# Improving Naturalness of Neural-based TTS system Trained with Limited Data

Layan Sawalha and Mohammed Salah Al-Radhi

## Introduction

- Study the process of generating speech waveform from textual input, called Text-To-Speech (TTS) synthesis, for artificially creating human voices.
- Develop synthetic speech that closely resembles human speech in multiple languages using limited data.
- Examine two TTS synthesis technologies and compare their approaches for multi-language support.

## Motivation

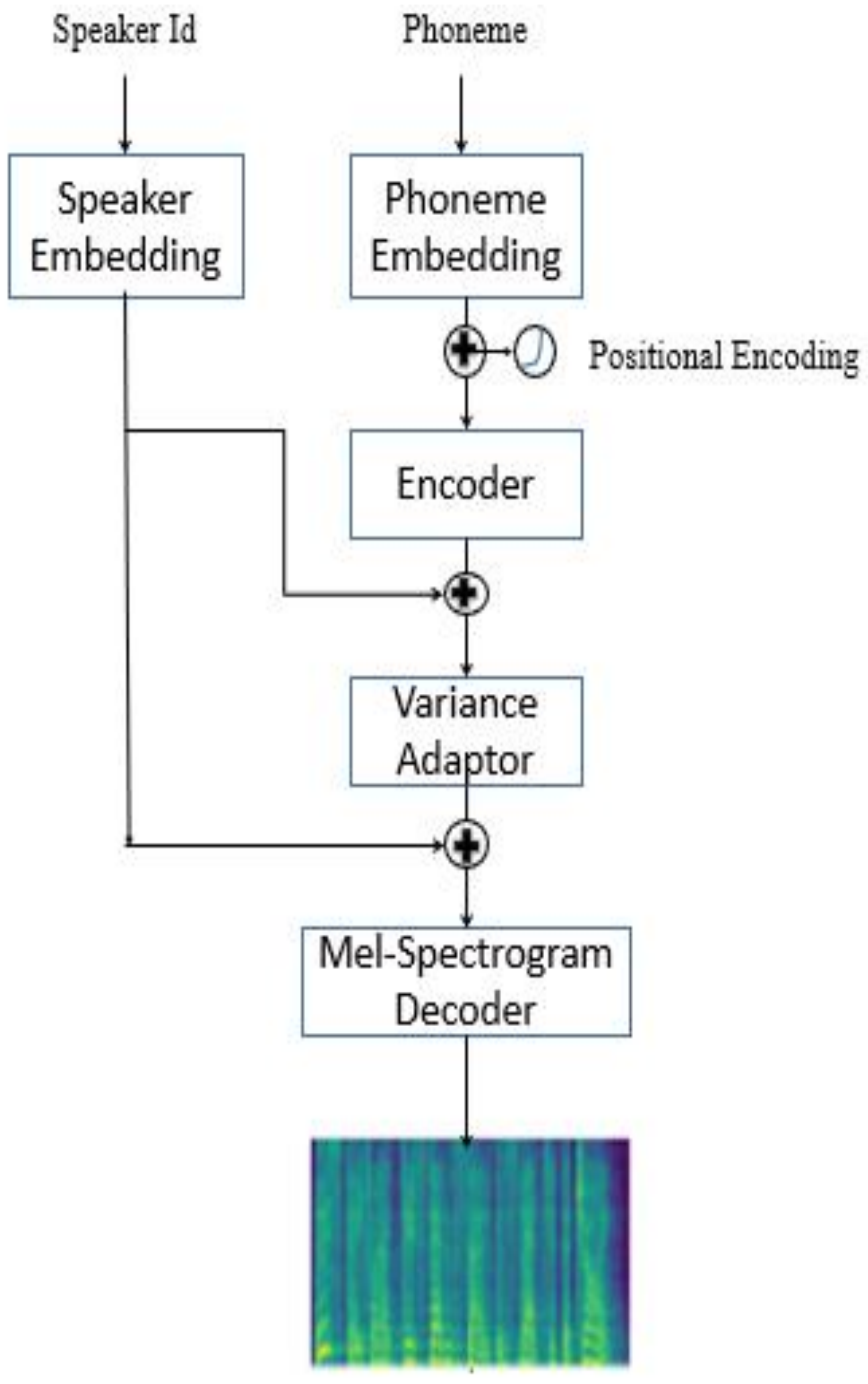
- Investigate two vocoders for different approaches using multiple datasets.
- Continuous vocoding overcomes discontinuity in speech parameters and reduces computational complexity.
- Ahocoder provides accurate, high-quality speech synthesis, ideal for speech manipulation.
- FastSpeech2, a non-autoregressive TTS, focuses on extracting pitch, energy, and duration for training and inference.
- Implementing a huge dataset in multiple languages can be a barrier, so a fast multi-language TTS with limited data is introduced.
- Overall, we explored different speech synthesis approaches to produce non-robotic, human-like speech for multiple datasets.

