MŰEGYETEM 1782

Implementing a Text-to-Speech Synthesis model on a Raspberry Pi for Industrial Applications

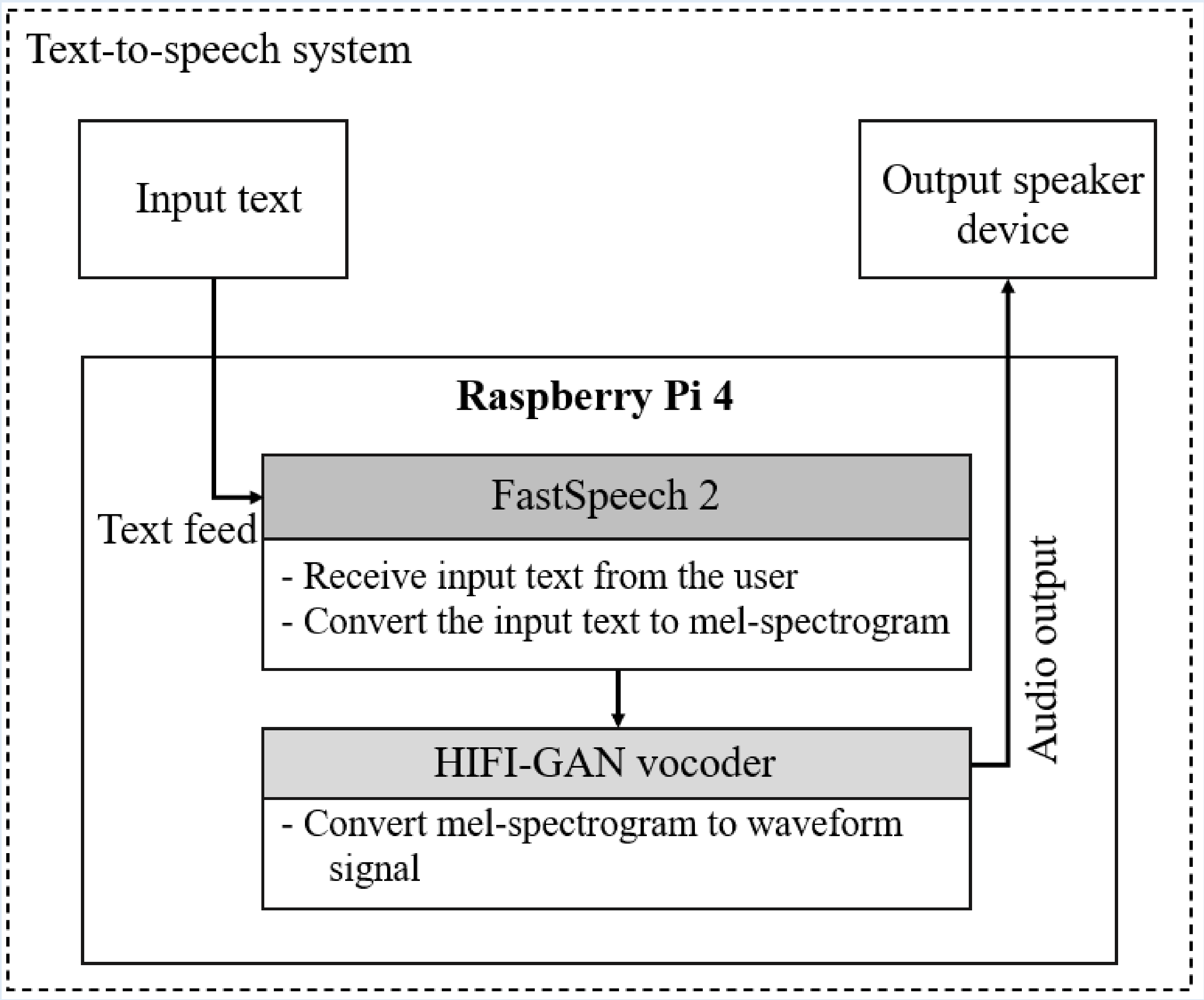
Ali Raheem Mandeel[1], Ammar Abdullah Aggar[2], Mohammed Salah Al-Radhi[1], Tamás Gábor Csapó[1]
[1]Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Hungary,
[2]Department of Computer Engineering, Ministry of Education, Iraq

1. Introduction

- ❖ Text-to-Speech (TTS) technology produces human-like speech from text.
- ❖ End-to-end TTS models produce highly natural synthesized speech but require extremely high computational resources.
- ❖ Deploying such high-quality TTS models in a real-time environment has been a challenging problem due to the limited resources of embedding systems,
- ❖ Our proposed model could be used in many real-life applications such as railway announcements and industrial purposes.

2. Methods

We built the overall system, as illustrated in Figure 1. We used the open-source pre-trained FastSpeech2 model and HiFi-GAN vocoder (V1) implemented on Raspberry Pi4 (RTF). Also, we did objective experiments to test the effectiveness of our model and compared the results to the same TTS model on Titan X GPU .



