ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

School of Information Sciences and Technology
Department of Informatics
Athens, Greece

Bachelor Thesis
in
Informatics

Automatic Speech Recognition for Greek Medical Dictation

Vardis Georgilas

Supervisor: Assoc. Prof. Themos Stafylakis
Department of Informatics
Athens University of Economics and Business

August 2025

Vardis Georgilas

Automatic Speech Recognition for Greek Medical Dictation

August 2025

Supervisors: Assoc. Prof. Themos Stafylakis

Athens University of Economics and Business

School of Information Sciences and Technology

Department of Informatics

Athens, Greece

Abstract

Medical dictation systems are essential tools in modern healthcare, enabling accurate and efficient conversion of speech into written medical documentation. The main objective of this thesis is to create a domain-specific system for Greek medical speech transcriptions. The ultimate goal is to assist healthcare professionals by reducing the overload of manual documentation and improving workflow efficiency. Towards this goal, we develop a system that combines automatic speech recognition techniques with text correction models, allowing better handling of domain-specific terminology and linguistic variations in Greek. Our approach leverages both acoustic and textual modeling to create more realistic and reliable transcriptions. We focused on adapting existing language and speech technologies to the Greek medical context, addressing challenges such as complex medical terminology and linguistic inconsistencies. Through domain-specific fine-tuning, our system achieves more accurate and coherent transcriptions, contributing to the development of practical language technologies for the Greek healthcare sector.

Περίληψη

Η ιατρική υπαγόρευση αποτελεί μια πρακτική λύση που βοηθά τους επαγγελματίες υγείας να μειώσουν τον χρόνο και την προσπάθεια που απαιτεί η γραπτή τεκμηρίωση. Η παρούσα πτυχιακή εργασία έχει ως αντικείμενο την κατασκευή ενός συστήματος για την αυτοματη μετατροπή ελληνικής ιατρικής ομιλίας σε κείμενο. Υλοποιήθηκε ένα σύστημα που συνδυάζει τεχνικές αυτόματης αναγνώρισης ομιλίας με μοντέλα γλωσσικής αξιολόγησης για βελτίωση της ακρίβειας. Με αυτόν τον τρόπο, το σύστημα διαχειρίζεται αποτελεσματικότερα την εξειδικευμένη ιατρική ορολογία και τις γλωσσικές ιδιαιτερότητες της ελληνικής γλώσσας, αντιμετωπίζοντας προβλήματα που προκύπτουν από την πολυπλοκότητα της ορολογίας και τη μεταβλητότητα της προφορικής ομιλίας. Για την αυτόματη αναγνώριση ομιλίας χρησιμοποιήθηκε το μοντέλο Whisper, το οποίο εκπαιδεύτηκε περαιτέρω σε ελληνικά δεδομένα ώστε να προσαρμοστεί καλύτερα στις ανάγκες του συγκεκριμένου τομέα. Επιπλέον, αξιοποιήθηκε ένα ειδικά προσαρμοσμένο ελληνικό GPT-2 μοντέλο, το οποίο λειτουργεί ως εργαλείο γλωσσικής αξιολόγησης, επιλέγοντας την καταλληλότερη πρόταση ανάμεσα σε πολλαπλές πιθανές μεταγραφές που παράγονται από το Whisper. Η ενσωμάτωση της ακουστικής και γλωσσικής πληροφορίας συμβάλλει σημαντικά στην αύξηση της ακρίβειας και της φυσικότητας των τελικών κειμένων. Με αυτή την προσέγγιση, το σύστημα στοχεύει στη δημιουργία αξιόπιστων και κατανοητών μεταγραφών, προσφέροντας ένα χρήσιμο εργαλείο για την υποστήριξη της καθημερινής εργασίας στον τομέα της υγείας.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Themos Stafylakis, for his continuous guidance, support, and insightful feedback throughout this thesis. His expertise was invaluable in helping me overcome challenges and deepen my understanding of the subject.

I would also like to thank Prof. Ion Androutsopoulos for his initial guidance and for introducing me to this very interesting research area. Working on this thesis allowed me to explore topics that I found both challenging and fascinating, and it has greatly expanded my knowledge in the field of speech recognition and language processing.

Finally, I am grateful to my friends and colleagues for their encouragement and support, and to my family for their patience, motivation, and constant belief in me throughout this journey.

Contents

1 Introduction

Medical dictation plays a pivotal role in modern healthcare workflows. Precise and timely documentation of medical information enhances diagnostic accuracy, ensures continuity of care, and supports legal protection. Traditional documentation is a very time-consuming process that places a significant burden on healthcare professionals, requiring them to spend more time in each case. Dictation systems provide a fast and natural alternative to manual data entry, reducing documentation workload and improving healthcare professionals efficiency.

Greek medical dictation is under-resourced compared to English systems. In our era, where speech technologies are rapidly advancing and widely adopted in healthcare, this gap highlights the urgent need for dedicated resources and development. Existing speech recognition systems perform poorly in Greek medical domain due to complex domain-specific terminology, the linguistic characteristics of the Greek language and possibly due to variations in individual speech patterns. It becomes very clear that there is a significant need for medical dictation systems capable of accurately transcribing Greek speech.

In this thesis, we develop a system that combines state-of-the-art automatic speech recognition techniques with domain adapted language models for Greek medical dictation. The speech recognition process component is based on the pre-trained Whisper model [Rad+23], further adapted to the Greek language. Additionaly, a Greek version of GPT-2 language model fine-tuned with Greek medical text is used to evaluate and select the best transcription hypothesis from the multiple candidates generated by the Automatic Speech Recognition (ASR) system. This approach integrates both acoustic and linguistic information to improve transcription accuracy and and better handle specialized medical terminology. Furthermore, we experimented with the mT5 model for automatic sentence correction. This application uses a different approach, where the model processes the entire sentence and attempts to correct all possible errors, instead of selecting the best transcription from multiple candidates. This method focuses on correcting the output by identifying and fixing mistakes within the text, aiming to improve overall transcription accuracy and fluency.

1.1 Motivation and Problem Statement

This thesis is motivated by the need of Greek medical dictation systems, which are essential in the modern healthcare environment. While English-language systems have seen significant progress, Greek medical dictation remains largely underdeveloped. Current speech recognition models have room for improvement in accurately handling domain-specific terminology and linguistic characteristics of the Greek language. The ultimate goal is to contribute to the development of Greek medical speech technologies by providing solutions that improve transcription accuracy, reduce documentation time, and enhance the efficiency of healthcare workflows.

1.2 Thesis Structure

Chapter 2: Background and Related Work

This chapter provides background on Greek automatic speech recognition and language modeling, focusing on their application to medical dictation. It reviews general ASR systems, domain adaptation techniques, and language models used for transcription refinement. The chapter also discusses challenges specific to Greek language processing, emphasizing the absence of dedicated Greek medical dictation models.

Chapter 3: Implemented methods and Systems

This chapter discusses the approaches and the models that was used as part of the thesis. The primary focus is on the fine-tuning of two pre-trained models, Whisper for automatic speech recognition (ASR) and GPT-2 for post-processing and correction of transcribed text.

Chapter 4: Data

This chapter presents an overview of the different datasets that was used for this task. It outlines their structure and key features, along with the preprocessing methods used to make the data suitable for training and evaluation.

Chapter 5: Experiments and Results

This chapter discusses the results of applying those models. It covers the training procedures, the evaluation metrics and the performance of different models.

Chapter 6: Conclusions and Future Work

This concluding chapter recaps the main findings and contributions of the thesis. It also proposes directions for future work, such as incorporating more speech data within the medical domain.

2

Background and Related Work

Automatic speech recognition and medical data processing are two important and rapidly evolving fields within artificial intelligence and machine learning. The implementation of trustworthy systems that can convert speech in the medical domain into accurate and readable text is crucial for improving the quality of healthcare services and supporting the work of medical professionals. This chapter briefly addresses the basics techniques related to automatic speech recognition, the uniqueness of medical speech and the language models that can be utilized to achieve more accurate results.

2.1 Automatic Speech Recognition

2.1.1 Traditional methods

The very first implementations of automatic speech recognition where statistical models, which aimed to capture the dynamics of human speech using probabilities and simplified assumptions. The most wide used technique was the Hidden Markov Models (HMMs), that often combined with Gaussian Mixture Models (GMMs) for acoustic modeling.

HMMs represent the sequence of spoken words using a limited set of hidden states, each corresponding to a specific phonetic unit. These states are linked by probabilities that describe how likely it is to move from one state to another, and how likely each state is to produce certain sounds. GMMs are used to estimate these sound probabilities, helping the system decide how well a set of audio features matches a specific phonetic state.

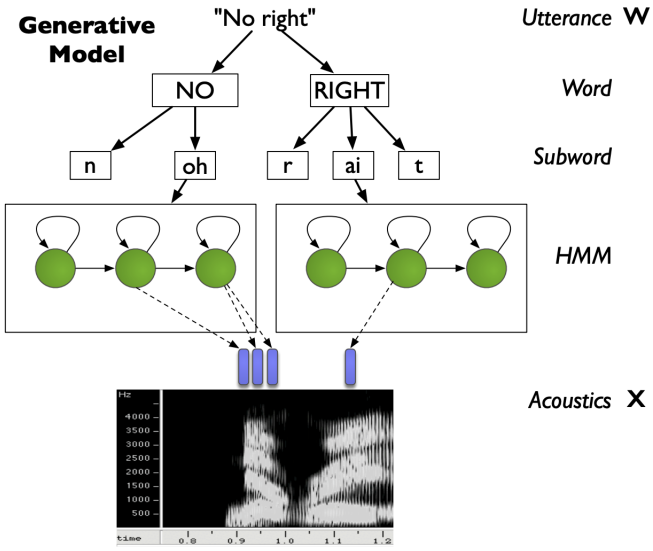

Fig. 2.1: An illustration of the HMM-GMM architecture. Each hidden state in the HMM is associated with a Gaussian Mixture Model that calculates the probability of observing a specific set of acoustic features for that state.

Although for many years HMM with GMM was the standard practice for ASR systems, they had a number of limitations. Those models relied on strong assumptions, such as the independence of observations and linear transitions between states, which do not fully reflect the nature of human speech. Moreover, their ability to learn complex phonetic and linguistic patterns was limited, fact that led to the need for more flexible and no linear models.

2.1.2 Neural network approaches

The rapid development in the field of machine learning, and especially deep learning, has led to the emergence of new approaches to speech recognition based on artificial neural networks. These models are able to learn more complex relationships between acoustic features and words, overcoming the limitations of statistical models.

Recurrent Neural Networks (RNNs) were adopted due to their ability to process sequential data, such as speech. Long Short-Term Memory (LSTMs) networks was a significant advancement, as they were able to keep long-term dependencies and avoid the vanishing gradient problem, which is common in standard RNNs. Furthermore, the emerge of the Transformer architecture changed the approach to sequence processing. Transformers rely on attention mechanisms which allow the model to focus on important parts of the input regardless of their position in the sequence. This makes it possible to process entire

sentences or even whole paragraphs simultaneously without the temporal dependence of RNNs, accomplishing faster training and better utilization of the information.

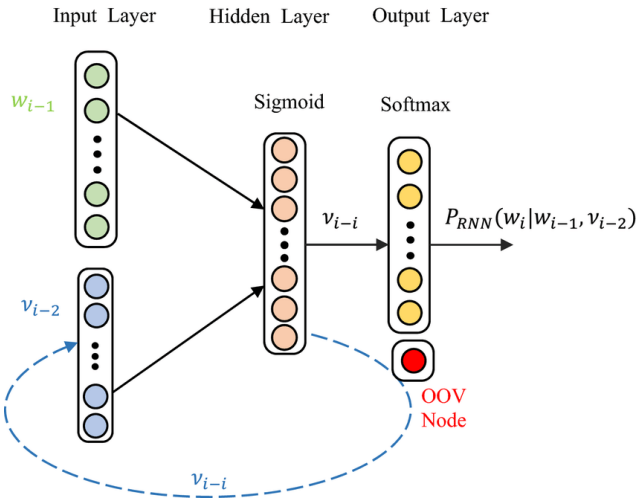

Fig. 2.2: An illustration of an RNN model used in speech recognition, showing sequential input processing, hidden state propagation and softmax based word prediction

Convolutional Neural Networks (CNNs) are often used to improve how features are extracted from raw audio, helping to identify important local patterns in the sound. Additionally, modern speech recognition systems tend to combine acoustic, pronunciation, and language models into one unified system that can be trained all at once. This approach simplifies the overall process, reduces the chance of errors, and leads to better results.

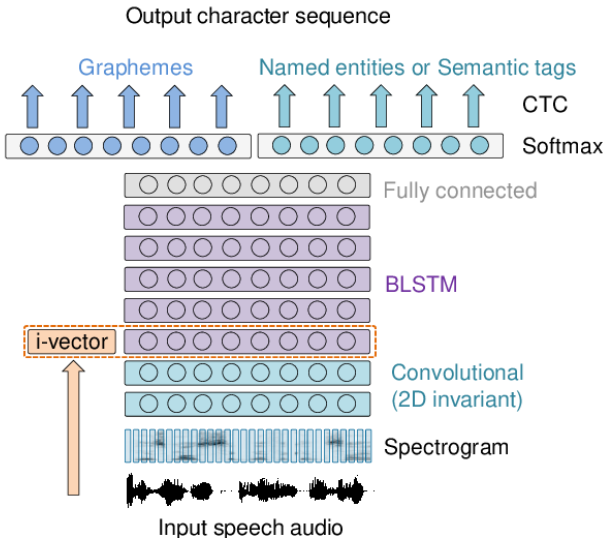

Fig. 2.3: A high-level view of an end-to-end ASR architecture. This approach uses a unified neural network to map audio features directly to text, simplifying the traditional pipeline by combining the acoustic and language models.

In medical dictation, adapting models to the specific domain and using specialized language models is essential. Recent advances in self-supervised learning enable models to first train on large amounts of unlabeled speech and then fine-tune on smaller, specialized datasets. This method is particularly useful for languages like Greek, where labeled medical speech data is limited. These models also handle different speakers, accents, and background noise better, which are common challenges in clinical environments.

Overall, neural network techniques provide effective solutions to improve medical dictation by increasing accuracy, speeding up processing, and better handling specialized terms and varied audio conditions.

2.2 Medical Speech Recognition

Automatic speech recognition in the medical domain poses unique challenges and demands due to the nature and complexity of the spoken content it processes. Medical language is characterized by specialized terminology, complex sentence structures, and a wide range of expressions that differ significantly from everyday speech.

