Interspeech 2024
1-5 September 2024, Kos, Greece

H4C-TTS: Leveraging Multi-Modal Historical Context for Conversational Text-to-Speech

Donghyun Seong[1], *Joon-Hyuk Chang*[1*]

[1]Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea
{sdh2382, jchang}@hanyang.ac.kr

Abstract

Conversational text-to-speech (TTS) aims to synthesize natural voices appropriate to a situation by considering the context of past conversations as well as the current text. However, analyzing and modeling the context of a conversation remains challenging. Most conversational TTS use the content of historical and recent conversations without distinguishing between them and often generate speech that does not fit the situation. Hence, we introduce a novel conversational TTS, H4C-TTS, that leverages multi-modal historical context to realize contextually appropriate natural speech synthesis. To facilitate conversational context modeling, we design a context encoder that incorporates historical and recent contexts and a multi-modal encoder that processes textual and acoustic inputs. Experimental results demonstrate that the proposed model significantly improves the naturalness and quality of speech in conversational contexts compared with existing conversational TTS.

Index Terms: Text-to-speech, conversational speech synthesis, multi-modal

1. Introduction

The rapid advancement of deep-learning technologies in recent years has led to significant developments in the field of speech synthesis [1]. Notably, expressive speech synthesis technology [2–5] has advanced significantly through training on a multi-emotional speech corpus, which includes multiple speakers and different emotions, making it increasingly challenging to differentiate it from human voices. Furthermore, conversational text-to-speech (TTS) aims to achieve more complex human-computer interaction (HCI) capabilities, focusing on appropriate emotional expression and natural dialogue in context. Such conversational TTS can be applied in a variety of fields, including voice interfaces, virtual assistants, interactive games, and conversational educational systems. Guo *et al.* [6] have modeled a GRU-based context by extracting semantic information from a historical context using a Tacotron2 model [7]. Nishimura *et al.* [8] utilize cross-modal attention to capture both long-term linguistic and prosodic contextual information, thereby enhancing the system's understanding of the conversational flows. M2-CTTS [9] improves natural voice generation by global and local context modeling of the historical context at both the utterance and phoneme levels. Meanwhile, graph-based approaches [10, 11] have been proposed to capture multi-scale context dependencies among different modalities, further enriching the contextual awareness of TTS systems. However, existing conversational TTS does not differentiate historical and recent contexts by considering them equally. This approach may fail to accurately reflect the recent situations or changes, potentially generating voices that do not align with the current context.

In this study, we introduce a novel context module architecture that leverages a multi-modal historical context for conversational TTS. The context module architecture comprises two fundamental components: a multi-modal encoder and a context encoder. First, the multi-modal encoder is composed of linguistic and acoustic encoders to utilize text and audio, respectively. Initially, the linguistic encoder processes historical text, capturing the meanings of words, grammatical structures, and contextual significance. Subsequently, the acoustic encoder extracts features from the historical audio, reflecting the emotional state or intent of the speaker. Second, the context encoder comprehensively analyzes the multi-modal information extracted from previous conversational content, providing the necessary information to understand the context of a recent utterance. This enables the identification of the relationship between past and recent conversations, enabling the synthesis of natural voices that align with the flow of conversations. In addition to the aforementioned methods, mel decoder with mixed context layer normalization (MCLN) and flow-based post-net with context are employed to enhance the context representation. Consequently, the proposed model has demonstrated its effectiveness in generating conversational speech that leverages a multi-modal historical context.

2. Background

2.1. Conversational dataset

TTS datasets are generally composed of a single speaker, multiple speakers, or a variety of emotions. For instance, the single speaker dataset LJSpeech [12], comprises recordings from a female voice actor and is widely used in TTS. Multi-speaker datasets such as VCTK [13] and LibriTTS [14] include voices from various speakers, enabling TTS models to learn different pronunciations and intonations. The multi-emotion dataset IEMOCAP [15] adds an emotional dimension to voice data, thereby enabling the development of TTS systems capable of reflecting emotions. These datasets are primarily composed of text and corresponding voices and are mainly used for TTS training. On the other hand, conversational datasets are used to develop conversational artificial intelligence (AI) systems, especially chatbots and interactive voice recognition systems. These datasets extend beyond simple text-audio pairs to include the flow of conversation, context, interactions between participants, and the intentions and emotions within the dialogue. Conversational datasets involve interactions between two or more participants, through which machines can learn to understand the context of a conversation and respond appropriately. They may also encompass specific situations or contexts in which the con-

[*] Corresponding author.

10.21437/Interspeech.2024-1480

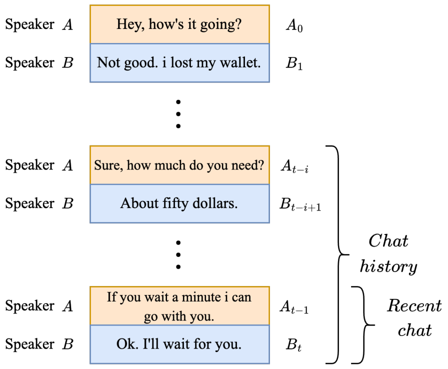


Figure 1: *The structure of a conversational dataset in DailyTalk.*

versation takes place, helping AI to generate suitable responses based on the situation. Examples of such datasets include the open-source Chinese conversational speech corpus[1], DailyDialog [16], and DailyTalk [17], which are structured in a dialogic format as shown in Figure 1. These datasets encompass dialogues across various scenarios and contexts, facilitating conversational TTS training.

2.2. Baseline TTS model

We adopt the FastSpeech2 (FS2) model [18], which is a non-autoregressive approach, as the baseline TTS model. The FS2 model comprises a text encoder, variance adaptor, mel decoder, and post-net. The text encoder uses the phoneme embeddings as inputs and generates hidden states h_p. The variance adaptor comprises a pitch predictor that predicts the pitch $\hat{p}$, an energy predictor that predicts the energy $\hat{e}$, a duration predictor that predicts the duration $\hat{d}$, and a length regulator that expands the phonemes to their corresponding duration lengths. To train the duration predictor, duration information corresponding to phonemes is extracted using the Montreal forced aligner (MFA) [19] as the target. However, the method requires retraining the MFA for accurate duration prediction, and if the duration information is not correctly predicted, it can lead to inaccurate mel generation. Therefore, we employ the unsupervised duration model [20] instead of an MFA to predict the durations, enabling duration learning during training in an end-to-end manner. Moreover, we use Gaussian upsampling [21] instead of vanilla upsampling with a length regulator to extend the phonemes to the frame level, which helps produce more natural-sounding speech. To train the pitch predictor, we extract the target pitch p by using PyWorld[2]. The mel decoder then transforms these frame-level representations into a mel spectrogram $\hat{y}$.

Consequently, the baseline TTS loss L_{fs2} is composed of L1 and L2 losses between the predicted values and the ground truth:

$$L_{var} = \left\| d - \hat{d} \right\|_2 + \left\| p - \hat{p} \right\|_2 + \left\| e - \hat{e} \right\|_2, \tag{1}$$

$$L_{fs2} = \left\| y - \hat{y} \right\|_1 + \lambda_{var} L_{var}, \tag{2}$$

where we use 1 for λ_{var} as the scaling factor of the variance adaptor, $\left\| \cdot \right\|_1$, and $\left\| \cdot \right\|_2$ denote the L1 and L2 losses, respectively.

3. Proposed method

In this section, we introduce the H4C-TTS by leveraging a multi-modal historical context for conversational TTS. The proposed model comprises a text encoder, variance adaptor, context module architecture, mel decoder with mixed context layer normalization (MCLN), and a flow-based post-net with context. The overall architecture is illustrated in Figure 2.

3.1. Context module architecture

3.1.1. Multi-modal encoder

Conversational data structure: We divide the conversational dataset into $chat\ history$, which contains the overall context of the conversation, and $recent\ chat$, which focuses on the latest conversations. We denote the $chat\ history$ and the $recent\ chat$ at the time stamp t as:

$$\begin{aligned} Chat\ history &= \{A_{t-i}, B_{t-i+1}, ..., A_{t-1}, B_t\}, \\ Recent\ chat &= \{A_{t-1}, B_t\}, \end{aligned} \tag{3}$$

where i is a parameter that determines the number of turns considered between the two speakers A and B.

Acoustic encoder & Linguistic encoder: To utilize text and audio from different tasks, we design a multi-modal encoder that contains acoustic and linguistic encoders. Although methods such as global style tokens (GST) [4] and vector quantization variational autoencoder (VQ-VAE) [22] have been used for extracting acoustic context embeddings from audio, we employ the Wav2vec 2.0 model [23], which is fine-tuned for the downstream task of speech emotion recognition (SER), to directly obtain acoustic context information at the utterance level from audio. The acoustic encoder uses a gated recurrent unit (GRU) layer to encode both the acoustic history embeddings $H^W_{t-i:t}$, which contain $chat\ history$, and the recent acoustic embeddings $H^W_{t-1:t}$, which contain information from the more $recent\ chat$.

To extract textual context embeddings from the text, we adpot sentence bidirectional encoder representations from transformers (BERT) [24] to derive semantic information at the utterance level from the text. Sentence BERT is specifically designed to capture the semantic similarity between sentences, aiding in generating speech that is suitable for conversational contexts. The linguistic encoder uses a GRU layer to encode both the linguistic history embeddings $H^T_{t-i:t}$, which contain information from $chat\ history$, and the recent linguistic embeddings $H^T_{t-1:t}$, which contain information from more $recent\ chat$.

