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 noise corrupted 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 (Kokoro) to convert medical scenarios into audio

2. Data Set Split

The resulting 850 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 Kokoro

- US Standard, Spanish, Indian, and Chinese
- Accent Map

Noise Injection

- Radio stressors (static, siren, wind, engine)
- Acoustic Discontinuity
- Light, Medium, and Heavy tiers

Baseline Model

Whisper Model (medium) to convert speech to text

- Control group

Fine-Tuned Model

Model chosen by lowest Training Loss and lowest WER

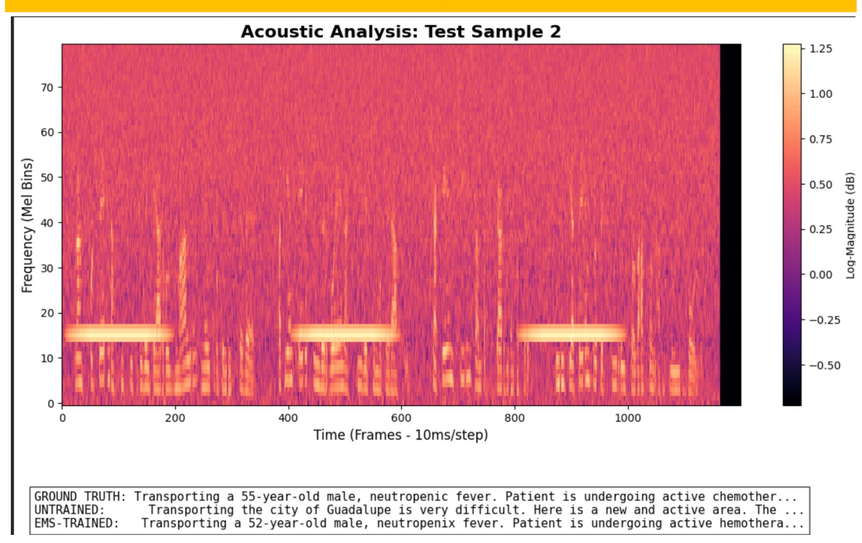
- Minor overfitting

Evaluation & Response

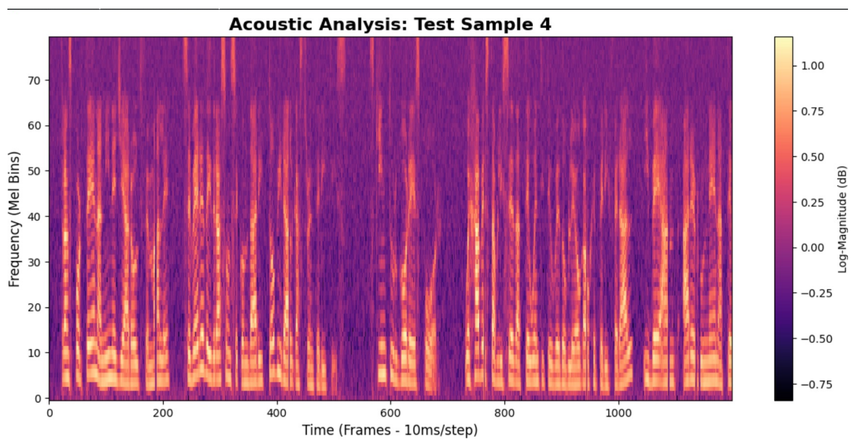
Measurement of **Word Error Rate (WER)** with a target reduction of **30–50%**

Step	Training Loss	Validation Loss	Wer
250	0.382184	0.612903	17.213842
500	0.313305	0.633918	18.855368
750	0.307207	0.640401	15.306122
1000	0.302503	0.643696	14.685004

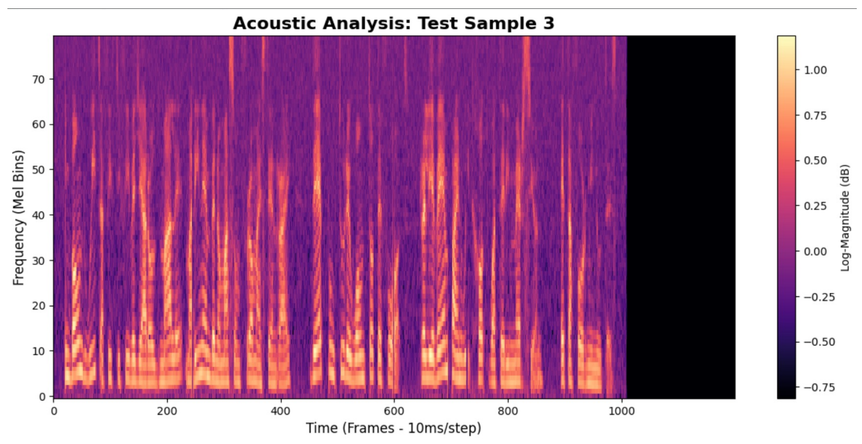
Samples




```
GROUND TRUTH: Transporting a 55-year-old male, neutropenic fever. Patient is undergoing active chemother...
UNTRAINED:      Transporting the city of Guadalupe is very difficult. Here is a new and active area. The  ...
EMS-TRAINED:   Transporting a 52-year-old male, neutropenix fever. Patient is undergoing active hemothera...
```




```
GROUND TRUTH: Transporting a 9-year-old female, severe dehydration secondary to viral gastroenteritis. S...
UNTRAINED:      Transporting to the 9-year-old family, you will see a secondary gas turbine, which is cal...
EMS-TRAINED:   Transporting a 9-year-old female, severe dehydration secondary to viral gastroenteritis. S...
```




```
GROUND TRUTH: Inbound with a 55-year-old male, marine envenomation from a Stingray. Deep puncture wound ...
UNTRAINED:      Inbound wheat at 55-yard-old Malé, marina en benomasom from Aistenrai, de punture eu onto...
EMS-TRAINED:   Inbound with a 55-year-old male, marine envenomation from a stingray. Deep puncture won't ...
```

Results

```
--- Testing UNTRAINED (Base) Model ---
Loading weights: 100%|██████████| 947/947 [00:00<0
Transcribing: 100%|██████████| 88/88 [02:05<00:00,  1.42s/it]

--- Testing TRAINED (EMS) Model ---
Transcribing: 100%|██████████| 88/88 [02:49<00:00,  1.92s/it]
==============================
RESULTS FOR YOUR PAPER:
Untrained Base WER: 74.76%
Fine-Tuned EMS WER: 14.69%
Improvement: 60.07 points
==============================
```

Domain-specific fine-tuning achieved a **60.07% reduction in WER** lowering transcription errors from 74.76% to 14.69%.

Conclusion

Fine-tuning Whisper-Medium on a synthetic-to-real EMS pipeline successfully bridged the gap between general-purpose ASR and mission-critical emergency telecommunications. By training specifically on "noise-corrupted" and "accented medical" data, the model transitioned from a baseline failure to a reliable clinical tool.

- **Robustness:** The model effectively isolated phonetic patterns from engine noise, sirens, radio static and discontinuity that previously caused total signal collapse.
- **Accuracy:** A final **14.69% WER** demonstrates that domain-specific fine-tuning is mandatory for high-stakes environments

FURI

Ira A. Fulton Schools of Engineering
Arizona State University