Integrating Audio, Visual, and Semantic Information for Enhanced Multimodal Speaker Diarization on Multi-party Conversation

Luyao Cheng[1], Hui Wang[1], Chong Deng[1], Siqi Zheng[1], Yafeng Chen[1], Rongjie Huang[2], Qinglin Zhang[1], Qian Chen [1], Xihao Li[3], Wen Wang[1],
[1]Tongyi Lab, [2]Zhejiang University, [3]University of North Carolina at Chapel Hill
{shuli.cly, tongmu.wh, dengchong.d}@alibaba-inc.com

Abstract

Speaker diarization aims to segment an audio stream into homogeneous partitions based on speaker identity, playing a crucial role in speech comprehension and analysis. Mainstream speaker diarization systems rely only on acoustic information, making the task particularly challenging in complex acoustic environments in real-world applications. Recently, significant efforts have been devoted to audio-visual or audio-semantic multimodal modeling to enhance speaker diarization performance; however, these approaches still struggle to address the complexities of speaker diarization on spontaneous and unstructured multi-party conversations. To fully exploit meaningful dialogue patterns, we propose a novel multimodal approach that jointly utilizes **audio**, **visual**, and **semantic** cues to enhance speaker diarization. Our approach structures visual cues among active speakers and semantic cues in spoken content into a cohesive format known as *pairwise constraints*, and employs a semi-supervised clustering technique based on **pairwise constrained propagation**. Extensive experiments conducted on multiple multimodal datasets demonstrate that our approach effectively integrates audio-visual-semantic information into the clustering process for acoustic speaker embeddings and consistently outperforms state-of-the-art speaker diarization methods, while largely preserving the overall system framework. The open-sourced details can be found in the project[1].

1 Introduction

Speaker diarization (SD) is the task of answering the question "who spoke when" by partitioning an audio stream into segments with timestamps and corresponding speaker labels. Speaker diarization is a crucial task in multi-party conversation scenarios, as it is important for speech comprehension and analysis to conduct automatic speech recognition (ASR) and also assign speaker labels to segments of audio or transcribed text. Many downstream natural language processing (NLP) tasks (Ganesh et al., 2023a; Shen et al., 2023; Le et al., 2019; Ganesh et al., 2023b) have been proven to benefit from speaker diarization results.

Traditional speaker diarization systems rely solely on acoustic information and they can be generally categorized into two types: clustering-based approaches and end-to-end (E2E) approaches. Clustering-based approaches typically comprise three stages: voice activity detection (VAD) to filter out non-speech frames, speaker embedding extractor to obtain acoustic embeddings from each short speech segment, and an unsupervised speaker clustering to assign these embeddings into speaker classes (Anguera et al., 2012; Park et al., 2022). E2E approaches treat speaker diarization as a sequence labeling task, tagging each speech frame with its speaker identity, known as End-to-end neural diarization (EEND) (Fujita et al., 2020, 2019b). Although this modeling approach can unify the modeling of silence, single speaker speech, and speaker overlap, the absence of clustering often leads to a significant performance degradation in multi-party meeting scenarios with an uncertain number of participants, particularly when there are more than 3 speakers. The most popular acoustic-only speaker diarization systems often rely on a clustering-based approach to determine the overall speaker results, while utilizing EEND as a sub-module to handle speaker changes and overlaps, such as Pyannote (Bredin, 2023) and DiariZen (Han et al., 2024). Acoustic-only speaker diarization approaches often suffer significant performance degradation in challenging acoustic environments characterized by noise, reverberation, and speech overlapping between multiple speak-

[1] https://github.com/GeekOrangeLuYao/multimodal_pairwise_constrained_speaker_diarization

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19914–19928
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

ers (Park et al., 2022). Recent studies have aimed to address this challenge by incorporating information from other modalities into the speaker diarization task. For instance, some works (Xu et al., 2022; Chung et al., 2020; Gebru et al., 2017) have integrated visual cues, such as facial features and lip movement, with audio to determine active speakers. Other studies (Flemotomos and Narayanan, 2022; Park and Georgiou, 2018; Zuluaga-Gomez et al., 2022) have utilized text data from automatic speech recognition (ASR) to identify speaker identity and detect speaker change points. Although combining acoustic information with a single additional modality has shown some benefits, there is currently no effective approach to integrate information from all three modalities—audio, visual, and textual—into the speaker diarization task.

In this paper, we propose a novel framework based on clustering-based speaker diarization, capable of simultaneously modeling speaker-related information from multiple modalities. Specifically, we incorporate visual information (e.g., face-tracking and lip movement) and textual information (e.g., dialogue and speaker-turn detection). These multimodal insights are integrated into *pairwise constraints* to enhance speaker clustering by replacing unsupervised clustering with a semi-supervised approach. This allows for effective multimodal fusion during the clustering stage. Our method is not limited by the absence of comprehensive multimodal datasets and maintains the structural integrity of traditional acoustic-only frameworks while benefiting from advancements in individual unimodal components. Experiments across multiple multimodal datasets have consistently demonstrated the effectiveness of our approach.

Our contributions can be summarized as follows:

- **We present a noval framework for speaker diarization, uniquely integrating audio, visual, and semantic information.** This is the first framework to leverage all these three modalities, enhancing the robustness and accuracy of speaker diarization.

- **We introduce a joint pairwise constraint propagation method into the speaker clustering process**, effectively enhancing speaker clustering performance through multimodal information-derived constraints.

- To comprehensively evaluate the effectiveness of our method, **we contribute a 6.3-hour video evaluation set sourced from in-the-wild scenarios**, which has been annotated with speaker identity labels, corresponding speech activity timestamps, and speech content.

2 Related Work

2.1 Multimodal Speaker Diarization

Acoustic-Only speaker diarization Audio-only speaker diarization has been studied extensively (Park et al., 2022). A typical speaker diarization systems employ a multi-stage framework, including VAD (Gelly and Gauvain, 2018), speech segmentation (Xia et al., 2022), acoustic embedding extraction (Snyder et al., 2018; Zheng et al., 2020; Chen et al., 2023) and unsupervised clustering such as agglomerative hierarchical clustering (AHC) (Day and Edelsbrunner, 1984) and spectral clustering(SC) (Wang et al., 2018). Recently, EEND where individual sub-modules in traditional systems can be replaced by one neural network has received more attention (Fujita et al., 2019a,c; Horiguchi et al., 2020) which treat speaker diarization as a frame-level sequence labeling task. Due to the absence of a clustering algorithm, the EEND method often experiences significant performance degradation in scenarios with a large number of speakers. Some approaches improve model performance by combining the global speaker predictions from clustering with the local speaker change and overlap detection results from EEND, such as EEND-VC (Kinoshita et al., 2021) and DiariZen (Han et al., 2024). Similar strategies have also been adopted by mainstream speaker diarization toolkits like Pyannote (Bredin, 2023).

