CASCA: Leveraging Role-Based Lexical Cues for Robust Multimodal Speaker Diarization via Large Language Models

William Nehrboss
wnehrboss@casca.ai

Abstract

In this paper, we introduce CASCA[1] a multimodal speech diarization framework that incorporates speaker role information. Motivated by the challenges of diarizing single-source customer-employee interactions in noisy environments, this framework utilizes a cascading sequence of fine-tuned large language models to characterize distinctions in speaker roles. Audio with linguistic content associated with particular roles is used to formulate acoustic speaker profiles; these profiles reduce the subsequent clustering task into a classification task. CASCA is robust to sparsity or low signal-to-noise ratios, conditions that tend to confound traditional clustering algorithms. Although intended for those domains with clear role distinction, e.g., doctor-patient, teller-customer, through topic segmentation, CASCA captures transient, topic-level speaker role information to reliably identify speaker profiles. This expands the domain of applicability. We validate the effectiveness of our approach on a benchmark of two-speaker conversations from a variety of domains, achieving an 80% reduction in word diarization error rates over our conventional baseline.

1 Introduction

Speaker diarization is the process of segmenting recorded audio according to the speaker source. It determines who spoke when by splicing audio into regions of homogeneous speaker source and applying a speaker tag to those regions. Accurate speaker diarization is crucial for effective conversation understanding, which is essential in myriad applications from customer service analytics to medical recordkeeping. Spoken conversations are rich in both linguistic and acoustic information. However, most current diarization systems utilize only acoustic information in speaker assignment. Some of the most popular diarization algorithms (Serafini et al., 2023), including Pyannote (Plaquet and Bredin, 2023), which we use as our baseline, are cluster-based. The general architecture of these systems is as follows:

- *Voice activity detection* isolates speech from non-speech.
- *Segmentation* splits regions of speech into smaller segments with a single active speaker.
- *Embedding extraction* yields vector representations capturing key audio characteristics.
- *Clustering* groups these embeddings to determine speaker assignment.

These systems, however, tend to generalize poorly to varied real-world situations. Embedding clusters are often imbalanced, non-Gaussian, or indistinct due to uneven speaker participation, shifts in tone or intonation, or background noise. These factors make accurately identifying cluster boundaries, and, in turn, speaker assignment unreliable. A prime example of a conversation that yields indistinct embedding clusters is presented in Table 3a. This work aims to solve these challenges by reformulating the clustering step into a classification step by incorporating a key source of speaker differentiation available in the linguistic content of the conversation: speaker roles.

Speaker roles within a given conversation tend to be distinct. The degree of this distinction can be high, for example, in conversations between a doctor and patient or salesperson and customer, or low, such as in casual conversations between two siblings or two roommates. Speaker roles can provide strong cues about the correct speaker assignment of certain speech segments within a conversation. For example, in a conversation between a doctor and a patient, the speech segments corresponding to the phrases “I am experiencing chest pain” and “I am going to recommend an X-ray” can be at-

[1] **C**ontext **A**ware **S**peech **C**lassification **A**rchitecture https://github.com/CASCA-Labs/CASCA

tributed to the roles of the patient and the doctor, respectively. This information can be used to isolate particular speech segments that correspond to each role. These segments, representing a sort of acoustic speech profile of each speaker, simplify the subsequent speaker assignment task into a classification task, avoiding the need for unreliable clustering algorithms.

1.1 Types of Speaker Role Distinction

In this paper, we will refer to two types of speaker role distinction: strong role distinction and weak role distinction. Strong role distinction is present when the role of each speaker is stable and pertains to the speaker themselves, more or less independent of the conversation. For example, in the case of a conversation between a bank teller and a customer, the roles of the speakers and what they might be expected to say are strongly determined by their relationship to the service being provided. Weak role distinction is present when the role of each speaker is fluid throughout the conversation. In these cases, there are no overriding contextual factors that explain the linguistic content. Importantly, however, even when roles are more fluid, speakers typically assume identifiable roles within certain segments of the conversation that relate to specific topics. For instance, in the conversation summary presented in Table 1, although the speakers are peers without apparent strong role distinction, they assume different roles within each topic segment: one informs the other about a promotion in the first segment and updates her about a mutual friend in the second. Leveraging this weak role distinction presents a challenge but is crucial to the robustness of our approach. This motivates the specification of the first stage of our pipeline, which extracts role distinctions on the topic level (see Sections 2.1–2.3).

