 2021. Transcribe-to-Diarize: Neural Speaker Diarization for Unlimited Number of Speakers Using End-to-End Speaker-Attributed ASR. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8082–8086.

Khare, A.; Han, E.; Yang, Y.; and Stolcke, A. 2022. ASR-Aware End-to-End Neural Diarization. *ICASSP 2022*, 8092–8096.

Kinoshita, K.; Delcroix, M.; and Tawara, N. 2021a. Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech. In *Interspeech*.

Kinoshita, K.; Delcroix, M.; and Tawara, N. 2021b. Integrating End-to-End Neural and Clustering-Based Diarization: Getting the Best of Both Worlds. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7198–7202.

kui He, M.; Du, J.; and Lee, C.-H. 2022. End-to-End Audio-Visual Neural Speaker Diarization. In *Interspeech*.

Landini, F.; Profant, J.; Diez, M.; and Burget, L. 2022. Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks. *Computer Speech & Language*, 71: 101254.

Liu, H.; Jia, Y.; Hou, J.; and Zhang, Q. 2019. Imbalance-aware Pairwise Constraint Propagation. In Amsaleg, L.; Huet, B.; Larson, M. A.; Gravier, G.; Hung, H.; Ngo, C.; and Ooi, W. T., eds., *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, 1605–1613. ACM.

Liu, S.; Huang, D.; and Wang, a. 2018. Receptive Field Block Net for Accurate and Fast Object Detection. In *The European Conference on Computer Vision (ECCV)*.

Lu, Z.; and Peng, Y. 2011. Exhaustive and Efficient Constraint Propagation: A Graph-Based Learning Approach and Its Applications. *International Journal of Computer Vision*, 103: 306–325.

Nagrani, A.; Chung, J. S.; Xie, W.; and Zisserman, A. 2020. Voxceleb: Large-scale speaker verification in the wild. *Comput. Speech Lang.*, 60.

Ouchi, H.; and Tsuboi, Y. 2016. Addressee and Response Selection for Multi-Party Conversation. In Su, J.; Carreras, X.; and Duh, K., eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2133–2143. The Association for Computational Linguistics.

Park, T. J.; Dhawan, K.; Koluguri, N. R.; and Balam, J. 2023. Enhancing Speaker Diarization with Large Language Models: A Contextual Beam Search Approach. *ArXiv*, abs/2309.05248.

Park, T. J.; and Georgiou, P. G. 2018. Multimodal Speaker Segmentation and Diarization using Lexical and Acoustic Cues via Sequence to Sequence Neural Networks. In *Interspeech*.

Park, T. J.; Han, K. J.; Kumar, M.; and Narayanan, S. S. 2020. Auto-Tuning Spectral Clustering for Speaker Diarization Using Normalized Maximum Eigengap. *IEEE Signal Processing Letters*, 27: 381–385.

Park, T. J.; Kanda, N.; Dimitriadis, D.; Han, K. J.; Watanabe, S.; and Narayanan, S. 2022. A review of speaker diarization: Recent advances with deep learning. *Comput. Speech Lang.*, 72(C).

Paturi, R.; Srinivasan, S.; and Li, X. 2023. Lexical Speaker Error Correction: Leveraging Language Models for Speaker Diarization Error Correction. *ArXiv*, abs/2306.09313.

Reynolds, D. A.; and Torres-Carrasquillo, P. A. 2005. Approaches and applications of audio diarization. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005*, 953–956. IEEE.

Ryant, N.; Church, K. W.; Cieri, C.; Cristia, A.; Du, J.; Ganapathy, S.; and Liberman, M. Y. 2019. The Second DIHARD Diarization Challenge: Dataset, task, and baselines. In *Interspeech*.

Sell, G.; and Garcia-Romero, D. 2014. Speaker diarization with plda i-vector scoring and unsupervised calibration. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, 413–417.

Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; and Khudanpur, S. 2018. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333.

Strehl, A.; and Ghosh, J. 2002. Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn. Res.*, 3: 583–617.

Tao, R.; Pan, Z.; Das, R. K.; and et al. 2021. Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3927–3935.

Tranter, S. E.; and Reynolds, D. A. 2006. An overview of automatic speaker diarization systems. *IEEE Trans. Speech Audio Process.*, 14(5): 1557–1565.

Von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17: 395–416.

Wang, H.; Zheng, S.; Chen, Y.; Cheng, L.; and Chen, Q. 2023. CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking.

Wang, Q.; Downey, C.; Wan, L.; Mansfield, P. A.; and López-Moreno, I. 2018. Speaker Diarization with LSTM. 5239–5243.

