

VIETNAMESE-TO-ENGLISH MUSIC CONVERSION SYSTEM

Nguyen Thien Bao¹

¹ University of Information Technology,
Vietnam National University, Ho Chi Minh City, VietNam

What ?

We propose a system to convert Vietnamese songs into English versions:

- From the original Vietnamese track, the system will perform source separation, lyric translation, and singing voice synthesis to create an English version of the song.
- It serves as a powerful support tool for language learners and content creators.

Why ?

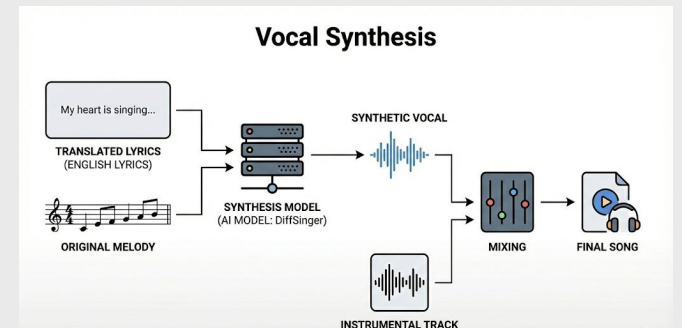
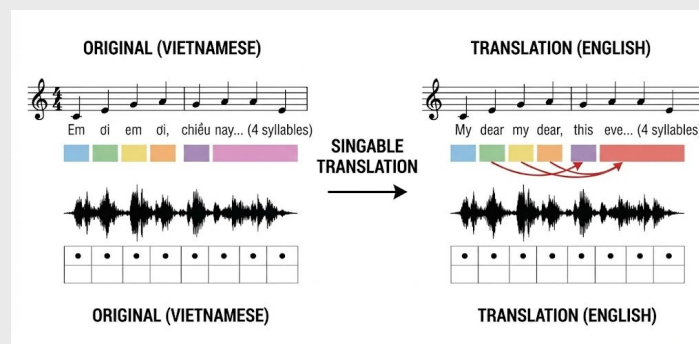
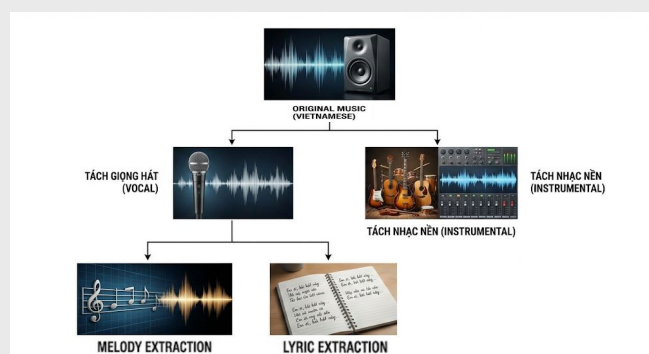
- Music serves as a cultural bridge and a powerful tool in language learning. Currently, many English learners want to listen to their favorite Vietnamese songs in English to both enjoy entertainment and learn vocabulary and intonation.
- The Vietnamese language has not yet been widely researched, especially in the field of musical translation.

Overview

Vocal - Instrumental

Lyrics Translation

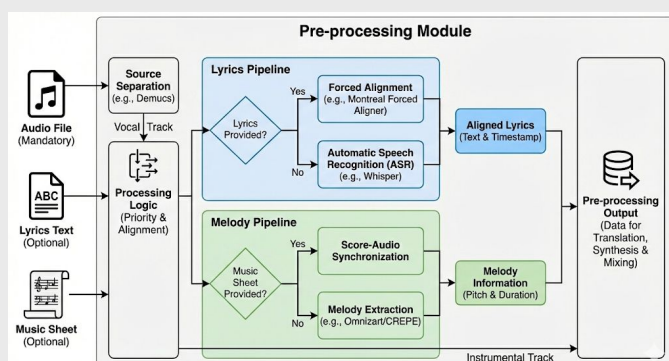
Singing Voice Synthesis



Description

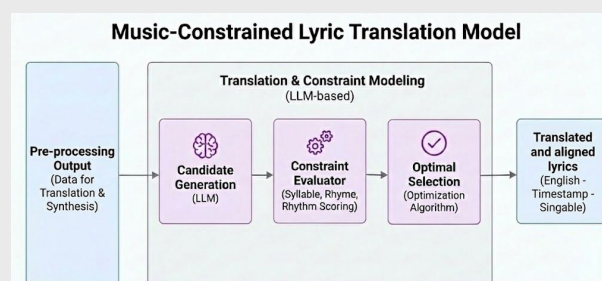
1. Input Pre-processing System

- Use an audio source separation model (e.g., Demucs) to separate the Vocal and Instrumental tracks from the original audio file.
- Apply a speech recognition model (e.g., Whisper) to convert Vocals into text (Lyrics) along with timestamp information.
- Use melody extraction algorithms (e.g., Omnizart or CREPE) to obtain pitch and duration information.
- Hybrid Method: Allows users to upload a Music Sheet or standard Lyrics to increase the accuracy of the input data.



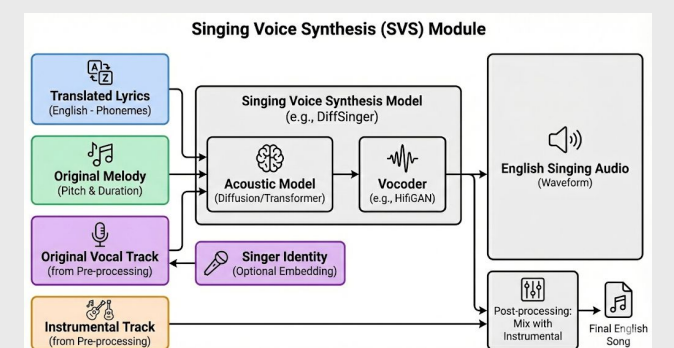
2. Music-Constrained Lyric Translation

- Use Large Language Models (LLM) to generate multiple translation candidates for each line of the song.
- Apply optimization algorithms to select the best translation based on Reward Functions for:
 - Syllable count
 - Rhyme & Rhythm
 - Meaning



3. Singing Voice Synthesis Module

- Build an Acoustic Model (e.g., based on Diffusion or Transformer architectures) to predict the Mel-spectrogram from the English lyrics and original melody information.
- Use a Vocoder (e.g., HiFiGAN) to convert the Mel-spectrogram into an audio Waveform.



4. Post-processing & Integration

- Merge the newly created English vocals with the instrumental track separated during pre-processing to create the final product. Integrate all processing steps into a complete, unified system.