Do We Still Need Audio? Rethinking Speaker Change Detection & Diarization with a Text-Based Approach

Peilin Wu and Jinho Choi
{peilin.wu, jinho.choi}@emory.edu
Department of Computer Science, Emory University

Introduction

Speaker Diarization (SD), essential for enhancing Speech-To-Text systems, identifies "who speaks what" in audio conversations jointly with Automatic Speech Recognition (ASR) systems. This technology is crucial for various recording applications and preparing AI training datasets, where distinguishing speakers enhances dialogue understanding and AI responses.

Problem

- The semantic features were often relied on outdated language models or were only used for a post-processing to rectify errors from preceding audio-based SD models, which did not fully exploit the semantic features.
- There has been an absence of research exploring the use of text as the sole input for SD. This gap highlights a missed opportunity to fully explore the capabilities of semantic features in SD.
- There are cases that researchers do not have original audio files but only the transcripts.

Contribution

The main contributions of this work includes:

1. Text-based sentence-level SD approaches using text as the only input that which achieves state-of-the-art result for short conversation.
2. Data processing pipeline tailored to optimize SD on ASR-generated transcripts.
3. Comprehensive analysis about performance and error types of the text-based approach.

Dataset

Standard of corpora used:

- Having both audio and transcript with ground truth speaker labels
- Open-domain
- Conversation need to be un-scripted

Name	Amount (h)	Conversation Num	Punctuation
AMI Corpus	100	171	Partial
CallFriend	20	41	No
CallHome	20	176	No
CHiME-5	50	20	Yes
DailyTalk	20	2541	Yes
ICSI Corpus	72	75	Yes
SBCSAE	23	60	No

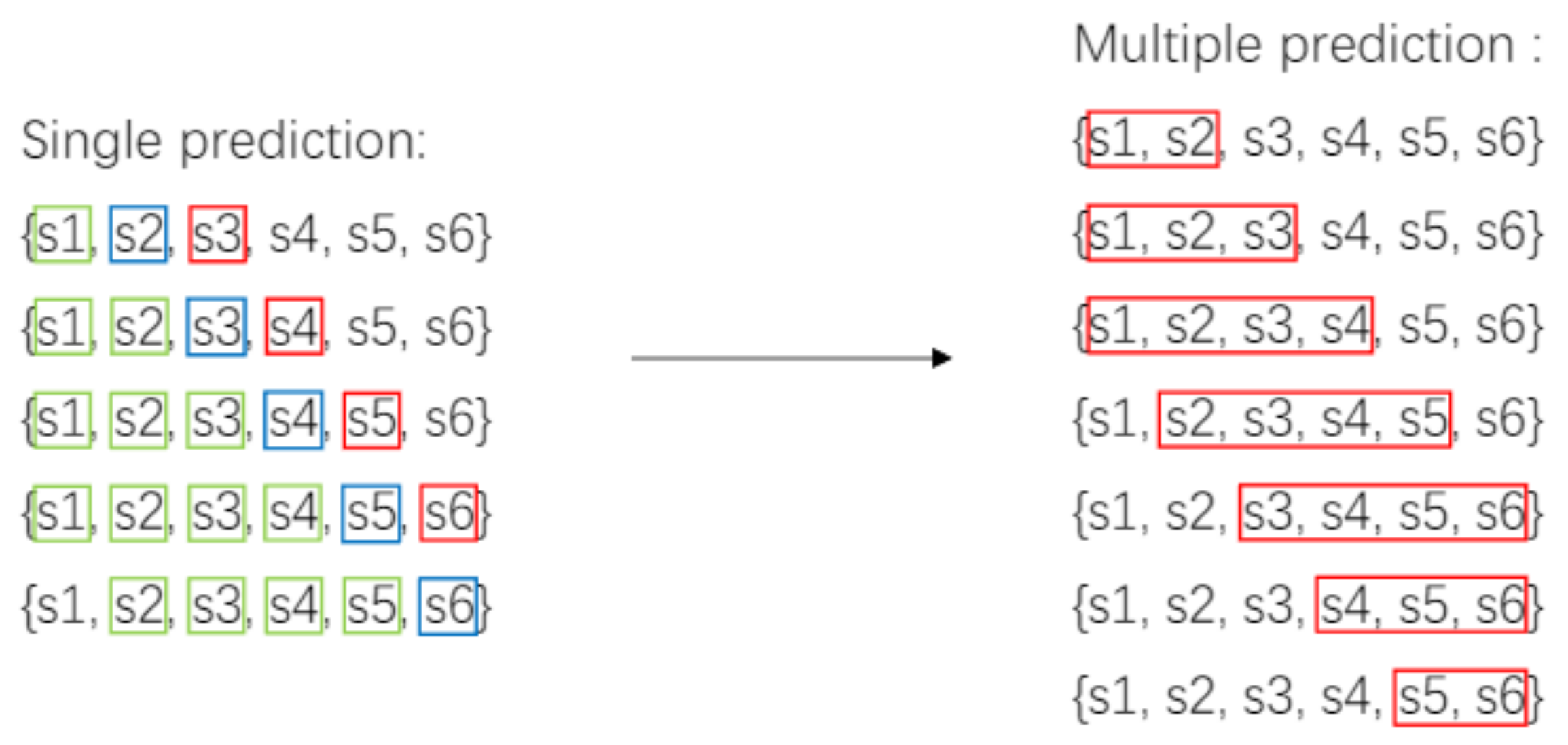
Methodology

Single Prediction Model (SPM)

The single prediction model (SPM) operates by evaluating the probability of a speaker change between sentences, using surrounding utterances as context.

