Advancing AI Voice Synthesis: Integrating Emotional Expression in Multi-Speaker Voice Generation

Shivani Sanjay Kolekar, David J. Richter, Md Ilias Bappi, Kyungbaek Kim
Department of Artificial Intelligence Convergence
Chonnam National University
Gwangju, South Korea
shivanikolekar@gmail.com, david_richter@jnu.ac.kr, i_bappi@jnu.ac.kr, kyungbaekkim@jnu.ac.kr

***Abstract*—Advancements in text-to-speech (TTS) synthesis have primarily focused on natural speech and speech intelligibility, but integrating nuanced emotional expressiveness and speaker variability remains a challenge, especially in dynamic environments such as customer service and in assistive speech technologies. This paper introduces a direct text input approach over conventional phoneme-first methods, such as FastSpeech, enhancing user experience. We integrate the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) along with pitch, energy, and duration in the variance adaptor of the FastSpeech 2 model to deepen the emotional expressiveness of speech. In this paper, we propose a Multi-speaker Emotional Text-to-speech Synthesis System (METTS) which allows users to input desired text, select from various speaker voices, and choose emotional tones ranging from happiness to sadness, surprise, neutrality, and anger. Unique to METTS is the feature that allows users to integrate personal voice datasets, making it highly customizable. We assess speech quality and naturalness with the NISQA model, achieving a 3.72 ± 0.78 MOS score for multi-speaker evaluation and 4.09 ± 0.65 for individual speaker voices. The paper details METTS's architecture, enhancements to FastSpeech2, and methods for embedding emotional and speaker variations.**

Index Terms–Medical AI, System Architecture, Microservices, Digital twin

1. Introduction

Text-to-speech (TTS) synthesis, the artificial production of human speech from text input, has undergone a remarkable transformation in recent years, becoming an indispensable tool in a variety of applications, ranging from virtual assistants to accessibility technologies. The core objective of TTS systems is to convert written text into spoken words in a way that is not only accurate but also natural-sounding. A key focus in the development of TTS systems has been enhancing the naturalness and expressiveness of the synthesized speech, aiming to make it more relatable and engaging for users.

In the field of TTS, significant strides have been made with the development of advanced neural network-based models such as Tacotron [1] and FastSpeech [2]. These models represent a leap forward in synthesizing high-quality, natural-sounding speech. Tacotron, a sequence-to-sequence model with attention, has been notable for its ability to produce highly natural speech. However, it requires considerable computational resources, especially in terms of processing time, which can be a limiting factor in real-time applications. On the other hand, FastSpeech emerged as a solution to the real-time speech synthesis challenge. By decoupling text analysis and speech synthesis, FastSpeech can generate speech significantly faster than real-time, making it more suitable for scenarios where inference speed is crucial. Despite these advancements, both models have their limitations, particularly when deployed in high-load environments such as large social networks or customer service systems where rapid response times are critical. Additionally, a significant challenge that persists in current TTS technologies, including Tacotron and FastSpeech, is their limited capability in delivering emotionally expressive speech and in adapting to diverse speaker voices, particularly in user-specific contexts.

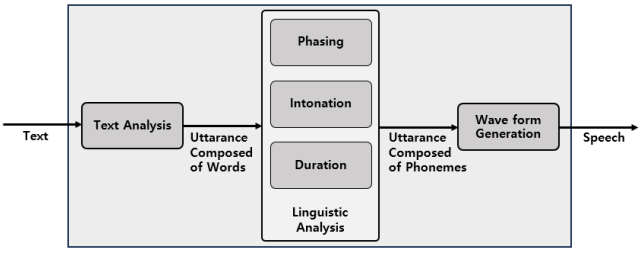

Figure 1. General TTS Reference Architecture.

