International Journal of Asian Language Processing

FastSpeechStyle: Fast, Emotion Controllable, and High-Quality Speech Synthesis

Received (Day Month Year)
Revised (Day Month Year)

Non-autoregressive text to speech models such as Fastspeech2 can fast synthesize high-quality speech. This model also allows explicit control of the speech signal's pitch, energy, and speed. However, controlling emotion while maintaining natural human-like speech is still a problem. In this work, we propose an expressive speech synthesis model that can synthesize high-quality speech with desired emotion. 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 used an Improved Conformer block, which replaces normal LayerNorm with Style Adaptive LayerNorm to "shift" and "scale" hidden features according to emotion embedding, instead of vanilla FFTBlock[1] to better model the local and global dependency in the acoustic model. We also propose a specific inference strategy to control the desired emotion of speech. The experimental results show that the proposed model with improved Conformer achieved higher scores than the baseline model in all naturalness and emotion similarity scores.

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

1. Introduction

With the advance of deep learning models, speech synthesis systems have created synthetic speech indistinguishable from the 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 or can be applied in critical situations[2] and storytelling[3]. However, controlling this expressive information in synthetic speech remains challenging for the current Text to Speech (TTS) systems.

The construction of an expressive speech synthesis model has been studied for a long time, from a synthesizing emotional speech by the Unit Selection method to HMM-Based methods by using the average emotion model[4] or model interpolation[5], and prominent in recent years is End to End models using deep neural networks[6].

The most common approach to emotional speech synthesis (ESS) using deep neural networks is to condition a TTS model with expressive features. In supervised learning, the emotion feature can be simply represented as a one-hot encoded vector[7] from small number of basic categories based on discrete emotion theory[8].

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 and depend on the homogeneity of emotions in the data samples. 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[9].

In an unsupervised manner that does not require emotion-labeled data, expressive information can be implicitly extracted by a reference encoder or by using a variational autoencoder[10]. Although this method can not interpret the emotion of speech, the prosody can be continuously controlled for each speaker and the model can acquire the ability to model a wide range of acoustic expressiveness[11]. Variational autoencoder (VAE) models try to model emotions in continuous latent space with Gaussian prior and manipulate these latent variables for emotional synthesis[12]. However, the drawback of such an approach is computation speed. Expressive information is conveyed in both text and speech: text representations can be obtained from pre-training[13][14] to capture the contextual information of the sentence[15][16], and emotional speech embedding can be extracted from the reference speech using a reference encoder. This model can well generate expressive speech using unseen tags. However speech quality of the style tag model is still lower than the reference model.

With the ability to explicitly control pitch, energy, and duration, Fastspeech2[17] architecture is perfectly tailored to the Text to Speech applications. Furthermore, the non-autoregressive property of Fastspeech2 proves more reliable and robust than other autoregressive models that often suffer from fail-alignment problems. For that reason, the proposed FastSpeechStyle used Fastspeech2 as the backbone model, the LayerNorm[18] layers were replaced by Style Adaptive Layer Norm (SALN)[19][20] to condition the output Mel-spectrogram by emotion embedding, and the FFTBlock[17] was replaced by Conformer Block to better model the local and global dependency in the acoustic model[21]. A Mel Emotion Encoder[19] was also used to generate emotion embedding from Mel-ground truth. The proposed FastSpeechStyle model can synthesize high-quality speech with a set of emotion tags by using a specific inference strategy.

Our main contributions are as follows:

- We propose a FastSpeechStyle model which uses Conformer Block instead of FFTBlock to better model the local and global dependencies. We also replace the LayerNorm with Style Adaptive Layer Norm to condition the output Mel-Spectrogram by the emotion embedding vector.
- The specific inference strategy of the FastSpeechStyle model was proposed to control the emotion of synthesized speech.
- Our proposed model achieved higher scores than the baseline model in all naturalness and emotion similarity evaluations.

The paper is organized as follows: We present an overview of emotional speech synthesis and related works in Section 1 before describing our proposed TTS system in Section 2. Then we show the experiment settings and evaluation results in Section 3. Finally, we conclude our paper in Section 4.

2. Emotional Speech Synthesis System

The Emotional Speech Synthesis System architecture is presented in Figure 1. This model consists of three main components: A Mel Emotional Encoder to extract information about prosody into an embedding vector, an Acoustic Model to generate Mel-spectrogram from input phonemes, and a Vocoder model to synthesize speech from Mel-spectrogram.

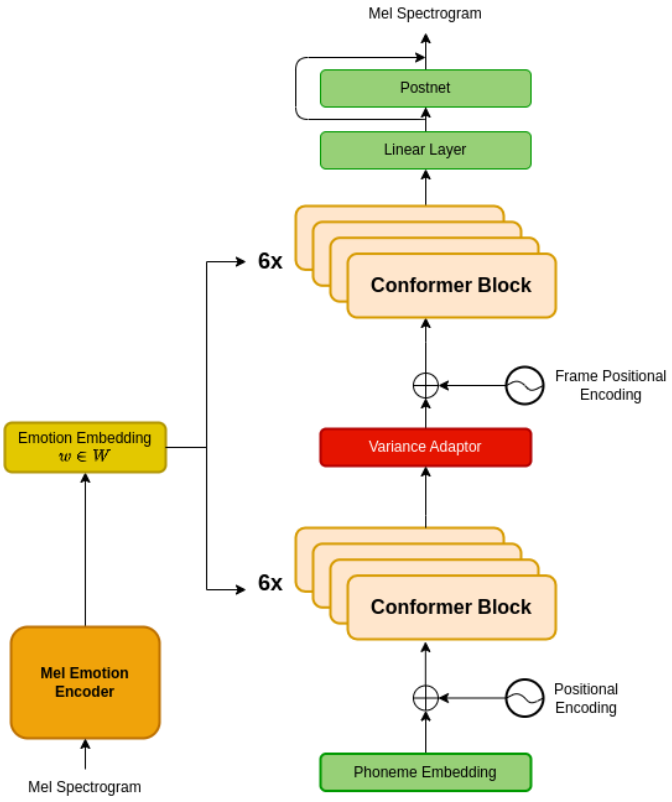

Fig. 1. FastSpeechStyle: Emotional Speech Synthesis Architecture.

2.1. *Mel Emotional Encoder*

The Mel Emotional Encoder (Emotion Encoder) is based on the idea of the Reference Encoder[6] to extract an emotion embedding vector that contains the speech's emotional information. The architecture is the same as[19], which comprises three stacked modules. 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. Finally, the multi-head self-attention mechanism with residual connection is used to encode global information. The output of self-attention was temporally averaged to get an one-dimensional emotion vector. At the training stage, the input of the Emotion Encoder is the ground truth Mel-spectrograms of the corresponding text script.

2.2. *FastSpeechStyle*

For faster generation and high stability, Fastspeech2 was chosen as the backbone model[17]. 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 in phoneme-level[22]. Finally, the Decoder to create Mel-spectrogram keeps the speaker's timbre consistent. The FFT block in Fastspeech2 was replaced by improved Conformer modules, which were conditioned by emotional embedding through the Style Adaptive LayerNorm, illustrated in Figure 2.

2.2.1. *Conformer*

Conformer is a combination of transformer and convolution modules. The Conformer for speech synthesis is slightly different from what is used for speech recognition models[21]. The order of self-attention depthwise convolution is switched to faster convergency, the convolution layer was used to replace the linear layer in Feed Forward Module, and ReLU was replaced by Mish[23]. Finally, the improved Conformer 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. With this architecture, the model can better model the global interaction with self-attention and the local correlations with the convolution layer in both the depthwise convolution module and the convolution feed-forward module.

2.2.2. *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

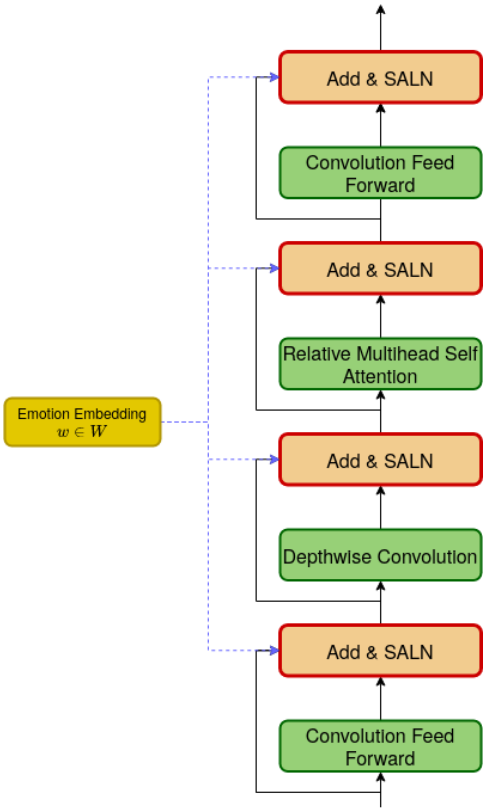

Fig. 2. The improved Conformer block and the integration of emotional embedding through Adaptive Layer Norm

layers of Conformer. These methods increase the number of parameters of the model and achieve low adaptation quality.

The main idea of the Style Adaptive Layer Norm (SALN[19]) is to "scale and shift" hidden features based on bias and gain conditioned by an emotional vector[19]. By adjusting the bias and gain values, the model can generate various speech styles, including emotions, and effectively synthesizing speech in the style/emotion of the target speaker with only one reference sample.

$$SALN(h, w) = g(w)y + b(w) \tag{1}$$

The affine layers, 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.

2.2.3. *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)[24].

$$L_{variation} = L_{pitch} + L_{duration} + L_{energy} \tag{2}$$

$$L_{total} = L_{variation} + L_{mel} + L_{ssim} \tag{3}$$

The loss values of the variation information $L_{variation}$ are calculated by the Mean Square Error (MSE) between the predicted and the ground truth pitch, energy, and duration. L_{mel} is the Mean Absolute Error (MAE) between the predicted and the ground truth Mel-spectrogram. For better audio fidelity, L_{ssim} used SSIM loss to measure the similarity between predicted and ground truth Mel- spectrogram.

2.3. *Hifi-gan Vocoder*

The Hifi-gan[25] was used to generate high-fidelity speech from the predicted mel-spectrogram, and the universal model was finetuned with the mel-spectrogram generated from FastSpeechStyle. The model noise was generated from the bias of the vocoder with zero input, and then it was subtracted from the output speech signal.

2.4. *Inference Strategy*

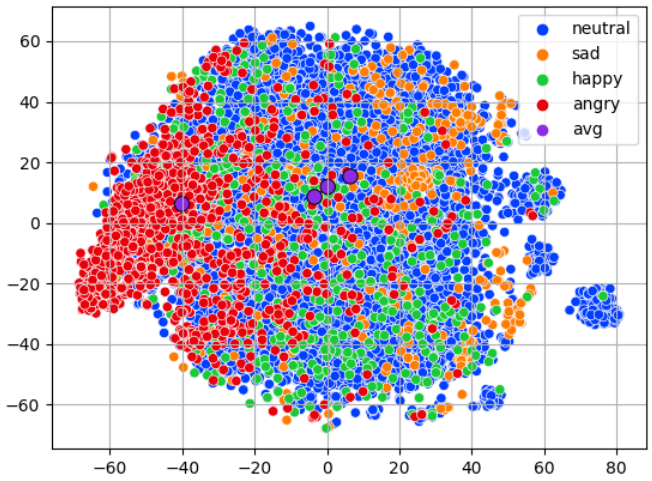

Fig. 3. T-SNE visualization of emotional embeddings of all data.

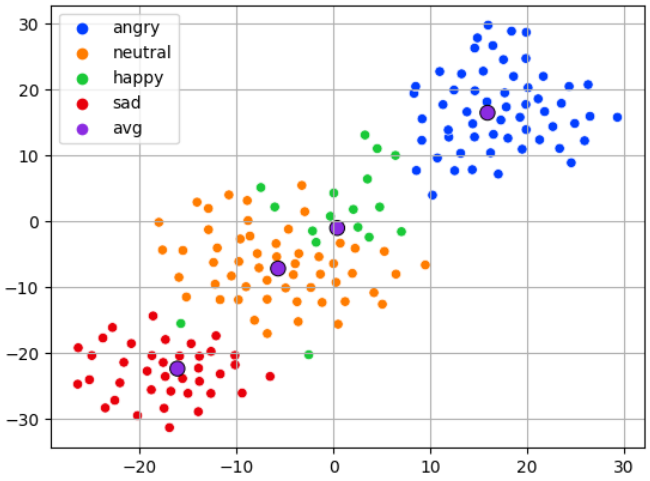

Fig. 4. T-SNE visualization of emotional embeddings of typical samples.

The emotion embedding vectors of all samples in the labeled dataset are visualized in (Figure 3) using the t-SNE algorithm[26]. The figure's four colors denote four emotion clusters: angry, sad, happy, and neutral. Each emotion embedding sample represented in the figure was generated by Mel Emotion Encoder with reference audio in training data as input. It shows that the Mel Emotion Encoder model cannot distinguish different types of emotion classes of all training data. The reason is that there are too many emotional variants in 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 4 visualizes the emotion embedding vectors of selected samples. It is true that the Mel 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 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 4 and Figure 3 denote the representation embedding vector of each emotion class.

3. Experiments

3.1. *Experimental Setup*

Dataset: the experiment dataset provided by the Vietnamese Language and Speech Processing (VLSP) which is VLSP-EMO: Emotional Speech Dataset, includes about 4.5 hours of a single speaker and four emotion labels: neutral, sad, happy, and angry.

Preprocess: The text scripts of data are traversed through a dictionary and converted to phonemes. Noise and breathing in the silence intervals of the audio are filtered by a kaiser filter. Kaldi Forced Aligner[27] is used to align phonemes and each audio segment. Samples containing background noise or mismatches between the script and audio will be removed. Explicit information such as pitch and energy is generated before training by using World Vocoder[28].

Model Configurations: We use the StyleSpeech[19] model as a baseline. The Encoder and Decoder of the baseline model are 6 FFTBlocks (Feed Forward Transformer Block[1]), and the Encoder, Decoder, and Variance Adaptor hidden dimensions are 384. The output dimension of emotion embedding is 128. For the Proposed FastSpeechStyle model, we use the same configuration as the baseline model. Six Conformer Blocks were also used.

Training Experiments: The baseline and proposed model were trained with the processed VLSP-EMO dataset on an NVIDIA Tesla A100 GPU. The batch size of 64 sentences was used during training. We use AdamW[29] with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$, and follow the same learning rate schedule as in Vaswani[30].

3.2. *Evaluation Metrics*

The evaluation of text to speech system is very challenging, especially for the emotional text to speech system. So, we conducted two subjective evaluations to measure the systems. The evaluation with **MOS** (Mean Opinion Score) was used for naturalness and the naturalness score for each sample is from 1 to 5. The evaluation with **ESS** (Emotion Similarity Score) was used for emotion similarity. The listeners were asked to choose the closest emotion with an audio sample and give an emotion similarity score from 1 to 100. If the selected emotion is different from the input emotion when inference, then the emotion similarity score will be zero. The mean value of all emotion similarity scores was reported as **ESS**. Both evaluations were conducted with 100 audio samples as a **test set** and evaluated by 38 listeners. Each listener had to evaluate 50 samples randomly selected from the **test set**.

3.3. *Naturalness Evaluation*

The naturalness was evaluated with the MOS metric on the baseline model, proposed model, and ground truth. The audio samples of the baseline and proposed models were generated with four emotion tags as input: angry, happy, neutral, and sad. All output emotion audio samples and ground truth samples were evaluated

by the MOS metric described above, and the results are shown in Table 1. It shows that the proposed model achieved a higher naturalness score for all emotions. The neutral emotion has the highest score in emotion classes, and the Ground truth has the highest score. This show that emotional features negatively affect naturalness.

Table 1. Mean Opinion Score of Naturalness with 95% confidence intervals.

Model	Angry	Happy	Neutral	Sad
Baseline	3.488 ± 0.169	4.09 ± 0.131	4.108 ± 0.115	3.586 ± 0.133
Proposed	3.608 ± 0.146	4.157 ± 0.111	4.376 ± 0.115	3.766 ± 0.135
Ground Truth		4.542 ± 0.096		

3.4. *Emotion Similarity Evaluation*

The evaluation of emotion similarity was completed with the ESS metric on the baseline and proposed model. The results are shown in Table 3. The emotion similarity score of the proposed model is slightly better than the baseline model. The similarity score of angry is highest in emotion classes for both models, which means this kind of emotion is easy to express. The emotion similarity score of happy is nearly zero. Furthermore, the emotion embedding of happy samples, represented in Figure 4, can be confused with neutral and angry samples. Those things lead to the same conclusion that the happy emotion samples are difficult to distinguish from angry and neutral emotions.

Table 2. Emotion Similarity Score with 95% confidence intervals.

Model	Angry	Happy	Neutral	Sad
Baseline	76.701 ± 3.697	3.886 ± 2.427	47.377 ± 5.05	48.141 ± 4.999
Proposed	77.232 ± 2.846	1.823 ± 1.619	48.988 ± 5.397	49.953 ± 5.157

3.5. *Performance in the VLSP Challenge 2022*

Our proposed TTS system was also submitted to the Emotional Speech Synthesis Shared tasks in VLSP Challenge 2022. The provided training dataset is only VLSP-EMO. The challenge uses two criteria for evaluation (shared task1 with only the VLSP-EMO dataset), the MOS for naturalness and SUS[31] (Semantically Unpredictable Sentences) for intelligibility. The naturalness test was conducted by 320 utterances; 64 listeners, including males/females and expert/non-expert, were asked to provide a score from 1 to 5. The intelligibility test was evaluated by 56 people on the test set of 245 utterances. The results in Table 3 show that our proposed system achieves the highest naturalness score while maintaining a high quality of intelligibility.

Table 3. VLSP TTS Challenge 2022 Results

System	MOS	SUS (%)
Proposed	**4.131**	44.5
A	3.403	52.7
B	3.941	49.4
C	3.875	38.1
D	3.822	39.0
E	2.719	72.3
F	3.938	42.9

4. Conclusion

We have proposed a FastSpeechSytle, a Fast and High-quality Emotional Speech Synthesis model which can fast generate high-quality and expressive speech for desired emotion. By applying improved Conformer to the FastSpeechStyle model, we achieve significantly improved quality of emotional speech. For future work, we plan to improve FastSpeechStyle to increase the naturalness of emotional speech, such as angry or sad emotions.

Acknowledgments

This research is supported by the computing infrastructure of Vinbigdata JSC. We are sincerely thank to the listeners for participating in data labeling and internal evaluation. Thanks to the VLSP Committee for organizing the VLSP challenge and providing an emotional dataset.

References

1. Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao and T.-Y. Liu, Fastspeech: Fast, robust and controllable text to speech, *Advances in Neural Information Processing Systems* **32** (2019).
2. M. Rusko, S. Darjaa *et al.*, Expressive speech synthesis for critical situations, *Computing and informatics* **33**(6) (2014) 1312–1332.
3. J. Přibil and A. Přibilová, Application of expressive speech in tts system with cepstral description, in *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, (Springer, 2008) pp. 200–212.
4. M. Tachibana, J. Yamagishi, K. Onishi, T. Masuko and T. Kobayashi, Hmm-based speech synthesis with various speaking styles using model interpolation, in *Speech Prosody 2004, International Conference*, 2004.
5. L. Qin, Z.-H. Ling, Y.-J. Wu, B.-F. Zhang and R.-H. Wang, Hmm-based emotional speech synthesis using average emotion model, in *International Symposium on Chinese Spoken Language Processing*, Springer2006, pp. 233–240.
6. R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark and R. A. Saurous, Towards end-to-end prosody transfer for expressive speech synthesis with tacotron, in *international conference on machine learning*, PMLR2018, pp. 4693–4702.

7. X. Zhu and L. Xue, Building a controllable expressive speech synthesis system with multiple emotion strengths, *Cognitive Systems Research* **59** (2020) 151–159.
8. H. Gunes, B. Schuller, M. Pantic and R. Cowie, Emotion representation, analysis and synthesis in continuous space: A survey, in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, IEEE2011, pp. 827–834.
9. D. Y. Choi and B. C. Song, Semi-supervised learning for continuous emotion recognition based on metric learning, *IEEE Access* **8** (2020) 113443–113455.
10. Y.-J. Zhang, S. Pan, L. He and Z.-H. Ling, 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)*, IEEE2019, pp. 6945–6949.
11. Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren and R. A. Saurous, Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis, in *International Conference on Machine Learning*, PMLR2018, pp. 5180–5189.
12. K. Akuzawa, Y. Iwasawa and Y. Matsuo, Expressive speech synthesis via modeling expressions with variational autoencoder, *arXiv preprint arXiv:1804.02135* (2018).
13. M. Kim, S. J. Cheon, B. J. Choi, J. J. Kim and N. S. Kim, Expressive text-to-speech using style tag, *arXiv preprint arXiv:2104.00436* (2021).
14. Y. Xiao, L. He, H. Ming and F. K. Soong, Improving prosody with linguistic and bert derived features in multi-speaker based mandarin chinese neural tts, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE2020, pp. 6704–6708.
15. J. Tu, Z. Cui, X. Zhou, S. Zheng, K. Hu, J. Fan and C. Zhou, Contextual expressive text-to-speech, *arXiv preprint arXiv:2211.14548* (2022).
16. A. Mukherjee, S. Bansal, S. Satpal and R. Mehta, Text aware emotional text-to-speech with bert, *Proc. Interspeech 2022* (2022) 4601–4605.
17. Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao and T.-Y. Liu, Fastspeech 2: Fast and high-quality end-to-end text to speech, *arXiv preprint arXiv:2006.04558* (2020).
18. J. L. Ba, J. R. Kiros and G. E. Hinton, Layer normalization, *arXiv preprint arXiv:1607.06450* (2016).
19. D. Min, D. B. Lee, E. Yang and S. J. Hwang, Meta-stylespeech: Multi-speaker adaptive text-to-speech generation, in *International Conference on Machine Learning*, PMLR2021, pp. 7748–7759.
20. S. Arik, J. Chen, K. Peng, W. Ping and Y. Zhou, Neural voice cloning with a few samples, *Advances in neural information processing systems* **31** (2018).
21. Y. Liu, Z. Xu, G. Wang, K. Chen, B. Li, X. Tan, J. Li, L. He and S. Zhao, Delightfultts: The microsoft speech synthesis system for blizzard challenge 2021, *arXiv preprint arXiv:2110.12612* (2021).
22. A. Łańcucki, Fastpitch: Parallel text-to-speech with pitch prediction, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE2021, pp. 6588–6592.
23. D. Misra, Mish: A self regularized non-monotonic activation function, *arXiv preprint arXiv:1908.08681* (2019).
24. Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE transactions on image processing* **13**(4) (2004) 600–612.
25. J. Kong, J. Kim and J. Bae, Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, *Advances in Neural Information Processing Systems* **33** (2020) 17022–17033.

26. L. van der Maaten and G. E. Hinton, Visualizing data using t-sne, *Journal of Machine Learning Research* **9** (2008) 2579–2605.
27. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, The kaldi speech recognition toolkit, in *IEEE 2011 workshop on automatic speech recognition and understanding*, (CONF), IEEE Signal Processing Society2011.
28. M. Morise, F. Yokomori and K. Ozawa, World: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE TRANSACTIONS on Information and Systems* **99**(7) (2016) 1877–1884.
29. I. Loshchilov and F. Hutter, Decoupled weight decay regularization, *arXiv preprint arXiv:1711.05101* (2017).
30. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* **30** (2017).
31. C. Benoît, M. Grice and V. Hazan, The sus test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences, *Speech communication* **18**(4) (1996) 381–392.