Audio-visual Speaker Diarization Facial activities and lip motion are highly related to speech (Yehia et al., 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. Some methods leverage the audio and visual cues for diarization using synchronization between talking faces and voice tracks (Chung et al., 2019). Other works (Xu et al., 2022; Wuerkaixi et al., 2022; Yin et al., 2024) utilized an attention-based network to perform middle-fusion and extract a unified representation of the two modalities. Recently, an interesting and promising direction is to use separate neural networks to process data streams of

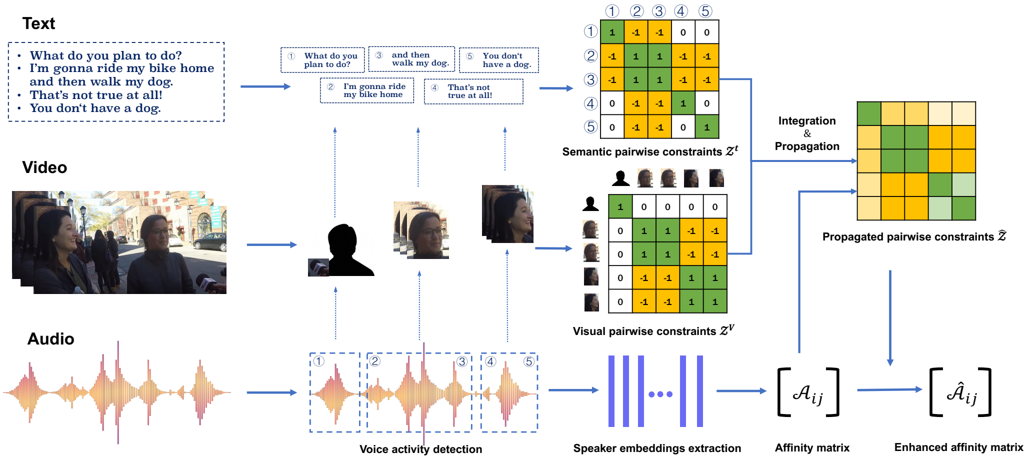

Text
What do you plan to do?
I'm gonna ride my bike home and then walk my dog.
That's not true at all!
You don't have a dog.
① What do you plan to do?
② I'm gonna ride my bike home
③ and then walk my dog.
④ That's not true at all!
⑤ You don't have a dog.
Semantic pairwise constraints $\mathcal{Z}^t$
Video
Visual pairwise constraints $\mathcal{Z}^V$
Integration & Propagation
Propagated pairwise constraints $\hat{\mathcal{Z}}$
Audio
Voice activity detection
Speaker embeddings extraction
$\mathcal{A}_{ij}$
Affinity matrix
$\hat{\mathcal{A}}_{ij}$
Enhanced affinity matrix

Figure 1: An overview of our proposed multimodal speaker diarization system.

two modalities and directly output speech probabilities for all speakers simultaneously (kui He et al., 2022), similar to audio-only EEND frameworks. All of these require expensive amounts of annotated audio-visual parallel data for training, which is expensive to acquire.

Audio-textual Speaker Diarization Some previous works (Zuluaga-Gomez et al., 2022; Flemotomos and Narayanan, 2022; Park and Georgiou, 2018; Paturi et al., 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., 2023) employed large language models as post-processing to correct word speaker-related boundaries according to local semantic context.

2.2 Pairwise Constrained Clustering

Speaker diarization systems typically rely on unsupervised clustering to handle an unknown number of speakers. When integrating multimodal information, direct cross-modal similarity comparisons are not feasible. Thus, incorporating semi-supervised signals into the clustering process becomes essential, a technique known as constrained clustering (Bibi et al., 2023). Pairwise constrained clustering is a common approach within this framework, where supplementary information defines pairwise relationships among samples through **Must-link constraints** (indicating two samples belong to the same class) and **Cannot-link constraints** (indicating they do not) (Davidson and Ravi, 2007). The process of refining the affinity matrix using these pairwise constraints is referred to as pairwise constrained propagation. Initially confined to data mining domain (Hoi et al., 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 Lyapunov equation (Lu and Peng, 2011), 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). Among them, E2CP (Exhaustive and Efficient Constraint Propagation) (Lu and Peng, 2011) is widely adopted due to its simple and effective hyperparameter configuration. In this paper, we employ E2CP as the core pairwise constrained clustering method to integrate multimodal constraints into speaker clustering.

3 Methodology

Figure 1 provides an overview of how our approach leverages multimodal information. In addtion to a clustering-based speaker diarization system, **video and text processing modules** are incorporated to independently extract visual and semantic information and derive pairwise constraints. Then a **joint propagation algorithm** will be employed to operate cross-modal pairwise constraints to enhance the affinity matrix constructed from acoustic speaker embeddings. The **enhanced affinity matrix** is subsequently integrated into the subsequent clustering procedure to assign speaker label for each speaker embedding. The following sections will present the joint propagation algorithm and the process of constructing visual and semantic constraints.

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 where D represents the dimension of the speaker embeddings and N denotes the number of speaker embeddings. Subsequently, we compute the affinity matrix $\boldsymbol{\mathcal{A}} = \{\mathcal{A}_{ij}\}_{N\times N}$, where $\mathcal{A}_{ij} = g(e_i, e_j)$ and $g(\cdot)$ represents the measurement of similarity.