Wang, Q.; Huang, Y.; Zhao, G.; Clark, E.; Xia, W.; and Liao, H. 2024. DiarizationLM: Speaker Diarization Post-Processing with Large Language Models. *ArXiv*, abs/2401.03506.

Watanabe, S.; Mandel, M.; Barker, J.; and Vincent, E. 2020. CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings. *ArXiv*, abs/2004.09249.

Xia, W.; Lu, H.; Wang, Q.; Tripathi, A.; Huang, Y.; Moreno, I. L.; and Sak, H. 2022. Turn-to-Diarize: Online Speaker Diarization Constrained by Transformer Transducer Speaker Turn Detection. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8077–8081.

Xu, E. Z.; Song, Z.; Tsutsui, S.; Feng, C.; Ye, M.; and Shou, M. Z. 2022. AVA-AVD: Audio-Visual Speaker Diarization in the Wild. MM '22, 3838–3847.

Yan, R.; Zhang, J.; Yang, J.; and Hauptmann, A. G. 2006. A Discriminative Learning Framework with Pairwise Constraints for Video Object Classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4): 578–593.

Yang, Z.; Hu, Y.; Liu, H.; Chen, H.; and Wu, Z. 2014. Matrix Completion for Cross-view Pairwise Constraint Propagation. In Hua, K. A.; Rui, Y.; Steinmetz, R.; Hanjalic, A.; Natsev, A.; and Zhu, W., eds., *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, 897–900. ACM.

Yehia, H. C.; Rubin, P.; and Vatikiotis-Bateson, E. 1998. Quantitative association of vocal-tract and facial behavior. *Speech Commun.*, 26: 23–43.

Yoshioka, T.; Abramovski, I.; Aksoylar, C.; Chen, Z.; David, M.; Dimitriadis, D.; Gong, Y.; Gurvich, I.; Huang, X.; Huang, Y.; Hurvitz, A.; Jiang, L.; Koubi, S.; Krupka, E.; Leichter, I.; Liu, C.; Parthasarathy, P.; Vinnikov, A.; Wu, L.; Xiao, X.; Xiong, W.; Wang, H.; Wang, Z.; Zhang, J.; Zhao, Y.; and Zhou, T. 2019. Advances in Online Audio-Visual Meeting Transcription. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 276–283.

Yu, F.; Zhang, S.; Fu, Y.; Xie, L.; Zheng, S.; Du, Z.; Huang, W.; Guo, P.; Yan, Z.; Ma, B.; Xu, X.; and Bu, H. 2022. M2Met: The Icassp 2022 Multi-Channel Multi-Party Meeting Transcription Challenge. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, 6167–6171. IEEE.

Yu, Y.; Zheng, S.; Suo, H.; Lei, Y.; and Li, W. 2021. Cam: Context-Aware Masking for Robust Speaker Verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, 6703–6707. IEEE.

Zhang, A.; Wang, Q.; Zhu, Z.; Paisley, J.; and Wang, C. 2019. Fully Supervised Speaker Diarization. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6301–6305.

Zhang, H.; Zhan, T.; Basu, S.; and Davidson, I. 2021a. A Framework for Deep Constrained Clustering. *CoRR*, abs/2101.02792.

Zhang, H.; Zhan, T.; Basu, S.; and Davidson, I. 2021b. A framework for deep constrained clustering. *Data Min. Knowl. Discov.*, 35(2): 593–620.

Zheng, S.; Huang, W.; Wang, X.; Suo, H.; Feng, J.; and Yan, Z. 2021. A Real-Time Speaker Diarization System Based on Spatial Spectrum. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, 7208–7212. IEEE.

Zheng, S.; Lei, Y.; and Suo, H. 2020. Phonetically-Aware Coupled Network For Short Duration Text-Independent Speaker Verification. In Meng, H.; Xu, B.; and Zheng, T. F., eds., *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, 926–930. ISCA.

Zheng, S.; and Suo, H. 2022. Reformulating Speaker Diarization As Community Detection With Emphasis On

Topological Structure. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8097–8101.

Zuluaga-Gomez, J.; Sarfjoo, S. S.; Prasad, A.; Nigmatulina, I.; Motlícek, P.; Ondrej, K.; Ohneiser, O.; and Helmke, H. 2022. Bertraffic: Bert-Based Joint Speaker Role and Speaker Change Detection for Air Traffic Control Communications. In *IEEE Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023*, 633–640. IEEE.

Appendix

A Constraints Sensitivity Analysis

In this section, we discuss the effects of constraint quality and quantity on experimental results. To facilitate variable control within our experiments, we employed a batch of simulated constraints. Specifically, the actual speaker labels for each extracted speaker embedding were determined using ground truth speaker timestamps, and the corresponding constraints were constructed by comparing these ground truth labels.

A.1 The Impact of Constraint Quantity

For a sequence of speaker embeddings $E = \{e_1, e_2, ..., e_N | e_i \in \mathbb{R}^D\}$, it should satisfy $|\mathcal{M}| + |\mathcal{C}| \leq N \times (N-1)$. Given that the number of speakers in a meeting scenario should be greater than or equal to 2, generally, $|\mathcal{C}| > |\mathcal{M}|$. To assess the impact of the quantity of constraints and the imbalance between must-link and cannot-link, we employ the following strategy for randomization: First, we randomly determine the must-link coverage coefficient p_{ml} and the cannot-link coverage coefficient p_{cl}, where $p_{ml} \in \{2\%, 4\%, 6\%, ..., 20\%\}$, and $p_{cl} = k_{imbalance} \times p_{ml}$ with $k_{imbalance} \in \{1, 2, 3, 4\}$. Ultimately, we select a proportion p_{ml} of must-links from all possible $\mathcal{M}$ and a proportion p_{cl} of cannot-links from all possible $\mathcal{C}$ to form a set of constraints.

Our experimental findings are presented in Figure 3. We observed that regardless of the ratio of constraint types, whether must-link or cannot-link (ranging from 1:1 to 1:4), there is a definitive enhancement in clustering performance as the constraints encompass an increasing number of embedding pairs.

A.2 The Impact of constraint Quality

Nearly all pairwise constrained clustering methods assume that the input constraints are entirely accurate; however, in practice, the constraints we obtain often contain many errors. This is especially common in multi-party meeting or interview scenarios, such as when there is audio-visual asynchrony or errors from transcript text decoded by ASR due to complex acoustic environments. In order to investigate the impact of incorrect constraints on our method, we have established the following randomization strategy: First, we randomly generate a completely correct set of constraints, including must-links and cannot-links. We then randomly alter the status of a proportion p_{err} of these constraints—turning must-links into cannot-links and vice-versa—thereby introducing a certain level of constraint errors while keeping the total number of constraints constant. In our experiments, $p_{err} \in \{5\%, 10\%, 15\%, 20\%, 25\%\}$.

The related results are reported in Figure 4, where it can be observed that erroneous constraints significantly degrade the clustering outcomes. This implies that utilizing our multimodal framework will benefit from the enhancement of multimodal model capabilities.

B Constrained Cluster Parameters Analysis

We conducted simulations of constraints to compare the optimal λ values when introducing errors in the constraints. The Figure 5 illustrate that the optimal E2CP parameter value λ for maximizing NMI depends on the error rate within the constraints. With 0% errors, the best performance is achieved at the lowest $\lambda = 0.1$, indicating that with highly accurate constraints, the algorithm benefits from a strong adherence to constraint guidance. However, for constraints with a 30% error rate, the peak NMI occurs at a higher $\lambda = 0.4$, suggesting that with less reliable constraints, the algorithm requires a more moderate constraint influence to balance error tolerance and performance. These results highlight the importance of adjusting λ in accordance with the fidelity of constraints to achieve optimal speaker diarization.
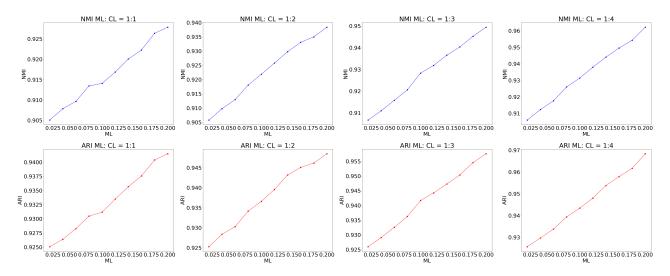
Figure 3: Results of constrained speaker cluster performance across various levels of constraints coverage, showcasing scenarios with imbalanced proportions of must-link and cannot-link constraints.
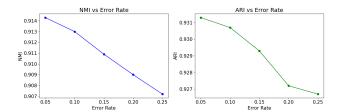

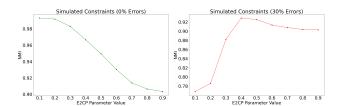Figure 4: Simulated constraints with errors and the effect for constrained clustering

Figure 5: Analysis of constrained clustering outcomes with varying λ values. It is observed that when constructed constraints contain errors, the peak of the optimal λ shifts towards 1.0.