Speech Recognition System Project Report for Medical Professionals

1. Project Overview

The goal of this project is to develop a speech recognition system tailored for medical professionals, capable of converting audio inputs (e.g., medical dictations or patient interactions) into accurate text outputs. The system must handle specialized medical terminology and perform reliably in real-world scenarios. This report analyzes the performance of various speech recognition models based on provided evaluation metrics and proposes a development strategy.

The dataset includes performance metrics for multiple models/methods across seven medical test cases (test_medikal_apandisit, test_medikal_dis, etc.), evaluated using similarity, BLEU, ROUGE-L, BERTScore F1, and WER (Word Error Rate) on development and test sets. Models include variants of Whisper, Whisper combined with other models (Deepseek, Gemini, GPT), and finetuned versions of Wav2Vec and Whisper on two datasets: TurkishCorpus (open-source) and GeneratedMed (created via OpenAI Audio API).

2. Dataset Details and Preprocessing

The system leverages two primary datasets:

- TurkishCorpus: An open-source dataset containing Turkish audio and text data, used for finetuning baseline models like Wav2Vec and Whisper.
- GeneratedMed: A custom dataset generated via the OpenAI Audio API, tailored for medical contexts. ,

Both includes audio-text pairs split into train, dev, and test.

Each subset contains paired .wav audio files and .txt transcription files, representing medical dictations or interactions in Turkish. For training efficiency, sample limits are applied (e.g., 5 for train, 2 for dev/test in initial tests, scalable to 1000/200/100).

2.1 Preprocessing Pipeline

Audio Preprocessing

- Loading: Audio files are loaded using torchaudio.load(), extracting waveforms and sampling rates.
- Resampling: Audio is resampled to 16kHz using librosa.resample() to match Whisper’s expected input format.
- Padding/Truncation: Mel spectrograms (generated via WhisperProcessor) are padded to 3000 time steps or truncated if longer, ensuring consistent input dimensions (batch_size, 80 mel bands, 3000 time steps).
- Normalization: The normalize_text() function processes transcriptions: Converts text to lowercase and strips whitespace. Maps Turkish characters (e.g., "ğ" → "g", "ı" → "i") to ASCII equivalents for consistency. Removes punctuation and special characters (e.g., commas, periods) using regex ([,\?\.\!\-\;:"“%‘”�]). Collapses multiple spaces into single spaces. Tokenization: WhisperProcessor.tokenizer converts normalized text into token IDs, padded/truncated to a max length of 448 tokens, with padding tokens replaced by -100 for loss computation.

- AudioTextDataset: A custom TorchDataset class pairs audio and text files, sorting them alphabetically to ensure alignment. It limits samples (e.g., max_samples=5) and provides items as dictionaries with speech (audio array), sentence (normalized text), and file paths.
- Batching: DataLoader with a custom data_collator batches data, applying audio and text preprocessing on-the-fly. Batch size is set to 8, with shuffling for training and gradient accumulation (steps=2) to manage memory.

2.2 Datasets

- Turkish Speech Corpus is an open source corpus dataset.
- However, GeneratedMed's audio-text pairs are designed for medical scenarios (e.g., "apandisit" for appendicitis), enhancing model familiarity with terminology.
- Turkish Language: Normalization of Turkish characters and forced decoder IDs (language="tr", task="transcribe") ensure accurate transcription in Turkish.

2.3 API

- Initial Transcription: Uses whisper.load_model("large") for baseline transcription of audio files (e.g., .mp3).
- Refinement: Integrates models like Gemini or GPT to refine transcriptions, focusing on medical accuracy via prompts (e.g., "Refined transcription in tr language").
- Flexibility: Supports multiple models (Gemini, GPT, Deepseek, Deepseek-R1) via command-line arguments, aligning with the hybrid approaches evaluated (e.g., Whisper + Gemini).

3. Performance Analysis

3.1 Metrics Overview

- Metrics Overview Similarity: Measures semantic similarity between generated and reference text (0 to 1, higher is better).
- BLEU: Assesses n-gram precision (0 to 1, higher is better).
- ROUGE-L: Evaluates longest common subsequence overlap (0 to 1, higher is better).
- BERTScore F1: Contextual embedding-based similarity (typically -1 to 1, higher is better).
- WER: Word Error Rate (lower is better; provided for finetuned models on dev/test sets).