2.2.1 Unique features of medical speech

Medical speech often includes technical terms, abbreviations, acronyms, and context-specific phrases that are rarely found in general language data. These elements require more advanced language models and domain-specific lexicons for accurate processing. Additionally, the clinical environment introduces variability in speaker roles, time-constrained communication, and background noise, all of which add complexity to the transcription task.

Accurate recognition of medical speech is essential for ensuring proper clinical documentation, improving workflow efficiency, and reducing the risk of medical errors. However, most general purpose ASR systems struggle in this context due to their lack of exposure to medical vocabulary and real world clinical conditions. The problem is even more pronounced in low-resource languages such as Greek, where annotated medical speech corpora are scarce. These challenges underline the importance of developing ASR systems specifically trained and adapted for the medical domain.

2.3 Language Models in Medical Transcription

The use of Language Models in speech recognition has become very important, especially in environments where accuracy and understanding of the content are critical, such as in the medical field. While ASR models convert spoken language into text, language models help process the generated text by correcting errors and improving coherence.

2.3.1 Language Models for Post-processing

Language Models play a crucial role in the post-processing of texts generated by automatic speech recognition systems. After the initial conversion from speech to text, the resulting text often contains errors, omissions or inconsistencies, especially when the original speech comes from demanding environments such as medical field. Language models help with the correction of those errors, resulting in a more coherent and well-structured transcription.

More specific, modern models that are based on transformer technologies, like GPT-2 and mT5, have the capability to understand the context of each sentence and provide improvements that make the text more natural and understandable. By training these models on large volumes of text, they can identify incorrect sentences or unusual expressions, and improve the flow and clarity.

Using these models for post-processing is especially important in this field, where accurate transcription and clear understanding of texts are critical. Moreover, the ability to adapt these models to specialized medical terminology and writing styles enhances the reliability and precision of the final transcript.

2.3.2 Importance in medical transcription contexts

In medical field, speech transcription is not just about capturing words but requires a high level of accuracy and adaptation to the specific context. Transcriptions are used official documents that impact patient health decisions making their reliability crucial. The presence of specialized terminology and complex phrasing poses challenges that general ASR systems cannot effectively handle.

The use of specialized language models that have been trained or fine-tuned on medical texts significantly helps maintain terminological consistency and reduces errors that could lead to negative consequences. Furthermore, improving the readability and clarity of the transcriptions can help communication among healthcare professionals and increase the efficiency of clinical documentation. The implementation of these technologies strengthens

the reliability of healthcare systems and contributes to better organize and manage medical records, thereby enhancing the quality of care provided.

2.4 Fine-Tuning and Adaptation Techniques

In modern machine learning systems the process of fine-tuning and adapting models to specific domains is a crucial step for improving performance. Instead of training models from scratch, a process that demands large amounts of data and computational resources, we use pretrained models and adapt them to the specific domain. This approach allows faster and more efficient training, because the model leverages the general knowledge it has already obtained.

2.4.1 Transfer Learning and Domain Adaptation

Transfer learning refers to the process where a model, trained on a large and general dataset, is reused and adapted to perform a specific task in a specialized domain. In the field of automatic speech recognition for medical applications transfer learning allows the model to leverage its knowledge of general language and apply it to the recognition of specialized medical terminology and phrasing.

This process plays a crucial factor because the language used in the medical field has specific uniqueness that general models do not cover. Through domain adaptation, the model is learning to recognize and transcribe specialized content more accurately, reducing errors and enhancing the overall quality of the transcription. This makes possible the utilization of existing powerful models, like Whisper, and adjusting them for the specific needs of medical speech, we can achieve better results without requiring extensive training from zero.

2.4.2 Parameter-Efficient Fine-Tuning (LoRA)

Low-Rank Adaptation (LoRA) [Hu+22] is a parameter-efficient fine-tuning technique designed to adapt large pre-trained models to specific tasks without updating all their parameters. Traditional fine-tuning retrains the full set of parameters, which becomes impractical for very large models due to high computational and storage costs.

To overcome this, LoRA introduces two learnable low-rank matrices, A and B, which approximate the weight updates within specific dense layers. Instead of modifying the original weight matrix $W \in R^{d \times d}$, LoRA keeps it frozen and injects a learnable low-rank update $\Delta W = BA$, applied to the input x. This modification is shown in Figure 2.1.

During training, only the small matrices A and B are optimized, significantly reducing the number of trainable parameters. At inference time, the learned adaptation can be merged with the original weights, introducing no additional latency. This makes LoRA an efficient approach for fine-tuning, especially when working with large models such as Whisper.

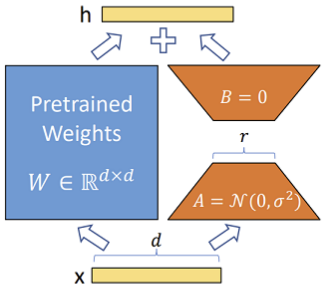

Fig. 2.4: We only train A and B [Hu+22].

In our work, we employed LoRA to fine-tune each model, enabling task-specific adaptation with minimal computational overhead.

3 System Design and Implementation

This chapter present the design and implementation of the Automatic Speech Recognition (ASR) system for Greek medical dictation. The core of this work involves adapting OpenAI's Whisper model to the specific features of the Greek language through a controlled fine-tuning process. A key aspect of our methodology is the comparative analysis of three different sizes of the Whisper model, small, medium, and large-v2. To further enhance transcription quality, we integrated a re-ranking mechanism based on a fine-tuned GPT-2 model, which was used to select the most contextually appropriate transcription among Whisper's alternatives. In this chapter, we present the design of these models, how we adapted them, and the steps involved in the transcription process.

3.1 Automatic Speech Recognition System

The foundation of our system is a advanced ASR model designed to accurately transcribe spoken Greek into text. We selected a state-of-the-art, pre-trained model and fine-tuned it on a extensive collection of Greek speech data. This process was repeated across three model scales to evaluate their performance and resource requirements.

3.1.1 Whisper Model

Whisper [Rad+23] is a state-of-the-art automatic speech recognition model developed by OpenAI. It follows an encoder-decoder architecture and has been trained on a wide range of labeled data from the web that supports multiple languages and various tasks. Due to this large variety of data, whisper shows excellent results in noisy environments and across different languages and voice tones.

This model is comprised of two main parts: the encoder, which receives the audio signals and turns it into feature representations, and the decoder, which uses these representations to generate the final text. Figure 3.1 shows an overview of Whisper's architecture and how data moves from the initial audio signal to the final text output. This figure helps to understand the key parts of the system and how they work together.

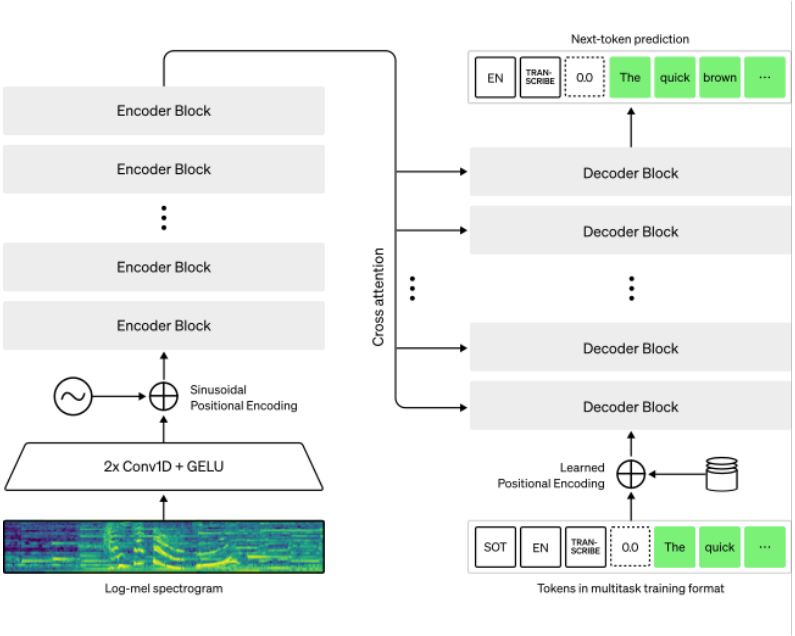

Fig. 3.1: The end-to-end architecture of the Whisper model, which is based on a transformer encoder-decoder framework. Audio is first converted to a log-Mel spectrogram, processed by convolutional layers and sinusoidal positional encodings, then passed through Transformer encoder blocks. The decoder uses learned positional encodings and cross-attention to generate text token-by-token, enabling tasks like transcription and translation in a multitask setup [Rad+23].

In this thesis, we examine three pre-trained versions of Whisper: small, medium, and large-v2. The primary motive for training three distinct models was to analyze the trade-off between performance and computational cost. While larger models like large-v2 are expected to have higher accuracy due to their increased parameters and greater representational capacity, they are also more computationally expensive and harder to deploy in resource-constrained environments. By fine-tuning and evaluating all three models we can determine the optimal one that meets our desired accuracy benchmarks while being practical and efficient for real-world medical dictation.

3.1.2 Adaptation to Greek Language

To adapt Whisper model to the Greek language we applied a fine-tuning methodology to each of the three models. This involved a extensive data preparation and training process.

A diverse set of different Greek speech datasets was aggregate to create a robust corpus of data, containing speech data with different domains, with varying acoustic environments, and from multiple speakers. While a detailed description of these datasets is provided in the next chapter, they include sources such as Mosel (Greek 2009) [Gai+24], Mozilla

Common Voice 11.0 (Greek) [Ard+19], and Google Fleurs (Greek) [Con+22]. The audio files were first resampled to 16 kHz, and the transcriptions were cleaned and normalized so that formatting and style were consistent across the dataset.

Each Whisper model variant was fine-tuned using Hugging Face's Seq2SeqTrainer. We set the learning rate to $5e-5$ so that the models could adapt gradually without overfitting. The batch size was 16 per device, and we used gradient accumulation over two steps to effectively reach a batch size of 32. This setup helped keep training stable while staying within memory limits. AdamW [LH19] optimizer with a weight decay of 0.1 was used for regularization and gradient checkpointing was enabled to reduce memory usage. Training was conducted using mixed precision (bf16) to reduce GPU memory requirements and accelerate computation. To make fine-tuning more efficient, Low-Rank Adaptation (LoRA) [Hu+22] was applied to reduce the number of trainable parameters. Specifically, LoRA was introduced to the **q_proj**, **k_proj**, **v_proj**, and **out_proj** layers using a rank of 32 and $\alpha = 64$. In this way, only about 2% of the full model's trainable parameters are trainable, allowing efficient fine-tuning under limited resources.

Tab. 3.1: Trainable Parameters Comparison of Whisper Models Using LoRA

Model	Parameters (Billion)	Trainable Params (Million)
Whisper Small	0.249	∼7.1 (2.8%)
Whisper Medium	0.783	∼18.9 (2.4%)
Whisper Large-v2	1.574	∼31.5 (2%)

To handle Greek transcription properly, we adjusted the processor so that decoding used Greek-specific prompt IDs. This way, the model was explicitly guided to generate text in Greek. For tokenization, we avoided adding extra special tokens and built the final label sequence by combining the decoder prompt IDs, the transcription tokens, and an end-of-sequence (EOS) token. This setup kept Whisper's decoding process intact while making it more suitable for Greek. During inference, we used greedy decoding with a maximum output length of 250 tokens, which was more than enough for typical Greek text. For training, we kept caching disabled so the model would process sequences consistently, though caching can be turned back on in deployment to speed up inference.

3.2 Greek GPT-2

The Greek GPT-2 model is based on OpenAI's Generative Pre-trained Transformer 2 (GPT-2), a wide recognized language model acknowledged for its ability produce contextually relevant text [Rad+19]. GPT-2 uses the Transformer architecture, which consists of stacked layers combining multi-head self-attention with feed-forward networks [Vas+17]. This

design enables the model to capture dependencies in text, making it adept at auto-regressive tasks like text generation and completion.

In our work we utilized a Greek specific vertion of GPT-2, developed by the the Hellenic Army Academy (SSE) and the Technical University of Crete (TUC). This variant, accessible as lighteternal/gpt2-finetuned-greek [1] on Hugging Face, was pre-trained on a huge corpus of Greek text, optimizing it for the linguistic characteristics of the Greek language. To tailor this model further for specialized applications, we fine-tuned it using Low-Rank Adapta-tion (LoRA), to minimize computational overhead.

3.2.1 GPT-2 Architecture

GPT-2 follows a transformer decoder architecture, consisting of multiple stacked layers of masked multi-head self-attention and point-wise feed-forward networks [Rad+19]. Unlike the original Transformer model [Vas+17], which includes both an encoder and a decoder (Figure 3.2), GPT-2 keeps only the decoder component, making it a unidirectional model that processes input sequences from left to right in an auto-regressive manner [Rad+19].

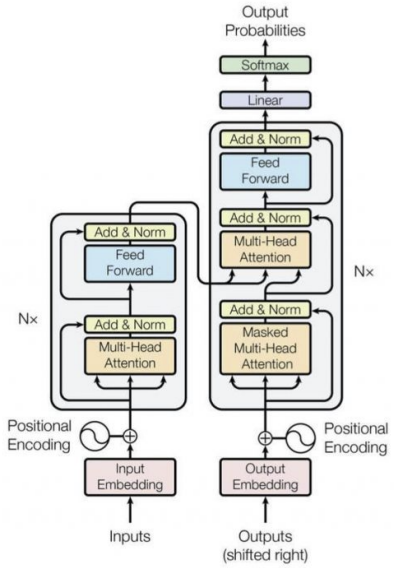

Fig. 3.2: The original Transformer architecture, the model is composed of an encoder (left) and a decoder (right), both consisting of N identical layers. Each encoder layer includes multi-head self-attention and a feed-forward network, while each decoder layer incorporates masked multi-head self-attention, encoder-decoder attention, and a feed-forward network. Residual connections and layer normalization are applied after each sub-layer. Positional encodings are added to the input and output embeddings to preserve the sequential order of tokens [Vas+17].

[1] Hugging Face – lighteternal/gpt2-finetuned-greek

In each decoder layer, masked multi-head self-attention ensures that the model can only attend to the current and previous tokens, enabling causal language modeling [Rad+19]. Residual connections and layer normalization are applied after each sub-layer to ensure training stability. Token embeddings are combined with positional encodings to capture word order, and the final hidden states are projected into the vocabulary space for next-token prediction.

The architecture allows GPT-2 to generate text iteratively, given an initial context, the model samples the most likely next token, appends it to the sequence, and repeats the process [Rad+19]. In this work, we used a Greek GPT-2 model with 124M parameters, consisting of 12 decoder layers, each with 12 attention heads and a hidden size of 768. This configuration provides a balance between computational efficiency and modeling capacity for domain-specific fine-tuning.

3.2.2 Fine-Tuning Process

Fine-tuning large language models like GPT-2 traditionally involves updating all parameters, which can be very resource-intensive. Low-Rank Adaptation (LoRA) offers a more efficient alternative by introducing low-rank matrices to selected layers, allowing the model to adapt with far fewer trainable parameters [Hu+22]. In our implementation, we applied LoRA to the attention (c attn) and projection (c proj) modules of the Greek GPT-2 model, using a rank of 16 and an alpha value of 32. This reduced the trainable parameters to 1,622,016, approximately 1.29% of the model's total 126,061,824 parameters.

We fine-tuned the model on a custom Greek medical text dataset. Hyperparameters are summarized in Table 3.2. Model performance was evaluated using perplexity on a validation set after each epoch, with lower perplexity indicating improved predictive accuracy.

By applying LoRA specifically to attention and projection layers, we were able to efficiently adapt the model to the domain without updating the full set of parameters. Across 30 epochs, perplexity steadily decreased, showing that the model became more confident in predicting the next token. This approach reduced computational cost and memory usage compared to full fine-tuning.

Tab. 3.2: Training Configuration and Hyperparameters

Parameter	Value
Learning Rate	5×10^{-5}
Batch Size	16
Gradient Accumulation Steps	2
Epochs	30
Optimizer	AdamW
Weight Decay	0.01
LoRA Rank	16
LoRA Alpha	32
Trainable Parameters	1,622,016 (1.29%)

3.3 mT5

3.3.1 Model Overview

The multilingual T5 (mT5) model is a sequence-to-sequence transformer architecture, extending the original T5 framework [Raf+20] to support over 100 languages. Unlike decoder-only models such as GPT-2, mT5 includes both an encoder and a decoder, allowing it to look at the entire input sequence before generating output. Its architecture makes it well-suited for text-to-text tasks such as translation, summarization, and grammatical error correction, even in low-resource languages like Greek.

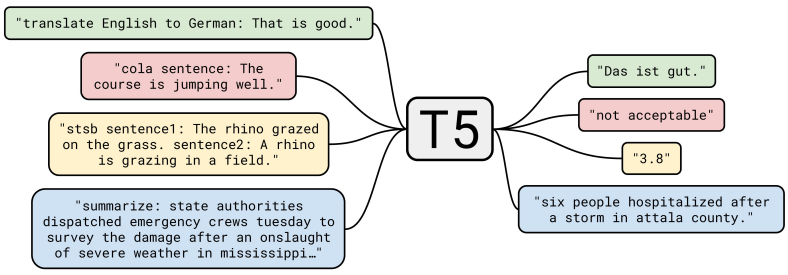

Fig. 3.3: Overview of the T5 model handling multiple NLP tasks using a unified text-to-text framework. Examples include translation, sentence acceptability classification, semantic textual similarity, and summarization, each represented with a specific input prompt and corresponding output[Raf+20].

3.3.2 Model Adaptation

We explored an alternative approach to automatic text correction using the multilingual T5 model [Xue+21]. The objective was to correct noisy Greek sentences generated by the

ASR model. These sentences typically contain a lot of misspellings, missing or incorrect punctuation, grammatical mistakes, and occasionally omitted or repeated words.

The task was framed as a supervised sequence-to-sequence problem, where each training sample consisted of a corrupted sentence as input and its corrected version as the target. More specific, we used an instruction-style prompting approach, where inputs were formatted as "correct: <corrupted sentence>", guiding the model to perform grammatical correction. An example of this input-output format is shown in Table 3.3. For this purpose, we fine-tuned the google/mt5-base, a transformer-based encoder-decoder model pretrained on a diverse multilingual corpus. By conditioning on the full input sequence before generating output, the model is able to capture long-range dependencies, which is especially important for effective grammatical correction.

Input (prompt)	**Model Output**
correct:Ποσο θα διαρκέσει η αποκατάσταση μιτα από το χιρουργειο.	Πόσο θα διαρκέσει η αποκατάσταση μετά από το χειρουργείο;

Tab. 3.3: Example of model input prompt and its corrected output.

3.4 Pipeline Design

Building a robust pipeline for automatic speech recognition (ASR) and transcription correction requires combining advanced models to process audio input and refine the output. This section describes the pipeline shown in Figure 3.4, which integrates Whisper ASR and GPT-2 re-ranking to produce an accurate final transcription.

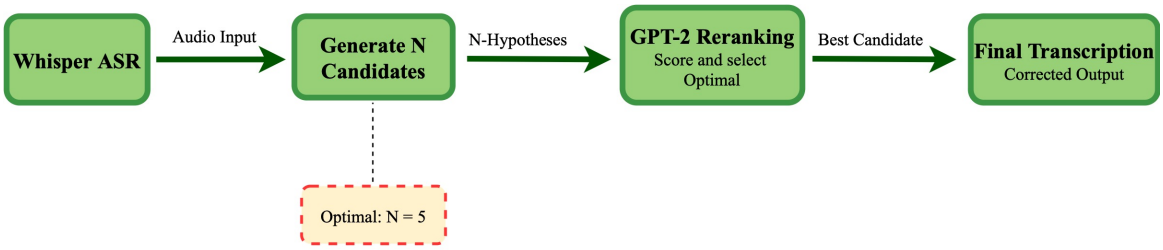

Fig. 3.4: Pipeline for generating and refining speech transcriptions

The pipeline begins with the Whisper ASR model [Rad+23] which processes the audio input. It generates a set of N candidate hypotheses, representing possible transcriptions. Whisper converts raw audio waveforms into text using its pre-trained encoder-decoder architecture and produces multiple transcription variants to capture potential ambiguities in the audio.

The Greek GPT-2 model [Rad+19] evaluates the N candidate hypotheses. Each candidate is scored based on grammatical correctness, contextual relevance, and semantic coherence.

Through this re-ranking process, the most accurate candidate is selected, ensuring that the final transcription aligns well with Greek language conventions. This step enhances the pipeline's accuracy by refining the initial Whisper outputs. The selected candidate from GPT-2 is used as the final transcription corrected output. The output of this step becomes the corrected transcription, representing the most reliable vertion of the original audio. The pipeline's design ensures that the final text is both linguistically accurate and contextually appropriate, making it suitable for Greek medical dictation and other Greek-language processing tasks.

A critical design choice is the selection of the optimal value of N. This parameter determines the number of hypotheses candidates generated by Whisper. The selection of N is a balance between computational load and the diversity of transcription options, enabling GPT-2 to effectively re-rank and select the best candidate. The optimal value of N was determined through empirical testing, ensuring robust performance for Greek audio inputs.

4 Data

This chapter describes the datasets used to train and evaluate the models developed in this thesis. We cover the preparation and processing of several Greek audio datasets used to fine-tune the Whisper model, a medical text corpus for adapting the Greek GPT-2 model, and a custom dataset with manually introduced errors for training the mT5 model in sentence correction. Each dataset was assembled with the needs of its respective task in mind, aiming to support effective training, reliable evaluation, and accurate representation of Greek linguistic features. Below we describe the origin, structure, and processing methods for each dataset.

4.1 Greek Speech Dataset for Whisper

For Whisper, we fine-tuned the model on a composite dataset of Greek speech audio paired with transcriptions. This dataset combined three publicly available sources to ensure a variety of speakers, accents, and acoustic conditions:

- **VoxPopuli & MOSEL**: VoxPopuli is a multilingual speech dataset released by Facebook AI, composed of European Parliament recordings across 23 languages, including Greek [Wan+21]. For Greek, we used the MOSEL dataset [Gai+24] , which provides transcriptions aligned with a subset of VoxPopuli audio. However, the alignments are not always precise, necessitating additional processing as described below.

- **Mozilla Common Voice 11.0**: A crowd-sourced dataset containing Greek speech samples from volunteer contributors, covering various accents and recording environments [Ard+20].

- **Google Fleurs** : A multilingual dataset including Greek speech, focusing on read-aloud sentences across diverse domains [Con+22].

The total duration and number of recordings for each dataset are shown in Table 4.1.

Dataset	Total Hours	Number of Recordings
Mosel	30.34	3876
Common Voice 11.0	6.06	5311
FLEURS	12.72	4136
Total	49.12	13323

Tab. 4.1: Total speech duration and number of recordings per Greek dataset.

The VoxPopuli subset [Wan+21] of the Mosel dataset had several challenges, such as missing transcriptions for some audio files and missing specific timestamps necessary for accurate alignment. These issues caused difficulty in creating a fully aligned speech-to-text dataset. To address this, we used the Aeneas toolkit [1], a forced alignment system that synchronize audio with text. While Aeneas helped create approximate alignments, background noise, speaker variability, and transcription errors sometimes led to poor or partially overlapping segments

To improve data quality, we applied the original Whisper model to validate the alignment quality by comparing Whisper's output with the existing transcriptions. Audio-transcription pairs that didn't match or agree with Whisper's output were discarded. Only pairs with high alignment accuracy and strong consistency between the original transcription and Whisper's output were kept. This filtering made the dataset of well-aligned speech-text pairings cleaner and more trustworthy, so it was ready for training and testing.

4.2 Greek Medical Text Dataset for GPT-2

The Greek GPT-2 model was trained on a custom dataset [2]. This dataset contains 20,430 samples and was constructed from three different sources to ensure a comprehensive representation of medical language in Greek:

- **Medical E-books**: These e-books provided detailed clinical terminology, covering topics such as medical procedures, diagnostics, and patient care. The texts were rich in domain-specific vocabulary, making them suitable for fine-tuning a model for medical text generation [Ιατ15; Σφη+15; Τσι+15].

- **QTLP Greek CC Corpus for the Medical Domain**: This corpus, sourced from Greek web documents automatically classified as medical included a variety of genres

[1] Aeneas is freely available at https://www.readbeyond.it/aeneas/
[2] Medical Text Dataset

such as reference materials, news/journalism, discussions, commercial content, and other medical texts [24].

- **Istorima Podcast Dialogues**: Dialogues from medical-domain podcasts, sourced from `istorima.org`, were included to capture conversational medical content. These dialogues introduced informal and contextually rich language, helping dataset's diversity.

This custom dataset helped the GPT-2 model become familiar with the specific vocabulary and terminology used in Greek medical sentences. As a result, it was better able to rank the candidate sentences produced by the Whisper model according to perplexity, allowing the pipeline to select the most accurate and contextually appropriate transcriptions.

4.3 Error-Augmented Dataset for mT5

In an initial experiment, we also trained a mT5 model for Greek text correction using a custom dataset of 56,000 sentence pairs. The dataset was created by introducing artificial errors, such as vowel swaps, duplicated letters, grammatical issues, and punctuation mistakes into clean Greek text sourced from the previous medical corpus and Wikipedia. Each example consisted of a corrupted sentence paired with its correct version, formatted as input-output pairs for sequence-to-sequence training. An example of this setup is shown in Table 4.2.

Corrupted Sentence	**Clean Sentence**
Ιδανηκο για προληψοη αλλαα και θεραπεια των κατακλισεων.	Ιδανικό για πρόληψη αλλα και θεραπεία των κατακλίσεων.
Η πιεση του αερα ρυθμιζετα ανυλιογα με το βαρος του ασθενη.	Η πίεση του αέρα ρυθμίζεται ανάλογα με το βάρος του ασθενή.

Tab. 4.2: Examples of corrupted–clean pairs from the error-augmented dataset.

While the model learned to correct a variety of errors effectively in this synthetic environment, the artificial nature of the data limited its ability to generalize to real-world mistakes. For this reason, the approach was ultimately not adopted in the final system.

5 Evaluation

This chapter presents the evaluation metrics and the results obtained from the models that we developed within the scope of this thesis. It first introduces the evaluation metrics used to assess the performance of the models, followed by detailed results for each model and the integrated pipeline, highlighting their effectiveness in processing Greek medical speech.

5.1 Evaluation Metrics

To evaluate the performance of our models we employed a set of standard metrics tailored to their respective tasks. Each metric is described below, providing the foundation for the results that we obtained.

5.1.1 WER and CER

Word Error Rate (WER) and Character Error Rate (CER) are standard metrics for evaluating the accuracy of automatic speech recognition systems, such as the Whisper models used in our work. WER measures the percentage of errors in the transcribed text at the word level, calculated as:

$$\text{WER} = \frac{S + D + I}{N} \times 100 \tag{5.1}$$

where S is the number of incorrect words, D is the number of missing words, I is the number of insertions extra words, and N is the total number of words in the reference transcription.

CER is similar but operates at the character level, making it more sensitive to fine-grained errors such as misspellings or punctuation mistakes:

$$\text{CER} = \frac{S_c + D_c + I_c}{N_c} \times 100 \tag{5.2}$$

where S_c, D_c, and I_c represent character-level substitutions, deletions, and insertions, and N_c is the total number of characters in the reference.

Both metrics were used to evaluate the original and fine-tuned Whisper models, as well as the Whisper-GPT-2 pipeline, on the test set of the dataset. Lower WER and CER values indicate higher transcription accuracy, critical for medical dictation applications where precise terminology is essential.

5.1.2 Perplexity

Perplexity is a key metric for evaluating language models, such as the Greek GPT-2 model. It measures the model's predictive uncertainty, quantifying how well it predicts the next token in a sequence. Perplexity is defined as:

$$\text{PPL(W)} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log p(w_i \mid w_{1:i-1})\right) \tag{5.3}$$

where $p(w_i|w_{1:i-1})$ is the probability of token w_i given the previous tokens, and N is the total number of tokens in the test set.

Lower perplexity indicates better model performance, reflecting stronger adaptation to the target domain, such as Greek medical text. Perplexity was used to assess GPT-2's ability to rank Whisper transcriptions by assigning lower perplexity scores to more contextually appropriate outputs. This metric was critical for evaluating the Whisper-GPT-2 pipeline's ranking effectiveness.

In practice, perplexity can also serve as a proxy for cross-entropy loss, since:

$$\text{PPL} = \exp(\text{CrossEntropy}) \tag{5.4}$$

This equivalence helps interpret model performance in terms of both information theory and practical error rates. For language modeling tasks, perplexity directly reflects how "surprised" the model is by the test data. A perfect model assigning a probability of 1 to the correct next token would have a perplexity of 1, whereas a model making uniform predictions would have a perplexity equal to the vocabulary size.

In our work, comparing perplexity across outputs allowed us to identify transcription variants that better aligned with the model's learned representations of fluent and domain-specific Greek.

5.1.3 BLUE

Bilingual Evaluation Understudy (BLEU) metric [Pap+02] is a widely used method for evaluating the quality of machine-generated text by comparing this text to reference texts. It measures the similarity by calculating the overlap of n-grams, which are sequences of n consecutive words. A higher number of matching n-grams indicates better quality, with BLEU scores ranging from 0 to 1, where a score of 1 represents a perfect match between the candidate and reference texts. BLEU is calculated for multiple values of n, and the scores are combined using the geometric mean. The precision p_n for each n-gram length is calculated as:

$$p_n = \frac{\text{Number of matched } n\text{-grams}}{\text{Total number of } n\text{-grams in candidate}} \tag{5.5}$$

The precision scores for different n-gram lengths are combined into a geometric mean to balance their contributions. The geometric mean is calculated as:

$$\text{Geometric Mean} = \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{5.6}$$

To account for differences in length between the candidate and reference texts, a brevity penalty (BP) is applied. The brevity penalty is defined as:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-\frac{r}{c}} & \text{if } c \leq r \end{cases} \tag{5.7}$$

where c is the length of the candidate text and r is the length of the reference text.

The final BLEU score is obtained by multiplying the geometric mean by the brevity penalty:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{5.8}$$

This formula ensures that both precision and length appropriateness are considered in the evaluation.

The BLEU metric is valued for its computational simplicity and interpretability, making it a standard in machine translation and text generation tasks. However, its reliance on exact n-gram matches limits its ability to capture semantic or contextual similarities. Moreover, the brevity penalty may penalize longer candidate texts that are semantically valid, potentially lowering scores despite high quality. These limitations suggest that BLEU is best used alongside other metrics for a comprehensive evaluation.

5.2 Experimental Results

In this section we are going to present the evaluation results for the Whisper and Greek GPT-2, with a focus on the final Whisper-GPT-2 pipeline. Each model was evaluated on its respective test dataset to assess its effectiveness in handling Greek medical speech tasks. The results justify the selection of the pipeline as the final system.

5.2.1 Whisper ASR Performance

The Whisper model was fine-tuned in three configurations (Small, Medium, Large-v2) on a composite dataset comprising Mosel, Mozilla Common Voice 11.0, and Google Fleurs, standardized to 16,000 Hz with an 80/10/10 split [Wan+21; Gai+24; Ard+19; Con+22]. To evaluate the performance of those models we compered the original pre-trained models with those that were fine-tuned on our data. The results are summarized in Table 5.1, showing the Word Error Rate (WER), the normalized Word Error Rate (nWER) and Character Error Rate (CER) across different sizes.

Tab. 5.1: Comparison of Whisper Model Configurations for Greek ASR

Model/Configuration	Model Size	WER (%)	nWER (%)	CER (%)
Original Whisper Small	242M	43.62	36.69	21.61
Original Whisper Medium	764M	34.71	27.21	19.30
Original Whisper Large-v2	1.54B	26.41	18.86	14.55
Fine-tuned Whisper Small	242M	30.31	26.54	13.28
Fine-tuned Whisper Medium	764M	19.45	16.17	8.96
Fine-tuned Whisper Large-v2	1.54B	**14.90**	**12.06**	**8.45**

It becomes visible that the fine-tuning process improves the performance across all model sizes. In particular, Whisper Small reduces the WER from 43.62% to 30.31%, while Whis-

per Medium achieves an even greater improvement from 34.71% to 19.45% and Whisper Large from 26.41% to 14.90%. Similarly, the CER values are nearly halved, confirming the effectiveness of fine-tuning.

In addition to the raw WER, we also report the normalized Word Error Rate (nWER), which is computed on transcripts that have been lowercased, with punctuation removed, and standardized for whitespace. This normalization reduces the impact of surface-level orthographic variations that do not affect clarity, ensuring that the evaluation focuses on real transcription errors rather than formatting differences. As shown in Table 5.1, the nWER values consistently follow the same improvement trends as WER and CER.

Overall, the experiments indicate that fine-tuning on Greek data is a key factor in achieving lower error rates and in improving Whisper's adaptation to the Greek language.

5.2.2 Greek GPT-2 Performance

The Greek GPT-2 model was fine-tuned on a domain-specific corpus designed to adapt the model to the medical context. Since no large-scale Greek medical speech corpus was available, we constructed a training set by combining two complementary types of data: medical texts sourced from books and other written resources, which provided the necessary domain-specific terminology, and transcribed speech data, which exposed the model to the idiomatic structures and variations of spoken Greek. This hybrid approach allowed the model to learn not only specialized vocabulary but also the stylistic and syntactic patterns typical of oral communication, which are highly relevant for correcting automatic speech recognition (ASR) outputs.

The evaluation of the fine-tuned model was conducted using perplexity on both the medical text dataset, the speech transcription dataset, and their combination. Perplexity measures how well a language model predicts a sequence of words, with lower values indicating stronger predictive capacity. Across all three evaluation settings, the fine-tuned model consistently outperformed the original pre-trained Greek GPT-2. More specifically, perplexity was substantially reduced on both medical texts and speech data, confirming that the model successfully learned domain-specific terminology as well as the idiomatic patterns of spoken language. The combined results further highlight the overall effectiveness of the fine-tuning strategy in adapting GPT-2 for error correction in Greek medical ASR (Table 5.2).

Tab. 5.2: Perplexity of Greek GPT-2 (Pre-trained vs Fine-tuned) on Medical and Speech Datasets

Dataset	Pre-trained GPT-2	Fine-tuned GPT-2	Improvement (%)
Medical Texts	45.73	35.36	22.7
Speech Transcriptions	103.21	67.67	34.4
Combined (All Data)	53.15	39.86	25.0

5.2.3 Whisper-GPT-2 Pipeline Performance

The performance of the Whisper-GPT-2 pipeline for Greek ASR was evaluated using WER, CER, and BLEU metrics, as shown in Table 5.3. The evaluation was conducted on the test dataset, which was considered representative of the overall distribution.

Tab. 5.3: Whisper-GPT-2 Pipeline Performance for Greek ASR

Pipeline Configuration	WER (%)	nWER (%)	CER (%)	BLEU (%)
Whisper Small (Baseline)	30.31	26.54	13.27	82.35
Whisper Small + GPT-2 (Reranked)	27.38	23.57	11.80	84.17
Whisper Medium (Baseline)	19.45	16.17	8.96	88.93
Whisper Medium + GPT-2 (Reranked)	18.23	14.86	8.35	89.60
Whisper Large-v2 (Baseline)	14.90	12.06	8.45	92.03
Whisper Large-v2 + GPT-2 (Reranked)	14.69	11.98	8.66	92.06

The pipeline incorporates a re-ranking step in which the Whisper model first generates N candidate transcriptions for each audio segment. Subsequently, the fine-tuned GPT-2 model evaluates these candidates and selects the most probable sentence. We tested several values for N (3, 5, and 8) and found that N=5 provides the best balance between transcription quality and computational efficiency. Increasing N further offered only marginal improvements while significantly increasing computation time, while smaller N values did not allow the GPT-2 model to fully leverage its re-ranking capabilities.

