Voice-Preserving Speech Machine Translation

List of expected participants:

- Nguyen Ngoc Bang Tam: bangtamnguyenn@gmail.com
- Do Tri Nhan: dotrinhan99@gmail.com
- Nguyen Ngoc Minh Khanh: nguyenngocminhkhanh4999@gmail.com

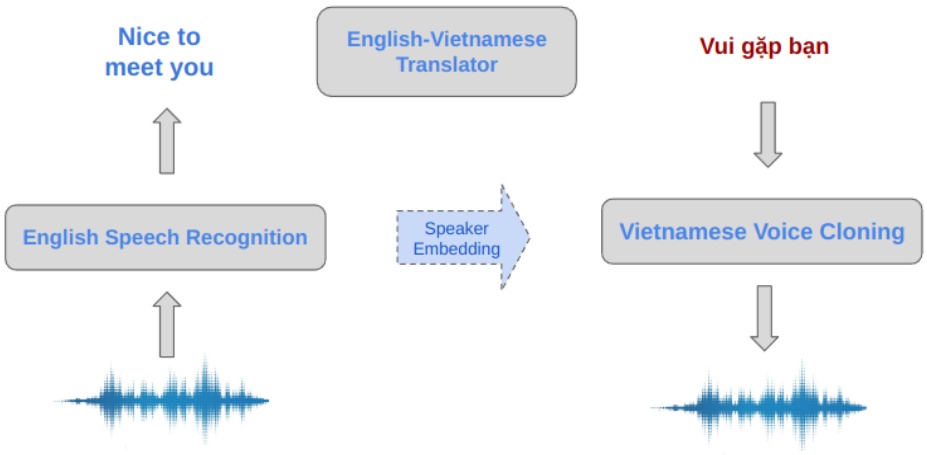
Keywords: Text to Speech, Automatic Speech Recognition, Machine Translation, Voice Cloning

Definition

Speech Translation for Low Resource Language with Voice I/O and Preserve the Characteristics of the Voice Input

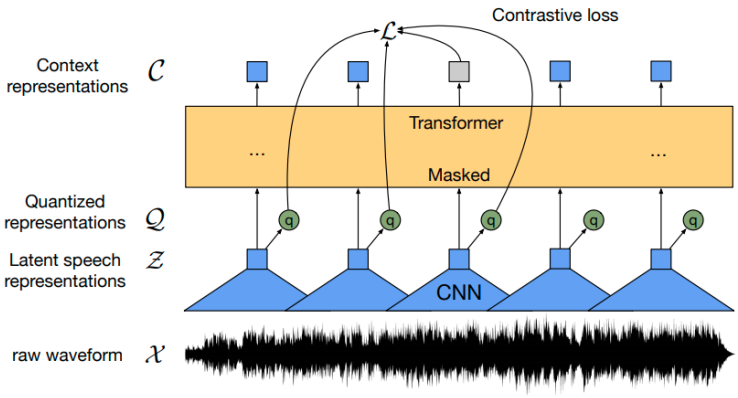
Abstract

The language barrier in spoken communication with locals is one of the significant challenges that foreign tourists encounter when traveling to Vietnam. Vietnamese is a tonal language with 6 different tones that dictate the meaning of a word, making it difficult for non-native speakers to quickly absorb in a short amount of time (especially English speakers). We would like to propose a prototype of a cross-lingual voice translator application, which allows English-speaking travelers to order food, ask for directions, and communicate with locals by tapping their phones. Simultaneously, local vendors can also benefit from our system as they can provide services to foreign tourists without dealing with unwanted language barriers. The proposed system consists of 3 core modules as illustrated below: (1) ASR to transcribe English audios to corresponding texts, (2) English-Vietnamese machine translation and (3) voice cloning-based TTS which can generate Vietnamese-spoken audios using the original speaker's voice. Our main contribution is the voice cloning-based TTS model for Vietnamese, which incorporates the speaker embedding of English input audio to synthesize Vietnamese-spoken speeches with the voice of original speakers.