3.2 Model Performance Summary

Model/Method		**similarity**	**bleu**	**rougeL**	**bert_score_f1**	**wer(dev)**	**wer(test)**
Whisper	test_medikal_apandisit	0,3357	0,5165	0,8073	0,6484		
	test_medikal_dis	0,1113	0,6147	0,8696	0,7556		

	test_medikal_femur	0,7505	0,5546	0,7995	0,6924		
	test_medikal_kolesistektomi	0,4899	0,6196	0,8792	0,6804		
	test_medikal_lichtenstein	0,0734	0,0017	0,1739	0,0229		
	test_medikal_lumpektomi	0,1396	0,0177	0,2991	0,1241		
	test_medikal_tah	0,4514	0,5023	0,7323	0,4961		
	Average	0,3359714286	0,4038714286	0,6515571429	0,4885571429		
Whisper + Deepseek	test_medikal_apandisit	0,1192	0,0951	0,2923	0,1476		
	test_medikal_dis	0,0735	0,033	0,1264	-0,2061		
	test_medikal_femur	0,0872	0,0559	0,2578	0,1889		
	test_medikal_kolesistektomi	0,0267	0,02	0,2033	0,202		
	test_medikal_lichtenstein	0,0516	0,0014	0,1104	-0,0157		
	test_medikal_lumpektomi	0,0581	0,0137	0,1434	-0,2137		
	test_medikal_tah	0,0317	0,0504	0,208	-0,1155		
	Average	0,064	0,0385	0,1916571429	-0,0017857142 86		
Whisper + DeepseekR1	test_medikal_apandisit	0,3211	0,4892	0,7933	0,6385		
	test_medikal_dis	0,1138	0,1804	0,4092	0,3408		

	test_medikal_femur	0,194	0,2363	0,5601	0,486		
	test_medikal_kolesistektomi	0,0847	0,2855	0,6271	0,5546		
	test_medikal_lichtenstein	0,083	0,0016	0,1518	0,045		
	test_medikal_lumpektomi	0,1024	0,0177	0,2216	0,0616		
	test_medikal_tah	0,2839	0,341	0,7045	0,3511		
	Average	0,1689857143	0,2216714286	0,4953714286	0,3539428571		
Whisper + Gemini	test_medikal_apandisit	0,8959	0,7443	0,9142	0,8075		
	test_medikal_dis	0,89	0,7212	0,9001	0,7364		
	test_medikal_femur	0,8083	0,7091	0,8824	0,7402		
	test_medikal_kolesistektomi	0,6351	0,7636	0,935	0,8073		
	test_medikal_lichtenstein	0,1811	0,0007	0,1831	0,1391		
	test_medikal_lumpektomi	0,2113	0,0367	0,3381	0,1259		
	test_medikal_tah	0,5569	0,6649	0,863	0,5918		
	Average	**0,5969428571**	**0,5200714286**	0,7165571429	**0,5640285714**		
Whisper + GPT	test_medikal_apandisit	0,3581	0,7092	0,9084	0,7819		
	test_medikal_dis	0,8818	0,7328	0,9206	0,6608		

	test_medikal_femur	0,8123	0,7384	0,9235	0,7182		
	test_medikal_kolesistektomi	0,6076	0,7196	0,9206	0,5996		
	test_medikal_lichtenstein	0,1693	0,0009	0,1769	0,1403		
	test_medikal_lumpektomi	0,1536	0,0341	0,3295	0,1285		
	test_medikal_tah	0,6511	0,646	0,8717	0,5661		
	Average	0,5191142857	0,5115714286	**0,7216**	0,5136285714		
Finetuned Wav2Vec on TurkishCorpus							
	test_medikal_apandisit	0,012	0,0869	0,0355	-0,8756		
	test_medikal_dis	0,0096	0,0994	0,0295	-0,8766		
	test_medikal_femur	0,0125	0	0,0306	-0,8332		
	test_medikal_kolesistektomi	0,0097	0,0868	0,0417	-0,8422		
	test_medikal_lichtenstein	0,0105	0,1045	0,0368	-0,9103		
	test_medikal_lumpektomi	0,0129	0,0853	0,0256	-0,8983		
	test_medikal_tah	0,0096	0,1012	0,0315	-0,8786		
	Average	0,0109714285 7	0,0805857142 9	0,0330285714 3	-0,8735428571	3,80024	3,54228

Finetuned Wav2Vec on GeneratedMed							
	test_medikal_apandisit	0,0135	0	0	-0,8485		
	test_medikal_dis	0,0158	0	0,0075	-0,8508		
	test_medikal_femur	0,008	0	0	-0,833		
	test_medikal_kolesistektomi	0,0148	0	0	-0,8339		
	test_medikal_lichtenstein	0,0117	0	0,0069	-0,8495		
	test_medikal_lumpektomi	0,0167	0	0	-0,823		
	test_medikal_tah	0,01	0	0	-0,8158		
	Average	0,012928 57143	0	0,00205 714285 7	-0,8363 571429	1	1
Finetuned Whisper on TurkishCorpus							
	test_medikal_apandisit	0,0388	0,0084	0,0561	0,4599		
	test_medikal_dis	0,008	0,004	0,0358	0,4188		
	test_medikal_femur	0,0234	0,0046	0,0619	0,44		
	test_medikal_kolesistektomi	0,0368	0,0033	0,0619	0,4698		
	test_medikal_lichtenstein	0,0459	0,0686	0,0364	0,3876		
	test_medikal_lumpektomi	0,0751	0,0012	0,0502	0,4372		

	test_medikal_tah	0,0218	0,0023	0,0485	0,4516		
	Average	0,03568571429	0,0132	0,05011428571	0,4378428571	5,98481	4,09318
Finetuned Whisper on GeneratedMed							
	test_medikal_apandisit	0,1344	0,0028	0,1592	0,5369		
	test_medikal_dis	0,1145	0,0012	0,141	0,4829		
	test_medikal_femur	0,1375	0,0054	0,199	0,5102		
	test_medikal_kolesistektomi	0,1505	0,0016	0,1678	0,5132		
	test_medikal_lichtenstein	0,0167	0	0,0076	0,2618		
	test_medikal_lumpektomi	0,1552	0,0005	0,1154	0,4677		
	test_medikal_tah	0,122	0,0012	0,1543	0,5332		
	Average	0,1186857143	0,001814285714	0,1349	0,4722714286	**0,94044**	**0,9172**

4. Discussion

The evaluation of various speech recognition models reveals significant performance disparities, with Whisper + Gemini and Whisper + GPT emerging as top performers. Whisper + Gemini achieves an average similarity score of 0.5969, a ROUGE-L of 0.7166, and a BERTScore F1 of 0.5640, while Whisper + GPT follows closely with 0.5191, 0.7216, and 0.5136, respectively. These metrics underscore their ability to produce accurate and contextually relevant transcriptions, making them well-suited for medical applications. In contrast, the standalone Whisper (Baseline) delivers moderate results (e.g., 0.6516 ROUGE-L) but falters on challenging cases like "lichtenstein" (0.0734 similarity), indicating limitations in handling specialized terminology without enhancement.

Other approaches fare less favorably. Whisper + Deepseek exhibits poor combination, reflected in its dismal -0.0018 BERTScore F1, suggesting incompatibility or inadequate integration. Similarly, Finetuned Wav2Vec models—whether on TurkishCorpus or GeneratedMed—display abysmal performance, with negative BERTScore F1 values (e.g., -0.8735 and -0.8364) and high WER (e.g., 3.8002 and 1.0000), pointing to ineffective training or a mismatch between the model and medical data. However, Finetuned Whisper on GeneratedMed outperforms its TurkishCorpus counterpart, achieving a BERTScore F1 of 0.4723 versus 0.4378 and a significantly lower WER (0.9404 vs. 5.9848). This highlights the superior quality and relevance of the GeneratedMed dataset, generated via the OpenAI Audio API, for medical speech recognition.