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:

$$\begin{aligned} \mathcal{M}^k &= \{(e_i, e_j) | l(e_i) = l(e_j)\}, \\ \mathcal{C}^k &= \{(e_i, e_j) | l(e_i) \neq l(e_j)\}, \end{aligned} \tag{1}$$

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 $\mathcal{C}$ are 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 $\boldsymbol{\mathcal{Z}}^k$:

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

A series of constraint matrix $\boldsymbol{\mathcal{Z}}^k$ are integrated into a final constraint matrix $\boldsymbol{\mathcal{Z}}$. 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:

$$\boldsymbol{\mathcal{Z}}' = \sum_k \alpha_k \boldsymbol{\mathcal{Z}}^k + \beta \boldsymbol{\mathcal{A}} - \theta \tag{3}$$

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

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

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

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

where $\mathbf{L}_e = \mathbf{D}_e^{-1/2} \boldsymbol{\mathcal{A}} \mathbf{D}_e^{-1/2}$ is the normalized Laplacian matrix, and $\mathbf{D}_e$ is the degree matrix of $\boldsymbol{\mathcal{A}}$ and $\mathbf{I}$ is a identity matrix. The parameter $\lambda \in [0, 1]$ modulates the impact degree of the constraints. The refined affinity matrix $\hat{\boldsymbol{\mathcal{A}}} \in \mathbb{R}^{N\times N}$ is then updated to incorporate the influences of the propagated constraints $\hat{\boldsymbol{\mathcal{Z}}}$:

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

Upon calculating the affinity matrix $\hat{\boldsymbol{\mathcal{A}}}$, 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 considered as prior knowledge, guiding the clustering focus towards a specific perspective of the scenario. In this paper, we fix k at 2, thereby extracting two distinct constraint types: visual constraint $\boldsymbol{\mathcal{Z}}^{\boldsymbol{v}}$ and semantic constraint $\boldsymbol{\mathcal{Z}}^{\boldsymbol{t}}$.

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): **(1) Face Tracking**. The first step involves detecting and tracking faces in video frames over time using a CNN-based face detector (Liu et al., 2018) and a position-based tracker. Only face tracks aligned with speech segments detected by VAD are retained for further processing. **(2) Active Speaker Detection**. This step determines whether tracked faces correspond to active speakers at any given moment. A two-stream network (Tao et al., 2021), comprising temporal encoders and an attention-based decoder, analyzes audio-visual synchrony to identify speaker activity. Low-confidence frames are filtered using a predefined threshold. **(3) Face Clustering**. A face recognition CNN (Huang et al., 2020) extracts embeddings from face tracks at uniform intervals (e.g., every 200 ms). These embeddings are then clustered with 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 along the time axis. If an acoustic embedding corresponds to multiple faces, we will select the speaker associated with the majority of those faces.

3.3 Semantic constraints construction

To extract speaker-related information from the transcriptions, we construct two Spoken Language Processing (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.

4 Experiments

4.1 Experimental Setup

Datasets. Our experiments are conducted on the AIShell-4 (Fu et al., 2021), Alimeeting (Yu et al., 2022), AVA-AVD (Xu et al., 2022), and our proposed datasets. The AVA-AVD dataset, which focuses on audio-visual diarization, provides diverse scenarios and face annotations but lacks ground truth transcripts. In contrast, AIShell-4 and Alimeeting are Mandarin datasets that include speaker-labeled transcripts, making them well-suited for audio-text-based tasks. Due to the absence of publicly available evaluation datasets with annotations for visual, semantic, and acoustic modalities, we construct a new dataset comprising 6.3 hours of video, manually annotated with speaker timestamps and speech content. Further details about this dataset are provided in the Appendix A. The combination of these diverse datasets further demonstrates the effectiveness of our methods across different domains.

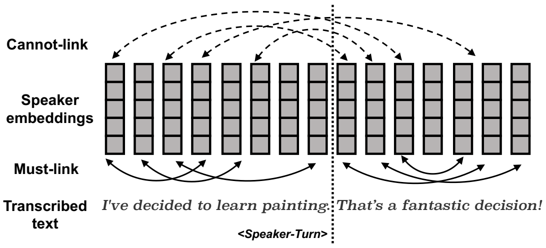


Figure 2: Semantic constraint construction is based on dialogue and speaker-turn detection. Text segments identified as non-dialogue imply that their embeddings are related through must-link constraints (solid connections). Conversely, detected transition points indicate that embeddings spanning these points should be connected via cannot-link constraints (dashed connections).

Implementation Details. In our system, the audio-based diarization modules follow the pipeline outlined in (Cheng et al., 2023). Our speaker embedding extractor is an adaptation of CAM++ (Wang et al., 2023), which has been trained on VoxCeleb dataset (Nagrani et al., 2020). The ASR we utilize is Paraformer (Gao et al., 2022), which has been trained with the aid of the FunASR (Gao et al., 2023) toolkits. For visual componets, we employ a series of pre-trained models for different tasks: RFB-Net (Liu et al., 2018) for face detection, TalkNet (Tao et al., 2021) for active speaker detection, and CurricularFace model (Huang et al., 2020) for extracting face embeddings. For semantic tasks, we train models on open-sourced meeting datasets designed for various scenarios. Specifically, we use separate datasets for English and Mandarin to train corresponding semantic models, ensuring language-specific adaptations. All that

Table 1: The results of speaker diarization on AIShell-4, Alimeeting and our proposed dataset.

Dataset	Methods	Modality	DER(%)↓	CpWER(%)↓
AIShell-4	Pyannote	Audio	12.2	-
	DiariZen	Audio	11.7	-
	Semantic-Aux SD	Audio + Semantic	-	15.23
	Proposed	Audio + Semantic	12.07	14.95
Alimeeting	Pyannote	Audio	24.4	-
	DiariZen	Audio	17.6	-
	Semantic-Aux SD	Audio + Semantic	-	36.15
	Proposed	Audio + Semantic	21.32	31.11
Proposed Dataset	Pyannote	Audio	15.57	-
	DiariZen	Audio	10.49	-
	CAM++ & VBx	Audio	10.31	18.03
	CAM++ & SC	Audio	9.37	17.04
	Proposed	Audio + Semantic	9.12	16.86
	Proposed	Audio + Visual	9.13	16.83
	Proposed	Audio + Semantic + Visual	9.01	16.36

training was conducted using a pre-trained BERT model (Devlin et al., 2019). We employ E2CP as the core algorithm for constraint propagation. In the post-clustering phase, our system adheres to the SC algorithm. Inspired by the work presented in (Park et al., 2020), our method incorporates refinement operations, such as row-wise thresholding and symmetrization, to enhance the performance of spectral clustering. More details of the implementation and hyperparameter settings can be found in the appendix B.

Evaluation Metrics. To demonstrate the impact of the speaker diarization system, we report the Diarization Error Rate (DER) (Fiscus et al., 2006), which generally composed of three parts: Missed Speech (MS), False Alarms (FA) and Speaker Error (SPKE). As the ASR and forced-alignment module have been used in the pipeline, we also report the Concatenated Minimum-permutation Word Error Rate (Watanabe et al., 2020). The cluster metrics, Normalized Mutual Information (NMI) (Strehl and Ghosh, 2002) and Adjusted Rand Index (ARI) (Chac'on and Rastrojo, 2020), will be reported in section 4.3.