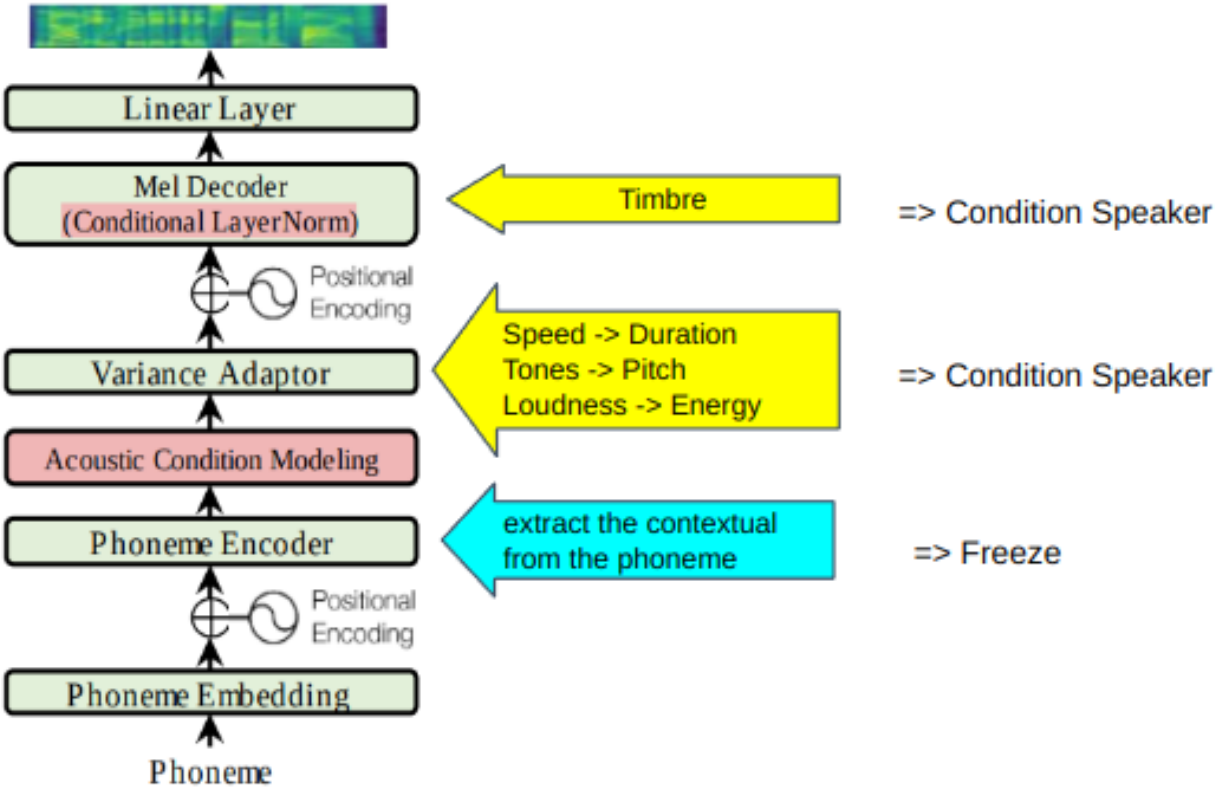
Method

English Automatic Speech Recognition: we intend to use the powerful large pretrained wav2vec.

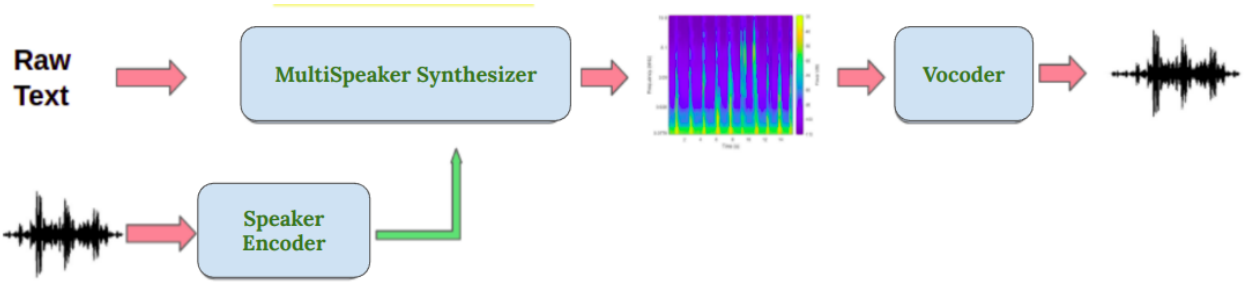


English-Vietnamese machine translation: we are going to use MTet: Multi-domain Translation for English-Vietnamese which uses a larger and better quality Machine Translation dataset.

Vietnamese Voice Cloning module: we build the model multi-speaker speech synthesis based on Fastspeech2 then inject the speaker embedding of English speakers to convert the voice output.



Multispeaker Fastspeech2



Voice Cloning Speech Synthesis

Dataset:

Speech to text: https://huggingface.co/tasks/automatic-speech-recognition
Machine Translation: https://huggingface.co/spaces/VietAI/MTet_Translation
Voice Cloning Text to speech for Vietnamese: https://huggingface.co/datasets/vivos

Expected Results

Our system can enable any non-Vietnamese-speaking person to “speak” Vietnamese with their own voice, and we aspire to demonstrate the system in the form of a cross-lingual voice translator application.

In the long term, we envision developing a real-time system for our demonstration, then providing a high-performance API to handle multiple concurrent requests.

Our Initial Demo: