Improving Speaker Diarization Using Semantic Information: Joint Pairwise Constraints Propagation

Anonymous ACL submission

Abstract

Speaker diarization, an important task in speech processing, has been predominantly relied on acoustic signal analysis to differentiate speakers. This reliance on acoustic features often overlooks the wealth of semantic content within speech that can provide additional clues regarding speaker identities. Addressing this gap, our study introduces a semantically enriched diarization approach that extends beyond the acoustic domain, tapping into the nuances of linguistic content. We present a novel method that employs advanced language understanding to extract semantic cues, which are integral to discerning speaker contributions within conversations. Our approach utilizes these cues to formulate pairwise constraints, introducing a multi-modal clustering process to segment and classify speakers and their spoken contents. By integrating these semantically derived constraints into the diarization pipeline, we achieve substantial improvements in accuracy. Extensive evaluations on public dataset illustrate that our method consistently outstrips acoustic-only systems, offering a more holistic perspective on speaker diarization by fully embracing the semantic information of speech.

1 Introduction

Speaker Diarization (SD) addresses the pivotal question of "who speaks when" by assigning speaker labels to segments of an audio stream (Wang et al., 2017; Du et al., 2022). In most application settings, these labels are integrated with the transcribed words or sentences from an Automatic Speech Recognition (ASR) system. Despite the wealth of transcribed text available, prevailing SD systems (Park et al., 2021) predominantly rely on acoustic signals. A traditional SD system typically comprises the following components: (1) a Voice Activity Detection (VAD) module; (2) a speaker embedding extractor employing technologies such as x-vector (Snyder et al., 2018), d-vector (Wan et al., 2017), ECAPA-TDNN (Dawalatabad et al., 2021), and ResNets (Zhou et al., 2021; Chen et al., 2023); and (3) a speaker clustering module like Agglomerative Hierarchical Clustering (AHC) (Han et al., 2008), Spectral Clustering (SC) (Wang et al., 2017), or UMAP-HDBSCAN (Zheng and Suo, 2022). While these components are effective, traditional SD systems overlook the semantic information embedded within speech content, which limit system's robustness. This oversight becomes particularly detrimental in challenging acoustic environments characterized by noise and reverberation, where relying on purely acoustic information can lead to degradation in performance.

Some previous works (Zuluaga-Gómez et al., 2021; Flemotomos and Narayanan, 2022; Park and Georgiou, 2018; Paturi et al., 2023) have utilized semantic information in speaker diarization tasks, but only limited to two-speaker scenarios with clearly distinctive semantic roles, such as job interviews and doctor-patient medical consultations. More recent works (Park et al., 2023; Wang et al., 2024; Cheng et al., 2023) leverage large language models to improve speaker-related text diarization. However, their use of semantic information are effective only for shorter context, primarily influencing boundary decisions within a few sentences. To have a better understanding of the global speaker structure of the entire conversations or multi-party meetings, the state-of-the-art approaches still resort to acoustic-only speaker clustering.

In this work, we address unrestricted multi-party meetings and conversations, characterized by random contents, multiple speakers and long duration, highlighting the need for comprehensive speaker profiling. We improve upon the acoustic-only speaker clustering methods by introducing a multi-modal clustering approach that fully considers the acoustic and semantic information in global context, namely the **Joint Pairwise Constraints Propagation (JPCP)** framework. The main contribu-

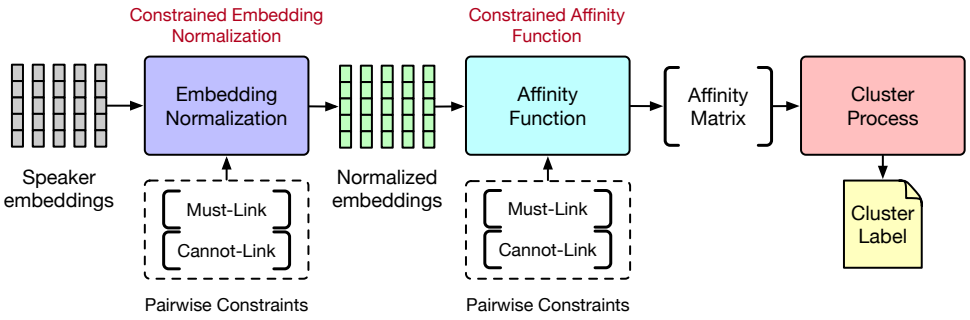

Figure 1: Semantic pairwise constraints are applied in both Embedding Normalization and the Affinity Function in the clustering process of speaker embeddings.

tions of this paper are as follows:

(1) We introduce additional spoken language processing (SLP) modules to extract speaker-related information from transcripts and develop a strategy to construct pairwise semantic constraints.

