Native Speech Processing with LLMs

Aaron Soh

Nanyang Technological University
College of Computing and Data Science
Singapore
asoh014@e.ntu.edu.sg

Abstract

Recent advances in Large Language Models (LLMs) have achieved state-of-the-art performance in Automatic Speech Recognition (ASR), surpassing ASR-only systems such as Whisper. However, their application to other speech processing tasks, particularly speaker diarisation (SD), remains underexplored. This work proposes extending existing speech-aware LLM architectures with diarisation-specific training and context-based prompting to enable joint transcription and segmentation of multi-speaker audio. By exploiting the semantic reasoning and multilingual capabilities of pretrained LLMs, the proposed approach aims to improve diarisation accuracy, enhancing accessibility for assistive technologies and real-time captioning applications that rely on accurate speaker-aware transcriptions.

Introduction

Speech processing, a field of interest for decades, has recently seen successful applications of the Transformer architecture (Vaswani et al. 2017) such as in the Whisper models (Radford et al. 2022) for Automatic Speech Recognition (ASR). Recent work has further demonstrated the use of LLMs for ASR tasks (Saon et al. 2025; Xu et al. 2024; Microsoft et al. 2025), leveraging existing LLM capabilities and bypassing the pre-training stage. Our work aims to explore how we can extend this to other speech processing tasks like speaker diarisation, combining these tasks into a single pipeline. **Speaker Diarisation** (SD) is the process of partitioning an audio stream containing human speech into homogeneous segments according to speaker identity. It enhances the readability of automatic speech transcriptions by structuring audio streams into speaker turns and, when combined with speaker recognition systems, by providing speaker identity (Wikipedia 2025). The task fundamentally addresses the question "who spoke when?". Without speaker diarisation, ASR transcripts would simply appear as joined sentences, leaving readers to interpret which sentence came from which speaker by themself.

Background

Traditional and current approaches to speaker diarisation utilise a combination of clustering and embedding models or end-to-end approaches with neural networks to output diarisation timestamps (Bredin et al. 2019). Recently, Granite-speech 3.3 (Saon et al. 2025) and Phi-4 Multimodal (Microsoft et al. 2025), both of which leverage a pre-trained LLM as its backbone, has shown promising results with using LLMs to perform ASR. An audio encoder is attached to the LLM backbone to produce acoustic embeddings, and a LoRA adapter is trained on top of the frozen LLM to process these inputs in order to generate a transcript. Such approaches have shown to outperform pure ASR models like Whisper across a variety of ASR benchmarks. Other approaches that involve LLMs in speaker diarisation tasks try to fix inaccuracies in a transcription model's output by correcting diarisation errors and autofilling speaker names (Wang et al. 2024; Efstathiadis, Yadav, and Abbas 2025), however this approach is more of a post-processing step on the outputs of an ASR/SD model, rather than using the LLM to generate diarisation labels natively. Such approaches also remain constrained to the accuracy of the underlying ASR/SD model, limiting generalisability.

Therefore, the goal of this research proposal is to explore **how we can further develop speech-aware LLMs to perform other speech processing tasks like speaker diarisation natively.**

Approach

Research has shown that LLMs possess strong semantic understanding capabilities as a result of its pre-training stage on a large amount of textual data, including conversational data that traditional acoustic models lack. Earlier work has also shown that joint ASR and SD models can leverage audio-lexical interdependencies to improve word diarisation performance (Mao et al. 2020). Taking inspiration from the model architecture and training recipe outlined by Granite-speech and Phi-4 Multimodal (Saon et al. 2025; Microsoft et al. 2025), we propose a similar architecture but with contextual prompting (Figure 1). Our proposed training strategy is split into two phases:

First Stage: We perform the joint optimisation of the audio encoder and speech modality adapter on large-scale SD data (Table 1). To speed up the training process, we propose using an existing speech-aware LLM (such as Granite-speech-3.3-2b) as the LLM backbone, and Whisper (Radford et al. 2022) for the audio encoder, since they are pre-

