#46 A Neural Speaker Diarization System for Doctors

Students: Adi Shenoy, Steven Li

Supervisors: Satwinder Singh, Reza Shahamiri

Introduction

Current Challenge

Doctors spend a large portion of their time transcribing medical notes into Electronic Health Records, leading to high levels of burnout, affecting **4 out of 10 doctors** [1]. A study has shown that computerised physician order entry is linked to a **29% higher rate of burnout** [2]. While medical scribes can help, they are difficult to train and retain.

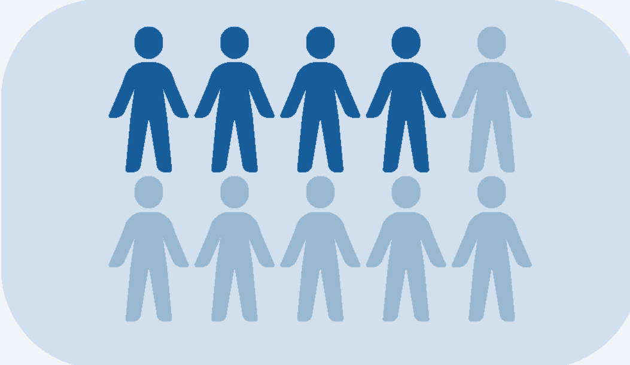

Figure 1: Doctor burnouts

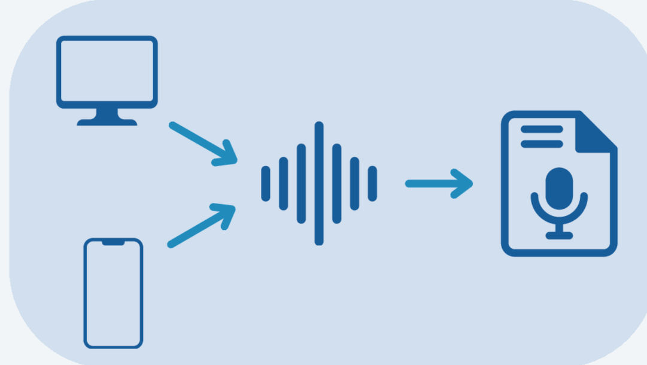

Figure 2: Audio to medical notes

Our Work

This project proposes a neural speaker diarization (SD) system tailored for patient-doctor conversations, designed to process audio and generate medical notes using fine-tuned (FT) models. Current SD systems lack training on medical datasets, and addressing this will improve efficiency while reducing doctors' burden.

