Proposing Ideas for Building a Vietnamese Text-to-Speech (TTS) Model

1. Input and Output

Input: Vietnamese character sequence

<u>Ex</u>: "Xin chào, tôi là một hệ thống tổng hợp giọng nói."

Output: An audio file (.wav, .mp3) containing the voice reading the input sentence

<u>Ex</u>:

🔊 "Xin chào, tôi là một hệ
thống tổng hợp giọng nói."

2. Pipeline

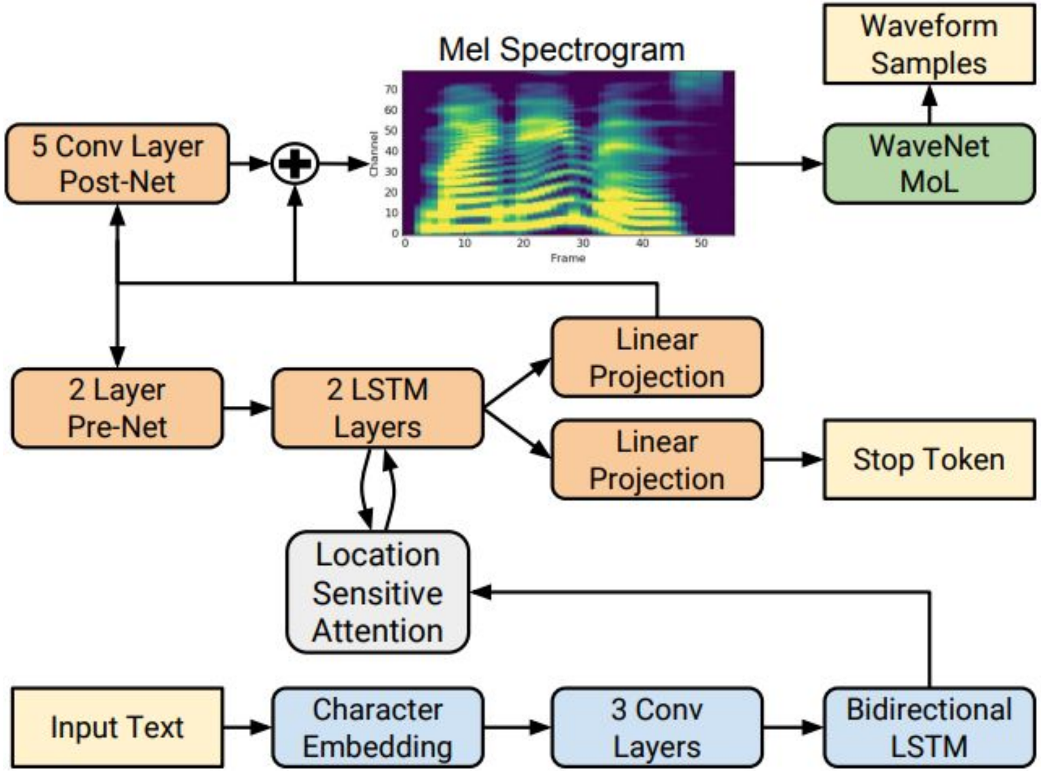


Fig. 1. Block diagram of the Tacotron 2 system architecture.

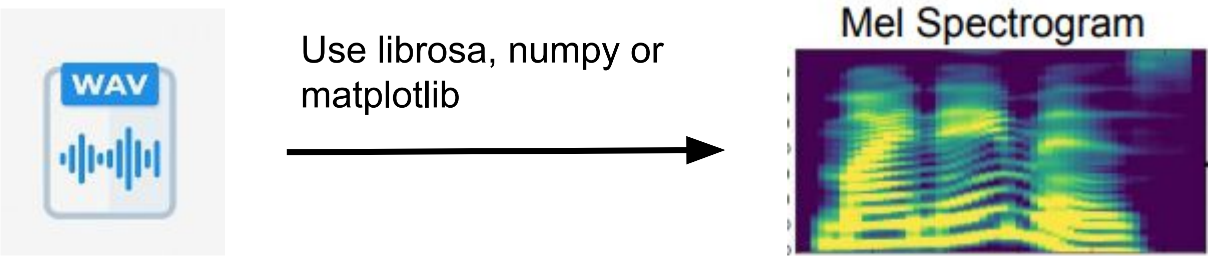
2. Pipeline

- **Data Preparation**: Dataset consists of pairs of (text sequences, audio files).

Ex:

Text sequences (.txt)	Audio files (.wav)
Xin chào, tôi là hệ thống TTS.	xin_chao_toi_la_he_thong_tts.wav
Tôi đang học về công nghệ giọng nói.	toi_dang_hoc_ve_cong_nghe_giong_noi.wav

- **Audio Processing**: Convert audio files into Mel-spectrograms to create a frequency representation of the sound for training Mel-spectrogram Prediction

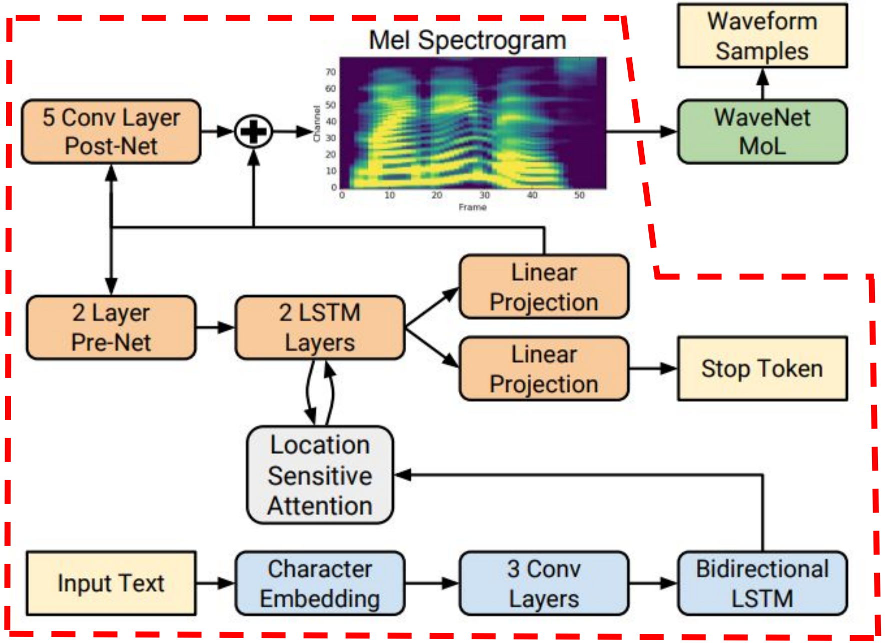


Use librosa, numpy or matplotlib

2. Pipeline

Model Training

Mel-spectrogram Prediction Model: Train a model to predict Mel-spectrograms from text sequences.



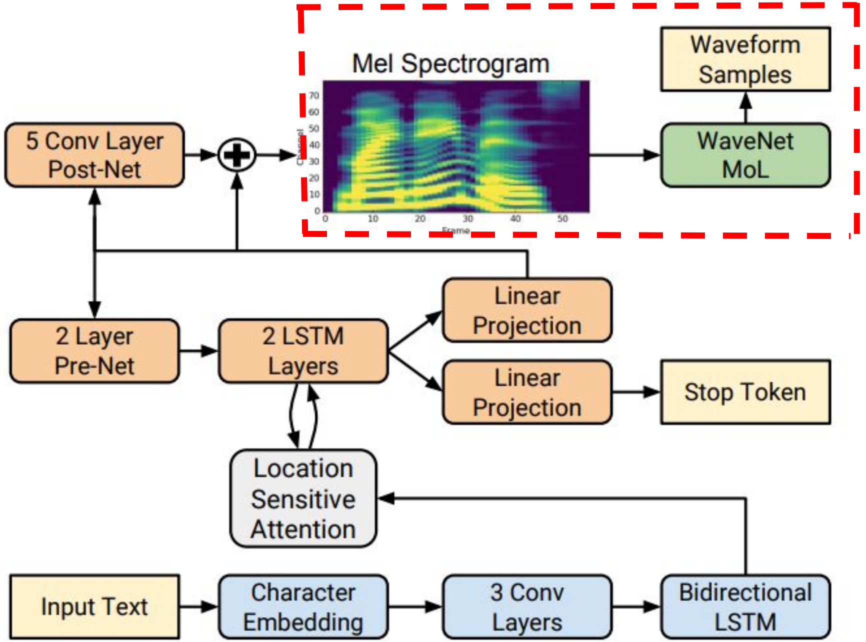
Mel-spectrogram Prediction Model

Fig. 1. Block diagram of the Tacotron 2 system architecture.

2. Pipeline

Model Training

Vocoding Model: Train a vocoder model to convert Mel-spectrograms into time-domain audio waveforms.

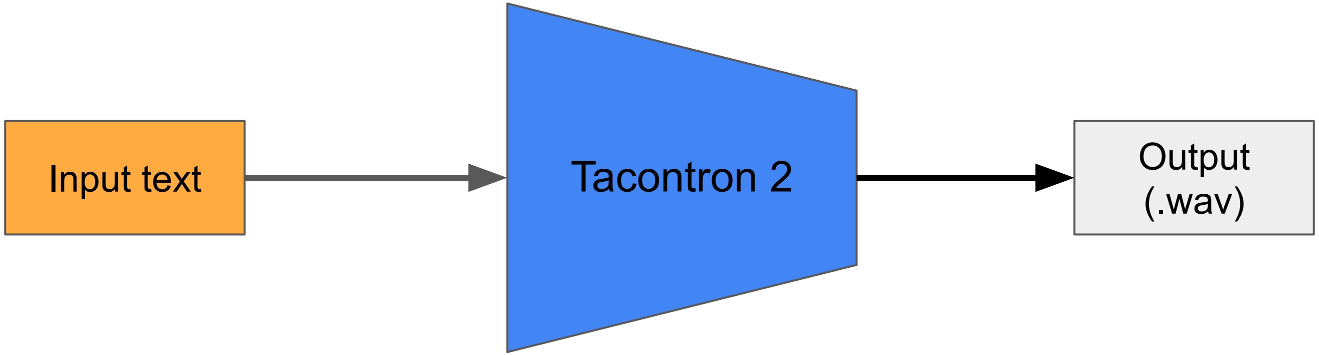


Vocoding Model

Fig. 1. Block diagram of the Tacotron 2 system architecture.

2. Pipeline

Speech Synthesis: Use the trained models to synthesize audio from input text sequences.



3. Problems and Solutions

- **Diverse Dialects and Accents**: Vietnamese has many regional dialects and accents. To address this, you can collect data from various voices and train the model to handle different accents.
- **Audio Quality**: Ensure high-quality audio data and use noise reduction and audio enhancement methods during the preprocessing stage.
- **Autoregressive**: Tacotron 2 is an autoregressive model, which results in slow processing. This can be improved by using non-autoregressive models like FastSpeech.

References

1. Tacotron 2
2. Tìm hiểu 1 số mô hình Text-To-Speech