Multilingual Speech Synthesis and Cross-lingual Voice Cloning for Personalized Applications

Aye Mya Hlaing, Win Pa Pa

Natural Language Processing Lab.,
University of Computer Studies, Yangon, Myanmar

Multilingual Speech Synthesis and Cross-lingual Voice Cloning for Personalized Applications

- *Multilingual speech synthesis* refers to the generation of human-like speech across multiple languages using a single multilingual model.
- In parallel with multilingual speech synthesis, *cross-lingual voice cloning* is also evolved at aiming to synthesize a user's voice in different languages while retaining the unique characteristics of the original voice.
- This allows for a personalized and consistent vocal identity across multiple languages.
- Focus on a multilingual text-to-speech (TTS) and cross-lingual voice cloning model for ASEAN languages that can be applied in personalized applications.

Problem Statement (1)

- Unlike resource-rich languages, most ASEAN languages are not commonly included in the benchmark datasets
- There are a few datasets such as FLURES[1], OpenSLR[2] that include Myanmar language speech data with paired transcribed text.
- The quality of TTS systems typically depends on the high-quality studio recordings of phonetically balanced utterances.
- There is no high-quality benchmark speech dataset for ASEAN languages to develop the multilingual speech synthesis that can generate production-quality natural sounding speech in ASEAN languages

[1] https://huggingface.co/datasets/google/fleurs

[2] https://openslr.org/80/

Problem Statement (2)

- Most current researches focus on zero-shot and few-shot cross-lingual voice cloning technologies for TTS that can replicate a speaker's unique vocal characteristics in different languages by using just little or few second audio clip.
- Research on this trend in ASEAN languages is relatively limited.
- Very few work was found, only on Indonesian and Vietnamese languages[1]

[1]Tran, C., Luong, C.M. and Sakti, S., 2023. STEN-TTS: Improving Zero-shot Cross-Lingual Transfer for Multi-Lingual TTS with Style-Enhanced Normalization Diffusion Framework. In *Proc. INTERSPEECH* (Vol. 2023, pp. 4464-4468).

Proposed Method

- To build **a high-quality multilingual speech dataset with transcribed text** for ASEAN languages especially designed for multilingual speech synthesis
  - ✓ parallel speech dataset in ASEAN languages built on parallel text
  - ✓ phoneme coverage of each language
  - ✓ recorded by multiple speakers
  - ✓ gender balance
- To apply **Unicode bytes representation**[1] as the shared representation in multilingual model to scale languages with large vocabularies
  - ✓ eliminate the need of grapheme-to-phoneme converters for each language

[1]Li, B., Zhang, Y., Sainath, T., Wu, Y. and Chan, W., 2019, May. Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes. In ICASSP 2019, (pp. 5621-5625). IEEE.

Proposed Method

- To train **multilingual speech synthesis model** by applying **VITS[1] architecture** with some modifications of speaker embedding and language embedding techniques
- To train **cross-lingual voice cloning model** that enables the cross-lingual voice transfer among ASEAN languages
- To develop **a website which provides multilingual Text-to-Speech services** for ASEAN languages and **lets the user clone his/her voices into different languages**

[1]Kim, J., Kong, J. and Son, J., 2021, July. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In International Conference on Machine Learning (pp. 5530-5540). PMLR.

Impact: Scientific and technological

- **Multilingual speech synthesis and cross-lingual voice cloning model** that allows individuals to synthesize speech and maintain their unique vocal identity across ASEAN languages
- **A benchmark speech dataset on ASEAN languages** especially for speech synthesis that can promote the speech processing research of ASEAN languages
- **Encourage personalized applications** in virtual assistants, language learning and accessibility tools especially in the context of ASEAN languages

Impact: Societal

- People with speech disorders can communicate more easily without losing their unique voice
- The user's satisfaction will increase when they get the synthesized voice that is similar to their voice
- Cross-lingual voice cloning facilitates applications in media, entertainment, and marketing, allowing voice actors to perform in multiple ASEAN languages without losing voice consistency.

Impact: Collaborative

- Enhancing communication in ASEAN regions
- Sharing datasets, computational resources, technologies, ideas across ASEAN partnerships
- Accessibility of larger datasets and computational resources is crucial for developing and fine-tuning multilingual models

Output/Outcome

- A multilingual text-to-speech with the ability of cross-lingual voice cloning for ASEAN languages that can be applied in personalized applications
- Training strategies and parameters that are more suitable for multilingual speech synthesis and cross-lingual voice cloning tasks for ASEAN languages.
- A more suitable representation scheme for each character in ASEAN languages
- A high-quality benchmark ASEAN speech dataset with paired transcribed text
- International Papers and Journals with the collaboration of colleagues from ASEAN countries

Conclusion

- Building a high-quality speech dataset for ASEAN languages that can be used in many speech processing researches such as Speech Synthesis, Automatic Speech Recognition, Speech to Speech Translation, etc.
- Modeling a multilingual TTS with cross-lingual voice cloning capability by applying state-of-the-art technologies.
- Applying the model in personalized application areas to bring benefits to social communities.
- Promoting the speech processing researches for ASEAN languages

Welcome Collaboration!
Thank you!

Contact us:

winpapa@ucsy.edu.mm
ayemyahlaing@ucsy.edu.mm