1.2 Prior Work

Utilizing linguistic information is recognized as a key opportunity to enhance diarization systems. Recent advances in n-gram models, particularly transformer-based models, have made the use of this information more accessible and valuable. Multimodal diarization approaches leveraging these models have proven effective. BERT-based models, for instance, have shown promise in post-processing transcribed dialogues and correcting errors from misaligned speaker turns (Paturi et al., 2023). Efforts have also been made to use a priori knowledge of speaker identities for downstream classification tasks in different contexts. One study (Flemotomos et al., 2020) involved training classifiers on sentence-level speech segments to construct speaker profiles in therapist-patient conversations. A subsequent investigation (Flemotomos and Narayanan, 2022) extended this approach to two domains, using linguistic information to constrain embedding clusters. Another study (Prasad et al., 2021) addressed problematic audio data in an aviation setting using a related method. Although these efforts are valuable, they are limited in certain respects, particularly in their dependence upon a priori information of speaker identities. In this paper, we tackle the more difficult problem of role-aided diarization without prior knowledge of speaker identities, where leveraging relational information is a central aspect of our methodology.

2 LLM Sequence

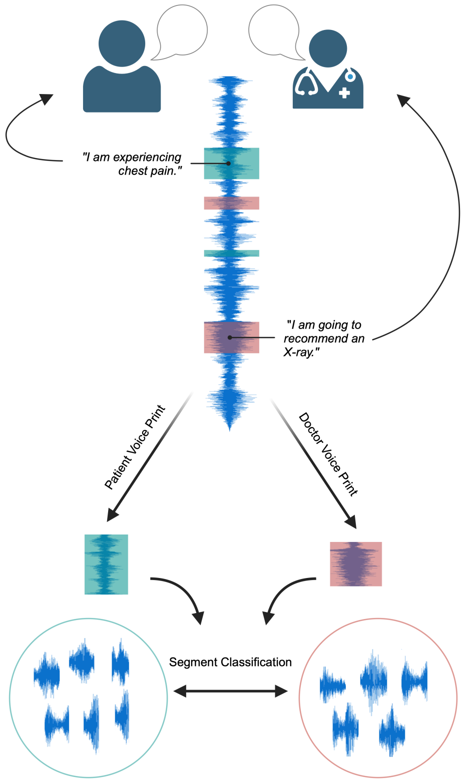

"I am experiencing chest pain."
"I am going to recommend an X-ray."
Patient Voice Print
Doctor Voice Print
Segment Classification

Figure 1: The motivation for CASCA. The roles of the doctor and patient can be used identify segments of speech that belong to each speaker. These segments compose acoustic speaker profiles against which can be used in speaker assignment of the remaining segments.

Characterizing speaker roles serves as the foundational step in isolating speech profiles. Firstly, transcripts are generated through an ASR model, which are then passed to three specialist LLMs: a summarization model, a topic segmentation model, and a role identification model. As each model processes a smaller segment of the conversation, information is passed downstream at each stage, allowing the maintenance of high-level context throughout. After roles are identified, a fourth specialist LLM identifies those segments of transcribed speech most likely to be associated with each role. The corresponding speech segments are then combined to form speech profiles for each speaker. Vector embeddings are generated for each speech profile as well as each speech segment; speech segments are then assigned a speaker source according to the maximum cosine similarity to the corresponding speech profiles. We highlight two conversations from our experiment: CALLHOME 0638 (see Table 1) from the CALLHOME (Canavan et al., 1997) dataset and MedData RES0102 (see Table 2) from the MedData (Farzandipour et al., 2022) dataset. CALLHOME 0638 is an example of a conversation with weak role distinction, elucidating the need for the topic segmentation stage. MedData RES0102 is an example of a conversation with fuzzy embeddings cluster boundaries (see Figure 3a) that cannot be accurately diarized using audio alone. Using role information, speech segments attributable to each speaker are used to build acoustic speech profiles (see Figure 3b), facilitating accurate speaker assignment (see Figure 4).

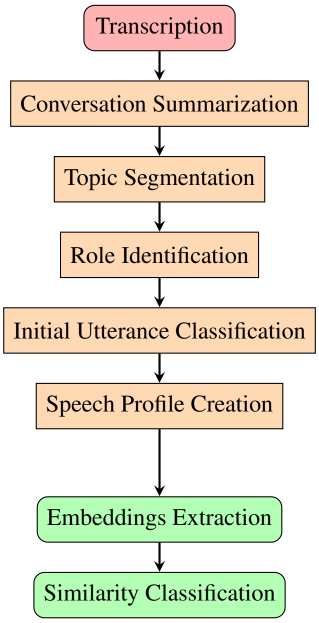


Figure 2: Architecture of CASCA

2.1 Summarization

The first model in our LLM pipeline summarizes the nature and type of the conversation, such as an interview, transaction, or friendly chat. This summary is passed downstream in the prompt of the role identification model to maintain context that could be lost during topic segmentation. Preserving the high-level context ensures that the assigned role profiles remain consistent with the nature of the conversation.

2.2 Topic Segmentation

Topic Segmentation of CallHome 0638, Chunk 01.