4.2 Main Results

Multimodal Speaker Diarization with Audio-Text Modalities. In Part 1 & 2 of Table 1, we compare our system with acoustic-only speaker diarization systems such as Pyannote and DiariZen, as well as with Semantic-Aux SD, an audio-text speaker diarization system, on the AIShell-4 and Alimeeting datasets. It can be observed that our system demonstrates a certain level of superiority over the classical speaker diarization toolkit, Pyannote, on both datasets. Specifically, our method achieves an absolute improvement of 0.13% in DER on the AIShell-4 dataset and 3.08% in DER on the Alimeeting dataset. However, our approach shows a slight disadvantage compared to DiariZen, which trains a frame-level EEND model using the training sets from AMI, Alimeeting, and AIShell-4 audio data, thus providing it with a noticeable advantage on these two homologous test sets.

Compared with Semantic-Aux SD (Cheng et al., 2023), another speaker diarization method that combines audio and text modalities, our experiments maintain consistent ASR results. Our proposed solution shows clear improvements on both datasets, with an absolute gain of 5.04% in CpWER on the Alimeeting dataset. Unlike Semantic-Aux SD, which primarily uses semantic information for boundary refinement of speaker diarization results, our approach **integrates semantic information into the speaker clustering process, leveraging semantic cues to correct more errors** that arise from relying solely on acoustic-only information.

Multimodal Speaker Diarization with Audio-Visual Modalities. We have also compared our approach with several audio-visual joint training speaker diarization methods on the AVA-AVD dataset, such as AVR-Net (Xu et al., 2022), AFL-Net (Yin et al., 2024), and DyViSE (Wuerkaixi et al., 2022), to demonstrate the effectiveness of our

Table 2: The results of audio-only and audio-visual speaker diarization experiments on AVA-AVD datasets.

Models	SPKE(%)↓	DER(%)↓
VBx + ResNet34	35.14	38.06
CAM++ + SC	19.85	22.11
AVR-Net	24.88	27.43
AFL-Net	21.10	23.65
AFL-Net + WavLM	19.57	22.12
DyViSE	20.86	23.46
Proposed	17.40	20.32

method. Due to the lack of annotated transcripts in the AVA-AVD dataset, only the SPKE and DER metrics are reported. Table 2 presents a comparison conducted on AVA-AVD, revealing the competitive performance of our method when utilizing only visual constraints. Compared to the baseline results of the AVA-AVD, AVR-Net, our model shows a 7.11% absolute improvement in DER. Additionally, when compared to audio-visual models such as AFL-Net and DyViSE, our model also exhibits a significant improvement in DER, with a relative 8.1% improvement over AFL-Net and a relative 13.3% improvement over DyViSE.

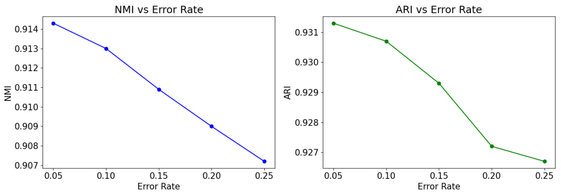


Figure 3: Simulated constraints with errors and the effect for constrained clustering

Multimodal Speaker Diarization with Audio-Visual-Text Modalities. In the last part of Table 1, we present the results of multiple speaker diarization systems on the proposed dataset. As previously mentioned, the acoustic-only speaker diarization SOTA(start-of-the-art) system, DiariZen, benefits from targeted training on the AIShell-4 and Alimeeting training sets, giving it a certain advantage over our method on these datasets. However, on the proposed dataset, our "CAM++ & SC" approach achieves a relative improvement of 10.7% in DER compared to DiariZen. Furthermore, when incorporating the semantic constraints proposed in this paper, the improvement over DiariZen reaches a relative 13.1% reduction in DER. In contrast, another well-known open-source speaker diarization toolkit, pyannote, exhibits a significantly higher DER of 15.57%, showing a larger gap compared to other approaches. When comparing the constrained propagation methods proposed in this paper, the results show that using only visual constraints or only semantic constraints result in very similar performance in terms of DER and CpWER. Compared to the acoustic-only baseline, incorporating visual constraints leads to a relative improvement of 2.56% in DER and 1.1% in CpWER, while incorporating semantic constraints yields a relative improvement of 2.66% in DER and 1.2% in CpWER. Further analysis reveals that combining **all three modalities provides even greater improvements over systems that combine only two modalities**. Specifically, the DER is relatively reduced by 3.8%, and the CpWER is relatively reduced by 4.0%. This indicates that **semantic and visual constraints are complementary, and integrating multiple modalities can lead to further performance gains.** Some decoding cases and visualizations can be found in the appendix F.

4.3 Analysis and Discussion

Constraint Construction. It should be noted that constraints constructed based on multimodal data often contain some errors, and at the same time, constraints cannot cover all embedding pairs. In this section, we discuss the impact of varying quantities and qualities of pairwise constraints on the results of speaker clustering. We employ several simulation strategies to generate pairwise constraints, which allows for better control over both the quantity and quality of these constraints. All experiments are conducted on our proposed dataset, utilizing speaker embeddings extracted by CAM++ that remain fixed throughout the experiments; only the pairwise constraints used in each experiment are varied.

(1) The Impact of constraint Quality. 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 transcipt 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

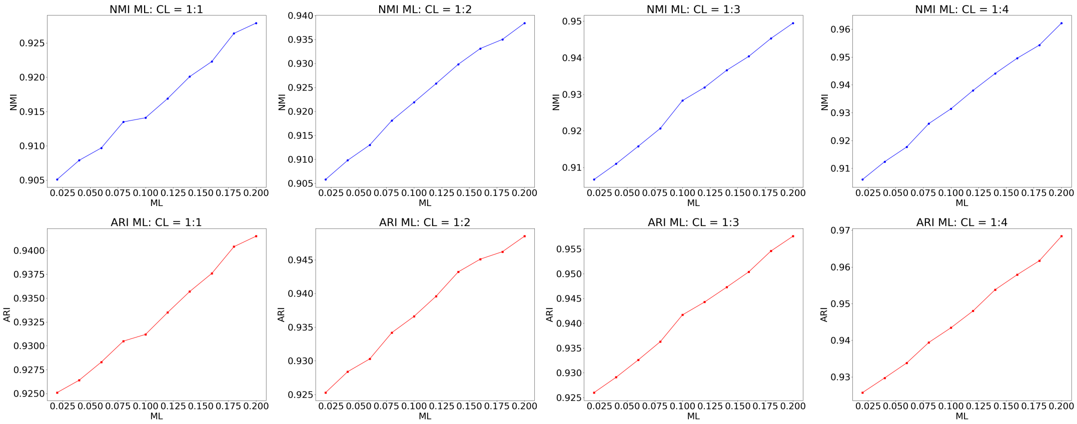


Figure 4: Results of constrained speaker cluster performance across various levels of constraints coverage, showcasing scenarios with imbalanced proportions of must-link and cannot-link constraints.

