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
VinBigData Joint Stock Company
thinhnv1811, dotrinhan99, cuongph2112, tuanvu.ksvt92, khoamd@gmail.com

Abstract

This paper presents the submitted text-to-speech system for VLSP 2022 Emotional Speech Synthesis (ESS) Challenge. In this year's challenge, participants need to construct an ESS system using emotional speech data from a single speaker (task 1) and another ESS system using data from a speaker with only neutral speech (task 2). To address these challenges, we propose an expressive speech synthesis system which can synthesize high-quality expressive speech. The proposed model includes two main components (1) Mel Emotion Encoder extracts emotion embedding from the Mel-spectrogram of audio, (2) the FastSpeechStyle, a non-autoregressive model, which is modified from vanilla Fastspeech2. The FastSpeechStyle replaces normal LayerNorm with Style Adaptive LayerNorm to "shift" and "scale" hidden features according to emotion embedding, the model also used an improved Conformer block instead of vanilla FFTBlock to better model the local and global dependency in the acoustic model.

Index Terms: text-to-speech, emotional speech synthesis, cross-speaker adaptation, style adaptive layer.

1 Introduction

In the VLSP 2022 Emotional Speech Synthesis Challenge, participants have to construct a synthetic voice in 4 different emotions using the shared training dataset. The challenge is divided into two tasks:

- Task 1: To build a Text-to-Speech (TTS) system that synthesizes four types of emotional speech: neutral, sad, happy, and angry. The training data consists of 4.5 hours of speech data from a single speaker crawled from a television drama.
- Task 2: To adapt the TTS system in Task 1 to a new speaker for whom the training data only includes neutral utterances.

Participants are not allowed to use external speech data or pre-trained TTS models for both tasks. However, open-source pre-trained vocoders are allowed to use due to the limited training data. The output from each task undergoes subjective evaluation through listening tests covering naturalness and emotion similarity.

With the advance of deep learning models, speech synthesis systems have created synthetic speech indistinguishable from human speech in terms of naturalness. Besides linguistic information, the speech also conveys information about speaking styles, such as speaker identity, emotion, and prosody. These types of information play a crucial role in effective verbal communication with a human. However, controlling this expressive information in synthetic speech remains challenging for the current TTS systems.

In the VLSP 2022 ESS challenge, we used the Fastspeech2 architecture as our backbone TTS model (Ren et al., 2020). With the ability to explicitly control pitch, energy, and duration, Fastspeech2 architecture is perfectly tailored to the task. Furthermore, Fastspeech2 non-autoregressive property proves more reliable and robust than other autoregressive models that often suffer from fail-alignment problems. To customize the Fastspeech2 architecture to the ESS task, we used the Style Adaptive Layer Norm (SALN) based on the Speaker Adaptive Layer Norm (Min et al., 2021) (Arik et al., 2018), which condition the output Mel-spectrogram by emotion embedding, we also replace the vanilla FFBlock (Ren et al., 2020) with an improved Conformer Block to better model the local and global dependency in the acoustic model (Liu et al., 2021). A Mel Emotion Encoder was

also used to generate emotion embedding from Mel ground truth. The paper is organized as follows: We present some related works in Section 2 before describing our proposed TTS system in Section 3. Then we show the experiment settings and evaluation results in Section 4. Finally, we conclude our paper in Section 5.

2 Related Work

The most common approach to emotional speech synthesis (ESS) is to condition a TTS model with expressive features. In supervised learning, the emotion trait can be simply represented as a one-hot encoded vector. Prosody features such as pitch, energy, and duration can be estimated from text and speech data before training the model to improve the controllability of emotional speech. However, due to the discrete values of the one-hot encoded vector, such approaches can only synthesize predefined emotions. Therefore, the limitation of this method is emotion ambiguity and cannot show properties such as degree of continuous emotion, multi-label emotion, and emotion context dependency

In an unsupervised manner that does not require emotion-labeled data, expressive information can be implicitly extracted by a reference encoder or using a variational autoencoder (Zhang et al., 2019). Although this method can not interpret the emotion of speech, the prosody can be continuously controlled for each speaker. Variational autoencoder (VAE) models try to model emotions in continuous latent space with Gaussian prior and manipulate these latent variables for emotional synthesis (Akuzawa et al., 2018). However, the drawback of such approach is computation speed. Expressive information is conveyed in both text and speech: text representations can be obtained from pre-training (Kim et al., 2021)

The ESS task can be done with the above methods. However, it is inadequate to generate emotion using only neutral training data in Task 2. Cross-speaker style transfer could be a good approach. (Ribeiro et al., 2022) were proposed a method using voice conversion to augment training data, but this approach needs a large multi-speaker dataset which is not allowed in task2.