Method

Synthesise Medical Data

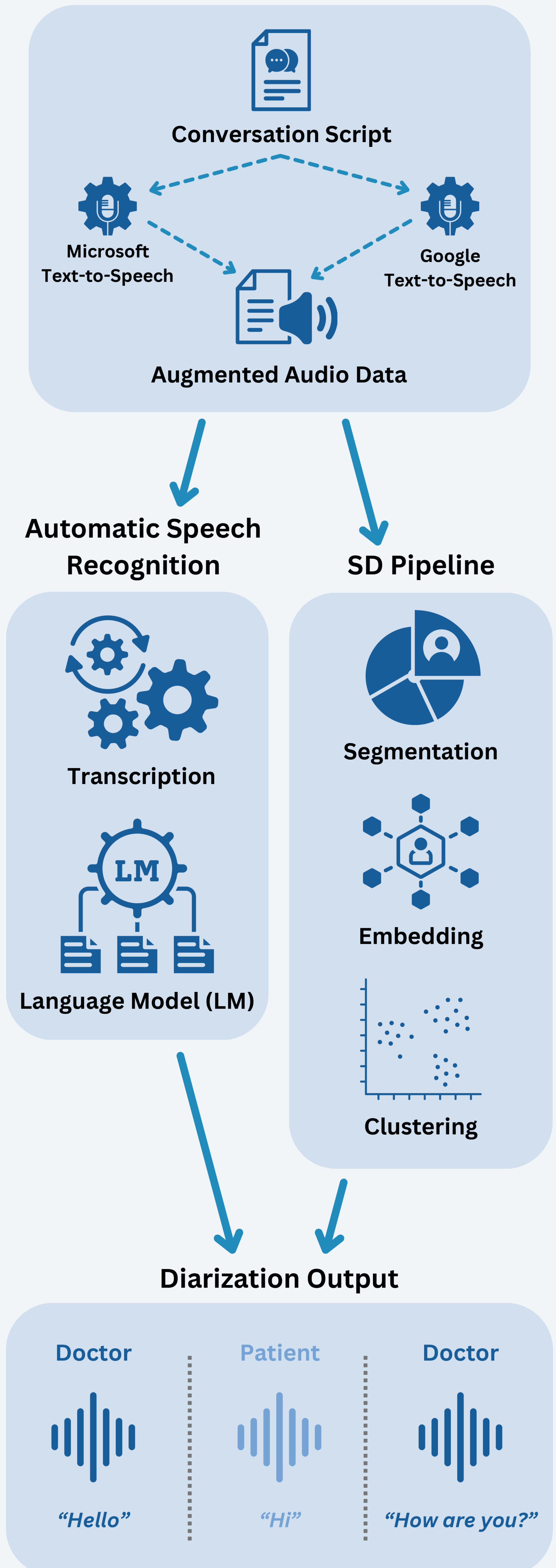

Figure 3: Overview of Research Method

Results

Synthethic Medical Dataset

- Synthetically generated with Text-to-Speech models.
- **9 hours** of audio data, **50** conversation scripts.
- Data augmentation applied: Pitch shift, Speed perturbation, Noise injection (NI).

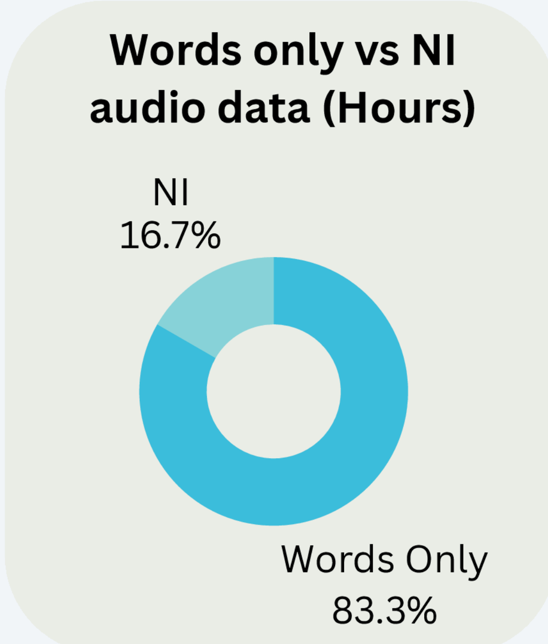

Figure 4: NI and Words only Audio Data Distribution

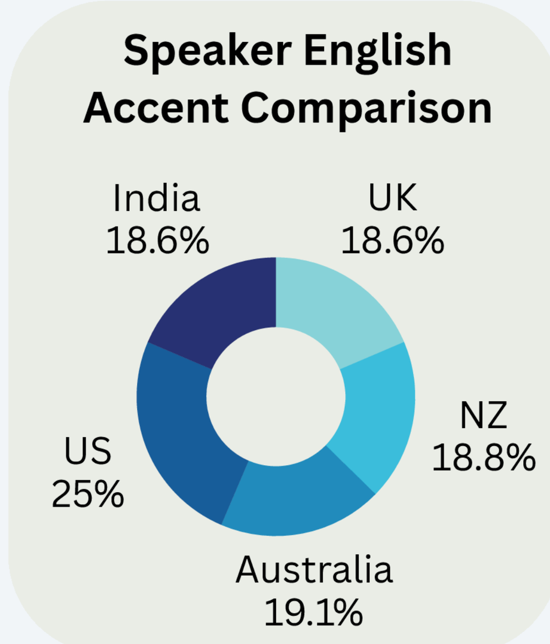

Figure 5: Speaker Accent Distribution

Mean Opinion Score (MOS) is used to evaluate if the audio is adequate, with a score of 3.5 as the baseline.

