Enhancing ASR Models for Reliable EMS Radio Transmissions

Lekha Shrivastava: Computer Science, Math
Mentor: Professor Nakul Gopalan
School of Computing and Augmented Intelligence

Can a domain-specific fine-tuned speech-to-text (ASR) model accurately transcribe noisy emergency medical services (EMS)-to-hospital radio communications to streamline hospital preparedness?

Abstract

This project addresses the critical gap in emergency medical services (EMS) communication where noisy radio transmissions often lead to misinterpreted patient data and delayed hospital preparedness. By fine-tuning an existing Automatic Speech Recognition (ASR) model with domain-specific medical jargon and simulated environment noise, this research aims to bridge the communication gap between field medics and Emergency Departments. The goal is to provide high-quality, real-time annotations that improve medical responsiveness and patient outcomes.

Current Gaps

1. Vocabulary Gap

Standard ASR models are trained on general and lack exposure to high-stakes medical jargon

- **Baseline Problem:** General models frequently misinterpret clinical terms → *"lorazepam"* to *"lorase pan."*
- **EMS Solution:** Domain-specific fine-tuning embeds 16 disciplines of medical terminology into the model's vocabulary and 4 different linguistic accents

2. Engineering Acoustic Robustness

Standard models expect "clean" speech

- **Baseline Problem:** treat sirens, engine rumble, and radio static as "signal corruption," leading to high WER rates
- **EMS Solution:** Training on a 5-Factor Noise Engine teaches the model to isolate phonetic patterns from persistent environmental stressors

3. Prioritizing Critical Information Retention

In emergency medicine, a 10% error rate is a safety risk

- **Baseline Problem:** models prioritize "fluency" (making the sentence sound natural) over "accuracy" (getting the specific medical values right).
- **EMS Solution:** Fine-tuning prioritizes Keyword Recall, ensuring patient vitals, medication dosages, and injury locations are transcribed with high precision, even if filler words are lost to noise

Data

1. Data Generation

- Real EMS-to-hospital recordings are scarce and mostly inaccessible due to privacy standards synthetic dataset
- A **synthetic-to-real pipeline** was created to model data after the BPC/CPD Corpus using text-to-speech tools (gTTS) to convert medical scenarios into audio

2. Data Set Split

The resulting 1,700+ audio files were divided

- **70% Training Set** (Light/Medium noise) to learn vocabulary
- **30% Test Set** (Heavy noise) to evaluate post fine-tuning performance

Acknowledgements

Thank you to Professor Gopalan and the Lab for guiding me through this process and being extremely supportive!

Source Transcription

16 domain-specific batches of EMS medical scenarios.

- Includes high-stakes medical jargon

Multi-Accent Synthesis

Conversion of text to audio using TTS engines (gTTS)

- 4 linguistic profiles: US Standard, Spanish, Indian, and US Southern

5 - Factor Noise Engine

Radio stressors: Static, Sirens, Engine Hum, Wind, and Scene Chaos

- Light, Medium, and Heavy tiers

Baseline Model

Fine-Tuned Model

Model Fine-Tuning

Training of pre-trained **Whisper** models using PyTorch/Hugging Face
Uses _____

Evaluation & Response

Measurement of **Word Error Rate (WER)** with a target reduction of **30–50%**

Baseline Results

--------- (insert table)
--------- (analysis + examples)

Fine -Tuned Results

--------- (insert table)
--------- (analysis + examples)

Conclusion

FURI

Ira A. Fulton Schools of Engineering
Arizona State University