Improving the expressiveness of TTS synthesis with non-autoregressive neural vocoding

Mohammed Salah Al-Radhi, Tamás Gábor Csapó, Géza Németh

malradhi@tmit.bme.hu

1. Research Question

Why are expressive speech important in delivering messages?

- conveys meaning, tone, and emotion.
- adds nonverbal cues that build connection and engagement.
- clarifies key points and highlights important ideas.
- captures attention and keeps listeners involved.

2. Problem Formulation

Flexible and appropriate rendering of expressivity in a synthetic voice is still out of reach:

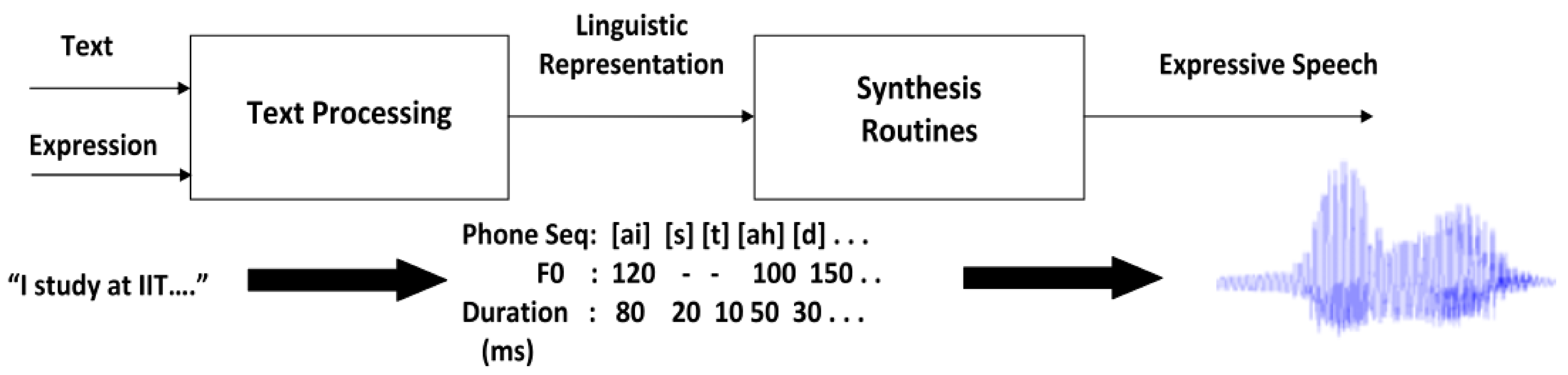
- making a voice sound happy or subdued, friendly or empathic, authoritative or uncertain is beyond what can be done today.

3. Goals

- Increase the flexibility in expression while maintaining the quality of state-of-the-art systems

