Proceedings of 2024 IEEE 12th International Conference on Intelligent Systems (IS)

FastSpeech2 Based Japanese Emotional Speech Synthesis

Masaki Ikeda
Information Systems Division
University of Aizu
Aizuwakamatsu, Japan
Email: m5261118@u-aizu.ac.jp

Konstantin Markov
Information Systems Division
University of Aizu
Aizuwakamatsu, Japan
Email: markov@u-aizu.ac.jp

***Abstract*—Modeling emotions is an important part of text-to-speech (TTS) research since its goal is to develop a technology for synthesizing naturally sounding speech. In this study, we aimed to build an end-to-end TTS system for the Japanese language that can synthesize emotional speech taking the FastSpeech2 as a base model. Most of the existing approaches use some form of conditioning the generation process on the emotion class label through a corresponding embedding. What separates such methods is the way this conditioning is implemented. Our approach is to add two new blocks which are essentially transformer multi-head self-attention blocks having an input combined with emotion embedding. They are inserted before and after FastSpeech2's variance adaptor and share parameters to ensure robust emotion conditioning when the amount of training data is small. The proposed model was objectively and subjectively evaluated using Mel-Cepstrum Distortion (MCD) and Mean Opinion Score (MOS) criteria respectively. The obtained results show that our approach performs better than a collection of emotion-specific models obtained by fine-tuning the base FastSpeech2.**

***Keywords*—Text-To-Speech; Emotional Speech Synthesis; FastSpeech2**

I. Introduction

Speech is the most primitive, modern, and common method of communication established by human evolution. The technology that can generate speech artificially is called speech synthesis and has become an important part of AI-powered applications. For example, it is used in AI assistants such as Apple's Siri or Amazon's smart speaker, the voices of video game characters, public announcements transportation, etc. Especially in Japan, where people tend to respect privacy and value anonymity, the use of VOCALOID [1], Softalk [2], and VOICEVOX [3], which are often utilized for commentary voices in streaming videos, is very active. Many users favor these tools because they can obtain natural speech easily without the need to record an individual's speech. In particular, there is a strong need for a synthetic speech that is not the speech of a specific person and is fluent and has natural inflection.

In recent years, Text-to-Speech (TTS) systems have made significant progress in the quality of synthesized speech thanks to deep learning methods such as Normalising Flows [4] [5], diffusion process [6], Transformer architecture [7] and others. While Transformer-based models, such as FastSpeech2 [8], have been outperformed by the latest diffusion models in terms of quality metrics [6], they are faster during the inference and thus more practical.

To synthesize naturally sounding speech, TTS models must account for some factors absent from simple text input, such as rhythm, intonation, and emotion [9]. Emotional speech is easier to perceive by the listeners and enriches the informational content of the spoken message.

Methods to enable emotions in synthesized speech have been studied since long ago first by introducing the style control vectors in the Hidden Markov Model (HMM) based TTS models [10]. Later, an emotion embedding was implemented in Recurrent Neural Network (RNN) based TTS [11] as well as in the hidden state of the Tacotron's decoder [12]. Such approaches are based on categorical labels to represent one or more emotions and various datasets have been collected to support those studies. Another way to control the synthesized speech's overall style including emotion is to use a reference utterance with the desired emotional state [13]. This method allows unlabeled training data to be used [14]. Embedding vectors for each emotion from reference speech and text are learned and their weighted sum is used to obtain utterance-level embedding. Some studies even attempt to introduce fine-grained emotional intensity control [15] [16]. However, estimating the intensity is challenging especially when it varies within a single utterance. Recently, following the growing popularity of prompt-based interaction with large language models (LLMs), some works have investigated and applied emotion control by a textual description of the desired emotion [17].

Research on Japanese emotional TTS has also been active with approaches to achieving affective synthesis by upgrading classical systems [18] or DNN-based solutions later on [19]. Our method is close to those applied when labeled data are available as it involves conditioning by emotion class embedding and is simple to implement.

II. Approach

A. FastSpeech2 baseline TTS system

Most TTS systems have three main blocks: text analysis, an acoustic model, and vocoder. Text analysis converts text into

979-8-3503-5098-2/24/$31.00 ©2024 IEEE

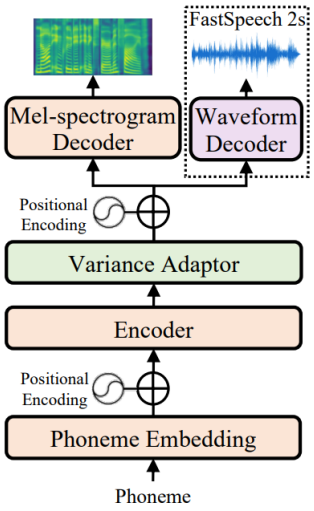

Fig. 1. The FastSpeech2 architecture [8]

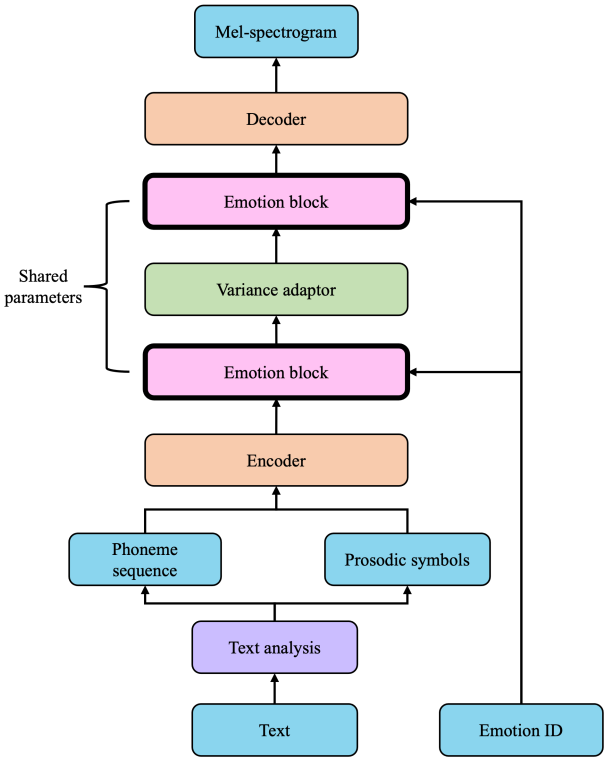

Fig. 2. The architecture of proposed model

linguistic features such as phonemes. Based on the linguistic features, the acoustic model produces acoustic features such as mel-spectrogram. Finally, the vocoder generates speech waveform from the mel-spectrogram.

We build our emotional TTS system using FastSpeech2 [8] as a base which is a non-autoregressive model for fast and high-quality speech synthesis. Its key components are the encoder, variance adaptor, and mel-spectrogram decoder as shown in Fig.1. The encoder extracts features from the input text converted into phonemes, the variance adaptor adds acoustic and duration information to the sequence, and the decoder generates the output mel-spectrogram features. Both the encoder and decoder consist of multiple feed-forward transformer (FFT) blocks each of which is a stack of multi-head attention layer and 1D-convolutional layer. The variance adaptor includes three predictors and a length regulator (LR). Predictors estimate phoneme duration, pitch, and energy for each token. LR adjusts the length of the phoneme sequence to the length of the mel-spectrogram by using the output of the duration predictor. Compared to conventional TTS, FastSpeech2's advantage is that it allows for parallel computation. Furthermore, it can synthesize high-quality speech faster by adding prosodic information from the pitch and energy predictors. Conventional autoregressive TTS models sometimes synthesize unsuitable speech by skipping and repeating text, but FastSpeech2 does not have such a problem because of the non-autoregressive nature of the model.

B. Model modification for emotional speech synthesis

Modifying model structure is the predominant approach not only to achieve emotional speech synthesis [11] [20] but also to generate multi-speaker [21] or custom voices [22]. Conditioning on a discrete parameter such as speaker ID or emotion type is usually implemented by learning a parameter embedding and inserting it in the processing pipeline. It can be included in the Layer Normalization [22], in the self-attention mechanism [23], or combined with the token representations in the encoder and decoder [24].

In a way similar to how the speaker conditioning is implemented in Transformer TTS [25], we modify our baseline FastSpeech2 model by inserting two new emotion blocks before and after the variance adaptor as shown in Fig.2. To achieve more stable and reliable output, we force these two blocks to share parameters. The architecture of each emotion block is shown in Fig. 3. It is inspired both by the way emotional factors are combined with LSTM models [26] and the transformer blocks used in the FastSpeech2 encoder and decoder. The embedding vector is added to the input token sequence representation effectively changing it in a different way for each emotion category.

III. Experiment

A. Dataset

In this research, we use the basic5000 part of the JSUT corpus [27] for model pre-training and the ITA part of the STUDIES corpus of emotional speech [28] for fine-tuning. Both datasets contain recorded human speech, text, and full-context labels. Full-context labels consist of linguistic features extracted from the text, such as phonemes, accents, and phoneme duration.