(2) We incorporate pairwise constraints propagation methods to embedding normalization and affinity function, and investigate the effectiveness of these semantic constraints.

2 Semantic Speaker Constraints

2.1 Semantic Speaker-related Tasks

In our study, we follow the extraction pipeline detailed in (Cheng et al., 2023) to obtain speaker embeddings from conversational speech signals. Utilizing a speaker embedding extractor, we obtain a series of embeddings $E = \{e_1, e_2, ..., e_N\}$ and N is the number of speaker embeddings. These embeddings are then mapped to the corresponding transcribed text using a Forced Alignment (FA) module (Shi et al., 2023), which aligns the text with the audio timestamps, ensuring accurate synchronization between audio and spoken words.

To extract speaker-related information from the transcriptions, we focus on 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.

Accounting for the ASR errors that can impact SLP tasks, we employ a hybrid approach that fuses acoustic and semantic information(Cheng et al., 2023) by integrating the outputs of semantic models with acoustic embeddings to further enhance the performance of identifying dialogue states and speaker turns, which are pivotal to SD system.

2.2 Semantic Pairwise Constraints

Prevailing speaker clustering systems extract speaker embeddings from acoustic signals only, making it difficult to incorporate semantic content. In order to leverage both modalities in the clustering process, we propose to use semantic information to establish constraints among embeddings.

We derive two types of constraints from semantic speaker-related information: must-link $\mathcal{M}$ and cannot-link $\mathcal{C}$, defined as:

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

where $l(\cdot)$ denotes the speaker label associated with an embedding.

The criteria for establishing $\mathcal{M}$ and $\mathcal{C}$ are as follows: $\mathcal{M}$ is formed between two embeddings if they originate from the same non-dialogue segment. Conversely, $\mathcal{C}$ is established between embeddings separated by a detected speaker-turn boundary.

2.3 Propagate and Normalize Semantic Constraints

The semantic constraints obtained by the aforementioned methods exhibit two shortcomings: (1) the SLP modules may introduce errors into the constraints; (2) due to the decoding text length limitations of SLP modules, the constructed constraints typically only apply to embeddings within a specific time duration, failing to affect embeddings separated by long intervals. In this paper, we design propagation and normalization mechanisms to mitigate these issues effectively.

We improve constraint reliability by integrating additional must-link and cannot-link pairs with cosine similarity. These pairs are identified based on their affinity scores relative to thresholds θ_m and

θ_c, indicating strong similarity and dissimilarity respectively.

$$\begin{aligned}\mathcal{M}_p &= \{(e_i, e_j)|g(e_i, e_j) > \theta_m\}, \\ \mathcal{C}_p &= \{(e_i, e_j)|g(e_i, e_j) < \theta_c\},\end{aligned} \tag{2}$$

where $g(\cdot)$ is the cosine similarity function. The semantic constraints can be filtered by cosine similarity results:

$$\begin{aligned}\mathcal{M}_n &= \{(e_i, e_j)|(e_i, e_j) \in \mathcal{M} \ \& \ (e_i, e_j) \notin \mathcal{C}_p\} \\ \mathcal{C}_n &= \{(e_i, e_j)|(e_i, e_j) \in \mathcal{C} \ \& \ (e_i, e_j) \notin \mathcal{M}_p\}\end{aligned} \tag{3}$$

The final must-link $\mathcal{M}_f$ and cannot-link $\mathcal{C}_f$ can be constructed as:

$$\begin{aligned}\mathcal{M}_f &= \mathcal{M}_p \cup \mathcal{M}_n \\ \mathcal{C}_f &= \mathcal{C}_p \cup \mathcal{C}_n\end{aligned} \tag{4}$$

For more details on the semantic constraints constructions process, please refer to the appendix A.

3 Semantic Constrained Diarization

The **JPCP** framework is illustrated in Figure 1. Similar to traditional acoustic-only SD systems, **JPCP** processes speaker embeddings through embedding normalization, affinity function and clustering to assign speaker labels. However, it distinguishes the acoustic-only SD systems by integrating semantic constraints into each of these stages, thereby improving the influence of semantic insights on the final clustering outcomes.