Text-to-speech system

Input text

Output speaker device

Raspberry Pi 4

Text feed

FastSpeech 2

- Receive input text from the user
- Convert the input text to mel-spectrogram

HIFI-GAN vocoder

- Convert mel-spectrogram to waveform signal

Audio output

Fig. 1. System Overview

3. Results

3.1 Computational Cost: The GPU-based TTS model is **1.9**, while the RTF of the RPi-based TTS model is **8.93**. The average run time of The GPU-based TTS model is (**10 sec**). In contrast, the average run time of the RPi-based TTS model is (**42-67 sec**) see Figure 2.

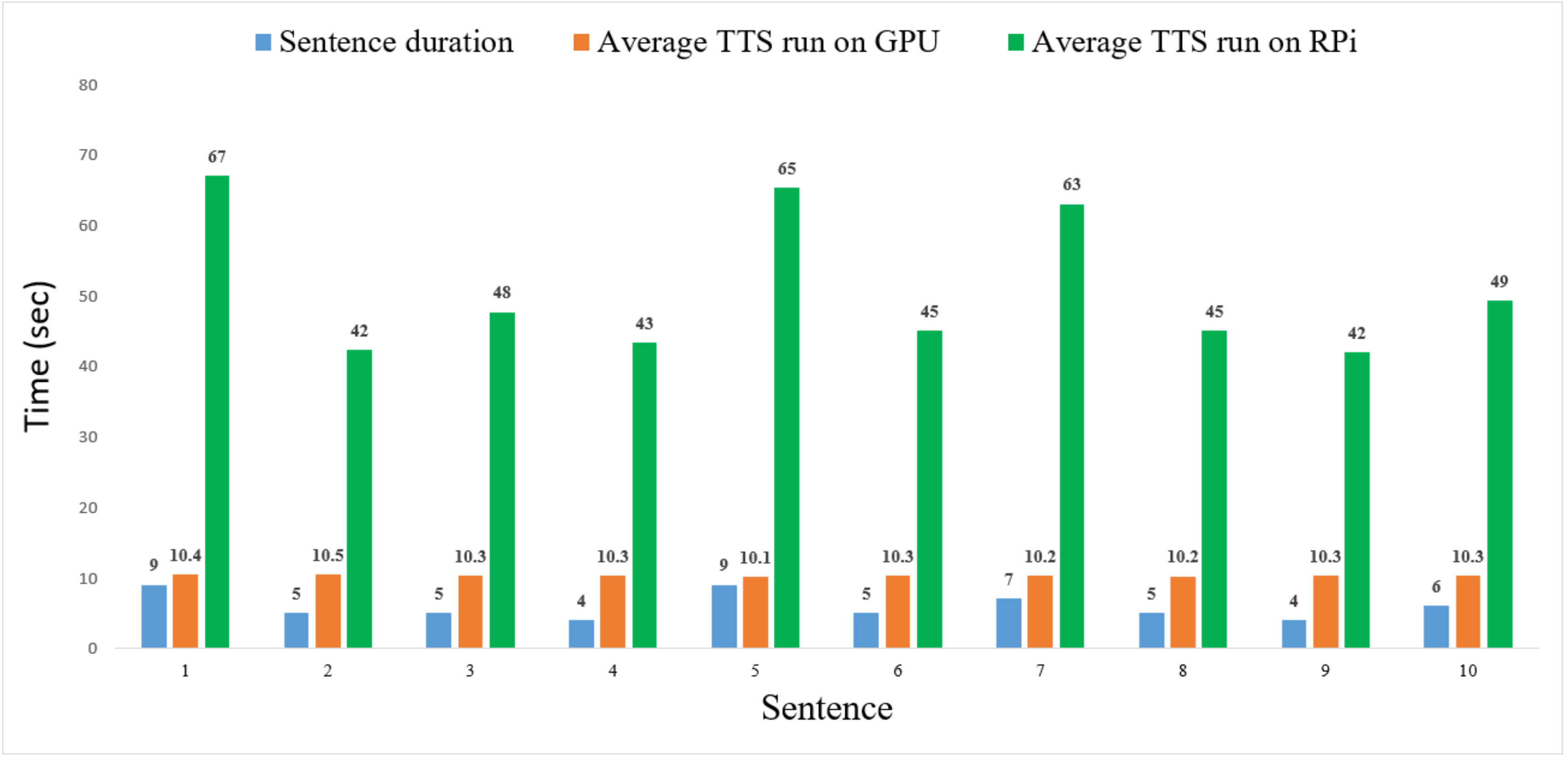


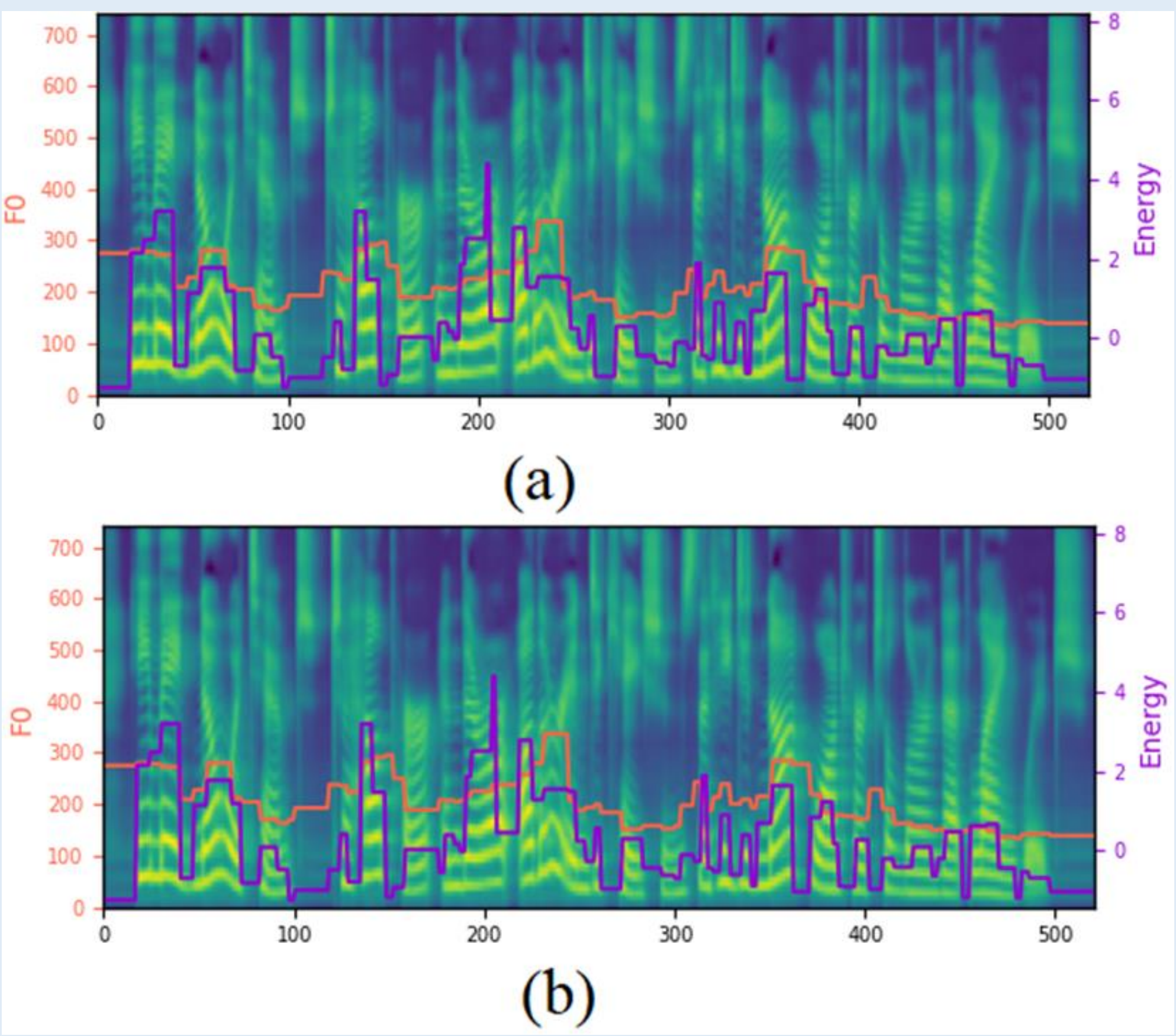
Fig. 2. The average run time for the two models.

3.2 Objective Evaluation: Normalized Covariance Metric (NCM) and frequency-weighted segmental SNR (fwSNRseg) were the same for the two models (Table 1). It means the RPi-based TTS model maintains the same quality of output speech.

Table 1. The objective metrices for the tow models.

System type	The objective metrics	
	NCM	*fwSNRseg*
RPi TTS model	0.03	0.857
GPU TTS model	0.03	0.857

3.3 Demonstration sample (spectrogram) : Also, no differences in the spectrograms of the two models (Figure 3).



(a)

(b)

Fig. 3.

(a) The synthesized speech of the RPi TTS model

(b) The synthesized speech of the GPU TTS model

4. Conclusions

- ❖ FastSpeech 2 and HiFiGAN (TTS model) was implemented on Raspberry Pi4,
- ❖ The proposed system's performance is adequate for standard audio-enabled industrial applications,
- ❖ The TTS model computation needs to be improved further and footprint size should be reduced.

5. Future work

We are looking forward to receiving other future work ideas or potential applications from visitors of the WINS 2023 workshop.

WINS 2023 workshop-02/07/2023