The basic5000 dataset includes 5000 pairs of texts and speech utterances recorded in neutral style by a single female speaker, designed to include readings of all commonly used

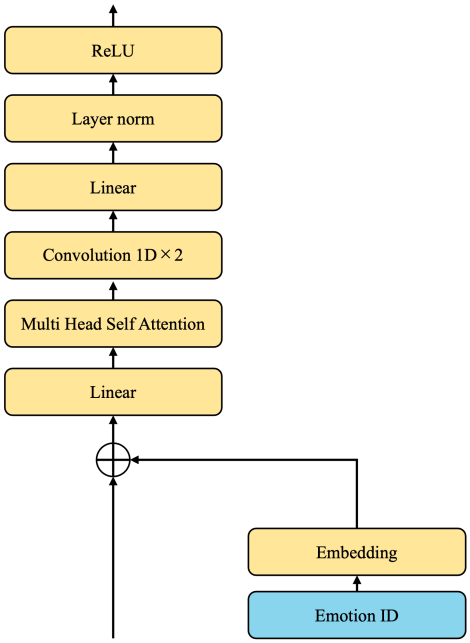

Fig. 3. The architecture of our emotion block

Japanese Kanji characters. This corpus is highly beneficial for training models, as a Japanese TTS system must be able to handle text containing a wide variety of Kanji characters. In contrast, the ITA dataset includes 400 pairs of texts and speech utterances recorded by a single female speaker, with each of the 100 sentences recorded in four different emotional states: neutral, angry, sad, and happy. It is designed to ensure a balanced inclusion of all Japanese phonemes.

As reference test data we selected two sentences uttered with four emotions: neutral, angry, sad, and happy. They were not used in the model training and served as ground truth for the evaluation.

B. Data pre-processing

Seven types of data are required to train our model: phonemes, prosodic symbols, phoneme duration, pitch, energy, mel-spectrogram, and emotion ID. All data except for emotion ID were prepared according to [29]. Emotion ID was extracted directly from the dataset emotion labels. Phonemes, prosodic symbols, and emotion ID are used as input to the model, and the mel-spectrogram, phoneme duration, pitch, and energy are used as targets during the supervised model training.

C. Evaluation metrics

All the models we created are tested using two commonly used evaluation metrics: objective Mel-Cepstrum Distortion (MCD) and subjective Mean Opinion Score (MOS).

MCD is a measure to evaluate the quality of synthesized speech. It quantitatively measures the difference between two utterances. Specifically, this metric is used to evaluate the similarity of the speech spectral features. The MCD is defined as:

$$MCD(X,Y) = \frac{10\sqrt{2}}{\ln 10} \sum_{t}^{T} \sqrt{\sum_{d}^{D} \left(x_t(d) - y_t(d)\right)^2} \quad (1)$$

where
X : output mel-cepstrum
Y : target mel-cepstrum
T : mel-cepstrum length
D : mel-cepstrum dimension
$x_t(d)$: d_{th} output me-cepstrum coefficient in t_{th} frame
$y_t(d)$: d_{th} target mel-cepstrum coefficient in t_{th} frame

The smaller the MCD is, the more similar the two mel-spectrograms are. One drawback of this metric is that it is not correlated with human sound perception in terms of intelligibility and clarity of the synthesized speech.

The Mean Opinion Score is a subjective evaluation metric that doesn't have MCD's problems but is time-consuming and expensive to perform. It is a numerical value of speech quality obtained by averaging scores that listeners assign to each utterance. In our MOS evaluation, listeners are asked to evaluate the test utterances by four criteria. The first is clarity, in which listeners answer how easy is to understand the utterance; the second is naturalness, where they evaluate how human-like the speech is; the third is emotion recognition, in which they are asked to recognize the emotion in the utterance; and the fourth is emotion intensity, in which they need to determine the emotion intensity in the utterance. All the listeners were native Japanese speakers.

D. Baseline models

In our experiments, we trained two additional FastSpeech2 models for performance comparison. The first model is pre-trained with basic5000 dataset and then fine-tuned with the emotional ITA data without any emotion labels. This model is further referred to as "Baseline". The second model is a collection of emotion-specific models obtained by fine-tuning the pre-trained FastSpeech2 on each emotion type data separately. This model we call "Emo-FT". To synthesize an utterance in this case, we just select the model corresponding to the required emotion type. In contrast to our proposed model, these don't have emotion blocks in their structure.

