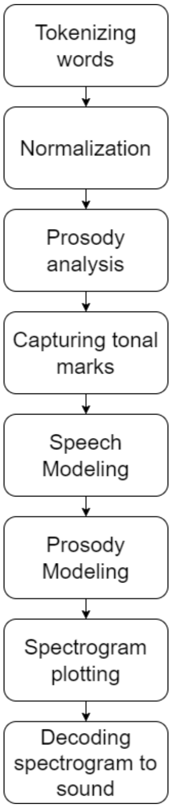
REPORT

Building a TTS model for Vietnamese language is a challenge. A lot of things must be considered, including the innotation and the accent marks on our words. Another task is to find the accent that when we use, it can be mutually understood by the Vietnamese.

Nevertheless, processing text in Latin language is an advantage that we receive since it does not require a separate dataset like in Japan, Korea, or China where the language system is semantically hard and challenging as well.

The pipeline that I propose that would make the processing more understandable:



Pipeline

Vietnamese is a tonal language hench the use of sounds and innotation in speech always happen. We first tokenize the words into small chunks of words or just a word only. Then we conduct the normalization task, in this case, we prioritize words that have signs or marks on them. We also need to identify the emphasis of a sentence and other properties such as bullet points, exclamation mark and so on.

The following step is to match the tonal marks to the correct word. We all know that when we apply a tone to a vowel, a new sound is made. However, some tones when we put them on, it is phonetically unpronounceable. That is why we must have a set of rules relating to using the correct ones to give the best results grammatically. Following that, we will build a model that focuses on the relationship between speech and linguistic features. The idea is that we want to generate a speech that is as close as to a human being in terms of sounds and coherence.

Another task we need to look at is prosody modeling. Since a sentence can have many meanings depending on how a person speaks. We must assess the “emotion” or topic of the sentence so that when the speech is taken place, it should sound at least smoothly and logical to the initial meaning of the sentence.

After finishing that, based on the result of the prosody model, we plot a graph resembles the sounds (spectrogram) to generate a speech waveform. Finally, we use a model to generate voice based on the waveform we produce.

Challenges and Solutions

The first and foremost challenge is collecting data. Such a big amount of speech data is hard to source and even if we have one, we are unsure of the variety of the speech. Speaking of the variety, Vietnamese language has many accents which differs within regions. To be able to tackle the problem is another task after we successfully gather the data.

A linguistic resource helps a lot in text analysis and as far as I know, the resources are still scarce and scietists are working on a project to build a huge resource to train.

We also have to understand the difficulties of training models. A solution might be using pre-trained models and fine-tune them. PhoBERT are pre-trained language models for Vietnamese is current the SOTA.

A solution is to include pitch contours and duration for natural-sounding speech synthesis. Incorporating linguistic and contextual cues can enhance prosody modeling.