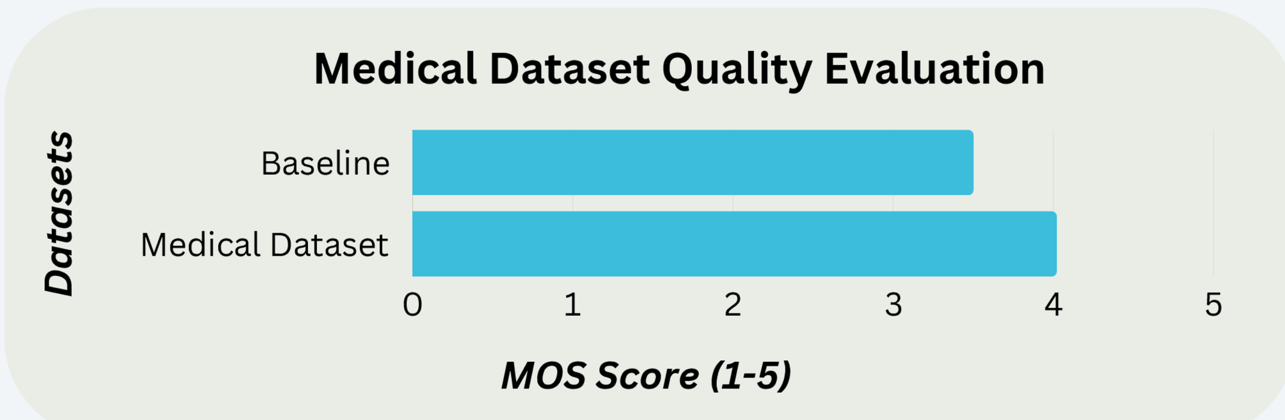

Figure 6: Medical Dataset Quality Evaluation

Speaker Diarization Fine-tuning

- Pyannote's SD pipeline is FT with medical data, particularly the **segmentation model**.
- The training process employs powerset multi-class encoding using SincNet and bi-directional Long Short-Term Memory networks [3].
- Accuracy was measured by Diarization Error Rate (DER) with AMI-SDM (non-synthetic) and medical datasets.