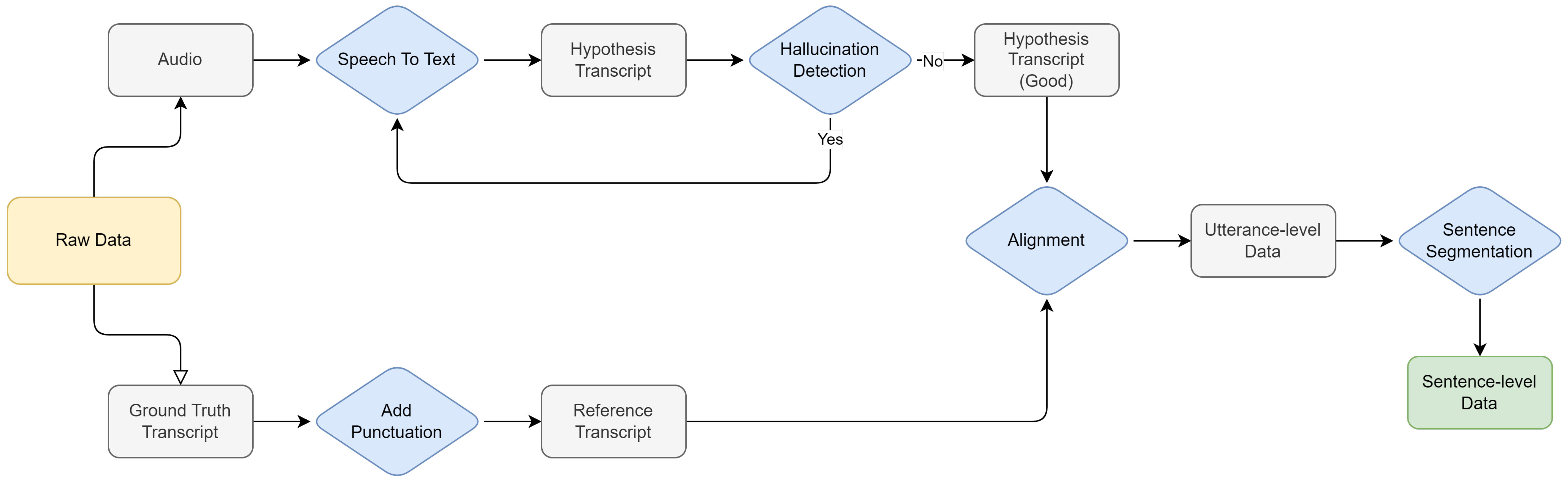
Multiple Prediction Model (MPM)

To enhance accuracy and robustness, this work introduces a multiple prediction model (MPM) that aggregates predictions over several windows within a dialogue so that more contextual information is leveraged on one prediction. The MPM extends the SPM by making predictions over multiple points within a sliding window across the conversation.



Data Processing

- Primary input of text-based model: ASR-generated transcripts
- Produce training and evaluation data with ASR-specific discrepancies.



Results

Our text-based models are compared to recent audio-based SD systems, including both modularized and end-to-end systems. The results indicate that text-based SD, especially with multiple predictions, offers a promising alternative to traditional audio-based methods, excelling in short conversational contexts.

Model	≤ 15 Min.		> 15 Min.		Overall	
	WDER	WDER-S	WDER	WDER-S	WDER	WDER-S
pyannote	0.269	0.233	0.137	0.127	0.225	0.187
x-vector+SC	0.378	0.339	0.150	0.175	0.302	0.184
x-vector+AHC	0.298	0.269	0.241	0.268	0.279	0.258
ECAPA+SC	0.402	0.371	0.199	0.152	0.334	0.278
ECAPA+AHC	0.291	0.256	0.166	0.267	0.249	0.239
NeMo-TitaNet	0.233	0.177	0.103	0.088	0.189	0.127
NeMo-MSDD	0.230	0.175	0.085	0.078	0.181	0.123
TOLD	0.206	0.129	**0.080**	**0.069**	0.164	**0.099**
T5-3B SPM	0.312	0.334	0.528	0.563	0.384	0.440
T5-3B MPM	**0.049**	**0.055**	0.114	0.129	**0.101**	0.104

Input Length Analysis

In order to assess the influence of amount of information on the performance, as well as the ability of the model to utilize the information, the text-based models are tested with different length of input for each sliding window.

Input Sentence	T5-3B SPM		T5-3B MPM	
	WDER	WDER-S	WDER	WDER-S
4	0.428	0.475	0.073	0.277
6	0.388	0.429	**0.056**	0.165
8	0.384	0.440	0.101	**0.104**

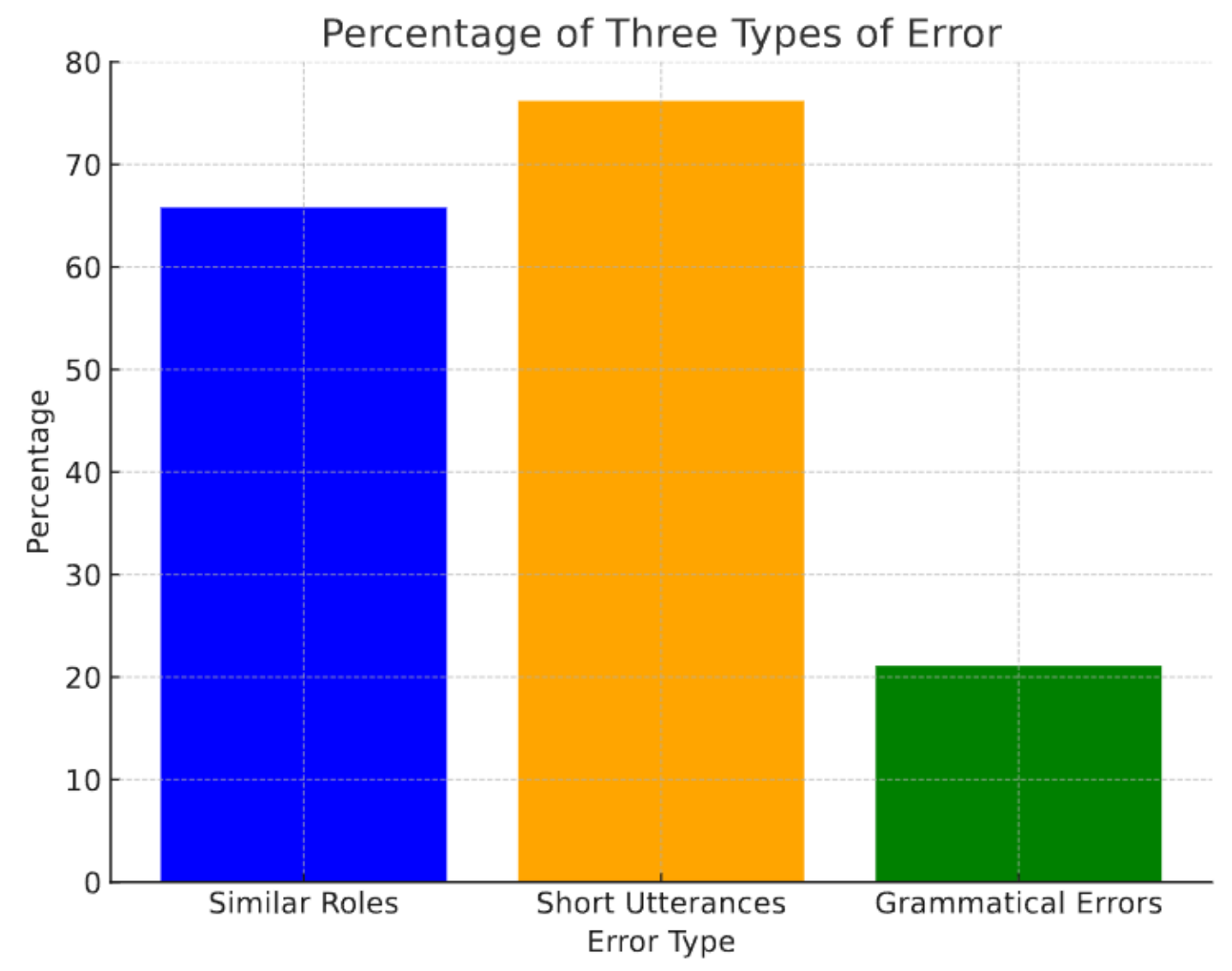
Error Types

In order to further determine the types of input where the text-based model makes mistakes, 50 single inputs with at least 1 incorrect prediction and unable to be recovered after aggregation were randomly selected and manually inspected. Three major types of error-prone input are concluded from this inspection:

Similar speaker roles
s1: They just said it was gonna be recorded whatever.
s2: So how's it going?
s3: Everything's going cool.
s4: When I first got here, things were kind of messed up, but I got your email.
Model Prediction: A, A, B, B
Correct Label: A, B, A, A

Short sentences
s1: Wow.
s2: What time is it there?
s3: What time is it?
s4: It's 3:40
Model Prediction: A, B, B, A
Correct Label: A, B, A, A

Grammatical Errors
s1: How things with you busy?
s2: I guess I sent you an email, but I suppose you haven't gotten it
Model Prediction: A, B
Correct Label: A, A



Conclusion

This work presents a novel approach to SD by using 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 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.

Future Work

- Develop text-based methodologies that can handle multiple speakers in a conversation.
- Include SCD models for fair comparison.
- Experiment with stronger base model