## Methodology

Three vocoders were implemented based on speech synthesis:

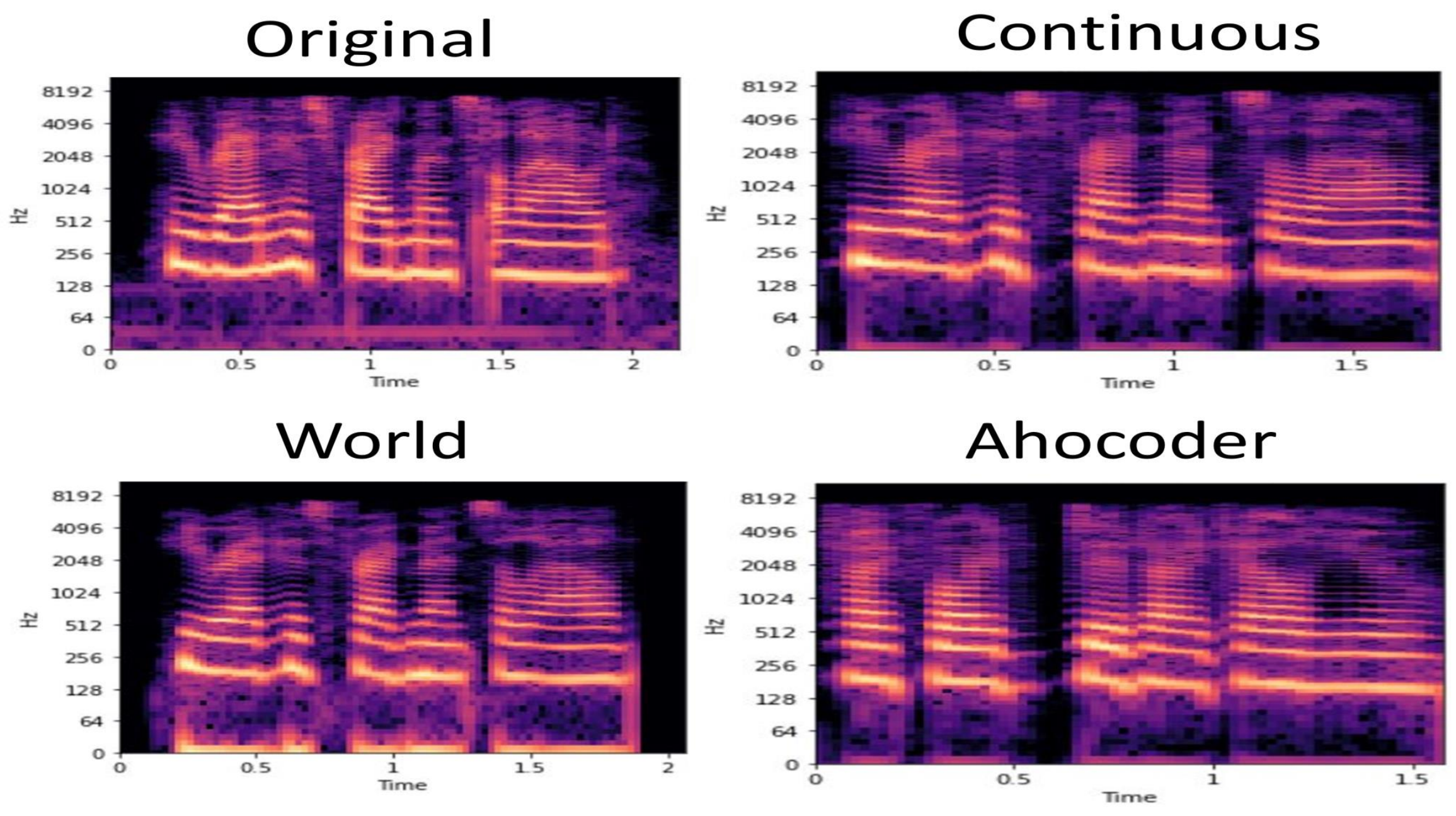
- ➢ World vocoder.
- ➢ Continuous vocoder.
- ➢ Ahocoder vocoder.

- FastSpeech2 improves training pipeline and eliminates information loss.
- Model successfully applied in both English and Arabic languages.
- High-quality and natural speech synthesis achieved using less than half of original dataset.

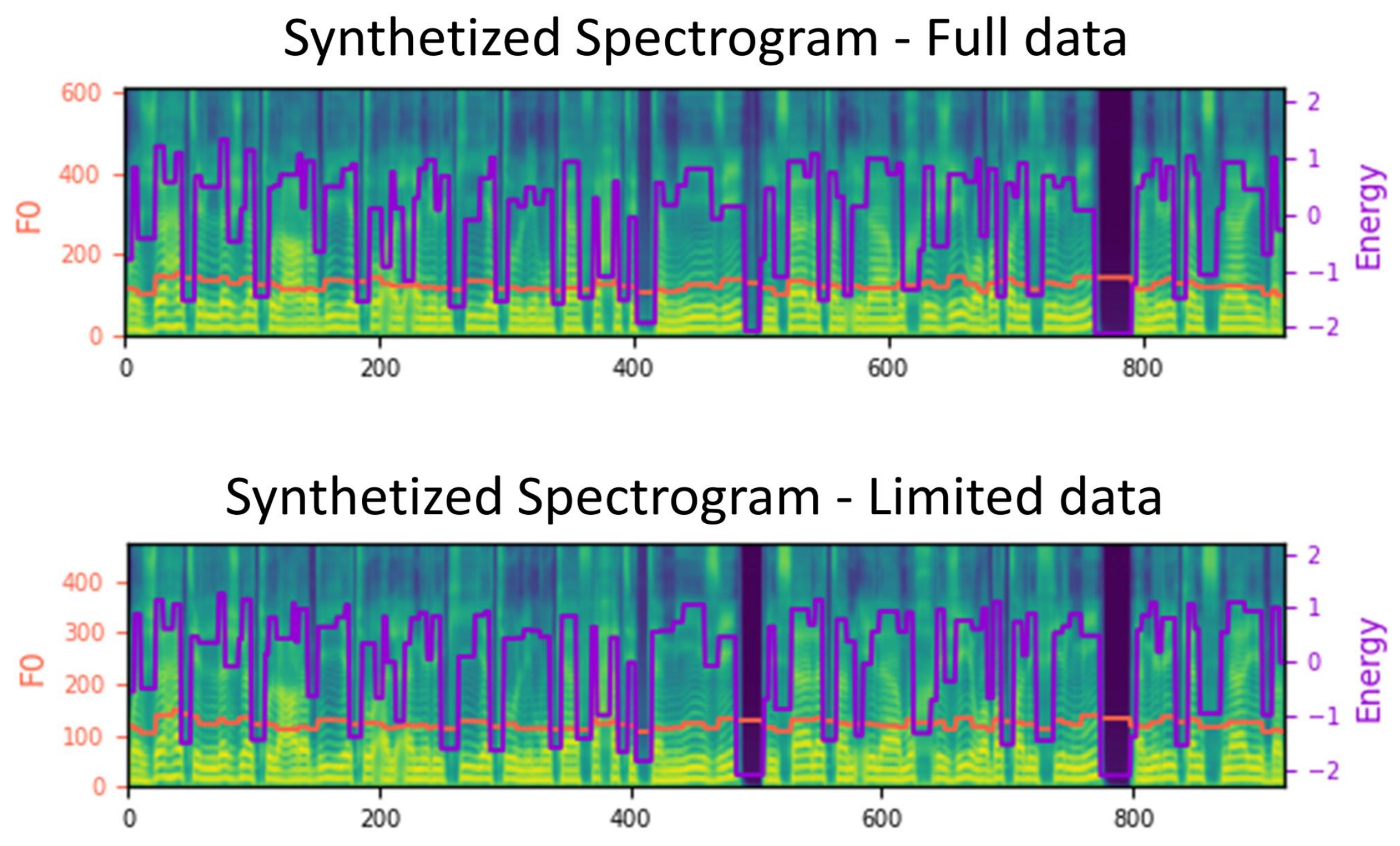


## Results

- The main goal was to integrate the Ahocoder as well as the continuous vocoder into the Merlin-based TTS toolkit.
- Continuous vocoder reduces alignment error by not requiring voiced/unvoiced decision. Ahocoder provides accurate, high-quality speech synthesis for manipulation and transformation.
- Continuous vocoder performs better than WORLD vocoder and Ahocoder in most cases. Ahocoder results in slightly higher scores than WORLD.



- Started with a FastSpeech2 baseline in English, then expanded to include Arabic language.
- Utilized less than half of original dataset while preserving high-quality speech synthesis, enabling the creation of a system that can generate speech with minimal data for a variety of languages.
- This approach provides a flexible and scalable solution for training and generating speech with limited data in multiple languages.



## Conclusions

- Investigated 2 significant TTS synthesis techniques for optimal voice quality.
- Evaluated World, Continuous, and Ahocoder vocoders, concluding Continuous vocoder produced best speech quality.
- Implemented FastSpeech2 for high-quality non-autoregressive TTS and integrated a different language using limited data while maintaining quality.

## Future Work

- Improving speech synthesis in underrepresented languages/accents.
- Adding realistic prosody/intonation to mimic human speech.
- Controlling specific aspects of speech synthesis (pitch, speed, emphasis).

Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics
Budapest, Hungary