IV. Results

A. MCD evaluation

Since we fine-tuned each of the Emo-FT models with emotion-dependent data, the amount available for training was 4 times less than we used for the proposed model fine-tuning. Therefore, it was possible to overtrain the Emo-TF models using the same number of training steps. Thus, we did an ablation experiment where we changed the training step number to assess its effect on the model's generalization ability. The result in terms of MCD is shown in Fig.4. It is clear that more fine-tuning only worsens the model's

Fig. 4. Average Mel-Cepstrum Distortion (MCD) of the test utterances for Emo-FT model fine-tuned with a different number of steps.

performance. For further comparisons, we selected the model fine-tuned with 30k steps.

Fig.5 summarizes the MCD results of the three types of models we have built. As we anticipated, the baseline model showed the worst result because it does not support emotional conditioning. The Emo-FT model performs much better since for the speech synthesis always the correct emotion model was selected. However, our proposed model achieved the smallest MCD score beating the Emo-FT. We believe that the modified architecture of our proposed model, the larger amount of fine-tuning data, and the bigger number of training steps have contributed to its good performance.

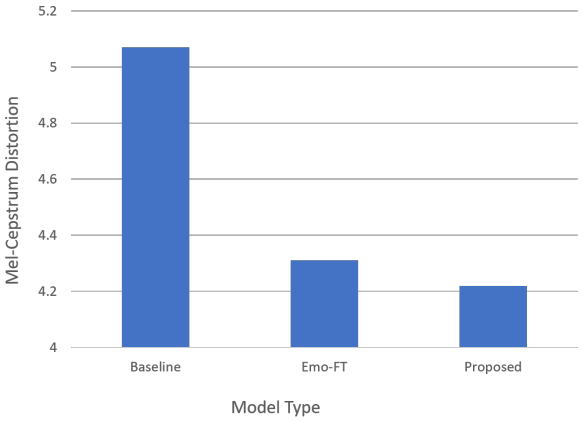

Fig. 5. Average mel-cepstrum distortion of the test utterances for each model

B. MOS evaluation

For the MOS evaluation, we recruited 10 native speakers of Japanese in their 20s. Two listening sessions were held with every listener and the results were averaged over sessions and listeners.

The MOS scores were obtained using clarity, naturalness, and emotion intensity criteria. The listeners' task was to assign a score between 1 and 5 for each utterance they listened to and for each of the three criteria. The synthesized utterances from all the models were pooled together with the ground truth (GT) and presented randomly to each listener. The scores mean and standard deviation are shown in Table I. Naturally, the GT scores were better than the other models in all criteria. The proposed model scores are the second best and exceed four points in emotion intensity. Furthermore, the standard deviation is also smaller than that of the other models.

TABLE I
THE RESULTS OF CLARITY, NATURALNESS AND EMOTION INTENSITY

GT / model	Clarity	Naturalness	Emotion intensity
GT	4.69 $\pm$ 0.23	4.45 $\pm$ 0.24	4.61 $\pm$ 0.27
Proposed	3.52 $\pm$ 0.35	3.20 $\pm$ 0.39	4.01 $\pm$ 0.27
Emo-FT	3.24 $\pm$ 0.5	2.58 $\pm$ 0.57	3.76 $\pm$ 0.76
Baseline	3.42 $\pm$ 0.41	2.78 $\pm$ 0.29	2.38 $\pm$ 0.89

We also did an experiment asking the listeners to identify the emotion in a given utterance. The averaged subjective emotion recognition accuracy for the synthesized speech and the ground truth utterances is given in Fig.6. It was not a

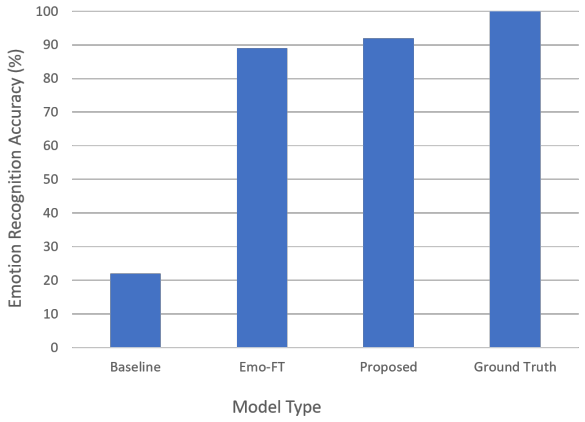

Fig. 6. Emotion recognition accuracy (%) by listeners for ground truth and test utterances synthesized by different models.

surprise that the listeners were able to determine the emotion in ground truth utterances perfectly. The proposed model got more than 90% accuracy, again slightly better than the Emo-FT. Overall, our model achieved the best performance among all the TTS models we evaluated and in most cases was close to the ground truth results.

