Golden Owl AI Intern Report

Exercise 2

Preparing Dataset

There is a dataset PhoAudiobook[1] (a 941-hour high-quality Vietnamese audiobook corpus released in 2025 for zero-shot TTS training) which is excellent for this problem.

The next step is to do some data preprocessing like normalizing the text: convert Convert to lowercase/uppercase as needed, expand abbreviations, handle punctuation for prosody (pauses, emphasis), and remove unwanted characters.

And for language like Vietnamese, which has a lot of loanwords or irregular words, we can map text to phonemes for precise pronunciation (e.g., "internet" pronounced as /in-tơ-nét/).

Feed into Acoustic Model and convert to sound.

We feed the processed text into an Acoustic Model to generate acoustic features such as mel-spectrograms representing pitch, duration, and energy.

Some common architectures are:

- Sequence-to-Sequence: Using RNN or Transformer to map text to spectrograms.
- Zero-Shot Models: Using models like XTTS-v2[2] to learn with minimal data.

To convert acoustic features into raw audio waveforms we can use models like HiFi-GAN[3], WaveGlow[4]

Post-Processing and Output

Finally, we can normalize the output audio, add sound effect or trim silence.

To evaluate the model, we can use metrics like Mean Opinion Score (MOS) for naturalness or Word Error Rate (WER) if comparing to references.

Some Challenges and Conclusion.

- Tone accuracy: Vietnamese is a tonal language with six distinct tones (level, rising, falling, dipping-rising, high-rising glottalized, low-falling glottalized). Each tone changes word meaning (e.g., "ma" can mean "ghost," "mother," depending on the tone). Incorrect tone rendering leads to misinterpretation or unnatural speech. To mitigate this problem,

we can incorporate explicit grapheme-to-phoneme (G2P) conversion with tools for precise tone mapping.

- Regional Accent Variations: Vietnamese language also distinct regional accents (Northern, Central, Southern), with differences in tone pronunciation, vowel quality, and vocabulary. Most dataset are biased toward Northern accents, making it hard to generalize the speech to Southern or Central accents. For this problem I think we can diversify our data by collecting more Southern or Central accents data or we can also normalize dialect-specific words (e.g., map "dĩa" to "đĩa" for Northern models)
- Loanword and Code-Switching: Vietnamese speakers often mix English or other languages (e.g., "Tôi enjoy nghe music"), and loanwords (e.g., "internet" pronounced /in-tơ-nét/) require special handling. Models trained on monolingual data may mispronounce these. We can use model that support code-switching like MMS-TTS, Include loanword dictionaries in preprocessing.

In conclusion, I have proposed a pipeline for Vietnamese Text-to-Speech problem. While there are many challenges, I think with the help of advanced technology, this problem can be solved in near future with the boom of AI.

References

[1] Vu, Thi, Linh The Nguyen, and Dat Quoc Nguyen. "Zero-Shot Text-to-Speech for Vietnamese." *arXiv preprint arXiv:2506.01322* (2025).

[2] Casanova, Edresson, et al. "Xtts: a massively multilingual zero-shot text-to-speech model." *arXiv preprint arXiv:2406.04904* (2024).

[3]Kong, Jungil, Jaehyeon Kim, and Jaekyoung Bae. "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis." *Advances in neural information processing systems* 33 (2020): 17022-17033.

[4]Prenger, Ryan, Rafael Valle, and Bryan Catanzaro. "Waveglow: A flow-based generative network for speech synthesis." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.