Adapting CSM 1B for Greek TTS: A Technical Report by Moira.AI

Moira.AI
moira.ai2024@gmail.com
www.moira-ai.com

September 22, 2025

Abstract

This report documents the process of adapting the CSM 1B conversational speech model for the Greek language. The primary motivation was to address the scarcity of high-quality, open-source Text-to-Speech (TTS) models for Greek. By leveraging Low-Rank Adaptation (LoRA), a parameter-efficient finetuning technique, we successfully retrained the model on a curated dataset of high-quality Greek speech. The resulting model generates highly intelligible and natural-sounding Greek speech with clear articulation. This work not only provides a valuable open-source TTS tool for the Greek-speaking community but also validates LoRA as a resource-efficient and effective method for adapting large speech models to less-resourced languages.

1 Introduction

Text-to-Speech (TTS) technology has made significant advancements[1] [3], yet high-quality, open-source models for less-resourced languages like Greek remain scarce. This scarcity limits the development of localized voice applications and hinders technological accessibility for millions of Greek speakers. To address this gap, Moira.AI initiated this project to create a publicly available, high-fidelity Greek voice by adapting a large, pre-trained foundational model.

Our approach utilizes the powerful CSM 1b conversational speech model [4] as a base. To make this adaptation process both rapid and computationally efficient, we employed Low-Rank Adaptation (LoRA) [2], a parameter-efficient finetuning (PEFT) technique [6]. This method allows for the specialization of massive models with a fraction of the resources typically required, making state-of-the-art speech synthesis more accessible. The result of this work is GreekTTS, a new open-source model, and this report details the methodology and outcomes of our initiative.

2 Methodology

Our approach is centered on the parameter-efficient finetuning of a large, pre-trained foundational speech model. This allowed us to produce a high-quality Greek voice without the prohibitive costs associated with training a model from scratch.

2.1 Base Model: CSM 1b

The foundation of our work is CSM 1b [4], a 1-billion parameter conversational speech model. It is designed to generate highly natural and expressive speech, making it an ideal candidate for adaptation to a new language. The base model is not able to generate coherent Greek at all even though it might have been exposed to it during training by data contamination.

2.2 Finetuning Technique: LoRA

To adapt CSM 1b for the Greek language, we employed Low-Rank Adaptation (LoRA) [2], a prominent parameter-efficient finetuning (PEFT) technique [6]. LoRA operates by freezing the vast number of original model weights and injecting small, trainable low-rank matrices into specific layers. This method dramatically reduces the number of trainable parameters, leading to a significant decrease in GPU memory requirements and training time.

The primary hyperparameters for our LoRA configuration were a **rank (r)** of 32 and an **alpha (α)** of 32. In brief, the rank determines the capacity and number of trainable parameters in the adapter matrices, while the alpha parameter acts as a scaling factor for the learned adaptations. For a comprehensive technical breakdown of these parameters and their interaction, the reader is encouraged to consult the foundational LoRA paper by Hu et al. [2].

2.3 Dataset and Preprocessing

The model was finetuned on a curated dataset of high-quality Greek speech, from diverse settings. A diverse mix of speakers and accents were used to create this proprietary dataset. To ensure data quality and compatibility with the model, a series of standard pre-processing steps were performed. All audio clips were resampled to a uniform sample rate of 24kHz, and the corresponding transcripts were normalized to clean and standardize the text, removing any extraneous characters or formatting.

2.4 Training Procedure

The finetuning process was significantly accelerated using the Unsloth library [5], a high-performance framework that integrates with the Hugging Face ecosystem to optimize memory usage and training speed. We wish to acknowledge the valuable support and guidance provided by the Unsloth team throughout this project.

The training was conducted on a single **GeForce RTX 4090** GPU with 24GB of VRAM. The model was trained for a total of **3 epochs**. To ensure stable training while maintaining a low memory footprint, we configured an **effective batch size of 4** by setting the training batch size to **1** and using **4 gradient accumulation steps**. The initial **learning rate** was set to **2e-4**, adjusted by a **linear scheduler** following a **5-step warmup**. We utilized the memory-efficient **8-bit AdamW optimizer** and enabled **automatic mixed-precision (FP16)** to further accelerate computation.

3 Evaluation and Results

The performance of the finetuned GreekTTS model was assessed by analyzing objective metrics from the training process. The successful convergence of the model is demonstrated by the training and validation loss curves, as shown in Figures 1 and 2. The steady decrease in both loss values indicates that the model was effectively learning the patterns of the Greek language from the training data. Crucially, the validation loss did not show signs of increasing, which suggests that the model generalized well to unseen data without significant overfitting.

3.1 Training Metrics

The successful convergence of the model during the finetuning process is demonstrated by the training and validation loss curves, as shown in Figure 1 and Figure 2 . A continuous improvement is observed in both curves, with a steady decrease in loss values throughout the training run. This indicates that the model was effectively learning the patterns of the Greek language from the training data. Crucially, the validation loss did not show signs of increasing, which suggests that the model generalized well to unseen data without significant overfitting. Furthermore, the validation loss curve had not completely flattened by the final epoch, suggesting that further training on the same dataset could potentially yield even better results.