V. CONCLUSION

In this study, taking the FastSpeech2 as basis we developed an emotional TTS model for the Japanese language. The modifications we introduced include the insertion of two new blocks after the encoder and before the decoder respectively. For stability and robustness, they share parameters. In each block, the emotion class conditioning is implemented through embedding which is combined with the token representations.

Evaluation experiments using objective MCD and subjective MOS metrics showed that our model can synthesize speech close to natural and can convey emotions to such an extent that they are correctly identified in more than 90% of the cases. It performed better than the other models including the baseline FastSpeech2 as well as a collection of emotion-specific models.

ACKNOWLEDGMENT

We would like to thank the listeners for their help in conducting the auditory experiment. Without their participation, we could not have completed our research.

REFERENCES

[1] Vocaloid, "Free your music production," https://www.vocaloid.com.
[2] Softalk, "Reading sentences containing Kanji and English in various voices," https://w.atwiki.jp/softalk.
[3] VoiceVox, "Free, medium-quality text-to-speech and singing voice synthesis software," https://voicevox.hiroshiba.jp.
[4] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.
[5] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.
[6] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He *et al.*, "Naturalspeech: End-to-end text-to-speech synthesis with human-level quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
[7] Y. Zheng, X. Li, F. Xie, and L. Lu, "Improving end-to-end speech synthesis with local recurrent neural network enhanced transformer," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6734–6738.
[8] 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.
[9] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A review of deep learning based speech synthesis," *Applied Sciences*, vol. 9, no. 19, p. 4050, 2019.
[10] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for hmm-based expressive speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 9, pp. 1406–1413, 2007.
[11] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional end-to-end neural speech synthesizer," *arXiv preprint arXiv:1711.05447*, 2017.
[12] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.
[13] 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*. PMLR, 2018, pp. 5180–5189.
[14] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5734–5738.
[15] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7254–7258.
[16] S. Wang, J. Gudnason, and D. Borth, "Fine-grained emotional control of text-to-speech: Learning to rank inter-and intra-class emotion intensities," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
[17] D. Yang, S. Liu, R. Huang, C. Weng, and H. Meng, "Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
[18] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li, "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages," in *Signal and information processing association annual summit and conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–10.
[19] K. Inoue, S. Hara, M. Abe, N. Hojo, and Y. Ijima, "Model architectures to extrapolate emotional expressions in dnn-based text-to-speech," *Speech Communication*, vol. 126, pp. 35–43, 2021.
[20] K. Lee, "Expressive-fastspeech2," https://github.com/keonlee9420/Expressive-FastSpeech2, 2021.
[21] J. Yang, J.-S. Bae, T. Bak, Y. Kim, and H.-Y. Cho, "Ganspeech: Adversarial training for high-fidelity multi-speaker speech synthesis," *arXiv preprint arXiv:2106.15153*, 2021.
[22] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T.-Y. Liu, "Adaspeech: Adaptive text to speech for custom voice," *arXiv preprint arXiv:2103.00993*, 2021.
[23] D. Diatlova and V. Shutov, "Emospeech: Guiding fastspeech2 towards emotional text to speech," *arXiv preprint arXiv:2307.00024*, 2023.
[24] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, T. Qin, and T.-Y. Liu, "Multispeech: Multi-speaker text to speech with transformer," *arXiv preprint arXiv:2006.04664*, 2020.
[25] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6184–6188.
[26] C. Cui, Y. Ren, J. Liu, F. Chen, R. Huang, M. Lei, and Z. Zhao, "Emovie: A Mandarin emotion speech dataset with a simple emotional text-to-speech model," *arXiv preprint arXiv:2106.09317*, 2021.
[27] R. Sonobe, S. Takamichi, and H. Saruwatari, "Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," *arXiv preprint arXiv:1711.00354*, 2017.
[28] Y. Saito, Y. Nishimura, S. Takamichi, K. Tachibana, and H. Saruwatari, "Studies: Corpus of japanese empathetic dialogue speech towards friendly voice agent," *arXiv preprint arXiv:2203.14757*, 2022.
[29] W. Nakata, "Fastspeech2 JSUT implementation," https://github.com/Wataru-Nakata/FastSpeech2-JSUT.