3.1 Constrained Embedding Normalization

We incorporate the semi-supervised dimension reduction (SSDR) algorithm (Zhang et al., 2007) into the speaker embedding normalization module to incorporate semantic constraints. We construct the following weight matrix $\mathbf{S}$:

$$\mathbf{S}_{ij} = \begin{cases} \frac{1}{N^2} + \frac{\alpha}{|\mathcal{M}|} & \text{if } (e_i, e_j) \in \mathcal{M}_f, \\ \frac{1}{N^2} - \frac{\beta}{|\mathcal{C}|} & \text{if } (e_i, e_j) \in \mathcal{C}_f, \\ \frac{1}{N^2} & \text{otherwise,} \end{cases} \tag{5}$$

where α and β are parameters that adjust the influence of must-link and cannot-link constraints.

The aim of constrained embedding normalization is to identify projection matrix $\mathbf{W}$ yield low-dimensional embeddings $\mathbf{W}^T e_k \in \mathbb{R}^d$ and d is the new dimension, aligning with both the manifold structure of the original embeddings and semantic constraints. We define the objective function $J(\mathbf{W})$ as:

$$J(\mathbf{W}) = \mathbf{W}^T E \mathbf{L}_s E^T \mathbf{W}, \tag{6}$$

with $\mathbf{L}_s = \mathbf{D}_s - \mathbf{S}$ being the Laplacian matrix and $\mathbf{D}_s$ is the degree matrix of $\mathbf{S}$. The optimization problem with $J(\mathbf{W})$ is a standard eigenproblem, solvable by determining the eigenvectors of $E\mathbf{L}E^T$ associated with the largest eigenvalues.

3.2 Constrained Affinity Function

After embedding normalization, we compute the affinity matrix $\mathcal{A} = \{\mathcal{A}_{ij}\}_{N\times N}$ where $\mathcal{A}_{ij} = d(e'_i, e'_j)$, with $d(\cdot)$ representing the cosine distance and e'_i representing the normalized embeddings. In traditional acoustic-only systems(Wang et al., 2017), $\mathcal{A}$ is subjected to a series of refinement steps like row-wise thresholding and symmetrization.

To further refine the affinity matrix with semantic speaker-related information, we introduce E^{2}CP (Lu and Peng, 2011), which effectively models the global relationships implied by the semantic constraints. The constraints are initially encoded in a matrix $\mathcal{Z}$:

$$\mathcal{Z}_{ij} = \begin{cases} +1 & \text{if } (e_i, e_j) \in \mathcal{M}_f, \\ -1 & \text{if } (e_i, e_j) \in \mathcal{C}_f, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Then the matrix is transformed to a propagated constraints matrix $\hat{\mathcal{Z}}$:

$$\hat{\mathcal{Z}} = (1-\lambda)^2(\mathbf{I} - \lambda\mathbf{L}_e)^{-1}\mathcal{Z}(\mathbf{I} - \lambda\mathbf{L}_e)^{-1}, \tag{8}$$

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 of the semantic 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 $\hat{\mathcal{Z}}$:

$$\hat{\mathcal{A}}_{ij} = \begin{cases} 1 - (1 - \hat{\mathcal{Z}}_{ij})(1 - \mathcal{A}_{ij}) & \text{if } \hat{\mathcal{Z}}_{ij} \geq 0, \\ (1 + \hat{\mathcal{Z}}_{ij})\mathcal{A}_{ij} & \text{if } \hat{\mathcal{Z}}_{ij} < 0. \end{cases} \tag{9}$$

Upon calculating the affinity matrix $\hat{\mathcal{A}}$, it is then fed into the spectral clustering (SC) process to derive the ultimate speaker clustering outcome.

4 Experimental setup

4.1 Dataset and Metrics

Our experiments are evaluated on AISHELL-4 (Fu et al., 2021) which focuses on multi-speaker meeting whose content is manually annotated.

We report the following clustering algorithm metrics: Normalized Mutual Information (NMI)

Diarization System	Constraints	Speaker Clustering Methods	Cluster Metrics		Speaker Metrics		
			ARI	NMI	SpkDiff #	CpWER (%)	TextDER (%)
Acoustic Only System	No Semantic Constraints	SC	-	-	11	26.1816	3.7723
Semantic Turn-Cut System	No Semantic Constraints	SC	0.8901	0.8616	11	25.6421	3.4636
JPCP System	Semantic Constraints	SSDR + SC	0.9010	0.8857	11	25.9185	3.8122
	Semantic Constraints	E2CP + SC	0.9162	0.8863	10	25.2774	3.0967
	Semantic Constraints	SSDR + E2CP + SC	**0.9171**	**0.8871**	**9**	**25.3168**	**3.0379**
	Simulation Constaints 6%	SSDR + E2CP + SC	0.9939	0.9879	4	24.5919	1.9810
	Simulation Constaints 12%	SSDR + E2CP + SC	**0.9961**	**0.9927**	**3**	**24.4809**	**1.9028**

Table 1: Performance evaluation of cluster metrics and speaker diarization results. SpkDiff # refers to difference in number of speakers between inference and ground truth.

and Adjusted Rand Index (ARI)(Chac'on and Rastrojo, 2020). As the transcribed text and FA module have been used in the pipeline, we directly report the Concatenated Minimum-permutation Word Error Rate (cpWER)(Watanabe et al., 2020). Additionally, we use the metric Text Diarization Error Rate (TextDER)(Gong et al., 2023), to evaluate the amount of text assigned to wrong speakers.

4.2 Configuration of Modules

Our system followed the pipeline outlined in (Cheng et al., 2023) while incorporating improvements to certain acoustic models. Our speaker embedding extractor is an adaptation of CAM++ (Wang et al., 2023)[1], which has been trained on a vast Mandarin speech corpus. For ASR modules, we utilize the Paraformer (Gao et al., 2022), which has been trained with the aid of the FunASR (Gao et al., 2023) toolkits[2]. These models are open-source and constant across all experiments.

For dialogue detection and speaker turn detection tasks, we rely on semantic models built upon a pre-trained BERT language model (Devlin et al., 2019). We generated training samples using a sliding window approach with a window length of 64 tokens and a shift of 16 tokens. The labels for semantic tasks stem from the manually annotated speaker information within the speech content.

The development of semantic constraints is outlined in Section 2.2. To evaluate the effectiveness of our proposed method, we also generated simulated pairwise constraints from ground truth speaker labels.

For the experiments on SSDR, we have set α to 10.0 and β to 2.0. The dimensionality of the embeddings extracted by CAM++ is 192, and we have selected a reduced dimension d of 180. In the E^{2}CP experiments where semantic constraints are employed, the parameter λ is set to 0.6. However, when utilizing simulated constraints, which are associated with a higher level of confidence, we have set λ to 0.2.

[1]The speaker embedding extractor came from https://github.com/alibaba-damo-academy/3D-Speaker

[2]The ASR and FA models came from https://github.com/alibaba-damo-academy/FunASR

5 Results and Disscussions

The results presented in Table 1 highlight the comparative analysis between our approach and acoustic-only system. We have employed the "Semantic Turn-Cut" strategy from (Cheng et al., 2023), which combines semantic timestamps with VAD results for more precise segmentation.

Our JPCP strategy demonstrates a noticeable improvement over the acoustic-only baseline. Specifically, our method achieved a 19% decrease in TextDER and 3.3% in cpWER. The decline in SpkDiff also substantiates that our method has a positive impact on the global clustering performance. When comparing against SSDR, the E^{2}CP variant of our JPCP approach shows even more pronounced effects on the clustering results.

The performance of these methods largely depends on the quality of the constraints, where the gains for the JPCP approach with conventionally inferred constraints are more modest due to their inherent limitations. However, the experiments with simulated constraints underscore the potential of our method. For a more detailed exploration of the results obtained using simulated constraints, readers are referred to the appendix B.

6 Conclusion

In this work, we have proposed the Joint Pairwise Constraint Propagation (JPCP) framework, a novel system that employs SLP modules to generate semantic constraints from ASR transcriptions and integrates them into the speaker clustering process. Our approach significantly enhances speaker diarization in multi-party meeting scenarios while preserving the integrity of the core system modules. Experimental results validate that incorporating semantic constraints markedly improves speaker diarization performance.

7 Limitations

In this work, the SLP modules employed, specifically Dialogue Detection and Speaker Turn Detection tasks, are not entirely accurate in capturing speaker-related semantic information in certain complex meeting scenarios. For instance, when meetings draw to a close, participants often express similar opinions, thereby blurring semantic speaker turn points. In future work, we aim to develop more suitable speaker-related SLP tasks to overcome these challenges.

Additionally, this study does not employ Large Language Models as done in works like (Park et al., 2023; Wang et al., 2024), instead opting for BERT. In our future work, we plan to utilize more advanced language models in the hopes of obtaining improved and more comprehensive speaker-related semantic information.

References

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, Haibo Wang, Luyao Cheng, Qian Chen, and Jiajun Qi. 2023. An enhanced res2net with local and global feature fusion for speaker verification. *ArXiv*, 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*.

Nauman Dawalatabad, Mirco Ravanelli, Franccois Grondin, Jenthe Thienpondt, Brecht Desplanques, and Hwidong Na. 2021. Ecapa-tdnn embeddings for speaker diarization. In *Interspeech*.

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*.

Zhihao Du, Shiliang Zhang, Siqi Zheng, and Zhijie Yan. 2022. Speaker overlap-aware neural diarization for multi-party meeting analysis. In *Conference on Empirical Methods in Natural Language Processing*.