At Kinko's? | Mm-hmm. | Is it the same job that you had when I left? | Well, I'm almost out of my job. | What does that mean? | They... They gave me a promotion to supervisor. | Wow. | Yeah, but I can't change out of it until the 10th of August. | Oh. | But I've been working on Sundays as a supervisor. | So I've sort of been the supervisor, but I still have my other job during the week. | Do you like being supervisor? | Yeah, I do. | Oh, good. | Okay. | Well, that's good. | And then I can work, like, evenings during school. | Uh-huh. | I don't have to work in the day. | Oh, very good. And how's Leroy? | He's okay. | He moved into Felipe's apartment. | Uh-huh. | And it was about, what, a week ago? | Uh-huh. | And I haven't seen much of him since. | See, I heard that they split up again.

Model Response: First Theme: Career Transition and Job Promotion. First Line: Just working. | Line That Transitions Away From Theme: And how's Leroy? Second Theme: Personal Life and Leroy's Move. First Line: And then I can work, like, evenings during school. | Line That Transitions Away From Theme: See, I heard that they split up again.

Table 1: This conversation lacks strong role distinction. The speakers are female friends, and the role each plays is fluid. However, their roles are distinct and stable within each topic segment. Topic 1: one speaker informs the other about a promotion in the first segment Topic 2: one speaker updates the other about a mutual friend.

The second model splits the conversation into regions with a homogeneous topic to facilitate the extraction of topic-level role information. This allows for the use of role information in conversations with weak role distinction (see Table 1). To do so, lines marking conversational boundaries are identified. Topic segmentation results in a list of conversation segments each pertaining to the same conversational object; the downstream role and line assignment models subsequently work with these thematically unified segments to identify and apply the role information.

2.3 Identification of Speaker Roles

Utterance	Speaker	Conf.
So yeah, it's the cough.	01	0.2
It started a couple of weeks ago.	02	1.0
A couple of weeks ago?	01	1.0
Okay.	01	0.13
And has it gone worse since then?	01	1.0
Or has it stayed about the same?	01	1.0
It got worse initially, but it's been about the... Well, actually, yeah, it's been getting worse since now.	02	0.95
I've started to... noticed blood in this freedom.	02	.34
I wasn't there at first.	02	0.45
Okay, when did you first notice that?	01	1.0
So I first saw some blood a few days ago.	02	1.0
It was a really small amount, so I didn't really see much, but I brought out blood.	02	0.95
Yesterday, and again this morning, it's been just about enough to cover 50, so it's not a lot of driving time, but it's pretty rough.	02	0.95
Okay, and before that, were you getting any production for your cult for the last few years?	01	0.15
Like, were you producing any music?	01	0.23
Uh, no.	02	0.3
No?	01	0.5
Okay.	01	0.1
Can you just describe your cough term?	01	1.0
Is it a wet cough or dry cough?	01	0.95
It's dry, but it's really with the exception of the blood.	02	0.95

Table 2: *MedData, Conversation RES0102* In this example, the two identified speaker roles, doctor and patient, are used to positively identify certain segments as belonging to each speaker.

The third model determines the roles of each speaker in a specific topic chunk, in the context of the broader conversation summary. The two following examples illustrate how speaker roles are found in cases of both strong and weak role distinction.

2.3.1 Weak Role Distinction

In the absence of clearly distinct, stable roles, the model relies on the differences in each speaker's relationship to the central topic of the conversation within the topic segment to define speaker roles. For instance, in the first topic segment of CallHome 0638 Table 1, the two speakers are distinct in their roles as informant and informee.

Model Response: *SPEAKER A is sharing updates about their job change and the current situation, seeking validation and support from SPEAKER B. SPEAKER B's role is that of a listener and supporter [showing] interest in SPEAKER A's well-being and the benefits of the job change, such as having more flexibility in working hours.*

2.3.2 Strong Role Distinction

The model can more easily extract role information when consistent distinctions are present. These are constrained by the conversation context established by the first model. For example, the model accurately characterizes the roles of the patient and doctor in MedData RES0102.

Model Response: *Speaker_01, who is sharing their symptoms with Speaker_02, who is likely a medical professional or seeking to understand Speaker_01's symptoms in a medical context. Speaker_01 is the individual experiencing and reporting their symptoms.*

2.4 Speech Profile Creation

Each transcribed utterance is passed to a fifth model, along with the surrounding conversational context and speaker roles identified in the previous step for classification. We assign the logarithmic probability (logprob) associated with the speaker label token as the confidence score for the classification. We explored several alternatives for this confidence score, including repeated prompt agreement (Portillo Wightman et al., 2023) and auxiliary fine-tuned models to determine confidence, but found these approaches to be too computationally expensive or unreliable. Utterances that are clearly associated with a particular role - to use the same example, "I am going to recommend an X-ray" which is clearly associated with the role of a doctor - tend to be classified with greater confidence. We then take the set of utterances with the highest confidence scores to form our speech profiles corresponding to the respective roles. The detailed algorithms used to mix these utterances are found in Section A.

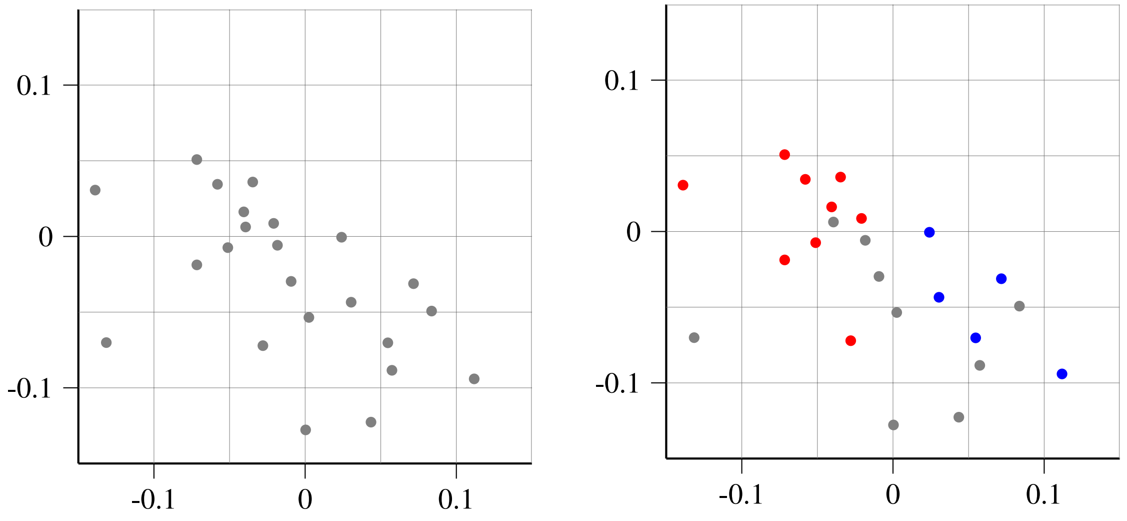
2.5 Final Classification

The vector embedding of each segment, σ_i, is calculated, and each segment is then classified according to similarity to each speech profile, α_j: $\max_j \in speakersetsim(\sigma_{x_i}, \alpha_j)$.

3 Experiment

3.1 Models

In the first step of automatic speech recognition (ASR), we use the Whisper Large V3 model with word-level alignment and segmentation using WhisperX (Radford et al., 2023; Bain et al., 2023). The tendency of Whisper to remove disfluencies, i.e. *"I I don't"*, *"uh"*, etc., significantly increased word error rates (WER) on verbatim transcripts. We chose WhisperX due to the reliability of generated timestamps. Transcripts are broken into utterances: each identified utterance is almost



(a) Cross cosine similarity of utterance embeddings MedData RES0102.

(b) Utterances labelled as Speaker 1 or Speaker 2 according to identified speaker roles.

Figure 3: This example highlights how role-aware diarization succeeds where traditional acoustic methods fail. The noisy audio of a short interaction results in embeddings with no identifiable clusters (Fig. 3a). However, the previously identified speaker roles of the doctor and patient inform the assignment of some of the utterances (Fig. 3b) to each speaker. This clarifies the acoustic distinctions between speakers. The subsequent speaker assignments using the speech profiles (see Figure 4) are nearly perfect.

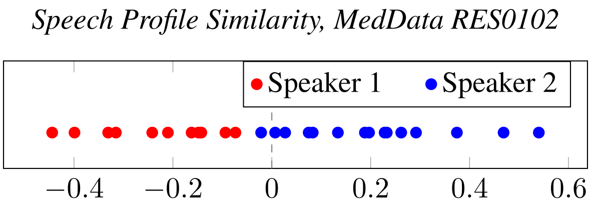


Figure 4: Difference in cosine similarity of each speech segment to the acoustic speaker profile of the patient and doctor respectively. The high accuracy of this classification far outperforms the audio-only clustering methods. Note: The one misclassified segment is Segment #19 (text: ‘No’, duration: 0.08 seconds) is extremely short; embeddings become unstable as speech segment length decreases.

always from a single spoken speaker. However, utterances are not separated by speaker turn; consecutive utterances may or may not be from the same speaker.

Each component model of our LLM sequence is a task-specific fine-tune of Mistral7B-Instruct-V1 (Jiang et al., 2023; Wang et al., 2024a). We chose this base model for a few reasons. Firstly, it is open source. Secondly, its small size lightens the computation burden of repeated LLM calls that the framework entails. Thirdly, it helps illustrate the potential of effectively fine-tuned, specialist small language models in diarization, a secondary contribution of this work. Current speech processing projects, for example (Wang et al., 2024b), are built upon extremely large and computationally costly models; CASCA demonstrates that limitations in reasoning associated with lower-parameter models can be overcome through careful delegation of tasks and fine-tuning.