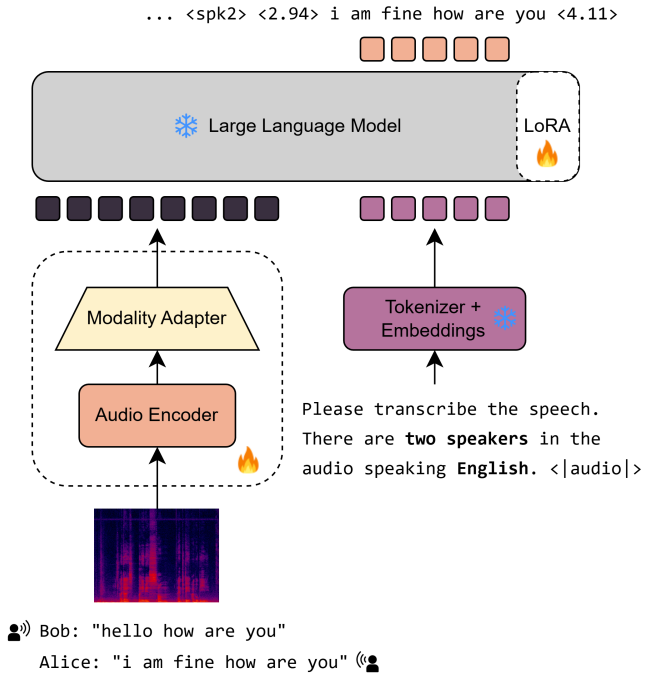

Figure 1: Proposed Model Architecture

trained models. We also propose Q-Former (Li et al. 2023a) for the modality adapter to extract features from our audio encoder to be fed into our LLM. The Q-Former uses learned queries that cross-attend to audio representations to extract important features. This stage aims to rapidly adapt existing speech-aware LLMs to multi-speaker characteristics for speaker diarisation.

Second Stage: We introduce **context-aware prompt training** by freezing the audio encoder and only updating the modality adapter and our LoRA adapter. We generate natural language prompts containing context on each audio clip (Figure 1), as well as annotated speaker-segmented transcripts for our model to learn. During training, we also introduce special `<|audio|>`, `<|timestamp|>`, `<|spk|>` tokens to denote the presence of audio embeddings, transcription segment timestamps and speaker change respectively, taking inspiration from the RTTM format to structure our model's outputs. In each prompt, context such as the language or the number of speakers can be provided in the system prompt. Prior work has shown success in improving the accuracy of ASR transcripts by biasing the language model towards the domain indicated by the prompt context, leveraging their in-context learning capabilities (Li et al. 2023b). This stage aims to improve the accuracy of transcripts by constraining the model and grounding its predictions, a strong advantage over traditional acoustic models. Our proposed approach allows the model to retain its original textual capabilities by simply disabling the adapters, allowing for further analysis of the generated transcripts. Note that in both stages, the LLM backbone is frozen.

Dataset	Hours	Language(s)
American Life Podcast	35	EN
ICSI Meeting	70	EN
AMI Meeting	100	EN
AliMeeting	118.75	ZH
CALLHOME	Unknown	ZH, EN, DE, JA, ES
CALLFRIEND	Unknown	EN, FR, JA
DIHARD-III	33	EN, ZH

Table 1: Open-Source SD Datasets

Evaluation

For ASR and SD, the metrics used are Word Error Rate (WER) and the Diarisation Error Rate (DER) respectively,

$$\text{WER} = \frac{\text{S} + \text{D} + \text{I}}{\text{N}} \tag{1}$$

$$\text{DER} = \frac{\text{FalseAlarm} + \text{Miss} + \text{Confusion}}{\text{Total}} \tag{2}$$

where S represents substitutions, D denotes the number of deletions, I corresponds to the number of insertions, and N is the total number of words in the reference transcript. For the diarisation metric, False Alarm refers to the duration of non-speech incorrectly classified as speech, Miss represents the amount of actual speech not detected by the system, Confusion denotes the duration of speech that is correctly detected but attributed to the wrong speaker, and Total refers to the total duration of speech (Fiscus et al. 2006).

Discussion

Unlike what prior work has demonstrated with ASR, applying a similar architecture to SD tasks poses a few key challenges. Training SD models require large amounts of annotated diarisation data, which are much more limited as compared to ASR data in terms of language variety and the hours of data available because they are expensive to obtain (Bredin et al. 2019; Dawalatabad et al. 2021). This training data imbalance could potentially result in poor generalisation to SD tasks and poor performance in languages other than English. To address this, self-supervised learning is a solution that has been applied to other speech domains (Mohamed et al. 2022) which could be adapted for our task. We believe that the difficulty in training LLM-enabled joint ASR and SD models can be addressed by leveraging the multilingual and semantic understanding capabilities of LLMs, reducing the amount of annotated data required.

Conclusion

This research proposal presents a novel approach to speaker diarisation by building on existing work that has demonstrated success with performing Automatic Speech Recognition (ASR) with speech-aware LLMs. By addressing the difficulties in training speaker diarisation (SD) models, this work could advance the field of speech processing while enhancing accessibility for assistive technologies and real-time captioning applications.

References

Bredin, H.; Yin, R.; Coria, J. M.; Gelly, G.; Korshunov, P.; Lavechin, M.; Fustes, D.; Titeux, H.; Bouaziz, W.; and Gill, M.-P. 2019. pyannote.audio: neural building blocks for speaker diarization. arXiv:1911.01255.

Dawalatabad, N.; Ravanelli, M.; Grondin, F.; Thienpondt, J.; Desplanques, B.; and Na, H. 2021. ECAPA-TDNN Embeddings for Speaker Diarization. In *Interspeech 2021*. ISCA.

Efstathiadis, G.; Yadav, V.; and Abbas, A. 2025. LLM-based speaker diarization correction: A generalizable approach. *Speech Communication*, 170: 103224.

Fiscus, J.; Ajot, J.; Michel, M.; and Garofolo, J. 2006. The Rich Transcription 2006 Spring Meeting Recognition Evaluation.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Li, Y.; Wu, Y.; Li, J.; and Liu, S. 2023b. Prompting Large Language Models for Zero-Shot Domain Adaptation in Speech Recognition. arXiv:2306.16007.

Mao, H. H.; Li, S.; McAuley, J.; and Cottrell, G. 2020. Speech Recognition and Multi-Speaker Diarization of Long Conversations. arXiv:2005.08072.

Microsoft; :; Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; Bach, N.; Bao, J.; Benhaim, A.; Cai, M.; Chaudhary, V.; Chen, C.; Chen, D.; Chen, D.; Chen, J.; Chen, W.; Chen, Y.-C.; ling Chen, Y.; Dai, Q.; Dai, X.; Fan, R.; Gao, M.; Gao, M.; Garg, A.; Goswami, A.; Hao, J.; Hendy, A.; Hu, Y.; Jin, X.; Khademi, M.; Kim, D.; Kim, Y. J.; Lee, G.; Li, J.; Li, Y.; Liang, C.; Lin, X.; Lin, Z.; Liu, M.; Liu, Y.; Lopez, G.; Luo, C.; Madan, P.; Mazalov, V.; Mitra, A.; Mousavi, A.; Nguyen, A.; Pan, J.; Perez-Becker, D.; Platin, J.; Portet, T.; Qiu, K.; Ren, B.; Ren, L.; Roy, S.; Shang, N.; Shen, Y.; Singhal, S.; Som, S.; Song, X.; Sych, T.; Vaddamanu, P.; Wang, S.; Wang, Y.; Wang, Z.; Wu, H.; Xu, H.; Xu, W.; Yang, Y.; Yang, Z.; Yu, D.; Zabir, I.; Zhang, J.; Zhang, L. L.; Zhang, Y.; and Zhou, X. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. arXiv:2503.01743.

Mohamed, A.; Lee, H.-y.; Borgholt, L.; Havtorn, J. D.; Edin, J.; Igel, C.; Kirchhoff, K.; Li, S.-W.; Livescu, K.; Maaløe, L.; Sainath, T. N.; and Watanabe, S. 2022. Self-Supervised Speech Representation Learning: A Review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1179–1210.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356.

Saon, G.; Dekel, A.; Brooks, A.; Nagano, T.; Daniels, A.; Satt, A.; Mittal, A.; Kingsbury, B.; Haws, D.; Morais, E.; Kurata, G.; Aronowitz, H.; Ibrahim, I.; Kuo, J.; Soule, K.; Lastras, L.; Suzuki, M.; Hoory, R.; Thomas, S.; Novitasari, S.; Fukuda, T.; Sunder, V.; Cui, X.; and Kons, Z. 2025. Granite-speech: open-source speech-aware LLMs with strong English ASR capabilities. arXiv:2505.08699.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.

Wang, Q.; Huang, Y.; Zhao, G.; Clark, E.; Xia, W.; and Liao, H. 2024. DiarizationLM: Speaker Diarization Post-Processing with Large Language Models. In *Interspeech 2024*, 3754–3758. ISCA.

Wikipedia. 2025. Speaker diarisation — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Speaker_diarisation&oldid=1310031905. [Online; accessed 7-October-2025].

Xu, T.; Huang, K.; Guo, P.; Zhou, Y.; Huang, L.; Xue, H.; and Xie, L. 2024. Towards Rehearsal-Free Multilingual ASR: A LoRA-based Case Study on Whisper. arXiv:2408.10680.