Will integrating F0 improve the naturalness of low-resource languages TTS? -- Cantonese as an example

Yifan Hu[1], Johan Sjons[1], Axel G. Ekström[2,3]
[1] Department of Linguistics and Philology, Uppsala University, Sweden
[2] Speech, Music & Hearing, KTH Royal Institute of Technology, Sweden
[3] Department of Psychology, Stockholm University, Sweden
yifan.hu.2298@student.uu.se, johan.sjons@lingfil.uu.se, axeleks@kth.se

Abstract

Background

Speech synthesis is now a mature technology with multiple well-established pipeline for implementing models and synthesising speech. Typically, a neural speech synthesis system includes a front end for analysing linguistic features, an acoustic model for extracting acoustic features, and a vocoder that generates waveforms from those features (Tan, Qin, Soong, & Liu, 2021). For most systems, the front end performs grapheme-to-phoneme conversion (G2P), and the primary acoustic feature predicted is duration.

However, G2P is difficult for low-resource languages, which often have rich linguistic or acoustic characteristics but lack well-labelled data and tools to build such data (Mak, Suen, & Lam, 2025). These limitations prevent the G2P front end from achieving its full potential. As for acoustic features, some models also consider features beyond duration, such as F0 and energy. FastPitch, based on FastSpeech, is one of the most widely known among these models. It uses ground-truth pitch conditioning during training and a pitch predictor during inference (Łańcucki, 2021).

Methods

Cantonese is mostly used in colloquial scenarios (Xiang, et al., 2024), but it receives little attention for even speech research, causing the difficulties in con-structing a perfect G2P front-end. We start to think if the issue can be overcome by intergrading the ground-truth data to the model. The architecture is similar to models like FastPitch for training.

However, first, such architectures have not been thoroughly tested on low-resource languages. Evaluations of FastPitch focused on English and Mandarin. Second, including a pitch predictor for inference increases system complexity and perplexity. The predictor is reasonable for supplying conditioning data, but it may not be the only solution.

In this project, we use GlowTTS as the base model with a FiLM layer to condition F0 data at the decoder, training and evaluating on a small dataset (~4 hours, 16 kHz). We are currently developing inference methods and plan to compare several approaches against the traditional pitch-prediction method.

Results and conclusion

Results are preliminary because inference methods are still under development. However, current results indicate the importance of G2P for improving the quality and reducing the data required, robustness of the F0 conditioning layer for training.

We are testing some alternatives to skip the pitch predictor for the inference, some of which will be presented at the conference.

References

Łańcucki, A. (2021). FastPitch: Parallel Text-to-speech with Pitch Prediction. *FastPitch: Parallel Text-to-speech with Pitch Prediction*. Retrieved from https://arxiv.org/abs/2006.06873

Mak, T. S., Suen, K. Y., & Lam, A. Y. (2025). Speech-guided Grapheme-to-Phoneme Conversion for Cantonese Text-to-Speech. *Interspeech 2025*, (pp. 2535–2539). doi:10.21437/Interspeech.2025-1428

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A Survey on Neural Speech Synthesis. *A Survey on Neural Speech Synthesis*. Retrieved from https://arxiv.org/abs/2106.15561

Xiang, R., Chersoni, E., Li, Y., Li, J., Huang, C.-R., Pan, Y., & Li, Y. (2024, June). Cantonese natural language processing in the transformers era: a survey and current challenges: Progress and Challenges in Cantonese... *Lang. Resour. Eval., 59*, 1747–1773. doi:10.1007/s10579-024-09744-w