Semantic-Aware Text-Based Speaker Diarization: Leveraging Language Models for Sentence-Level Speaker Change Detection

Peilin Wu
Computer Science
Emory University
Atlanta, GA 30322 USA
peilin.wu@emory.edu

Jinho D. Choi
Computer Science
Emory University
Atlanta, GA 30322 USA
jinho.choi@emory.edu

Abstract

This paper introduces a novel text-based Speaker Diarization (SD) method leveraging semantic information through advanced language models, diverging from traditional audio-based approaches. Our model, focusing on Sentence-level Speaker Change Detection (SCD) in two-speaker conversations, employs single and multiple prediction mechanisms to improve accuracy in identifying speaker turns. The methodology is validated on a dataset processed by Automatic Speech Recognition (ASR), demonstrating competitive performance against conventional audio-based SD systems, especially in short conversational contexts.

1 Introduction

Speaker Diarization (SD), the task to determine speakers for audio segments (Park et al., 2022a), is essential for parsing conversational audio, especially when combined with Automatic Speech Recognition (ASR) to determine "who speaks what" for conversational AI data preparation. While traditional SD methods have progressed from combining segmentation (Bredin and Laurent, 2021) and clustering (Bredin, 2023) to End-to-End Neural Diarization (EEND) systems (Landini et al., 2023), effectively integrating semantic features to enhance SD has been challenging. Previous methods either underutilized modern language models (Flemotomos et al., 2020) or only applied them as a post-processing error corrector (Paturi et al., 2023). This paper presents a groundbreaking SD approach that uses text as the primary input, harnessing advanced language models to seamlessly incorporate semantic information into the diarization process.

2 Related Work

Existing literature on SD largely focuses on audio-only methods (Horiguchi et al., 2020), with varying degrees of integration of semantic features, either directly within the diarization pipeline (Park et al., 2023) or indirectly as a means of post-processing (Wang et al., 2024). Joint ASR+SD efforts (Kanda et al., 2022) and multimodal approaches (Cheng et al., 2023) have been explored, but text-based or semantic-focused SD remains underdeveloped. Our work aims to fill this gap by presenting a text-based SD model that harnesses the full potential of semantic features.

3 Text-based Speaker Diarization

3.1 Task Overview

This paper tackles the task of text-based SD as Sentence-level Speaker Change Detection (SCD) in two-speaker conversations. This approach prioritizes sentence-level analysis over word-level for its richer contextual information, which is more conducive to accurate speaker identification. Two-speaker conversations represent a common and pragmatically significant scenario, making it an ideal focus for demonstrating the capabilities of text-based SD.

3.2 Model Design

3.2.1 Single Prediction Model

The single prediction model operates by evaluating the probability of a speaker change between sentences, using surrounding utterances as context. Formally, let $S = \{s_1, s_2, ..., s_n\}$ as be a sequence of n sentences in a conversation. The objective is to predict a binary variable y_i for each pair of consecutive sentences (s_i, s_{i+1}), where $y_i = 1$ if a speaker change occurs between s_i and s_{i+1}, or $y_i = 0$ otherwise. The model utilizes a context window of certain number of sentences at front and at back of sentence s_{i+1} as the input. The change prediction result for S will be a sequence of speaker change predictions $R = \{y_1, y_2, ..., y_{n-1}\}$, which can be used to deduce the final speaker information.

3.2.2 Multiple Prediction Model

While straightforward, the single prediction approach is prone to errors due to its reliance on a limited contextual window. To enhance accuracy and robustness, we introduced a multiple prediction model that aggregates predictions over several points within a dialogue. Let $W = \{w_1, w_2, ..., w_m\}$ be a sequence of windows, where each window w_j consists of a subsequence of sentences from S, and m is the total number of windows covering the conversation. Each window w_j overlaps with its predecessors and successors, ensuring comprehensive coverage of the conversation. The objective is to predict a sequence of binary variables y_i for each window w_j, where each element of y_i corresponds to a potential speaker change within w_j. An aggregation mechanism for all windows W, which can be a majority vote or a weighted average based on confidence scores, is applied to form a robust final result that extends contextual insights and mitigates the impact of isolated prediction errors.

3.3 Training Data Processing

Acknowledging the primary application of text-based SD on ASR-generated transcripts, the training data is produced using state-of-the-art ASR to simulate real-world ASR discrepancies, which were then aligned with ground-truth annotations to produce transcripts with speaker information. This method ensures that the model is fine-tuned for practical applications, particularly in improving the fidelity of ASR-generated transcripts for SD tasks.

4 Experiment

4.1 Data

4.1.1 Dataset

This paper uses a curated dataset from seven diverse, open-domain sources. The dataset is split with a ratio of 8:1:1 for train:development:test set on conversation level. The details of the curated dataset can be seen at Table 2 in Appendix A.1.

4.1.2 Data Processing

We prepared the training data using state-of-the-art ASR from OpenAI Whisper (Radford et al., 2022), enriched punctuation with GPT-4, and employed spaCy (Honnibal et al., 2020) for sentence segmentation to closely simulate real-world ASR conditions.

4.2 Methodology

For both Single Prediction and Multiple Prediction Model, the T5-3B (Raffel et al., 2020) is used for actual implementation. The T5-11B was also tested, but its performance did not exceed that of the T5-3B version.

4.3 Result

The Word Diarization Error Rate (WDER) (Shafey et al., 2019) is used as the evaluation metric. To adapt for different lengths of conversations in the curated dataset, WDER-S and WDER-W are introduced as the weighted averages of WDER according to the number of sentences and words in the conversation. The results in Table 1 indicate that text-based SD, especially with multiple predictions, offers a promising alternative to traditional audio-based methods, excelling in short conversational contexts.

Model	WDER	WDER-S	WDER-W
pyannote (Bredin, 2023)	0.253	0.157	0.146
x-vector+SC	0.349	0.144	0.118
x-vector+AHC	0.292	0.279	0.278
ECAPA-TDNN+SC	0.374	0.145	0.115
ECAPA-TDNN+AHC	0.286	0.268	0.267
NeMo-TitaNet (Koluguri et al., 2021)	0.220	0.129	0.103
NeMo-MSDD (Park et al., 2022b)	0.207	0.120	0.100
TOLD (Wang et al., 2023)	0.164	**0.099**	**0.069**
T5-3B Single Prediction	0.384	0.440	0.444
T5-3B Multiple Prediction	**0.103**	0.104	0.115

Table 1: Performance comparison with audio-based SD systems (the lower the better). The x-vector (Snyder et al., 2018) and ECAPA-TDNN (Desplanques et al., 2020) are two speech embedding extraction methods. The SC and AHC in the table refer to spectral clustering (Wang et al., 2022) and agglomerative hierarchical clustering (Pedregosa et al., 2011).

5 Conclusion

This paper presents a novel approach to SD by integrating semantic features into the diarization process, offering a viable alternative to audio-based methods. The proposed text-based SD model, employing sentence-level analysis for speaker change detection, significantly outperforms traditional systems in terms of Word Diarization Error Rates (WDER). This advancement highlights the potential of semantic information in enhancing diarization accuracy and opens new avenues for research in conversational AI, suggesting further exploration into complex conversational scenarios and model refinements for broader application.

References

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines. In *Proc. Interspeech 2018*, pages 1561–1565.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Interspeech*.

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

Alexandra Canavan, David Graff, and George Zipperlen. 1997. CALLHOME American English Speech LDC97S42. Web Download.

Alexandra Canavan and George Zipperlen. 1996. CALLFRIEND American English-Non-Southern Dialect LDC96S46. Web Download.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael 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 *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, page 28–39, Berlin, Heidelberg. Springer-Verlag.

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 *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14068–14077, Toronto, Canada. Association for Computational Linguistics.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, pages 3830–3834.

John DuBois, Wallace L. Chafe, Charles Meyer, and Sandra A. Thompson. 2000-2020. Santa Barbara Corpus of Spoken American English. Web Download.

Nikolaos Flemotomos, Panayiotis Georgiou, and Shrikanth Narayanan. 2020. Linguistically aided speaker diarization using speaker role information. In *The Speaker and Language Recognition Workshop (Odyssey 2020)*, odyssey$_2$020.$ISCA$.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

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.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Naoyuki Kanda, Xiong Xiao, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. 2022. Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed asr.

Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. 2021. Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context.

Federico Landini, Mireia Diez, Themos Stafylakis, and Lukáš Burget. 2023. Diaper: End-to-end neural diarization with perceiver-based attractors. *arXiv preprint arXiv:2312.04324*.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

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

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. 2022a. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317.

Tae Jin Park, Nithin Rao Koluguri, Jagadeesh Balam, and Boris Ginsburg. 2022b. Multi-scale speaker diarization with dynamic scale weighting.

Rohit Paturi, Sundararajan Srinivasan, and Xiang Li. 2023. Lexical Speaker Error Correction: Leveraging Language Models for Speaker Diarization Error Correction. In *Proc. INTERSPEECH 2023*, pages 3567–3571.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Laurent El Shafey, Hagen Soltau, and Izhak Shafran. 2019. Joint speech recognition and speaker diarization via sequence transduction.

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.

Jiaming Wang, Zhihao Du, and Shiliang Zhang. 2023. Told: A novel two-stage overlap-aware framework for speaker diarization.

Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopez Moreno. 2022. Speaker diarization with lstm.

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

A Appendix

A.1 Dataset Details

The sources of the dataset are as follows:

Corpus	Hour	# of Dialogue
AMI Corpus (Carletta et al., 2005)	100	171
CallFriend (Canavan and Zipperlen, 1996)	20	41
CallHome English (Canavan et al., 1997)	20	176
CHiME-5 (Barker et al., 2018)	50	20
DailyTalk (Lee et al., 2023)	20	2541
ICSI Corpus (Janin et al., 2003)	72	75
SBCSAE (DuBois et al., 2000-2020)	23	60

Table 2: Corpora used for the curated dataset.

This table 2 details the composition of our curated dataset, indicating the volume (in hours) and the number of dialogues sourced from each corpus. The selection was made to ensure a wide range of conversational contexts and settings, from formal meetings (AMI Corpus, ICSI Corpus) to casual conversations (CallFriend, CallHome English), and challenging acoustic environments (CHiME-5), providing a comprehensive base for training and evaluating conversational AI systems.