$$DER = \frac{false\ alarm + missed\ detection + confusion}{total\ duration}$$

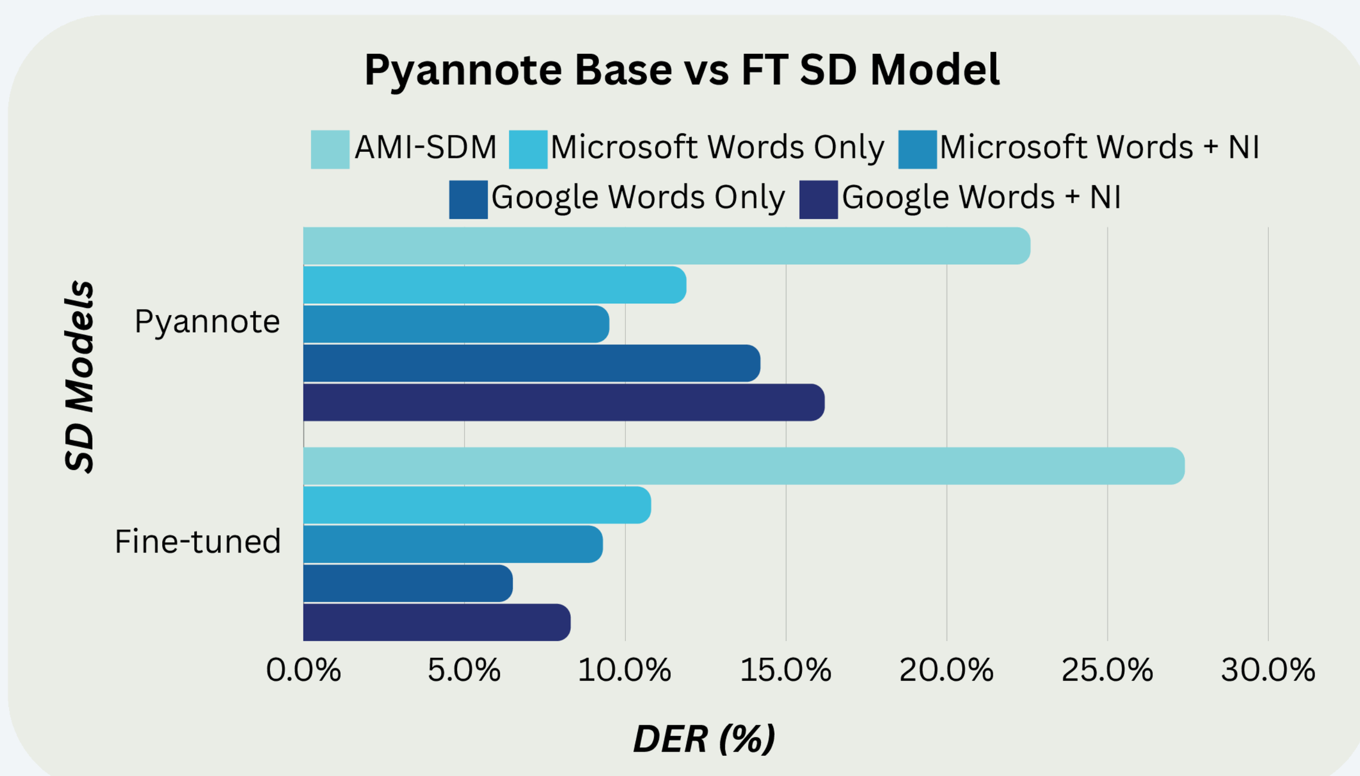

Figure 7: Pyannote Base vs FT SD Model

Automatic Speech Recognition

- FT four existing Automatic Speech Recognition (ASR) models on the synthesised medical data using short utterances from the full conversations.
- Accuracy assessed via:
  - Word Error Rate (WER): Overall transcription accuracy.
  - Medical WER (mWER): Accuracy in medical jargon-heavy segments.

$$WER = \frac{num\ substitutions + num\ deletions + num\ insertions}{total\ number\ of\ words}$$

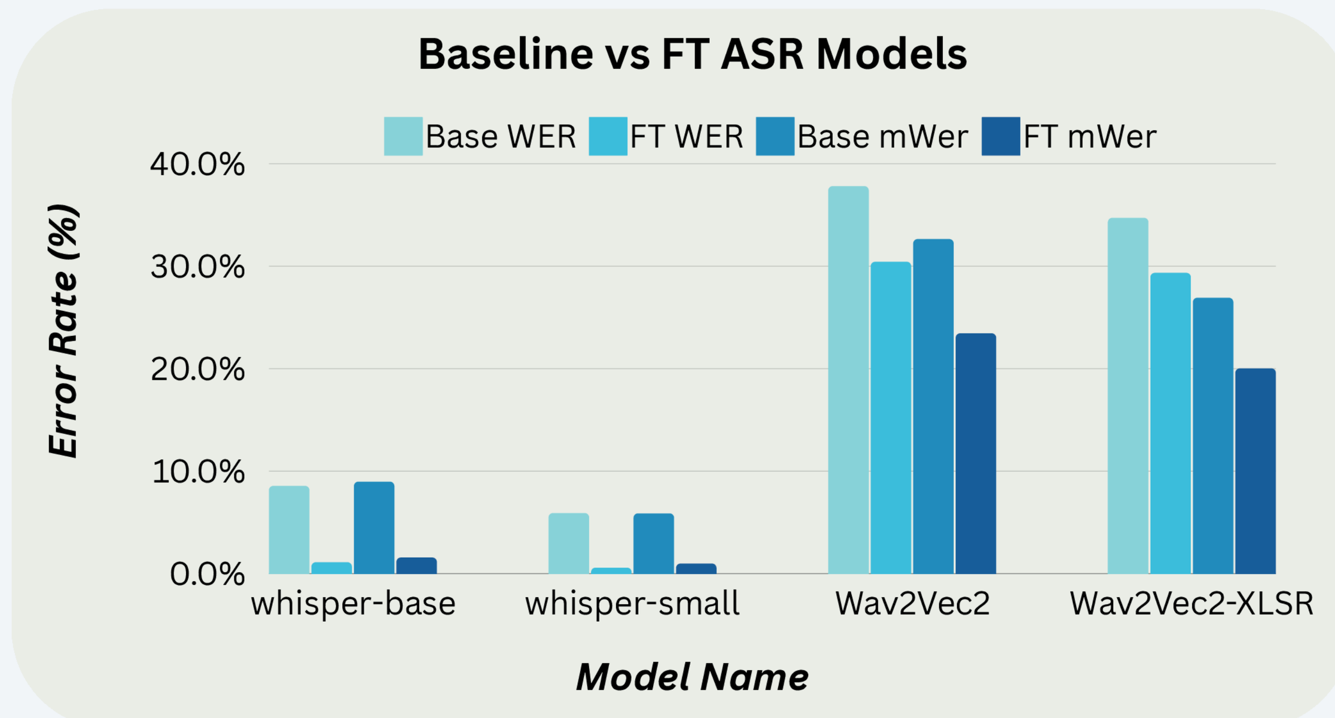

Figure 8: Error Rate Comparison of FT and Baseline models

Language Model Integration

- Integrated LM with the FT models, to improve transcription accuracy and correct semantic errors.
- Improved mWER and WER for Wav2Vec2 and XLSR. However, Whisper models show reduced accuracy.

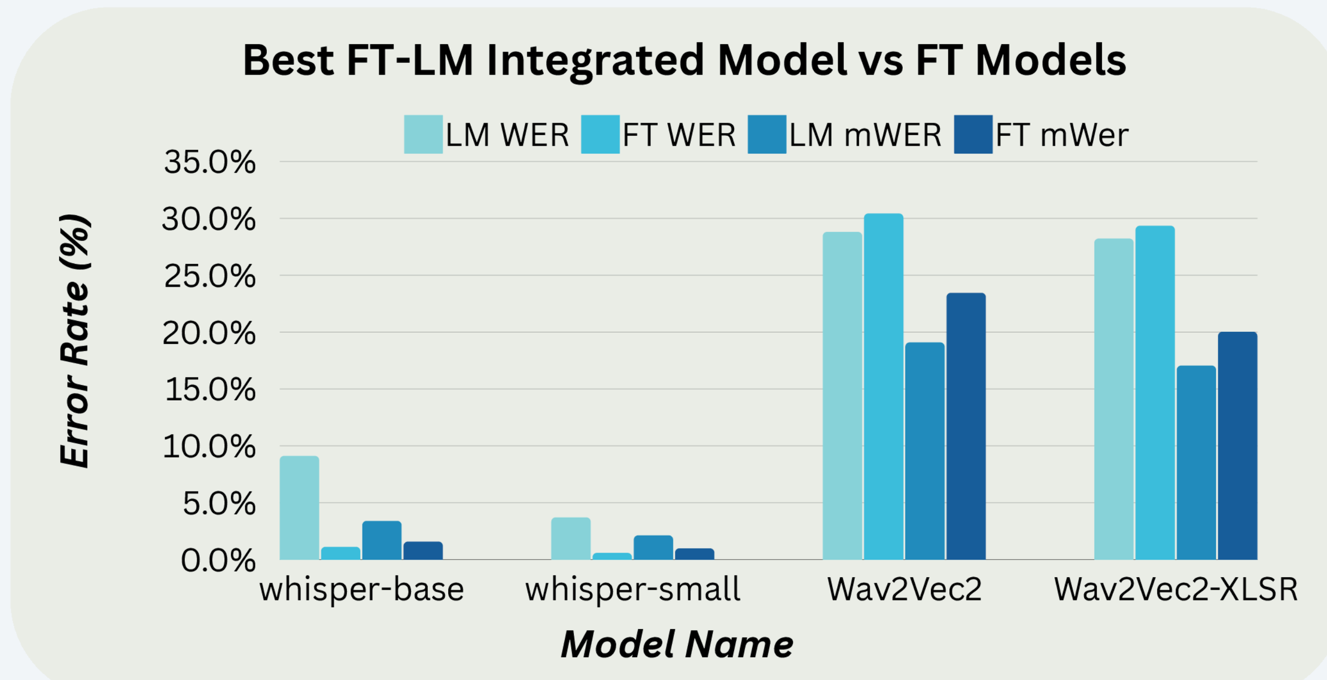

Figure 9: Error Rate Comparison of FT and FT-LM Integrated models

Best Model Variants Comparison

Table 1: Comparison of WER and mWER for Best Model Variants

Model Name	WER (%)	mWER (%)
whisper-base-FT	1.11	1.57
whisper-small-FT	**0.58**	**0.98**
Wav2Vec2-LM	28.81	19.09
Wav2Vec2-XLSR-LM	28.22	17.05

Proof-of-Concept

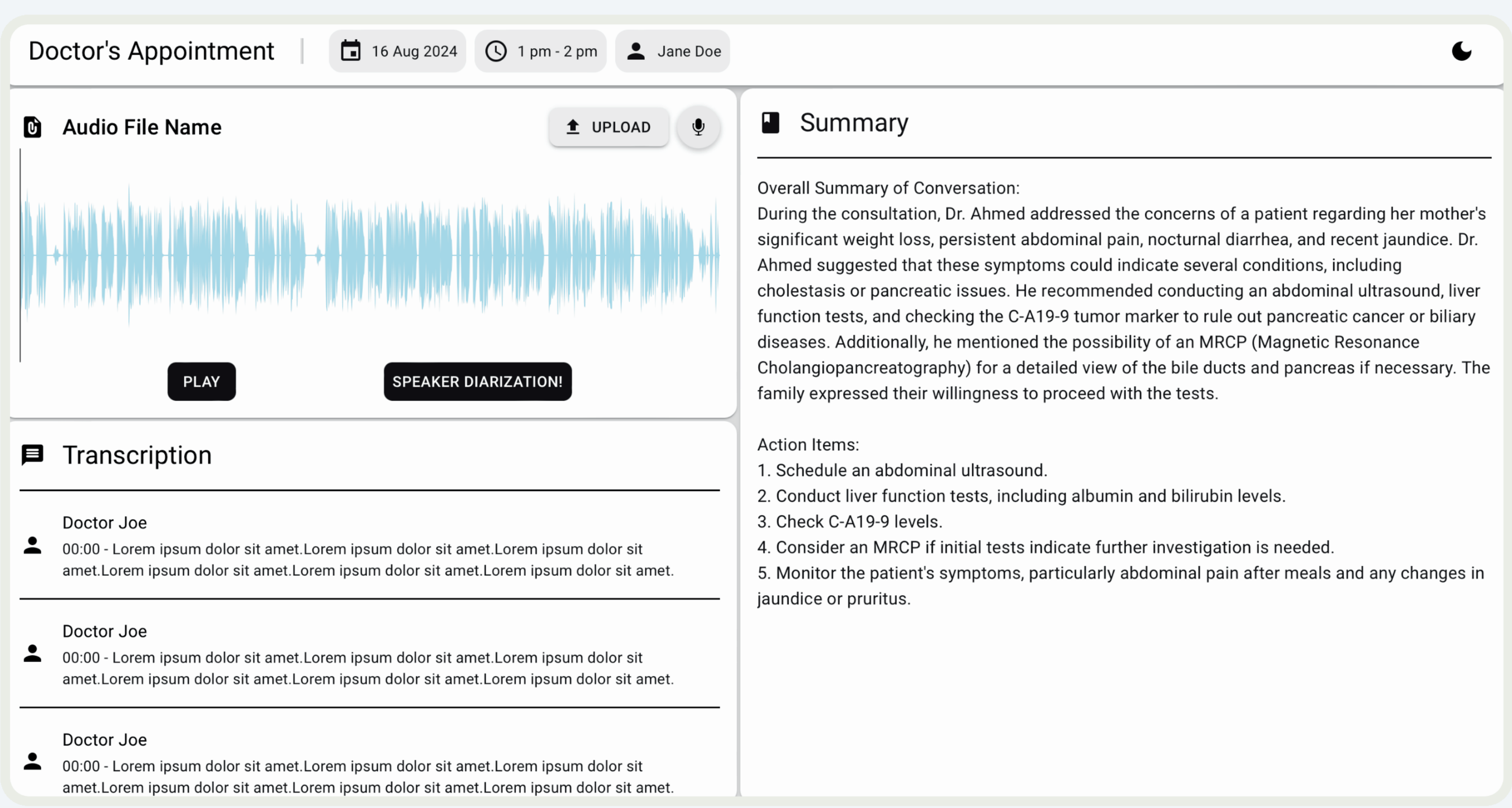

Figure 10: Proof-of-Concept Front-end

Conclusions

- Diverse synthetic medical datasets were generated and evaluated with MOS.
- FT SD model is more accurate on medical datasets.
- FT ASR models boost transcription accuracy for medical vocabulary and conversations.
- NLP improves accuracy for Wav2Vec2 and XLSR models but reduces it for Whisper.
- Created functional Proof of Concept, with full SD, ASR and LM integration.

Future Work

- Test with authentic medical conversations to determine model accuracy and real-world applicability.
- FT SD and ASR models using real medical conversations to further boost performance.
- Include more diverse accents and medical noises for a more general and robust model.
- Experiment with larger model sizes to evaluate impact on performance.

[1] T. D. Shanafelt et al., "Changes in Burnout and Satisfaction With Work-Life Integration in Physicians and the General US Working Population Between 2011 and 2020," Mayo Clin. Proc., Mar. 2022
[2] West CP et al., Physician burnout: contributors, consequences and solutions. J Intern Med. 2018
[3] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in Proc. INTERSPEECH, 2023

UNIVERSITY OF AUCKLAND Waipapa Taumata Rau NEW ZEALAND

ENGINEERING DEPARTMENT OF ELECTRICAL, COMPUTER, AND SOFTWARE ENGINEERING