In this paper, we introduce METTSpeech, a novel FastSpeech2-based TTS system, distinguished by its integration of the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [3]. eGeMAPS plays a crucial role in our system by providing a comprehensive set of acoustic parameters specifically designed for effective and relevant emotion feature extraction. This allows our TTS system to capture and reproduce a wide range of emotional states in speech, significantly enhancing the emotional expressive-

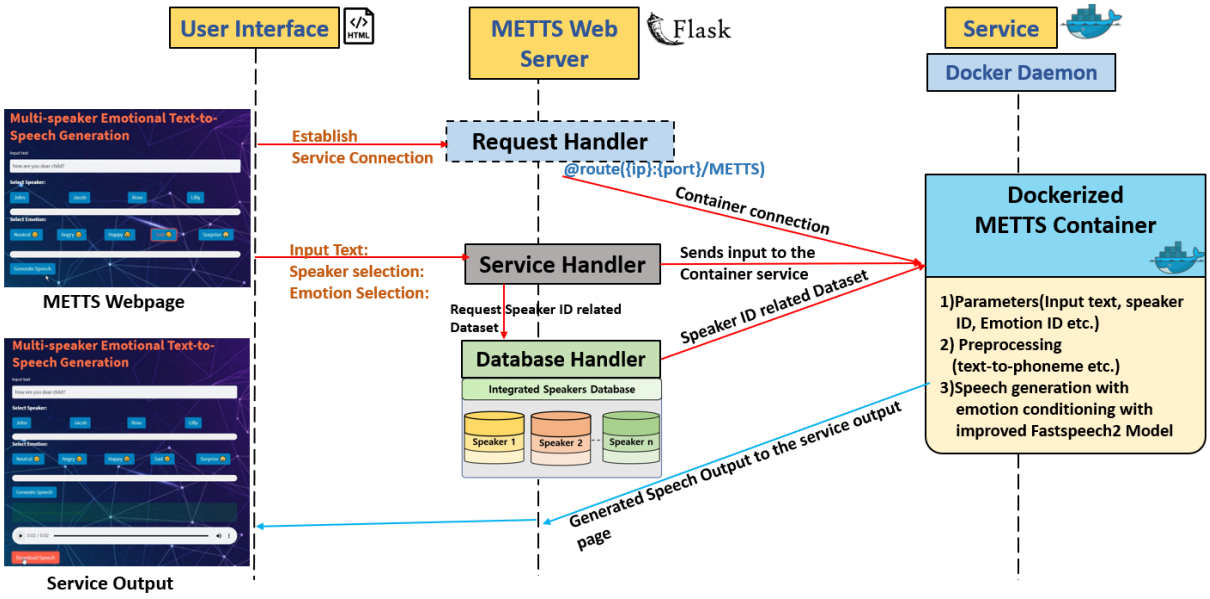

Figure 2. METTS Request-Response processing sequence diagram.

ness and naturalness of the output. Our proposed system also offers customizable user-specific voice style integration, enabling users to train their personal voice datasets with our Multi-speaker Text-to-Speech System (METTS). We focus on providing fast processing speed for voice synthesis by implementing istftnet melspectrogram vocoder [4] in METTSpeech architecture which reduces redundant estimations of high-dimensional spectrograms. This feature, combined with the proposed METTSpeech model, ensures that our system not only maintains high-quality speech synthesis but also excels in processing speed, making it ideal for demanding environments. By facilitating direct text input and leveraging eGeMAPS for advanced emotional feature extraction, our model innovatively enriches the speech synthesis process. This advancement extends the potential applications of TTS systems, accommodating personalized and emotionally nuanced voice interactions in areas ranging from tailored customer service to bespoke content creation on social media platforms.

2. Related Work

In recent years, machines have managed to master the art of generating speech that is understandable by humans. However, the linguistic content of an utterance encompasses only a part of its meaning. Affect, or expressivity, has the capacity to turn speech into a medium capable of conveying intimate thoughts, feelings, and emotional—aspects that are essential for engaging and naturalistic interpersonal communication [5]. Current state-of-the-art deep learning methodologies focus on development of TTS systems and have been enhancing the naturalness and expressiveness of the synthesized speech.

Statistical parametric speech synthesis (SPSS) [5] adopts the three-stage model presented in Fig. 1 (inspired from [6]), namely, the use of text analysis to suitable linguistic representations of the target utterance, the prediction of speech parameters using an acoustic model, and the final waveform synthesis (vocoding). In particular, the text analysis module includes necessary preprocessing steps (text normalization, graphemeto-phoneme conversion etc.) followed by the extraction of all relevant features which are composed in lingustic analysis, such as phonemes, duration, or part-of-speech tags. Those features, along with the accompanying speech parameters, are fed to a statistical Machine Learning (ML) model that learns a mapping from linguistic to acoustic feature wave form generation (e.g., the fundamental frequency, spectrum, or cepstrum), finally generating the speech.

2.1. Emotional Speech Synthesis (ESS) Pipeline

A traditional approach of TTS system follows three steps:

1) a text analysis module that converts the input text to appropriate linguistic features
2) an acoustic model that converts those features to acoustic features
3) a vocoder, which generates the final utterance [6]

Incorporating emotions into this pipeline is primarily done in two ways: either an emotional voice conversion module is tasked with adapting the emotion of the synthesized speech, or the transformation is made as an intermediate step before vocoder processing [7]. Following a similar trend as TTS, ESS (Emotional Speech Synthesis) transitioned to a data-driven paradigm with the advent of SPSS. In this context, ESS is primarily envisioned as an intervention on acoustic features before the vocoding step: the relevant features

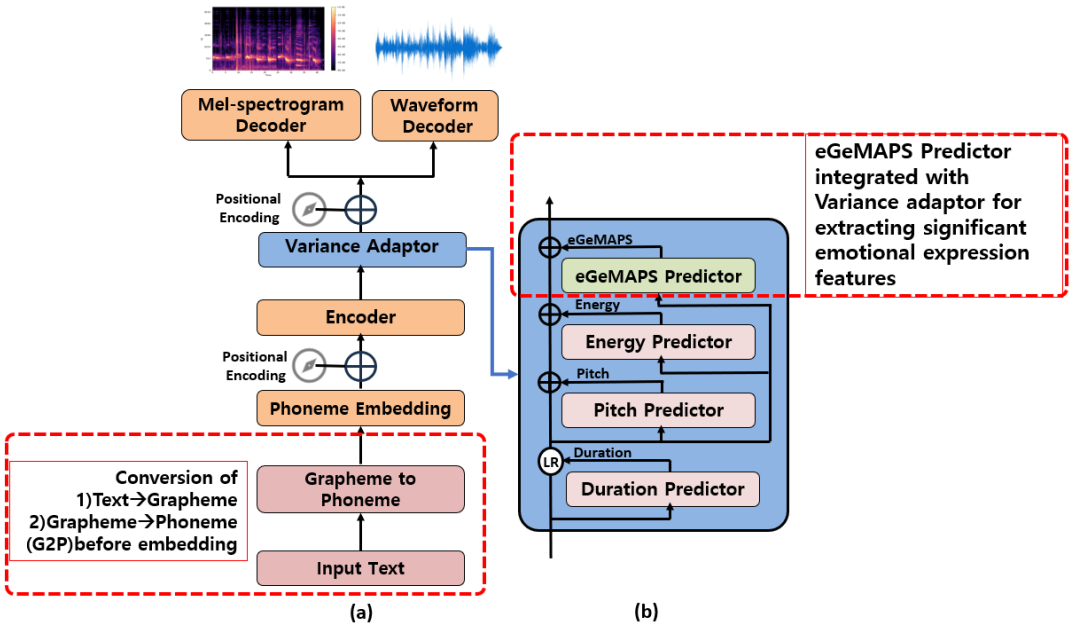

Figure 3. The overall architecture for METTSpeech model (a) with Direct text Input and (b) Integrated eGeMAPS predictor in Variance Adaptor. LR in subfigure (b) denotes the length regulator