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

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 *Interspeech*.

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*.

Chen Gong, Peilin Wu, and Jinho D. Choi. 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)*, pages 778–783.

Kyu Jeong Han, Samuel Kim, and Shrikanth S. Narayanan. 2008. Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:1590–1601.

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.

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, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth S. Narayanan. 2021. A review of speaker diarization: Recent advances with deep learning. *ArXiv*, abs/2101.09624.

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

Xian Shi, Yanni Chen, Shiliang Zhang, and Zhijie Yan. 2023. Achieving timestamp prediction while recognizing with non-autoregressive end-to-end asr model. *ArXiv*, abs/2301.12343.

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

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez-Moreno. 2017. Generalized end-to-end loss for speaker verification. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883.

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, P. A. Mansfield, and Ignacio Lopez-Moreno. 2017. Speaker diarization with lstm. *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.

Daoqiang Zhang, Zhi-Hua Zhou, and Songcan Chen. 2007. Semi-supervised dimensionality reduction. In *SDM*.

Siqi Zheng and Hongbin Suo. 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)*, pages 8097–8101.

Tianyan Zhou, Yong Zhao, and Jian Wu. 2021. Resnext and res2net structures for speaker verification. *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 301–307.

Juan Zuluaga-Gómez, Seyyed Saeed Sarfjoo, Amrutha Prasad, Iuliia Nigmatulina, Petr Motlícek, Karel Ondrej, Oliver Ohneiser, and Hartmut Helmke. 2021. Bertraffic: Bert-based joint speaker role and speaker change detection for air traffic control communications. *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 633–640.

A Constraints Construction Cases and Illustrations

Figure 2 presents a concrete example of how constraints are constructed from the two SLP tasks: For the Dialogue Detection task, a text segment judged as non-dialogue suggests that all corresponding embeddings are related by must-link constraints, akin to the green connections in Figure 3. Conversely, for the Speaker-Turn Detection task, the presence of a transition point dictates that embeddings spanning this point should be connected with cannot-link constraints, represented by the red connections in the figure.

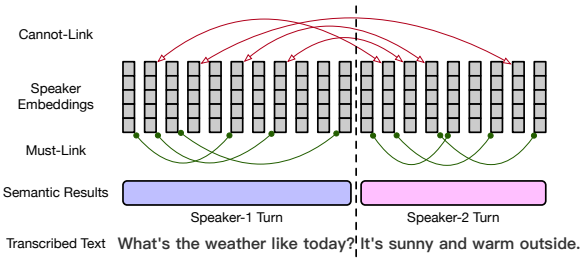

Figure 2: A case of strategy for constructing constraints.

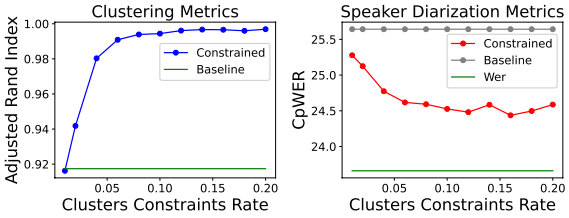

Figure 3: The impact of pairwise constraints rate on both clustering metrics and the effectiveness of the overall speaker diarization system.

There is an inherent transitivity associated with must-link constraints. That is, if $(e_i, e_j) \in \mathcal{M}$ and $(e_j, e_k) \in \mathcal{M}$, then it can be inferred that $(e_i, e_k) \in \mathcal{M}$. Unfortunately, such a property does not extend to cannot-link constraints. If $(e_i, e_j) \in \mathcal{C}$ and $(e_j, e_k) \in \mathcal{C}$, 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, constraints derived solely from semantic information can only determine the relationship between embeddings adjacent to a speaker-turn. This limitation underlines the need to supplement constraints with additional information, such as cosine scores, as described in Section 2.3.

B Experiments on Simulation Constraints

In this section, we will detail the construction of simulation constraints to explore the upper-bound performance of our proposed system. By utilizing the actual speaker labels derived from the ground truth timestamps, we can assign each speaker embedding with its corresponding speaker label. Following the definitions of must-link and cannot-link constraints in Section 1, we randomly select a proportion p of constraints from the $N \times (N-1)/2$ possible embedding pairs (e_i, e_j) to serve as the experimental constraints for our study.

Figure 3 presents the results of our experiments,

illustrating a significant improvement in system performance with an increasing proportion of constraints. Remarkably, the use of a minimal amount of constraints (approximately 6%) achieves results that approach the theoretical upper bound, compared to systems not employing any constrained clustering techniques.