The preprocessing pipeline further enhances system efficacy, with audio resampling to 16kHz and text normalization of Turkish characters ensuring compatibility and consistency. Based on this analysis, Whisper + Gemini is selected as the primary model due to its exceptional performance across key metrics (similarity: 0.5969, ROUGE-L: 0.7166, BERTScore F1: 0.5640), supported by its integration with Gemini for transcription refinement. Whisper + GPT serves as a robust fallback, offering competitive results (e.g., 0.7216 ROUGE-L). For optimal outcomes, finetuning on the GeneratedMed dataset is recommended, as it demonstrably enhances Whisper's performance (e.g., 0.4723 BERTScore F1 vs. 0.4378 on TurkishCorpus), aligning the system with the medical domain's unique demands.

5. Conclusion

The development of a speech recognition system for medical professionals hinges on achieving high accuracy and contextual fidelity, particularly for Turkish medical terminology. The Whisper + Gemini model, finetuned on the GeneratedMed dataset, emerges as the strongest candidate, delivering a balanced performance with a similarity score of 0.5969, a ROUGE-L of 0.7166, and a BERTScore F1 of 0.5640. This superiority is reinforced by the GeneratedMed dataset's proven value—reducing WER to 0.9404 compared to 5.9848 on TurkishCorpus—and the robust preprocessing pipeline, which ensures reliable audio and text inputs through resampling and normalization.

While challenges persist, such as the system's struggles with specific cases (e.g., "lichtenstein") and the need for lower WER in finetuned models, these can be mitigated through targeted data augmentation and optimization. The integration of Gemini for refinement enhances transcription quality, making the system practical for real-world medical dictations and patient interactions. With a web-based deployment leveraging Whisper + Gemini and GeneratedMed finetuning, this solution meets the target goals of high similarity (>0.5), low WER (<1.0), and strong contextual accuracy (BERTScore F1 >0.5). Moving forward, development should focus on refining edge cases and real-time performance, with a prototype targeted for April 30, 2025, to advance medical documentation efficiency.

6. Future Improvements

While the proposed speech recognition system, leveraging Whisper + Gemini and finetuned on the GeneratedMed dataset, demonstrates strong potential for medical applications, several areas remain for further improvement of its accuracy and usability.

- Improved Dataset: Models consistently underperform on specific test cases like "lichtenstein" (e.g., 0.0734 similarity for Whisper baseline, 0.0450 BERTScore F1 for

Whisper + DeepseekR1), indicating difficulties with less common medical procedures or terms. Augmenting GeneratedMed dataset with additional audio-text pairs for underrepresented medical terms and procedures. This could involve recording or synthesizing more diverse dictations (e.g., using your voice or AI-generated samples) and incorporating them into training. Techniques like data augmentation (e.g., adding noise or varying pitch) could further improve generalization.

- Reduction of Word Error Rate (WER): Finetuned models, such as Whisper on TurkishCorpus (WER 5.9848 dev, 4.0932 test) and even GeneratedMed (0.9404 dev, 0.9172 test), exhibit higher WER than desired for real-world reliability (<1.0 is the target, but ideally closer to 0.5). Optimize the finetuning process by increasing training epochs, experimenting with learning rate schedules (e.g., cosine annealing), or using larger batch sizes with gradient accumulation.
- Real-Time Processing: The current system, while effective for offline transcription, lacks explicit optimization for real-time use, which is critical for live patient interactions or dictations. Implement model quantization (e.g., using ONNX or PyTorch's quantization tools) to reduce inference latency, and explore streaming audio input with chunked processing.
- Dialect Support: The system is tailored to Turkish, with preprocessing focused on normalizing Turkish characters (e.g., "ğ" → "g"), but it may not generalize to regional dialects or multilingual medical environments. Expand the dataset to include Turkish dialects (e.g., Black Sea or Southeastern variants) and consider multilingual finetuning with datasets like Common Voice for other languages common in medical settings (e.g., English, Arabic).
- User Interface: A web-based interface to meet basic needs for medical professionals.

Rabia Eda Yılmaz
03.04.2025