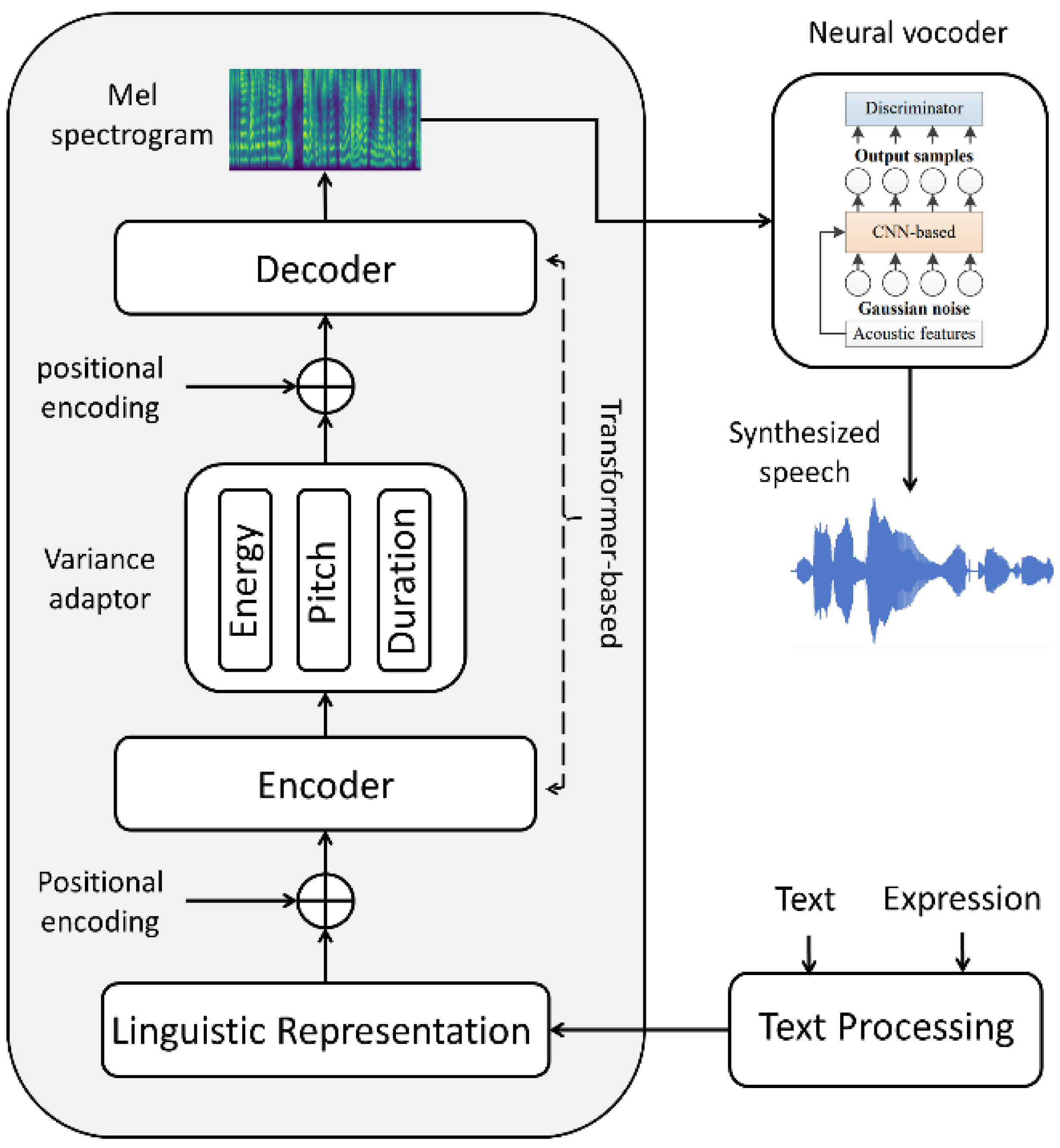
4. Expressive Speech Synthesis

- In **linguistics**, expressivity may change the choice of words or syntactic structures.
- In **acoustics**, it impacts various characteristics like energy, pitch, duration, etc.



5. Methods

- Propose a high-quality and expressive multi-speaker TTS model, which can flexibly synthesize speech with the style extracted from a target speaker.
- Used a non-autoregressive Mel-spectrogram prediction model (i.e., FastSpeech2), which has demonstrated improved speed and robustness compared to traditional autoregressive models.



6. Experimental conditions

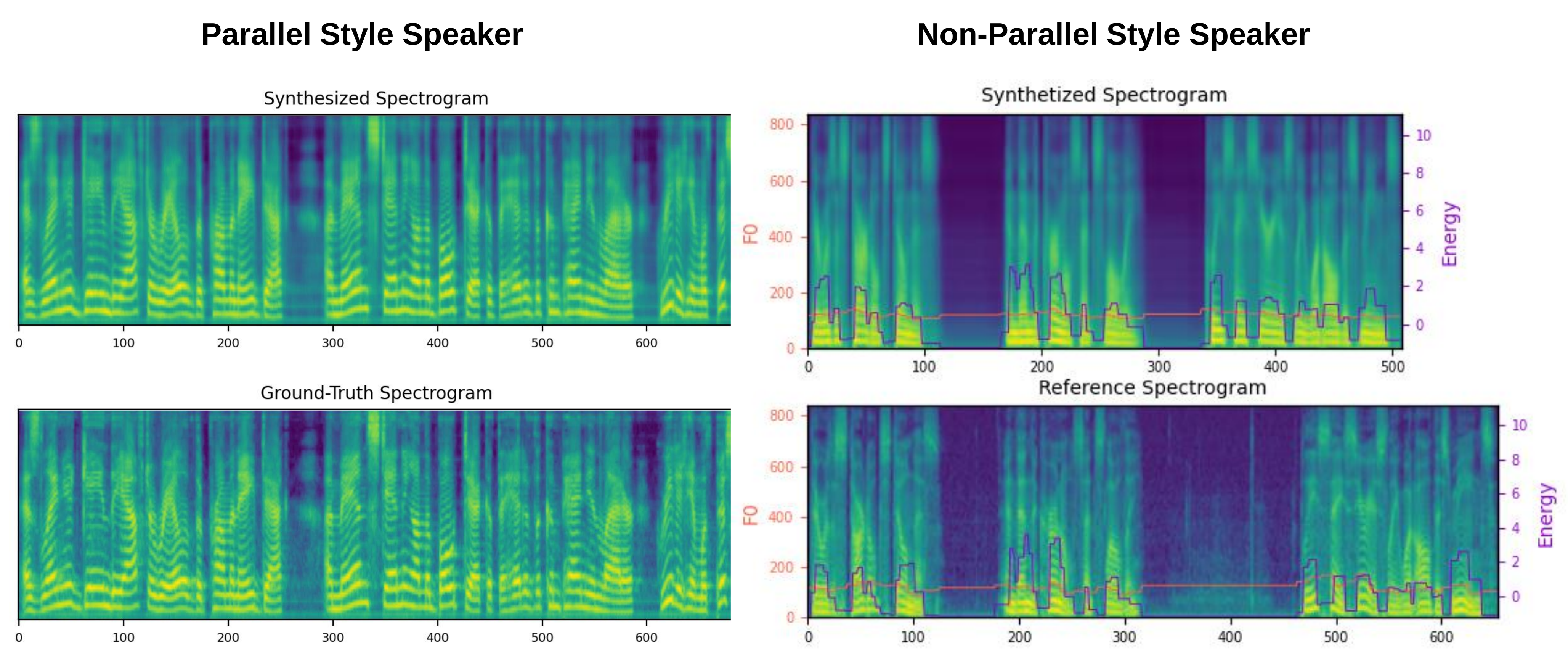
Datasets

- LibriTTS dataset contains 110 hours speech with 1151 reading-style speakers. We convert the speech sampling rate to 16KHz.
- use 12.5ms hop size, 50ms window size to extract mel-spectrogram.
- convert text into phoneme using grapheme-to-phoneme conversion and take phoneme as the encoder input.

Model

- 4 FFT blocks on both phoneme encoder and mel-spectrogram decoder following FastSpeech2.
- The architecture of pitch, energy and duration predictor in the variance adaptor are the same as those of FastSpeech2.
- The phoneme discriminator consist of fully connected layers.

7. Results



Loss	Training	Validation
Mel Loss	0.5752	0.6324
Pitch Loss	0.1455	0.2545
Energy Loss	0.1719	0.2076
Duration Loss	0.0867	0.0987
Total Loss	0.9792	1.1933

- Figures show the example of generated speech from the reference speaker.
- We observe that our model generates high quality mel-spectogram which is comparable to ground-truth mel-spectrogram.
- Based on the losses, model is performing well on the training and validation data.

8. Conclusion and Future work

- Propose an expressive TTS model to generate various styles of speech of multiple speakers.
- Confirmed through experiments that our model synthesize high-quality spectrogram given the reference audios from both parallel and non-parallel speakers.
- We will evaluate subjective naturalness of synthesized speech
- We will extend the TTS model to more languages and Enable multi-lingual speech style transfer

9. Acknowledgment

- APH-ALARM project (contract 2019-2.1.2-NEMZ-2020-00012) funded by the European Commission and the National Research, Development and Innovation Office of Hungary.

References

1. Changhwan Kim, Se-yun Um, Hyungchan Yoon, Hong-Goo Kan, FluentTTS: Text-dependent Fine-grained Style Control for Multi-style TTS, Interspeech 2022.
2. Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet, Analysis of expressivity transfer in non-autoregressive end-to-end multispeaker TTS systems, Interspeech 2022.
3. Xiao, Y., He, L., Ming, H., Soong, F.K., Improving Prosody with Linguistic and Bert Derived Features in Multi-Speaker Based Mandarin Chinese Neural TTS. IEEE ICASSP, pp. 6704-6708.

TMIT SmartLab Intelligent Interactions

Budapest University of Technology and Economics
Department of Telecommunications and Media Informatics
Budapest, Hungary

MŰEGYETEM 1782