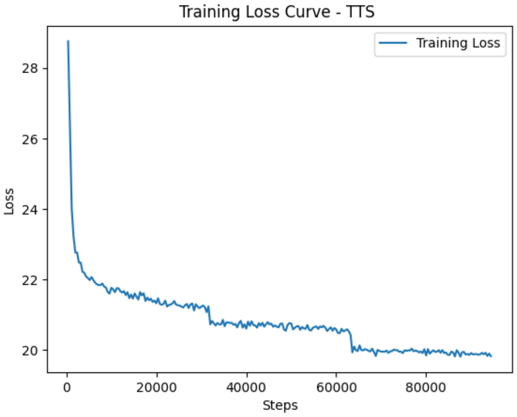

Figure 1: Training loss over 3 epochs

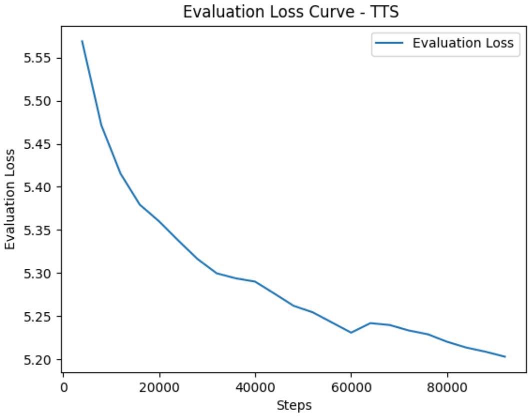

Figure 2: Validation loss over 3 epochs

3.2 Qualitative Monitoring

Alongside the quantitative metrics, qualitative monitoring was conducted throughout the training cycle. At regular intervals, audio samples were generated from checkpoints to informally assess the model's progress. A clear and steady improvement in speech quality was observed as training progressed. Early-stage checkpoints produced less intelligible or robotic-sounding speech, while later checkpoints demonstrated significant gains in clarity, pronunciation accuracy, and overall naturalness. This observational evidence of improvement correlates directly with the decreasing loss values shown in the training metrics.

3.3 Subjective Evaluation

For direct assessment, a collection of generated audio samples is available online. These samples showcase the model's performance across a variety of texts. This includes small clips as well as long-form generations.

- **Audio Samples:** `https://moiraai2024.github.io/GreekTTS-demo/`

3.4 Discussion and Future Work

The results of this project validate our core hypothesis: that parameter-efficient finetuning of a large foundational speech model is a highly effective and resource-efficient strategy for developing high-quality TTS voices for less-resourced languages. The successful training run, evidenced by the steadily decreasing loss curves, confirms that the CSM 1b model has the capacity to adapt to the phonetic and prosodic patterns of Greek.

The choice of LoRA was instrumental in the project's success, allowing for rapid iteration and training on consumer-grade hardware. Furthermore, the use of the Unsloth library provided a significant performance uplift, making the entire process more feasible. The final GreekTTS model, with its clear and intelligible output, stands as a strong proof-of-concept for this methodology. Although it's not without its pitfalls.

It is important to acknowledge the limitations. The 30-hour dataset, while high-quality, is relatively small. This may limit the model's ability to capture a wide range of prosodic diversity, and the resulting voice likely has a singular, consistent style reflective of the training data.

3.5 Future Work

Building upon the success of this initial project, several avenues for future work could further enhance the GreekTTS model and its utility for the Greek-speaking community.

- **Dataset Expansion:** The most impactful next step would be to expand the training dataset. Incorporating more hours of speech from a wider variety of speakers would help to improve the model's naturalness, prosody, and ability to generalize.
- **Hyperparameter Optimization:** While the chosen LoRA parameters (r=32, α=32) yielded excellent results, a systematic exploration of different ranks and alpha values could lead to further improvements in voice quality or even smaller model adapter sizes.
- **Adding Expressive Capabilities:** The current model produces speech with a neutral, narrative tone. A future project could focus on finetuning the model on a dataset containing emotional speech to enable the generation of more expressive and context-aware audio.
- **Exploring Alternative Foundational Models:** This project has provided us with a robust Greek base dataset and the expertise required to rapidly adapt high-performance TTS models. Moira.ai is committed to monitoring future open-source releases and leverage them to provide state-of-the-art (SOTA) Text-to-Speech capabilities for the Greek language.

4 Conclusion

In this project we set out to address the scarcity of high-quality, open-source Text-to-Speech models for the Greek language. By employing a parameter-efficient finetuning approach, we successfully adapted the CSM 1b foundational model using Low-Rank Adaptation (LoRA). The process, accelerated by the Unsloth library, proved to be a resource-efficient and highly effective method for this task. The result of this work is GreekTTS, a new open-source model that generates clear and intelligible Greek speech. This project not only delivers a valuable tool to the Greek-speaking community but also serves as a successful case study for the rapid development of speech technologies for less-resourced languages. Moira.ai remains committed to building upon this foundation and will continue to advance the state-of-the-art for Greek speech synthesis.

References

[1] hexgrad. Kokoro-tts. https://huggingface.co/hexgrad/kokoro-tts, 2024. Accessed: 2025-09-21.

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[3] Resemble AI. Chatterbox Multilingual: Open Source TTS for 23 Languages. https://www.resemble.ai/introducing-chatterbox-multilingual-open-source-tts-for-23-languages/, September 2025. Accessed: 2025-09-21.

[4] SesameAILabs. CSM (Conversational Speech Model) 1b. https://github.com/SesameAILabs/csm, 2025. Accessed: 2025-09-21.

[5] Unsloth AI. Unsloth: 5x faster, 70% less memory finetuning. https://github.com/unsloth/unsloth, 2024. Accessed: 2025-09-21.

[6] Lingling Xu and et al. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*, 2023.