From the results, can be understood that re-ranking consistently improves performance across all tested Whisper model sizes. The WER reduction is approximately 9.66% for the Whisper Small model, 6.27% for the Whisper Medium model, and 1.41% for the Whisper Large-v2 model. Corresponding gains are also observed in CER and BLEU scores, highlighting that re-ranking enhances both word-level accuracy and overall sentence

quality. Looking at the full pipeline compared to the original models, the full Whisper-GPT-2 pipeline achieved WER reductions of 37.23% for Whisper Small, 47.45% for Whisper Medium, and 44.38% for Whisper Large-v2.

Another important consideration is choosing the ideal model for practical usage. In environments where transcription needs to be fast and computational resources are limited, it is important to select the model with the best balance between performance and efficiency. As presented in the Table 5.3 while Whisper Large-v2 achieves a lower WER, it requires significantly more computational resources, making it less practical for routine deployment. Whisper Medium, on the other hand, provides strong performance with substantially reduced WER and CER, while being faster than Large-v2, making it the optimal model for real-world applications.

These findings highlight the effectiveness of combining a strong pre-trained ASR model with a domain-adapted language model. The re-ranking stage is very important especially for medical dictation, where even small improvements in transcription accuracy can be vital for understanding specialized terminology and avoiding critical misinterpretations. Overall, this demonstrates that the Whisper-GPT-2 pipeline is an effective approach for improving transcription accuracy and producing higher quality outputs in Greek medical dictation.

6 Conclusions and Future Work

6.1 Conclusions

The primary objective of this thesis was to develop and evaluate an effective pipeline for Automatic Speech Recognition (ASR) tailored to Greek medical dictation. The proposed Whisper-GPT-2 pipeline integrates fine-tuned Whisper models (Small, Medium, and Large-v2) for transcription with a fine-tuned Greek GPT-2 model for re-ranking, leveraging domain-specific medical texts and transcribed speech data to improve transcription accuracy.

The fine-tuned Whisper models achieved major reductions in Word Error Rate (WER) and Character Error Rate (CER) as shown in (Table 5.1). The Greek GPT-2 model, fine-tuned on a hybrid corpus of medical texts and speech transcriptions, showed reduced perplexity indicating enhanced predictive capability for medical domain-specific language. The integrated Whisper-GPT-2 pipeline further improved performance by incorporating a re-ranking step. This approach yielded consistent gains across all tested model sizes (Table 5.3), highlighting the value of combining a robust ASR model with a domain-adapted language model to address challenges in Greek medical dictation, such as specialized terminology, homophones, and spoken-language variations. The exclusion of the mT5 model from the final pipeline was justified by its underperformance.

A key contribution of this work is the curation of a high-quality speech-to-text dataset, addressing issues in the VoxPopuli subset of the Mosel dataset [Wan+21], such as missing transcriptions and timestamps. This dataset, released openly to promote reproducibility [1], enhanced the pipeline's performance and supports its potential to reduce the workload of medical professionals by improving the accuracy and fluency of clinical documentation. The Whisper-GPT-2 pipelines (Small, Medium, and Large-v2 configurations) are publicly hosted on Hugging Face Spaces [2], where they currently run on CPU. For optimal real-time transcription, deployment on GPU hardware such as an NVIDIA GPU with at least 16GB memory is recommended.

[1] Hugging Face – Vardis/Greek_Mosel
[2] Hugging Face Spaces

6.2 Future Work

The main priority for future research is to incorporate actual medical speech data, which was not available for this study. The current pipeline relied on general speech transcriptions and medical texts rather than real-world clinical dialogue. Incorporating authentic Greek medical speech data, such as recordings from doctor-patient interactions or clinical dictations, is essential for capturing the specialized terminology and acoustic variations in medical environments. This would significantly enhance the pipeline's robustness and accuracy, enabling it to better manage the complexities of real-world medical dictation and reduce errors in critical contexts, such as diagnostic reports or treatment plans. Collecting and annotating such data would be a transformative step toward practical deployment in healthcare environments.

Bibliography

[24] *QTLP Greek CC Corpus for the Medical Domain*. Greek web documents automatically classified as Medical domain and licensed under Creative Commons. Includes genre classification: Reference, News/Journalism, Discussion, Commercial, Other. Version 1.0.0. 2024.

[Ard+19] Rosana Ardila, Megan Branson, Kelly Davis, et al. "Common Voice: A Massively-Multilingual Speech Corpus". In: *arXiv preprint* arXiv:1912.06670 (2019).

[Ard+20] R. Ardila, M. Branson, K. Davis, et al. "Common Voice: A Massively-Multilingual Speech Corpus". In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. 2020, pp. 4211–4215.

[Con+22] Alexis Conneau, Min Ma, Simran Khanuja, et al. "FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech". In: *arXiv preprint arXiv:2205.12446* (2022).

[Gai+24] Marco Gaido, Sara Papi, Luisa Bentivogli, et al. "MOSEL: 950,000 Hours of Speech Data for Open-Source Speech Foundation Model Training on EU Languages". In: *arXiv preprint* arXiv:2410.01036 (2024).

[Hu+22] Edward J. Hu, Yelong Shen, Phillip Wallis, et al. "LoRA: Low-Rank Adaptation of Large Language Models". In: *Proceedings of the 10th International Conference on Learning Representations (ICLR)*. 2022.

[LH19] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations (ICLR)* (2019).

[Pap+02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics. 2002, pp. 311–318.

[Rad+19] Alec Radford, Jeffrey Wu, Rewon Child, et al. *Language Models are Unsupervised Multitask Learners*. OpenAI Blog. Technical report. 2019.

[Rad+23] Alec Radford, Jong Wook Kim, Tao Xu, et al. “Robust Speech Recognition via Large-Scale Weak Supervision”. In: *Proceedings of the 40th International Conference on Machine Learning (ICML)*. 2023.

[Raf+20] Colin Raffel, Noam Shazeer, Adam Roberts, et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *Journal of Machine Learning Research* (2020).

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017, pp. 5998–6008.

[Wan+21] Changhan Wang, Morgane Riviere, Ann Lee, et al. “VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003.

[Xue+21] Linting Xue, Noah Constant, Adam Roberts, et al. “mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021.

[Ιατ15] Γεώργιος Ιατράκης. *Γυναικολογικά προβλήματα και λύσεις*. 2015.

[Σφη+15] Π. Σφηκάκης, Α. Κόκκινος, Μ.Χ. Κυρτσώνη, et al. *Ασκήσεις σημειολογίας και διαφορικής διαγνωστικής στην παθολογία*. 2015.

[Τσι+15] Μάρκος Τσιπούρας, Νικόλαος Γιαννακέας, Ευάγγελος Καρβούνης, and Αλέξανδρος Τζάλλας. *Ιατρική Πληροφορική*. 2015.

List of Acronyms

ASR Automatic Speech Recognition

WER Word Error Rate

BLEU Bilingual Evaluation Understudy

CER Character Error Rate

GPU Graphics Processing Unit

GMM Gaussian Mixture Model

GPT-2 Generative Pre-trained Transformer 2

mT5 Multilingual Text-to-Text Transfer Transformer

HMM Hidden Markov Model

LoRA Low-Rank Adaptation

RNN Recurrent Neural Network

CNN Convolutional Neural Network

LSTM Long Short-Term Memory

BLSTM Bidirectional Long Short-Term Memory

EOS End Of Sequence

List of Figures

List of Tables