We use the WavLM-Large embedding model (Chen et al., 2022) for embedding speech segments and each speech profile.

3.2 Fine-Tuning

3.2.1 Generation of Training Data

The raw material for our fine-tuning data was sourced from open-source datasets of real-world dialogue, including DailyDialogue (Li et al., 2017), SWDA: Switchboard Dialogue Act (Jurafsky et al., 1997), and CallFriend (Canavan and Zipperlen, 1996). We used GPT-4 (OpenAI, 2023) to generate task-appropriate responses corresponding to conversations drawn from these datasets. The data generation methodology was specific to each task, depending on its complexity.

- **Conversation Summarization:** Straightforward, single-step prompting proved sufficient to generate accurate conversation summaries.

- **Topic Segmentation:** We utilized a two-stage chain-of-thought (CoT) prompting approach to assist the model in identifying major topics and subsequently determining the boundary phrases of each topic.

- **Role Identification:** We paired conversation summaries created earlier with conversation transcripts to identify speaker roles, emphasizing distinctions between the roles.

- **Utterance Assignment:** Iterating through each utterance in the conversations, we provided the model with the identified speaker roles, the target utterance, the surrounding context, and the correct utterance label. We prompted the model to explain the logical progression from the provided information to the correct answer. This produced a data set that provided effective logical instruction for this task, as the base model initially performed poorly on this type of reasoning-based task.

Task	**Source**	**Entries**	**Method**
Summarization	DD, SWDA	~2000	Few-Shot
Topic Segmentation	CallFriend	~500	CoT
Role Extraction	DD, SWDA	~2000	Few-Shot
Line Assignment	DD, SWDA	~3000	CoT

Table 3: DD: DailyDialogue; SWDA: Switchboard Dialogue Act; CallFriend

We fine-tuned using a LoRA adapter with a learning rate of 0.0002 and cosine decay. Training lasted for three epochs using 4-bit quantization for computational optimization. Data preprocessing included a random split (95%) - (5%) training-validation. The fine-tuning data is made publicly available. [2]

3.3 Test Data

In this section, we present the results of our approach on 96 hours of out-of-domain conversation data from various settings, collected mainly from TalkBank (MacWhinney, 2023). We constructed our evaluation set from selected subsets of two-speaker conversations chosen from available TalkBank data, without any prior knowledge of the audio. These data sets include CORAAL (Kendall and Farrington, 2023), featuring interviews with African-American participants; CALLHOME (Canavan et al., 1997), comprising telephonic conversations between friends and family; and an open-source set of simulated doctor-patient conversations (Farzandipour et al., 2022), which we mix with background noise to simulate challenging real-world conditions[3]. A few conversations from two miscellaneous sources, MICASE (Simpson et al., 2002), containing academic dialogue, and SBCBASE (Du Bois et al., 2000–2005), containing a mix of informal dialogue, were also included to explore different conversation scenarios. Selections from MICASE and SBCBASE were limited to the few two-speaker conversations available in these datasets. This experiment took about 4 hours of active GPU usage on an A100.

3.4 Evaluation Metrics

Unlike most diarization systems, ASR transcription is an integral part of our framework. CASCA is oriented towards the classification of already transcribed speech segments; therefore, we can use the word error rate to measure the accuracy of our system. The fidelity of the final transcripts effectively conveys how well conversational information is preserved during the entire pipeline of speech processing. Our metric of interest is Word Diarization Error Rate (**WDER**), which is used to evaluate diarization systems that incorporate ASR (Shafey et al., 2019; Tran et al., 2022). To define WDER, we first specify its two component metrics, Word Error Rate (**WER**) and Concatenated Permutation Word Error Rate (**cpWER**).

WER captures the accuracy of the transcription without considering the speaker identification error. It is calculated as:

$$WER = \frac{lev(R, H)}{|R|} \tag{1}$$

cpWER for two speakers accounts errors from both transcription and incorrect speaker speaker attributions (diarization errors). Given that hypothesis H_1 corresponds directly to reference R_1, and hypothesis H_2 to reference R_2, the cpWER is calculated as:

$$cpWER = \frac{lev(R_1, H_1) + lev(R_2, H_2)}{|R_1| + |R_2|} \tag{2}$$

$lev(R, H)$ represents the Levenshtein distance between the reference transcript R and the hypothesis transcript H, and $|R|$ is the total number of words in the reference transcript. Finally, WDER is specified as:

$$\textbf{WDER = cpWER - WER} \tag{3}$$

[2] https://github.com/CASCA-Labs/CASCA

[3] Background noise is mixed to achieve a signal-to-noise ratio of 11.

Source	# Dialogues	Total (m)	# Words	# Segs	Avg Seg Length (s)	Avg Words/Seg
CALLHOME	90	1795.10	289785	23277	4.6	12.45
MedData	266	3138.56	620210	50195	3.7	12.35
CORAAL	23	628.78	176037	14129	2.7	12.46
MISC	7	224.80	68072	6012	2.2	11.32

Table 4: Composition of Evaluation Set

This is less forgiving than other specifications of WDER or the analogous time-based Diarization Error Rate (DER). Under this specification, confused speaker assignments are counted as both an insertion in the stream where they are erroneously added and a deletion in the stream from which they are missing.

We calculate these metrics using the MeetEval toolkit (von Neumann et al., 2023). Technically, we employ a time-constrained optimal reference combination word error rate to calculate WER and a time-constrained concatenated minimum permutation word error rate to cpWER. Time constraints reduce the computational burden and result in only a negligible overestimate of the true cpWER.

3.5 Baseline

To contextualize the marginal value of role distinction in diarization, we present a baseline audio-only diarization system. For this purpose, we employ Pyannote (Plaquet and Bredin, 2023), which is integrated into the WhisperX framework. Pyannote is one of the most popular diarization frameworks and achieves competitive performance on most diarization tasks. Its integration with WhisperX is advantageous as it enables an equitable comparison of the two methods, each utilizing the same ASR output and attempting to classify speaker segments bounded by the same timestamps.

3.6 Results

3.6.1 ASR

Whisper-V3 yields an ASR error rate of 16.6% across all conversations. The ground truth transcripts are verbatim transcripts, which contain disfluencies, nonstandard nomenclature, or names; this is the source of much of the ASR error. This error rate is in line with benchmarks for the model; our reported CALLHOME ASR word error rate of 19.75% is within 2% of the standard achieved in OpenAI's technical report (Radford et al., 2023). This difference is partially or wholly explained by less robust word standardization. Note that more linguistic information is retained than this error rate suggests, as incorrect transcription of disfluencies tends to have little impact on meaning.

3.6.2 Baseline Performance

Our baseline achieves a mean WDER error rate of 22%. The distribution of errors is somewhat bimodal (see Figure 5). This is due to the presence of conversations in which the differentiation in the acoustic characteristics of each speaker's voice is insufficient to clearly define clusters in the utterance embeddings (e.g., Figure 1). This causes the clustering algorithm to go awry and, in turn, results in extremely inaccurate diarization.

3.6.3 CASCA

CASCA exhibits markedly improved performance across the evaluation set. The mean WDER of 4.2% represents an 80% improvement in accuracy. As expected, CASCA performs best in the presence of strong role distinction, such as in the professional MedData conversations, and worst in the presence of weak role distinction, such as in the casual CALLHOME conversations. However, even in those cases, CASCA still outperforms the baseline, a result that validates the utility of the topic segmentation stage.

4 Conclusions and Future Work

In this paper, we offer a conceptual framework for the dynamic utilization of speaker role distinction in speaker diarization through a sequence of specialized LLMs. We demonstrate that this framework far surpasses acoustic only diarizaiton for a variety of conversation types. Performance varies with the strength of the distinction between speaker roles. These results highlight the potential of leveraging the rich role information contained within the conversational text for the task of speaker diarization. Tracking our original motivation for this project, our expectation is that this framework will offer the most value in physical commercial settings, where speaker roles are very distinct but hostile recording environments make

Source	ASR WER	Baseline WDER	CASCA WDER	Improvement	Role Distinction
CALLHOME	19.75%	38.8%	13.5%	25.3%	Weak
MedData	14.44%	16.1%	1.6%	14.5%	Strong
CORAAL	22.53%	48.6%	7.3%	41.3%	Strong
MISC[4]	24.18%	56.4%	4.7%	51.7%	Moderate

Table 5: CASCA achieves a 4.2% average WDER, outperforming the baseline of 22%.

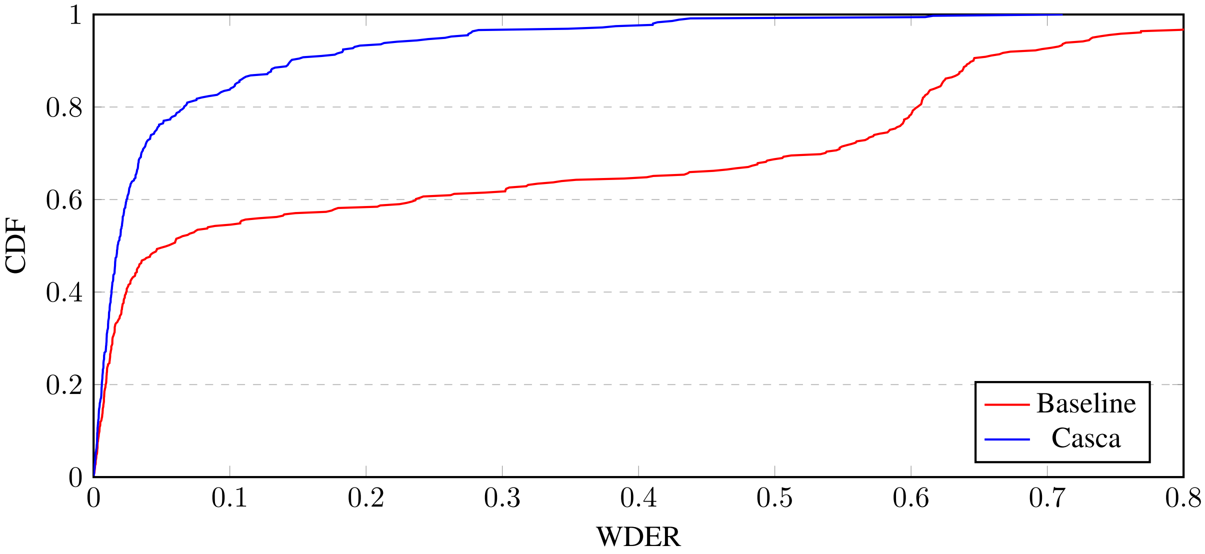


Figure 5: Comparison of cumulative distribution functions of errors for both CASCA and baseline. CASCA displays higher accuracy, especially in cases where the vocal characteristics of participant voices are similar.

acoustic clustering challenging. Our use of seven billion parameter LLMs is also notable as it reduces the cost of the system while illustrating the potential for downsizing speech processing models to fine-tuned specialists. An attractive next step of research is to explore other methods of utilizing the identified speech segments in speaker assignment. Our current method of classifying speech segments according to cosine-similarity speech profiles is simplistic. Other methods, such as the use of reference segments to constrain relationships between embeddings in the definition of clusters, could be more reliable. Additionally, the linguistic content of the speech profiles could be used to develop more sophisticated speaker profiles by identifying speaker's pronunciation of particular words. This would further simplify speaker assignment into a type of text-dependent speaker verification.

5 Limitations

A significant limitation of this study is its exclusive focus on dyadic conversations. In two-speaker interactions, role distinctions are generally apparent and informative. However, with additional participants, these distinctions become increasingly ambiguous. Discerning unique roles in multiparty conversations without prior information is exceptionally challenging, barring specific contextual factors such as commercial interactions where participants have distinct relationships to the subject matter. One potential approach for multiparty conversations could involve progressively identifying roles - establishing one speaker's role, using that context to inform another, and iteratively uncovering roles until the set is fully specified. The feasibility of this method, along with alternative approaches for extending this framework to conversations with more participants, remains a topic for future research. Another limitation stems from the framework's reliance on the accuracy of initial ASR transcription. If sufficiently severe as to affect meaning, errors in this stage could confound the downstream role analysis, undermining the entire diarization process. Finally, the computational demands of sequential specialized LLM processing present a practical limitation. Although the use of smaller language models mitigates this issue to some extent, the computational cost still substantially exceeds that of audio-based diarization systems. Current audio-based systems can achieve processing speeds exceeding 60 times real-time, whereas our system averages only 5 times real-time using an A100 GPU.

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-accurate speech transcription of long-form audio. In *Proceedings of INTERSPEECH 2023*.

Alexandra Canavan, David Graff, and George Zipperlen. 1997. CALLHOME american english speech (LDC97S42).

Alexandra Canavan and George Zipperlen. 1996. CALLFRIEND american english–non-southern dialect (LDC96S46).

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

John W. Du Bois, Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000–2005. Santa barbara corpus of spoken american english, parts 1–4.

Fareez Farzandipour, Tapan Parikh, Charles Wavell, and et al. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9:313.

Nikolaos Flemotomos, Panayiotis Georgiou, and Shrikanth Narayanan. 2020. Linguistically aided speaker diarization using speaker role information. In *Proceedings of Odyssey 2020: The Speaker and Language Recognition Workshop*, pages 119–126. ISCA.

Nikolaos Flemotomos and Shrikanth Narayanan. 2022. Multimodal clustering with role induced constraints for speaker diarization. In *Proceedings of INTERSPEECH 2022*, pages 3518–3522.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical report, University of Colorado at Boulder and SRI International.

Tyler Kendall and Charlie Farrington. 2023. The corpus of regional african american language. Version 2023.06.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Brian MacWhinney. 2023. Talkbank. Accessed: 2023-06-16.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Rohit Paturi, Sundararajan Srinivasan, and Xiang Li. 2023. Lexical speaker error correction: Leveraging language models for speaker diarization error correction. In *Proceedings of INTERSPEECH 2023*.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proceedings of INTERSPEECH 2023*.