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 experimental results are illustrated in Figure 3. It can be observed that errors in the constraints do indeed lead to a decline in clustering performance. However, even when the error rate reaches 25%, the NMI experiences only a 0.7% relative decrease compared to the NMI at a 5% error rate. **This indicates that our method exhibits a certain degree of robustness to erroneous constraints.**

(2) Impact of Constraint Quantity and Ratios We investigate how the number of constraints and the ratio of must-link to cannot-link sets affect speaker clustering in our approach. We formulate a simulation strategy: for a sequence of speaker embeddings $E = \{e_1, e_2, ..., e_N | e_i \in \mathbb{R}^D\}$, we vary the must-link coverage (p_{ml}) and cannot-link coverage (p_{cl}) proportions. Specifically, $p_{ml} \in \{2\%, 4\%, 6\%, ..., 20\%\}$, and $p_{cl} = k_{ratio} \times p_{ml}$ with $k_{ratio} \in \{1, 2, 3, 4\}$. We select $p_{ml}\%$ of must-links and $p_{cl}\%$ of cannot-links from all possible pairs. As shown in Figure 4, it can be observed that as the number of constraints increases, the clustering performance of the algorithm consistently improves. For instance, in the case of ML:CL = 1:1, the NMI increases from 0.905 to 0.925, and the ARI improves from 0.925 to 0.940. Additionally, our method demonstrates that **an imbalance between the quantities of ML and CL does not hinder performance gains.**

(3) Impact of ASR for Semantic Constraints. In practice, visual and audio data are often collected independently using different devices, with semantic information extracted from audio via an ASR system. Complex acoustic environments can affect both speaker embedding extraction and ASR accuracy. In this section, we evaluate the performance of our method under varying ASR accuracy levels. Specifically, we first partition the Alimeeting test set into "Easy" and "hard" subsets based on whether the ASR Word Error Rate (WER) exceeds 21%. We then test both the acoustic-only solutions (CAM++ & SC) and our proposed method, which incorporates semantic constraints, on these two subsets. Table 3 presents our experimental results. It can be observed that on the "hard" subset, the DER is relatively higher, indicating that both speaker embedding extraction and the ASR system encounter certain errors in complex acoustic environments. Nevertheless, our method achieves notable improvements on both subsets. On the "Easy" subset, where the acoustic-only solution already performs well (DER = 7.31%), our approach achieves a relative improvement of 8.9%. On the "hard" subset, our method demonstrates a significant relative improvement of 17.4%, showcasing its **robustness in challenging acoustic conditions.**

Table 3: The gain achieved by our proposed method using acoustic and semantic information, in comparison to the acoustic-only approach, varies across cases with different ASR levels in Alimeeting.

Test Subsets	DER(%) acoustic only	DER(%) proposed models	DER relative gain(%)
Easy subsets (WER <21%)	7.31	6.65	8.9
Hard subsets (WER >21%)	38.02	31.45	17.4

5 Conclusions

In this study, we propose a novel multimodal approach that leverages audio, visual, and semantic information to enhance speaker diarization. By incorporating additional visual and textual processing modules, we generate complementary pairwise constraints that are integrated into the clustering process through a joint pairwise constraint propagation method. Experimental results demonstrate significant performance improvements. This research contributes to the advancement of more sophisticated systems for the speaker diarization task, providing potential directions for future exploration.

6 Limitations

The semantic information we utilize is derived from the Dialogue Detection and Speaker-Turn Detection models, which are trained based on the BERT architecture, rather than employing more advanced Large Language Models. Additionally, we observe that the embeddings corresponding to the constraints constructed for the current semantic tasks—Dialogue Detection and Speaker-Turn Detection—are relatively close to each other in the embedding space. While this proximity aids in identifying more precise speaker transition points, it also limits our ability to extract long-term semantic information from the text. In theory, our approach can be adapted to incorporate various multimodal sources of information. However, another modality that could significantly assist in speaker identity determination—speaker location information—has not been integrated into our experiments. We plan to explore this further in future work.

References

Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370.

Adel Bibi, Ali Alqahtani, and Bernard Ghanem. 2023. Constrained clustering: General pairwise and cardinality constraints. *IEEE Access*, 11:5824–5836.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, volume 3869 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

Jos'e E. Chac'on and Ana I. Rastrojo. 2020. Minimum adjusted rand index for two clusterings of a given size. *Advances in Data Analysis and Classification*, 17:125–133.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, and Jiajun Qi. 2023. An enhanced res2net with local and global feature fusion for speaker verification. *CoRR*, abs/2305.12838.

Luyao Cheng, Siqi Zheng, Zhang Qinglin, Hui Wang, Yafeng Chen, and Qian Chen. 2023. Exploring speaker-related information in spoken language understanding for better speaker diarization. In *Annual Meeting of the Association for Computational Linguistics*.

Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman. 2020. Spot the conversation: speaker diarisation in the wild. In *Interspeech*.

Joon Son Chung, Bong-Jin Lee, and Icksang Han. 2019. Who said that?: Audio-visual speaker diarisation of real-world meetings. In *Interspeech*.

Ian Davidson and S. S. Ravi. 2007. Intractability and clustering with constraints. In *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*, pages 201–208. ACM.

William H. E. Day and Herbert Edelsbrunner. 1984. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1:7–24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Jonathan G. Fiscus, Jerome Ajot, Martial Michel, and John S. Garofolo. 2006. The rich transcription 2006 spring meeting recognition evaluation. In *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*, pages 309–322. Springer.

Nikolaos Flemotomos and Shrikanth Narayanan. 2022. Multimodal clustering with role induced constraints for speaker diarization.

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, Xin Xu, Jun Du, and Jingdong Chen. 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*, pages 3665–3669. ISCA.

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

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe. 2019a. End-to-End Neural Speaker Diarization with Permutation-free Objectives. In *Interspeech*, pages 4300–4304.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. 2019b. End-to-end neural speaker diarization with self-attention. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 296–303.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. 2019c. End-to-end neural speaker diarization with self-attention. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 296–303.

Yusuke Fujita, Shinji Watanabe, Shota Horiguchi, Yawen Xue, and Kenji Nagamatsu. 2020. End-to-end neural diarization: Reformulating speaker diarization as simple multi-label classification. *ArXiv*, abs/2003.02966.

Ananya Ganesh, Martha Palmer, and Katharina Kann. 2023a. A survey of challenges and methods in the computational modeling of multi-party dialog. *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*.

Ananya Ganesh, Martha Palmer, and Katharina Kann. 2023b. A survey of challenges and methods in the computational modeling of multi-party dialog. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 140–154, Toronto, Canada. Association for Computational Linguistics.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. *ArXiv*, abs/2305.11013.

Zhifu Gao, Shiliang Zhang, Ian Mcloughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Interspeech*.

Israel D Gebru, Sileye Ba, Xiaofei Li, and Radu Horaud. 2017. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1086–1099.

Gregory Gelly and Jean-Luc Gauvain. 2018. Optimization of rnn-based speech activity detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3):646–656.

Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, and Lukas Burget. 2024. Leveraging self-supervised learning for speaker diarization. *arXiv preprint arXiv:2409.09408*.

Steven C. H. Hoi, Rong Jin, and Michael R. Lyu. 2007. Learning nonparametric kernel matrices from pairwise constraints. In *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*, pages 361–368. ACM.

Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu. 2020. End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors. In *Interspeech*.

Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. 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*, pages 5900–5909. Computer Vision Foundation / IEEE.

Adam L. Janin, Don Baron, Jane Edwards, Daniel P. W. Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The icsi meeting corpus. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*, 1:I–I.

Naoyuki Kanda, Xiong Xiao, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. 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)*, pages 8082–8086.

Aparna Khare, Eunjung Han, Yuguang Yang, and Andreas Stolcke. 2022. Asr-aware end-to-end neural diarization. *ICASSP 2022*, pages 8092–8096.

Keisuke Kinoshita, Marc Delcroix, and Naohiro Tawara. 2021. 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)*, pages 7198–7202.

Mao kui He, Jun Du, and Chin-Hui Lee. 2022. End-to-end audio-visual neural speaker diarization. In *Interspeech*.

Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget. 2022. Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks. *Computer Speech & Language*, 71:101254.

Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. In *Conference on Empirical Methods in Natural Language Processing*.

Hui Liu, Yuheng Jia, Junhui Hou, and Qingfu Zhang. 2019. Imbalance-aware pairwise constraint propagation. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 1605–1613. ACM.

Songtao Liu, Di Huang, and andYunhong Wang. 2018. Receptive field block net for accurate and fast object detection. In *The European Conference on Computer Vision (ECCV)*.

Zhiwu Lu and Yuxin Peng. 2011. Exhaustive and efficient constraint propagation: A graph-based learning approach and its applications. *International Journal of Computer Vision*, 103:306–325.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Comput. Speech Lang.*, 60.

Tae Jin Park, Kunal Dhawan, Nithin Rao Koluguri, and Jagadeesh Balam. 2023. Enhancing speaker diarization with large language models: A contextual beam search approach. *ArXiv*, abs/2309.05248.

Tae Jin Park and Panayiotis G. Georgiou. 2018. Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks. In *Interspeech*.

Tae Jin Park, Kyu J. Han, Manoj Kumar, and Shrikanth S. Narayanan. 2020. Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap. *IEEE Signal Processing Letters*, 27:381–385.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Comput. Speech Lang.*, 72(C).

Rohit Paturi, Sundararajan Srinivasan, and Xiang Li. 2023. Lexical speaker error correction: Leveraging language models for speaker diarization error correction. *ArXiv*, abs/2306.09313.

Weizhou Shen, Xiaojun Quan, and Ke Yang. 2023. Generic dependency modeling for multi-party conversation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, 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*, page 3927–3935.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

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

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023. Cam++: A fast and efficient network for speaker verification using context-aware masking.

Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. 2018. Speaker diarization with lstm. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5239–5243.

Quan Wang, Yiling Huang, Guanlong Zhao, Evan Clark, Wei Xia, and Hank Liao. 2024. Diarizationlm: Speaker diarization post-processing with large language models. *ArXiv*, abs/2401.03506.

Shinji Watanabe, Michael Mandel, Jon Barker, and Emmanuel Vincent. 2020. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *ArXiv*, abs/2004.09249.

Abudukelimu Wuerkaixi, Kunda Yan, You Zhang, Zhiyao Duan, and Changshui Zhang. 2022. Dyvise: Dynamic vision-guided speaker embedding for audio-visual speaker diarization. In *24th IEEE International Workshop on Multimedia Signal Processing, MMSP 2022, Shanghai, China, September 26-28, 2022*, pages 1–6. IEEE.

Wei Xia, Han Lu, Quan Wang, Anshuman Tripathi, Yiling Huang, Ignacio Lopez Moreno, and Hasim Sak. 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)*, pages 8077–8081.

Eric Zhongcong Xu, Zeyang Song, Satoshi Tsutsui, Chao Feng, Mang Ye, and Mike Zheng Shou. 2022. Ava-avd: Audio-visual speaker diarization in the wild. MM '22, page 3838–3847.

Rong Yan, Jian Zhang, Jie Yang, and Alexander G. Hauptmann. 2006. A discriminative learning framework with pairwise constraints for video object classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):578–593.

Zheng Yang, Yao Hu, Haifeng Liu, Huajun Chen, and Zhaohui Wu. 2014. Matrix completion for cross-view pairwise constraint propagation. In *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, pages 897–900. ACM.

Hani C. Yehia, Philip Rubin, and Eric Vatikiotis-Bateson. 1998. Quantitative association of vocal-tract and facial behavior. *Speech Commun.*, 26:23–43.

Yongkang Yin, Xu Li, Ying Shan, and Yuexian Zou. 2024. Afl-net: Integrating audio, facial, and lip modalities with cross-attention for robust speaker diarization in the wild.

Takuya Yoshioka, Igor Abramovski, Cem Aksoylar, Zhuo Chen, Moshe David, Dimitrios Dimitriadis, Yifan Gong, Ilya Gurvich, Xuedong Huang, Yan Huang, Aviv Hurvitz, Li Jiang, Sharon Koubi, Eyal Krupka, Ido Leichter, Changliang Liu, Partha Parthasarathy, Alon Vinnikov, Lingfeng Wu, Xiong Xiao, Wayne Xiong, Huaming Wang, Zhenghao Wang, Jun Zhang, Yong Zhao, and Tianyan Zhou. 2019. Advances in online audio-visual meeting transcription. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 276–283.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. 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*, pages 6167–6171. IEEE.

Hongjing Zhang, Tianyang Zhan, Sugato Basu, and Ian Davidson. 2021a. A framework for deep constrained clustering. *CoRR*, abs/2101.02792.

Hongjing Zhang, Tianyang Zhan, Sugato Basu, and Ian Davidson. 2021b. A framework for deep constrained clustering. *Data Min. Knowl. Discov.*, 35(2):593–620.

Siqi Zheng, Yun Lei, and Hongbin Suo. 2020. Phonetically-aware coupled network for short duration text-independent speaker verification. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 926–930. ISCA.

Juan Zuluaga-Gomez, Seyyed Saeed Sarfjoo, Amrutha Prasad, Iuliia Nigmatulina, Petr Motlícek, Karel Ondrej, Oliver Ohneiser, and Hartmut Helmke. 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*, pages 633–640. IEEE.

A Proposed Dataset

In this section, we provide an overview of the proposed dataset. The dataset includes a variety of acoustic and visual scenarios, sourced from in-the-wild videos. The dataset includes a total of 92 speakers, with each meeting involving 2 to 10 participants. The data exhibits significant variability in both content and environmental conditions. The total duration of the dataset is approximately 6.3 hours, with individual video clips ranging from 7 to 29 minutes. The dataset covers a wide range of scenarios, including interviews, talk shows, meetings, press conferences, round-table discussions, and TV programs. It has been meticulously annotated with speaker identity labels, corresponding speech activity timestamps, and transcribed speech content.

These manual annotations come from annotators at external data company. Before annotation, we only ask for the speaker content of each speaker in the video, and the timestamp of each sentence. We provided annotators with a sample annotation from Alimeeting, and the results they returned were consistent with this batch of data. The speaker id is completely anonymized in the annotation and labeled as $\{c1, c2, c3, ...\}$

B Implementation details

In this section, we provide the implementation details of our experiments.

In our system, the audio-based diarization modules follow the pipeline outlined in (Cheng et al., 2023). Our speaker embedding extractor is an adaptation of CAM++ (Wang et al., 2023)[2], which has been trained on VoxCeleb dataset (Nagrani et al., 2020). 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[3].

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

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

Table 4: Constraints derived from various modalities. We separately evaluate 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

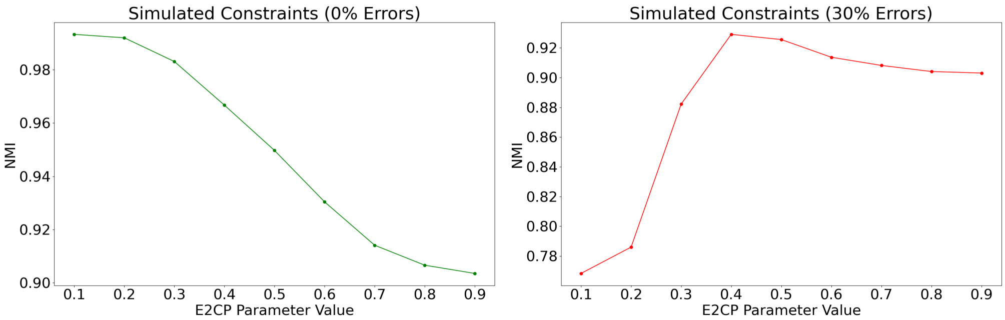


Figure 5: Analysis of constrained clustering outcomes with varying λ values. It is observed that when constructed constraints contain errors, the peak of the optimal λ shifts towards 1.0.

For the visual componets, we employ a series of pretrained models for different tasks: RFB-Net (Liu et al., 2018)[4] for face detection, TalkNet (Tao et al., 2021)[5] for active speaker detection, and CurricularFace model (Huang et al., 2020)[6] for extracting face embeddings.

For semantic tasks, we train muliple models with open-source meeting datasets for different scenarios. Specifically, we employ 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 employ the methods described in Section 3.3 to construct the semantic constraints.

The VBx approach (Landini et al., 2022) is a canonical method for speaker diarization, where the original paper utilizes speaker embeddings based on the x-vector model. We replace this with the more robust CAM++ model. Additionally, since the post-processing step of the E2CP method incorporates SC (Von Luxburg, 2007), we also investigate the performance of a method that relies solely on speaker embeddings and SC. These 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-clustering. 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. In the post-clustering phase, we adhere to the Spectral Clustering (SC) algorithm, consistent with the baseline. Inspired by the work presented in (Park et al., 2020), our method incorporates refinement operations such as row-wise thresholding and symmetrization to enhance the performance of spectral clustering. For the row-wise thresholding step in SC, the p-percentile parameter is set to 0.982.

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

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

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

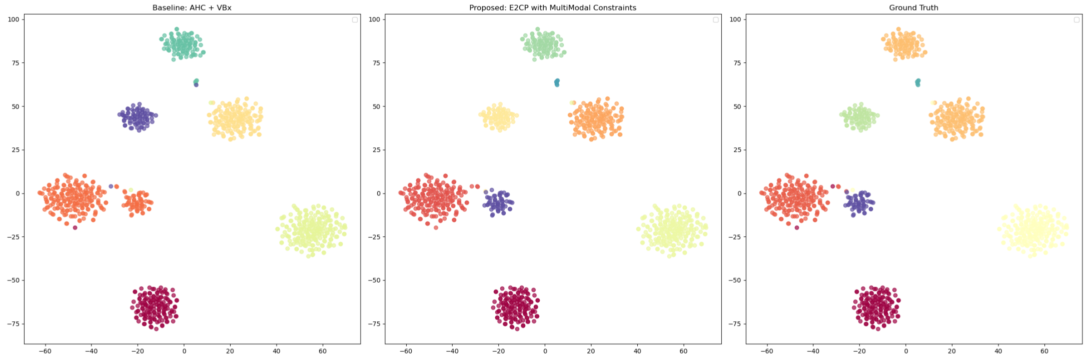


Figure 6: The t-SNE for cluster cases

C Constraints Statistics

Table 4 illustrates the statistical information of the visual constraints and semantic constraints constructed on our proposed dataset. Upon analysis, it is evident that visual constraints significantly outperform semantic constraints in terms of coverage. This disparity is attributed to the fact that the two semantic tasks employed in our semantic model are only capable of evaluating the relationships between embeddings within adjacent speaker turns, whereas visual constraints are assessed across embedding pairs with substantial temporal intervals. Furthermore, the method we designed to combine constraints from different modalities successfully merges them.