would be mapped to their emotional equivalents before being used to synthesize speech. In particular, mappings between both prosodic and spectral features were learned using data Deep neural network-based synthesis employ neural networks as the model of choice to substitute one or more components of a traditional emotional speech synthesis pipeline. Initially, sequential models (Recurrent Neural Networks (RNNs) or long short-term memory networks (LSTMs)) were used for acoustic modeling, such as the early DeepVoice systems [8]. WaveNet was the first DL model to directly generate the waveform from linguistic features [9]. In Tacotron [1] which is a auto-regressive model,WaveNet [10] is included as an attempt to directly synthesize phoneme sequences to speech inference. Although, in case of high-load environment solutions such as large social networks where inference speed is essential, FastSpeech 2 [11] a non-autoregressive TTS model, shows higher accuracy in terms of Mean Opinion Score (MOS).

2.2. FastSpeech 2 model

FastSpeech2 functions as a non-autoregressive acoustic model, designed for swift and high-fidelity speech synthesis. The model processes a series of input tokens to produce mel spectrograms. These spectrograms are subsequently converted into waveforms using a vocoder. The primary elements of FastSpeech2 include its encoder, variance adapter, and decoder. Each component of the model serves a specific purpose:

- *The Encoder* processes textual information, extracting features that determine the content of speech.
- *The Variance Adaptor* then enriches this input sequence with acoustic properties and timing details.
- *The Decoder* synthesizes all this information to produce the mel spectrogram features.

Both the encoder and decoder consist of a feed-forward transformer block, incorporating a series of multi-head self-attention layers and 1D-convolution.

Existing TTS models inlcuding FastSpeech2 are limited while providing enhanced quality of speech when it comes to multi-speaker speech generation with emotional expression. Along with this, these models are not user-friendly considering the phoneme input is directly provided. In our proposed METTSpeech system we intend to overcome these limitations of previous models.

3. METTSpeech Model Formulation

In this paper, we propose the METTSpeech model, which incorporates and adapts the FastSpeech 2 model with significant modifications such as increased feature predictors(eGeMAPS) in the variance adopter and direct text to speech conversion availability for efficient usability. Our proposed METTS system is an end-to-end multi-speaker text-to-speech system which provides a novel platform for users to convert desired text into natural sounding user-specific voice.

The base model for METTSpeech is FastSpeech2 [11]. We make several modifications to FastSpeech2 to achieve our proposed system. In this section, first we discuss about the FastSpeech2 model, then we explain the significant changes made to achieve METTSpeech.

The MettSpeech architecture incorporates an encoder, variance adaptor and decoder which are described as follows. An encoder converts a token embedding sequence

into the token's hidden representation of $h \in \mathbb{R}^{n\times \text{hid}}$ and output the pitch, energy, and durations (p, e, d) for each token. Then, the length regulators "upsample" $h \in \mathbb{R}^{n\times \text{hid}}$, accumulated with $p, e \in \mathbb{R}^n$ according to $d \in \mathbb{R}^n$. Token duration is measured by the number of mel spectrogram frames, which leads to length regulator output $h \in \mathbb{R}^{m\times \text{hid}}$, where $m = \sum_{i=0}^{n} d_i$. Later, the upsampled tokens are passed through a decoder, and the hidden dimension is reflected in mel channels using a linear layer. The final output is a predicted mel spectrogram $h \in \mathbb{R}^{m\times c}$. The model learns to generate a mel spectrogram from the input text sequence using reconstruction loss:
$L_{\text{rec}} = ||y - \hat{y}|| + ||d - \hat{d}||^2 + ||e - \hat{e}||^2 + ||p - \hat{p}||^2$, where $\hat{y}, \hat{d}, \hat{e}, \hat{p}$, are predicted mel spectrogram, duration, pitch, and energy [12].

3.1. Emotion Conditioning

Embedding lookup tables are employed in the development of METTSpeech from FastSpeech2 to facilitate initial speaker and emotion conditioning. By merging speaker and emotion embeddings, we create a conditioning vector c. This method of concatenation enables the generation of 50 distinct embeddings from an original set of 15 in the lookup table, enhancing the conditioning process. Our initial alterations involve incorporating the conditioning vector c into the encoder's output, which is then supplied to the variance adaptor.

3.2. eGeMAPS Predictor

The variance adaptor is extended by integrating additional parameters. The architecture of METTSpeech model is referred from FastSpeech2. The derived modifications including the variance adoptor integrated with eGeMAPS predictor can be observed in Fig. 3. In speech synthesis with METTSpeech, we incorporate an eGeMAPS predictor (EMP) into the variance adaptor to ascertain k parameters from the Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). Although eGeMAPS includes 88 low-level descriptors commonly utilized in diverse voice analysis tasks, we focus on those most pertinent for conveying emotion. By analyzing all eGeMAPS features within the English subset of the ESD dataset, and utilizing a CatBoost classifier for emotion classification, we strive to isolate the most influential features for emotional expression. This selection method is aimed at enriching the speech output with low-level descriptors that significantly correlate with the emotional states we aim to synthesize.

3.3. Conditional Layer Normalization

Due to the inherent difficulty in accurately estimating speaker characteristics for new, unseen speakers, using imprecise speaker representations as input to the decoder can result in a disparity between the training of the source model and zero-shot synthesis. Consequently, we investigate an enhanced conditioning method that utilizes speaker representation as model input, aiming to boost the TTS model's generalizability in zero-shot situations.

Additionally, we consider the application of meta-learning techniques to fine-tune our model on a small subset of data from unseen speakers, potentially improving its ability to generalize from limited information. This adaptation, aimed at refining the model's performance on novel speaker voices, is an added component to our methodology to further mitigate the issues of speaker variability in zero-shot TTS tasks.

The scale and bias vectors in layer normalization are determined by leveraging a small conditional network to process the extracted speaker representation. Specifically, a linear layer W^γ for scale and W^β for bias are employed, referring to conditional layer normalization (CLN) in AdaSpeech4 [13]. These layers, utilizing the extracted speaker representation E as input, output adaptive scale and bias vectors as follows:

$$\gamma = E \times W^\gamma, \quad \beta = E \times W^\beta. \tag{1}$$

For the purpose of emotional speech synthesis, we replace the standard layer normalization in both the self-attention and feed-forward networks with this conditional layer normalization.

To enhance the fidelity of the synthesized speech, we adopt an adversarial training approach similar to that employed by GANspeech [14], [15], integrating it into the METTS framework. The JCU (joint conditional and unconditional loss [16]) discriminator's conditional architecture, is well-aligned with the requirements of our system that accommodates multiple speakers and a range of emotions. Throughout a unified training phase, both the discriminator and METTS are concurrently trained. The conditioning discriminator utilizes a composite embedding c, which merges both speaker and emotion embeddings, mirroring the conditioning strategy of the generator.

4. Evaluation

4.1. Data Preprocessing

To train the METTSpeech model, we focus on providing a lightweight solution for speech synthesis with multiple speakers and a fixed set of emotions, so we opt for the Emotional Speech Dataset (ESD) [17].

The dataset boasts a diverse array of speakers and encompasses a broad lexical scope. Comprising 350 spoken expressions by 10 different speakers and encapsulating five distinct emotional states — Neutral, Angry, Happy, Sad, and Surprised — the collection amasses a total of 1750 spoken expressions per speaker. The dataset features a comprehensive word count of 11,015. The vocabulary of the utterances is varied, as is the tonal expression within the sentences. Accompanying each audio file is a textual transcript and a single label denoting the expressed emotion.

The dataset is split into training, testing and validation subsets, whereas the validation and test subsets consists of

20 and 30 utterances, each consisting of five emotions and ten speakers (totalling 1000 and 1500 utterances for testing and validation respectively).

For Feature extraction, we consider utterance, text transcript, phonemes, mel-spectrograms, durations, pitch and other acoustic features such as eGeMAPS. The Grapheme-to-phoneme (G2P) model from the Montreal-forced-aligner (MFA) [18] toolkit is used to extract phonemes, duration of extracted phonemes, punctuation and silence tokens from text annotations. To extract pitch from ground truth waveforms, the pyworld library was employed. The energy feature was extracted by normalizing spectrograms by frequency dimension and finally eGeMAPS features were extracted using the openSMILE toolkit [19].

4.2. Model Configuration and Performance

In this paper, we propose the METTSpeech model architecture, while employing analogous hyperparameters, we have customized specific parameters: the phoneme embedding, along with the encoder and decoder hidden dimensions, are adjusted to 512, and the Conv1D filter sizes for both encoder and decoder are also set to 512. The encoder and decoder are constructed with six layers each. The embeddings for speaker and emotion have a hidden dimension of 256. For the eGeMAPS predictor, which is integrated into the Variance Adaptor block, we maintain consistency with the architecture of existing predictors: it comprises two 1D convolution layers with a kernel size of 3 and stride of 1, accompanied by a ReLU activation function and a dropout rate of 0.5. The iSTFTNet vocoder, trained on the English portion of the ESD database. Mel-spectrograms were derived from waveforms using a filter length of 48 milliseconds, a hop length of 12 milliseconds, and encompassing 80 mel frequency channels inspired from EmoSpeech [12].

The training was done on a Nvidia GeForce RTX 3090 graphics card, accommodating a collective batch size of 256 for a total of 25,000 training steps. For optimization and scheduling the protocol was aligned with FastSpeech2. Training sessions were orchestrated applying the Adam optimization algorithm with a learning rate of 0.0001.

We observe the Mean Squared Error(MSE) loss [20] value for extracted features while training. In Fig. 4, The MSE loss trend is depicted for Mel-spectrograms, pitch, energy, eGeMaps and duration features for 25,000 training steps. We can observe a decreasing trend for all features. At the final step,the average MSE loss values are: Mel-spectrograms(0.58), pitch (0.018), energy(0.022), eGeMaps(0.065) and duration(0.010).

We evaluated the performance by calculating Mean Opinion Score (MOS) with the NISQA-TTS model [21] and predicted a MOS score on a 5-point scale (higher score denotes superior perceived quality) instead of the traditional human-based Mean Opinion Score evaluation, which requires extensive human resources. The NISQA-TTS model provides efficiency and consistency in automatically assessing the naturalness in emotion transmission for our proposed system. We compared the performance of the proposed METTSpeech model with the general FastSpeech2 model. We compare the MOS value of original utterances (ground truth) and utterances generated by these two AI models.The result can be observed in Table 1.

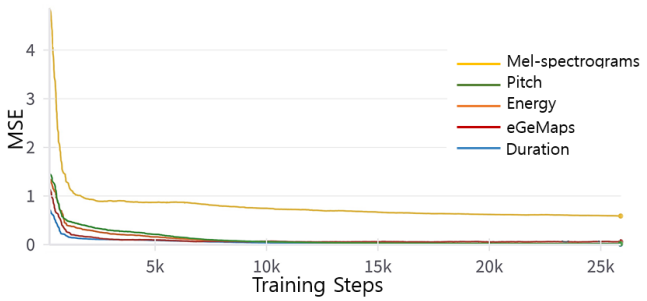

Figure 4. MSE loss trends for synthesized speech feature prediction over training steps

TABLE 1. THE COMPARISON OF MOS VALUE CALCULATED WITH NISQA.

Model	MOS (NISQA)	
	Multi-speaker	Individual Speaker
FastSpeech2	3.12±0.6	3.76±0.78
METTSpeech (Proposed)	3.72±0.78	4.09±0.65

Our model achieved a 3.72 ± 0.78 MOS score for multi-speaker (all different speakers in the dataset) evaluation and a 4.09 ± 0.65 for individual speaker voices, whereas FastSpeech2 showed 3.12 ± 0.6 and 3.76 ± 0.78 . The multi-speaker MOS score reflects the quality and naturalness of the synthesized voices when considering the variation and diversity across all speakers as a whole which may lead to lower a MOS value than in the individual speaker scope. To understand and improve the quality of the multi-speaker score, different speaker styles (accents etc.) of all speakers considering slightest differences needs to be considered for feature extraction.

5. METTS Web Server Integration

To create an end-to-end system, the METTS web server was created. We divided the request-response flow into three main modules, namely: the user interface, the METTS web server and the dockerized service as shown in Fig. 2. We deploy the METTS docker container which consists of a preprocessing method for text to phoneme conversion and speech generation with an emotion conditioning module using the METTSpeech Model. Users can generate desired speech by providing 3 input parameters, including - typing the text input, selecting the speaker voice (extracted from set of ESD dataset speakers) and emotion choice out of 5 emotions - neutral, angry, happy, sad and surprised. The input parameters are shared with the service handler which proceeds to get the selected speaker related information from the database and transfers it to docker service. The

response from the METTS service after the speech synthesizing process is received by the web server which transfers the generated output to the user interface. We allow the synthesized audio to be downloaded by the user for ease of usability. Since we focus on providing a lightweight model, the inference time is observed as average 3 seconds per sentence.

6. Conclusion and Future Work

In this paper, AI Voice Synthesis integrating Emotional Expression and a Multi-Speaker Voice Generation system has been proposed. The system envelops the METTSpeech model based on FastSpeech2. Evaluation is conducted by calculating the MOS value performance. The proposed system shows enhanced naturalness in the quality of inference. The implemented system successfully synthesizes text into emotion embedded speech, while the naturalness can be improved further by training on larger datasets with extensive speaker styles and vocabulary. In the future, we intend to explore range of English accents, speaker styles and evaluate the system with multiple language datasets.

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program(IITP-2023-RS-2022-00156287) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government(MSIT).

References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[2] 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*, vol. 32, 2019.

[3] W. Treemongkolchok, P. Punyabukkana, D. Wanvarie, and P. N. Pratanwanich, "An analysis of acoustic features for attention score in thai moca assessment," in *2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2022, pp. 1–6.

[4] T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, "Istftnet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6207–6211.

[5] A. Triantafyllopoulos, B. W. Schuller, G. İymen, M. Sezgin, X. He, Z. Yang, P. Tzirakis, S. Liu, S. Mertes, E. André, R. Fu, and J. Tao, "An overview of affective speech synthesis and conversion in the deep learning era," *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1355–1381, 2023.

[6] I. Isewon, J. Oyelade, and O. Oladipupo, "Design and implementation of text to speech conversion for visually impaired people," *International Journal of Applied Information Systems*, vol. 7, no. 2, pp. 25–30, 2014.

[7] S. Lei, Y. Zhou, L. Chen, Z. Wu, S. Kang, and H. Meng, "Towards expressive speaking style modelling with hierarchical context information for mandarin speech synthesis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7922–7926.

[8] S. Ö. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *International conference on machine learning*. PMLR, 2017, pp. 195–204.

[9] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[10] 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.

[11] 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.

[12] D. Diatlova and V. Shutov, "Emospeech: Guiding fastspeech2 towards emotional text to speech," *arXiv preprint arXiv:2307.00024*, 2023.

[13] Y. Wu, X. Tan, B. Li, L. He, S. Zhao, R. Song, T. Qin, and T.-Y. Liu, "Adaspeech 4: Adaptive text to speech in zero-shot scenarios," *arXiv preprint arXiv:2204.00436*, 2022.

[14] J. Yang, J. Lee, Y. Kim, H. Cho, and I. Kim, "Vocgan: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network," *arXiv preprint arXiv:2007.15256*, 2020.

[15] 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.

[16] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018.

[17] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 920–924.

[18] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[20] Y. Lee, J. Yang, and K. Jung, "Varianceflow: High-quality and controllable text-to-speech using variance information via normalizing flow," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7477–7481.

[21] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *arXiv preprint arXiv:2104.09494*, 2021.