Arnab Poddar, Md Sahidullah, and Goutam Saha. 2018. Speaker verification with short utterances: A review of challenges, trends and opportunities. *IET Biometrics*, 7(2):91–101.

Gwenyth Portillo Wightman, Alexandra DeLucia, and Mark Dredze. 2023. Strength in numbers: Estimating confidence of large language models by prompt agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 326–332, Toronto, Canada. Association for Computational Linguistics.

Amrutha Prasad, Juan Zuluaga-Gomez, Petr Motlicek, Saeed Sarfjoo, Iuliia Nigmatulina, Oliver Ohneiser, and Hartmut Helmke. 2021. Grammar based speaker role identification for air traffic control speech recognition. *Preprint*, arXiv:2108.12175.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR.

Luca Serafini, Samuele Cornell, Giovanni Morrone, Enrico Zovato, Alessio Brutti, and Stefano Squartini. 2023. An experimental review of speaker diarization methods with application to two-speaker conversational telephone speech recordings. *Computer Speech
Language*, 82:101534.

Laurent Shafey, Hagen Soltau, and Izhak Shafran. 2019. Joint speech recognition and speaker diarization via sequence transduction. In *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 396–403.

R. C. Simpson, S. L. Briggs, J. Ovens, and J. M. Swales. 2002. The michigan corpus of academic spoken english.

Bao D. Tran, Ramesh Mangu, Ming Tai-Seale, Jennifer E. Lafata, and Kai Zheng. 2022. Automatic speech recognition performance for digital scribes: A performance comparison between general-purpose and specialized models tuned for patient-clinician conversations. *AMIA Annual Symposium Proceedings*, 2022:1072–1080.

Thilo von Neumann, Christoph Boeddeker, Marc Delcroix, and Reinhold Haeb-Umbach. 2023. MeetEval: A toolkit for computation of word error rates for meeting transcription systems. *Preprint*, arXiv:2307.11394.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. *Preprint*, arXiv:2401.00368.

Quan Wang, Yiling Huang, Guanlong Zhao, Evan Clark, Wei Xia, and Hank Liao. 2024b. DiarizationLM: Speaker diarization post-processing with large language models. *Preprint*, arXiv:2401.03506.

A Algorithms for Creation of Speech Profiles from Utterances

A.1 Speech Profile Mixing

Each assigned utterance with a confidence score above the a threshold τ_c is appended to the topic-level speech profile for the assigned speaker. Because we only need a small fraction of all classified utterances for reference speech profiles, we can tolerate a high rate of false negatives; we use a conservative $\tau_c = 0.99$

The end result is two sets X of segments for each topic, each set containing speech utterances from opposing speakers. The relation between labels within the same topic is known, but the assignment of speaker labels "A" or "B" is arbitrary between topics. Therefore, we mixed these segments using embedding similarity. We clean and mix the topic-level segment sets according to the following algorithms:

A.1.1 Clean Topic Sets

Let $X = \{\sigma_{x_{ij}} \mid i \in topics, j \in speakers\}$ be the set of segments of each topic of opposing speakers. For each pair of segments $\sigma_{x_{ij}}$ and $\sigma_{x_{kl}}$ in X, calculate the cosine similarity:
$sim(\sigma_{x_{ij}}, \sigma_{x_{kl}}) = \cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{|\vec{A}||\vec{B}|}$ where $\vec{A}$ and $\vec{B}$ are the embedding vectors of $\sigma_{x_{ij}}$ and $\sigma_{x_{kl}}$, respectively.

For each segment $\sigma_{x_{ij}}$, calculate the average cross similarity by averaging the similarities of $\sigma_{x_{ij}}$ with all other segments. The centroid segment x_c is the one with the maximum average similarity for $\sigma_{x_{ij}}$. Retain a segment if its similarity to x_c exceeds $.2 \times \tilde{x}$, where $\tilde{x}$ represents the median similarity to the centroid x_c.

A.1.2 Mix Topic Sets

The resulting homogenized pairs are then mixed according to the combination that maximizes the joint cosine similarity of the mixed pairs. This process is highly reliable due to the length of the audio in each subtopic speech profile; longer speech strings yield more reliable embeddings (Paturi et al., 2023). The richer phonetic information available allows the embedding model to more effectively capture the characteristics of the speaker's voice; indeed, (Poddar et al., 2018) showed that there is a monotonic relationship between the length of the speech segment and the accuracy of the embeddings. This fact makes the successive merging of the topic-level speech profiles highly reliable.

Acknowledgements

Briefly, we would like to extend our gratitude to Max Bain for the excellent timestamp alignment in his WhisperX system. We also thank the MeetEval team for their exceptional work in developing a user-friendly and powerful error measurement system. Without those systems, this project would have been significantly more difficult.