D Constrained Cluster Parameters Analysis

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 $\hat{\boldsymbol{\mathcal{Z}}}$ will be closer to $\boldsymbol{\mathcal{Z}}$, whereas when λ approaches 1, the resulting $\hat{\boldsymbol{\mathcal{A}}}$ will be closer to $\boldsymbol{\mathcal{A}}$.

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.

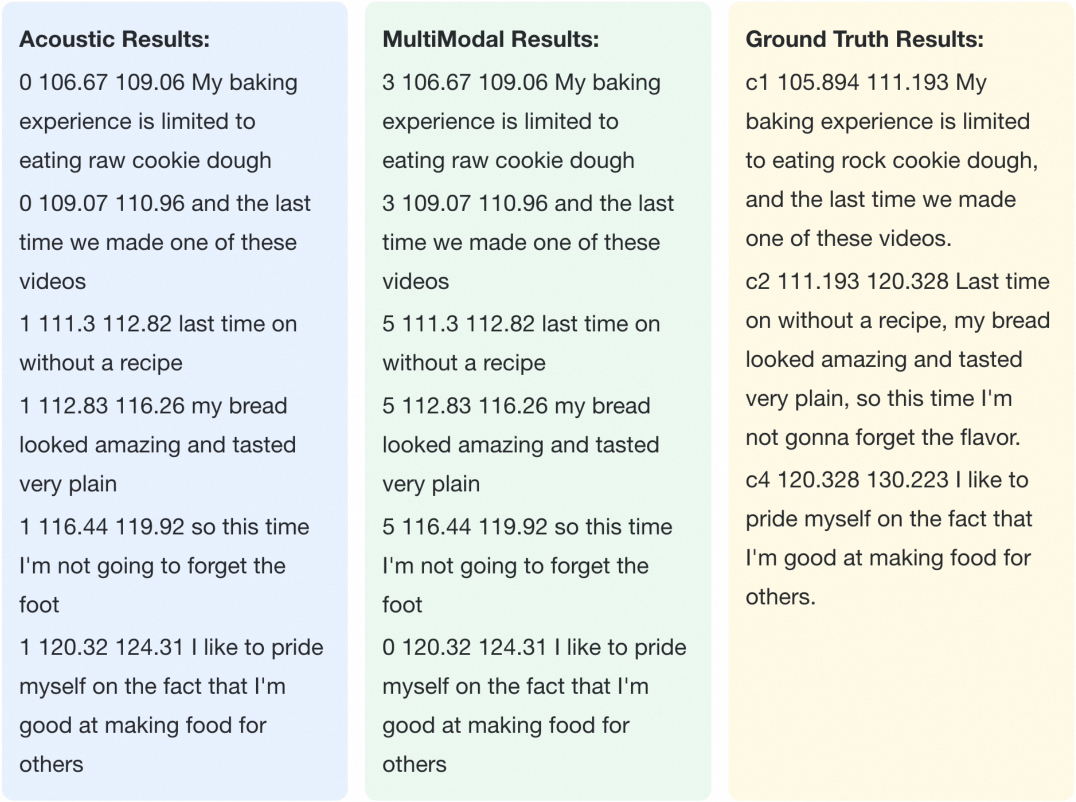
We conducted simulations of constraints to compare the optimal λ values when introducing errors in the constraints. The Figure 5 illustrate that the optimal E2CP parameter value λ for maximizing NMI depends on the error rate within the constraints. With 0% errors, the best performance is achieved at 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.

E Sensitivity Analysis for α_k and β

We analyzed the sensitivity of the parameters α_k and β in Section 3.1 and find that the system exhibits low sensitivity to these parameters, as shown in Table 5. This behavior can be attributed to the discretization process in Eq. 4, which ensures that minor numerical fluctuations will not result in significant changes to the final constraints. Additionally, the row-wise thresholding strategy (Wang et al., 2018) employed in our spectral clustering algorithm further enhances the robustness and stability of our approach. Regarding the fusion strategy in Eq. 3, note that the constraints derived from individual modalities already achieve high accuracy in Table 4, so the primary purpose of the fusion strategy is to enable complementary interactions between constraints from different modalities.

F Decoding cases and Cluster Visualization

We utilized the t-SNE (van der Maaten and Hinton, 2008) algorithm to demonstrate the results of our clustering method, as shown in Figure 6. We compared the results of VBx, E2CP with semantic and visual constraints, and ground-truth, and observed that the E2CP with semantic and visual constraints



Acoustic Results:

0 106.67 109.06 My baking experience is limited to eating raw cookie dough

0 109.07 110.96 and the last time we made one of these videos

1 111.3 112.82 last time on without a recipe

1 112.83 116.26 my bread looked amazing and tasted very plain

1 116.44 119.92 so this time I'm not going to forget the foot

1 120.32 124.31 I like to pride myself on the fact that I'm good at making food for others

MultiModal Results:

3 106.67 109.06 My baking experience is limited to eating raw cookie dough

3 109.07 110.96 and the last time we made one of these videos

5 111.3 112.82 last time on without a recipe

5 112.83 116.26 my bread looked amazing and tasted very plain

5 116.44 119.92 so this time I'm not going to forget the foot

0 120.32 124.31 I like to pride myself on the fact that I'm good at making food for others

Ground Truth Results:

c1 105.894 111.193 My baking experience is limited to eating rock cookie dough, and the last time we made one of these videos.

c2 111.193 120.328 Last time on without a recipe, my bread looked amazing and tasted very plain, so this time I'm not gonna forget the flavor.

c4 120.328 130.223 I like to pride myself on the fact that I'm good at making food for others.

Figure 7: Decoding case

Table 5: The sensitivity of the system to the α_k and β.

α_1	α_2	β	Constraints Acc	DER(%)
1	1	0.4	0.9906	9.01
1.2	1	0.4	0.9905	9.01
0.8	1	0.4	0.9906	9.01
1	1.2	0.4	0.9906	9.01
1	0.8	0.4	0.9905	9.01
1	1	0.3	0.9934	9.00

method effectively improved the clustering results compared to VBx, especially in terms of clustering the points at the edges of clusters, after introducing constraints.

In Figure 7, we present a decoding case, where each row follows the format: speaker-ID, start-time, end-time and content. For the convenience of aligning timestamps with textual information, the decoding results presented here do not include punctuation marks such as periods. In both the acoustic-only and multimodal results, the same force-alignment results was applied, resulting in identical timestamp values. It can be observed that there is a clear semantic transition point near 111.3 seconds. The acoustic-only result fails to correctly segment this speaker change point; however, by leveraging semantic constraints, our method successfully separates the speakers.