To handle all issues above, we proposed the same expressive speech synthesis model architecture for both tasks to synthesize emotional speech with a set of emotion tags and a specific inference strategy for each task to strengthen the emotion of speech.

3 Emotional Speech Synthesis System

Our architecture for Vietnamese ESS consists of 3 main components, a Mel Emotional Encoder to extract information about prosody into an embedding vector, an Acoustic Model to synthesize Mel-spectrogram from input phonemes, and a Vocoder for synthesizing speech from Mel-spectrogram.

3.1 Mel Emotional Encoder

The Mel Emotional Encoder (Emotion Encoder) is based on the idea of the Reference Encoder (Skerry-Ryan et al., 2018) to extract an emotion embedding vector that contains emotional information of the given speech. The architecture comprises three parts (Min et al., 2021). The first module is spectral processing with fully-connected layers to create hidden features, the temporal processing module is convolutional layers with residual connections to learn the context information of the speech segments, and finally, the multi-head self-attention mechanism is used to extract emotional information. At the training stage, the input of the Emotion Encoder is the ground truth Mel-spectrograms of the corresponding text script.

3.2 FastSpeechStyle

For faster generation and improved stability, the authors chose Fastspeech2 as the backbone model (Ren et al., 2020). This non-autoregressive acoustic model consists of an Encoder to extract the contextual information from the phoneme and a Variance Adaptor with explicit variation information modeling, including duration, pitch, and energy predictor, which adjusts the speed, tones, and loudness of the voice. Finally, the Decoder to create Mel-spectrogram, keeps the speaker's timbre consistent. The encoder and decoder of this customized Fastspeech2 are Conformer modules conditioned by emotional embedding through the Style Adaptive Layer Norm.

3.2.1 Conformer

Conformer is a combination of transformer and convolution modules. The Conformer for TTS is slightly different from what is used for speech recognition models (Liu et al., 2021). It is composed of four stacked modules, a convolutional feed-forward module, a depthwise convolution module, a self-attention module and a second convolutional feed-forward module. The self-attention

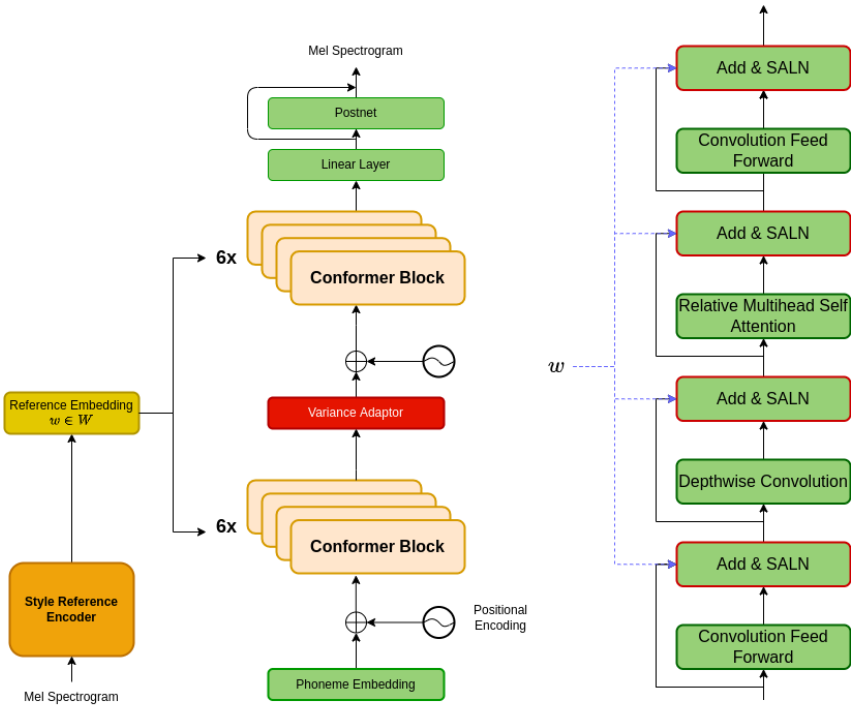

Figure 1: Emotional Speech Synthesis Architecture. Figure (a) shows the overall pipeline for FastSpeechStyle. Figure (b) shows the improved Conformer block and the integration of emotional embedding through Adaptive Layer Norm

of transformer extracts the global interaction, and convolutions in CNNs capture the local correlations.

3.2.2 Style Adaptive Layer Norm

There are many ways to integrate emotional embedding into the encoder and decoder of the backbone model, such as concatenation or element-wise addition with layers of Conformer. These methods increase the number of parameters of the model and achieve low adaptation quality.

The main idea of Style Adaptive Layer Norm (SALN) is to "scale and shift" hidden features based on bias and gain conditioned by an emotional vector (Min et al., 2021). By adjusting the bias and gain values, the model can generate various styles of speech and effectively synthesize speech in the style of the target speaker with only one reference sample.

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

The affine layers, which is a single fully connected layer, convert the emotion embedding **w** to bias **b** and gain **g** respectively for each hidden feature y in the formula 1. The LayerNorm in the Conformer blocks will be replaced with the SALN layer to change the style of the synthesized speech.

3.2.3 Loss Function

The loss function for the proposed acoustic model includes the popular fastspeech2 loss functions combined with a structural similarity index measure loss (SSIM) (Wang et al., 2004).

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

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

The loss values of the variation information are MSE between the predicted and the ground truth pitch, energy, and duration. Loss for predicted Mel-spectrogram is calculated with MAE and SSIM to measure the similarity for better audio fidelity.

3.3 Hifigan Vocoder

With limited data in the competition, training a new vocoder model is not possible. Therefore, the authors use a pretrain of the Hifigan model for English that has been published (Kong et al., 2020), then finetune with a ground-truth Mel-spectrogram generated from the acoustic model on the provided

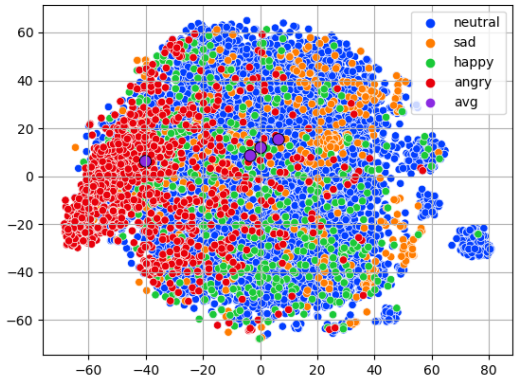

Figure 2: t-SNE visualization of emotional embeddings of all data.

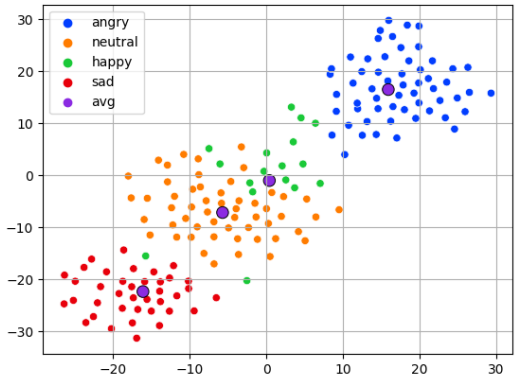

Figure 3: t-SNE visualization of emotional embeddings of typical samples.

data. The noise generated by the model is removed by subtracting the generated waveform from the bias of the vocoder model with zero input.

3.4 Inference Strategy

The emotion embedding vectors of all samples in Task1 dataset are visualized in (Figure 2) by using the t-SNE algorithm (van der Maaten and Hinton, 2008). Four different colors in the figure denote four emotion clusters: angry, sad, happy and neutral. Each emotion embedding sample represented in the figure was generated by Emotion Encoder, using reference audio in training data. It shows that the Emotion Encoder model cannot distinguish different types of emotion classes of all training data. The reason is there are too many emotional variants of each class. Therefore, the most typical samples for each emotion, which clearly express the emotional level, were selected to extract the distribution of emotion embedding.

Figure 3 visualizes the emotion embedding vectors of selected samples. It's true that the Emotion Encoder model has the capability of distinguishing different types of emotion classes if we can utilize the distribution of selected emotion embedding vectors in the emotional vector space. So we conclude that it is possible to control the emotions of the FastSpeechStyle synthesis system if we establish a connection between the emotion and corresponding distribution. The simplest way is to create a representation embedding vector for each emotion class by the element-wise average of emotion embedding vectors included in each emotion cluster. During inference, these vectors are used to synthesize desired emotions. The avg label in Figure 2 and Figure 3 denote the representation embedding vector of each emotion class.

4 Experiment and Analysis

4.1 Data Analysis and Processing

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.

Text scripts of data are traversed through a dictionary and converted to phonemes. Noise and breathing in the silence intervals of the audios are filtered by a kaiser filter. ForceAlignKaldi is used to perform alignment between phonemes and each audio segment (McAuliffe et al., 2017). Samples containing background noise or have a mismatch between the script and audio will be removed. Explicit information such as pitch and energy is preprocessed before training by using PyWorld to estimate fundamental frequency.

4.2 Experiment

The authors use the same configuration and model architecture for both tasks and train two tasks on two corresponding datasets from scratch. The Encoder and Decoder of customized Fastspeech2 are 6 Conformer blocks, and multi-head attention at each block is set to 2. The Variance Adapter uses 3 predictors of pitch, energy, and duration with a

Task		Emotion			
		Angry	Neutral	Sad	Happy
Task1	MOS	3.593	4.118	3.43	3.759
	ESS	63.95	49.509	42.333	7.294
Task2	MOS	3.66	3.374	3.486	3.219
	ESS	42.267	43.11	12.257	15.532

Table 1: The Mean Opinion Score (MOS) and Emotion Similarity Score (ESS) results.

convolutional hidden size of 384. With Reference Encoder, 2 LinearNorm classes are used for spectral processing. The temporal processing is two 1D convolutional layers combined with residual skip connections and 2 heads for the multi-head self-attention module. The output emotion embedding dimension is 128.

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.

In the inference stage, we adapt the inference strategy from the original, which was described in **??**, for two tasks, Task1 and Task2:

- Task1: We adjust the speed of speech to increase emotion strength by modifying duration predictor output in Variance Adaptor with an experiment factor.
- Task2: The angry emotion is expressed clearly, but two emotions, happy and sad, were mixed with neutral, leading to a high false negative for these emotions in prediction. Therefore, during the inference of these emotions, the output of pitch, energy, and duration in model Task1 is used to replace the pitch, energy and duration output of model Task2 and is then added to the Variance Adaptor output. The pitch and speed adjustment strategies are also applied to increase emotion strength.

4.3 Internal Evaluation

We conducted an internal evaluation before submission. The models were evaluated on two main criteria: naturalness and emotion similarity. A set of 200 sentences was used in this evaluation. Twenty Vietnamese listeners participated in this experiment via a web-based interface. Each subject listened to 50 samples randomly selected from synthesized audio. After listening, they were asked to give a naturalness score for each sample from 1 to 5, and then they needed to choose the closest emotion and give an emotion similarity score from 1 to 100 to complete the test. If the selected emotion is different from the input emotion when inference, the emotion similarity score will be zero. The Mean Opinion Score of naturalness and Emotion Similarity Score is represented in Table 1. The results show that the emotion similarity of angry, neutral, and sad for both tasks is extremely strong. Meanwhile, happy can not be distinguished from other emotions. The result also shows comparable quality in the aspect of naturalness.

5 Conclusion

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.

Acknowledgments

We are deeply grateful to the listeners who participated in data labeling and internal evaluation.

References

Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. 2018. Expressive speech synthesis via modeling expressions with variational autoencoder. *arXiv preprint arXiv:1804.02135*.

Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural voice cloning with a few samples. *Advances in neural information processing systems*, 31.

Minchan Kim, Sung Jun Cheon, Byoung Jin Choi, Jong Jin Kim, and Nam Soo Kim. 2021. Expressive text-to-speech using style tag. *arXiv preprint arXiv:2104.00436*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in*

Neural Information Processing Systems, 33:17022–17033.

Yanqing Liu, Zhihang Xu, Gang Wang, Kuan Chen, Bohan Li, Xu Tan, Jinzhu Li, Lei He, and Sheng Zhao. 2021. Delightfultts: The microsoft speech synthesis system for blizzard challenge 2021. *arXiv preprint arXiv:2110.12612*.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. 2021. Meta-stylespeech: Multi-speaker adaptive text-to-speech generation. In *International Conference on Machine Learning*, pages 7748–7759. PMLR.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Manuel Sam Ribeiro, Julian Roth, Giulia Comini, Goeric Huybrechts, Adam Gabryś, and Jaime Lorenzo-Trueba. 2022. Cross-speaker style transfer for text-to-speech using data augmentation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6797–6801. IEEE.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE.