VinBigData - Center for Virtual Assistant Technology

FastSpeechStyle : Vietnamese Emotional Speech Synthesis for VLSP 2022 Shared Task

Van Thinh Nguyen, Tri-Nhan Do, Hung-Cuong Pham, Tuan Vu Ho, Dang-Khoa Mac

1. Overview

In the VLSP 2022 Emotional Speech Synthesis Challenge, participants have to construct a synthetic voice in 4 different emotions:

- ESS system using emotional speech data from a single speaker (task 1)
- ESS system using data from a speaker with only neutral speech (task 2).

2. ESS System

Our architecture consists of 3 main components:

- Mel Emotional Encoder: to extract information about prosody into an embedding vector
- Acoustic Model: to synthesize Mel-spectrogram from input phonemes
- Vocoder: to generate speech from Mel-spectrogram

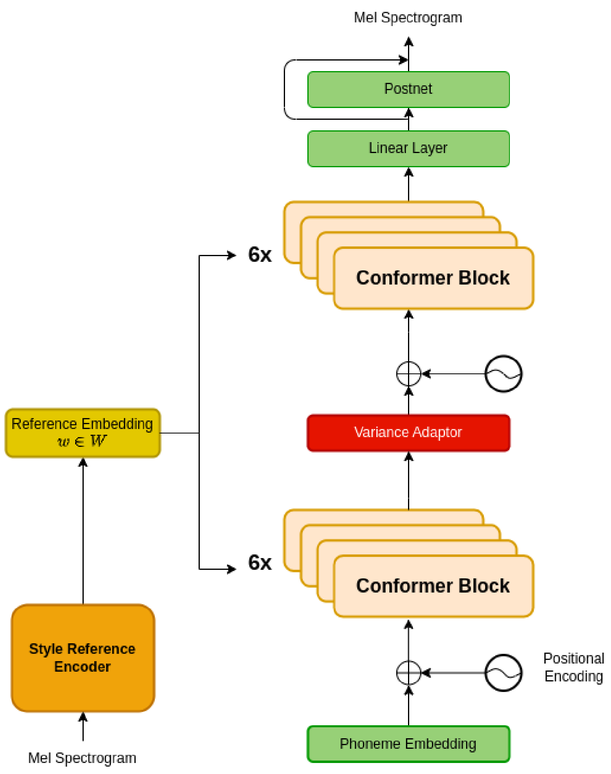


Figure 1: Emotional Speech Synthesis Architecture

2.1 Style Adaptive Layer Norm

SALN is to ”scale and shift” hidden features based on bias and gain conditioned by an emotional vector.

By adjusting the bias and gain values, the model can generate various styles of speech in the style of the target speaker with only one reference sample.

$$SALN(h, w) = g(w) \cdot y + b(w) \quad (1)$$

2.2 Conformer

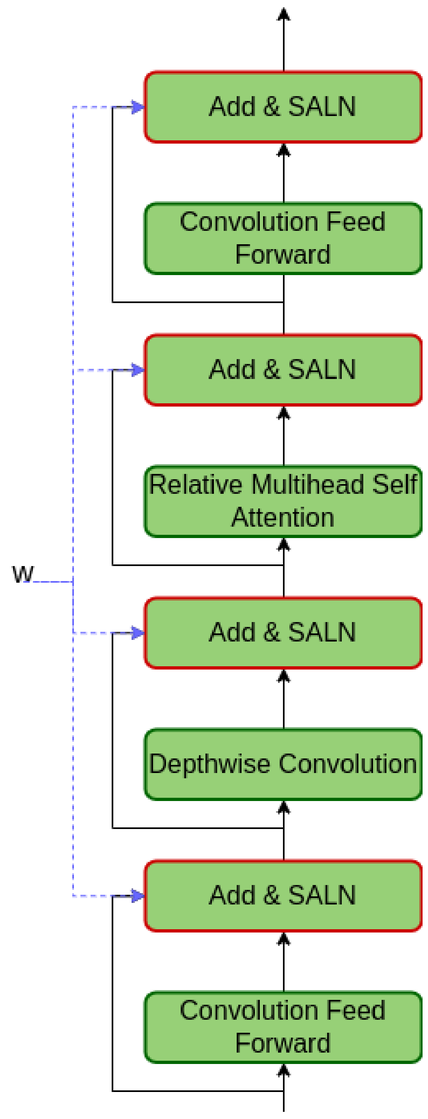


Figure 2: Improved Conformer Block for TTS

Conformer is a combination of transformer and convolution modules which is composed of:

- Convolutional feed-forward module
- Depthwise convolution module
- Self-attention module
- Second convolutional feed-forward module

2.3 Loss Function

Fastspeech2 loss functions combined with a structural similarity index measure loss.

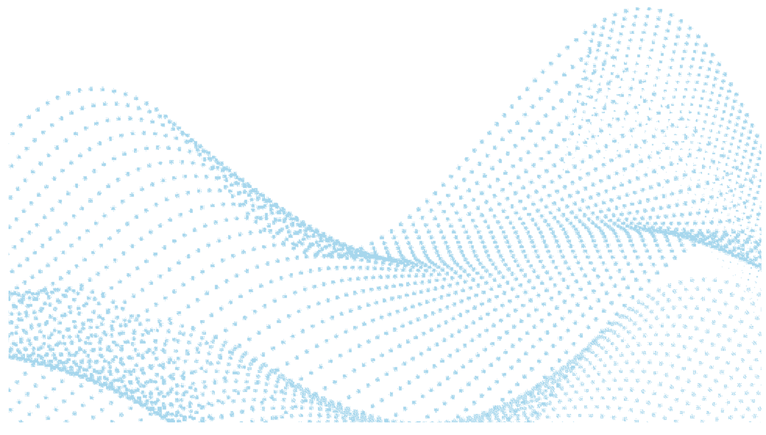
$$L_{variation} = L_{pitch} + L_{duration} + L_{energy} \quad (2)$$

$$L_{total} = L_{variation} + L_{mel} + L_{ssim} \quad (3)$$

3. Experiment and Result

Dataset for two tasks provided by the organizers include:

- VLSP-EMO for task 1: Emotional Speech Dataset includes about 5 hours of a single speaker and four emotional labels: neutral, sad, happy, and angry.
- VLSP-NEU for task 2: Neutral Speech Dataset includes 4 hours of another speaker.



The proposed models were trained with a batch size of 64 for subtask 1 and 24 for subtask 2, on a Tesla A100 NVIDIA GPU. The model converges after 100,000 iterations.

	Sub-Task 1	Sub-Task 2
Naturalness (5)	4.131	4.047
Intelligibility (%)	44.5	18.8
Speaker Similarity (4)	/	2.466

Table 1: VLSP TTS 2022 Result for FastSpeechStyle

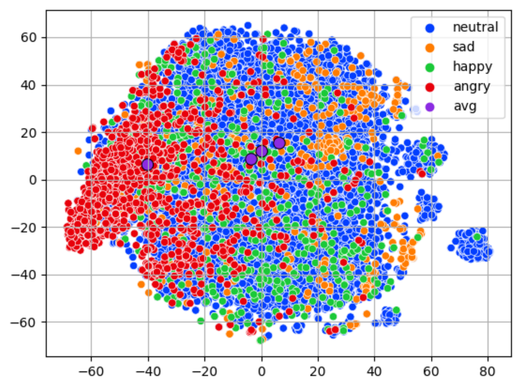


Figure 3: t-SNE visualization for embedding of all data

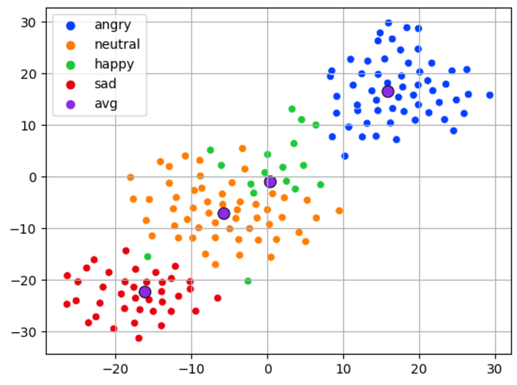


Figure 4: t-SNE visualization for embedding of typical samples

4. Conclusions

Thanks to the VLSP competition, a new approach to the Emotional Speech Synthesis problem was proposed when applied to Vietnamese. The authors demonstrate how to extract emotional vectors in an unsupervised manner using the Reference Encoder. The model can generate various emotions, thus allowing us to control the influence of emotions in a speech while retaining naturalness.