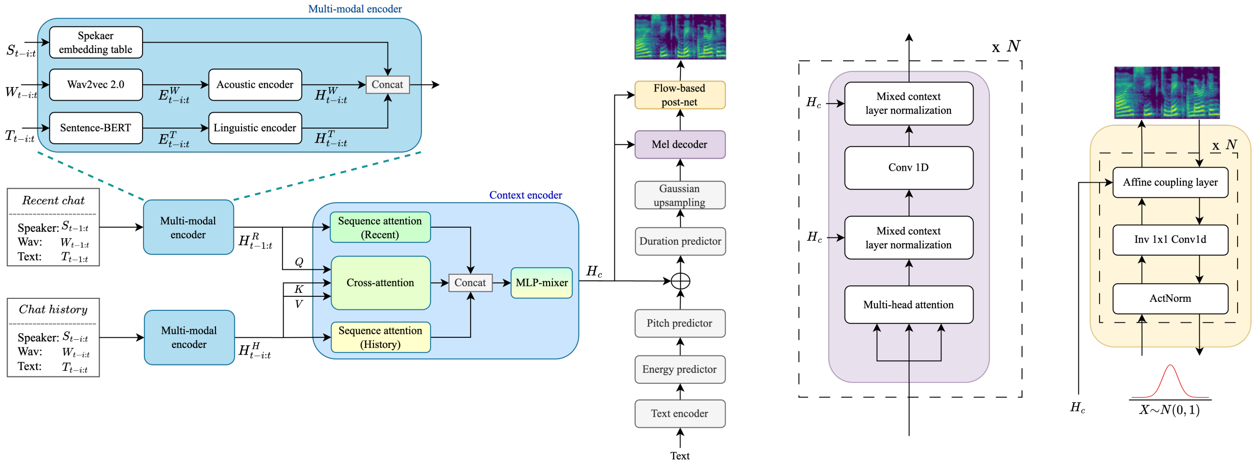
3.1.2. Context encoder

Sequence attention: We exploit a sequence attention mechanism to handle data across different time steps. It computes attention weights to determine the importance of each element within a sequence. Generally, it involves calculating the attention scores for each time step of the input sequence and then normalizing the scores using a softmax function to convert them into weights. These weights are then utilized to compute the weighted average of each element in the input sequence, allowing for the calculation of recent context embeddings $H^R_{t-1:t}$ and history context embeddings $H^H_{t-i:t}$ to enhance the model's understanding of temporal dynamics.

Cross attention: We employ a cross-attention mechanism using scaled dot-product attention [25] to effectively model both recent and historical contexts. The recent context embeddings $H^R_{t-1:t}$ are utilized as the query, whereas the historical context

[1] https://magichub.com/datasets/mandarin-chinese-conversational-speech-corpus-mutiple-devices/

[2] https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder



(a) Proposed model

(b) Mel decoder

(c) Flow-based post-net

Figure 2: *Overall architecture for the proposed model.*

embeddings $H^H_{t-i:t}$ serve as both the key and value. We calculate the attention scores by measuring the similarity between the query and each key, then normalize these scores using a softmax function and apply the resulting weights to each value, summing them up afterward. Through this cross-attention mechanism, we can discern the information within the historical context that is most relevant to the recent context, enabling a more nuanced modeling of the recent context.

MLP-mixer: Upon concatenating the outputs of the sequence attention (recent, history) and the cross attention, we employ an MLP-mixer [26] to generate context vector H_c that is modeled from both recent and historical contexts. The token mixer layer aggregates and blends information across various positions within the context vector, learning the spatial correlations and patterns. However, the channel-mixing layer models the interactions among different channels at the same position, integrating channel-specific information and enabling a deeper understanding of the features.

3.2. Mel decoder with mixed context layer normalization

The mel decoder is composed of feed-forward transformers (FFTs), that contain multi-head attention, 1D convolution, and layer normalization. Style adaptive layer normalization (SALN) [5] has been proposed to effectively capture and reflect style variations. Unlike traditional layer normalization, which uses fixed gain and bias, SALN uses a style vector as the gain and bias, allowing the input features to be scaled and shifted based on style, achieving normalization that is sensitive to the style context. On the other hand, mix style layer normalization (MSLN) [27] achieves the normalization effect on style by shuffling style information in SALN, using it as gain and bias to prevent overfitting. Therefore, we design a mixed context layer normalization (MCLN) to effectively incorporate the context. The MCLN uses the context vector H_c generated by the context module as a condition, as follows:

$$\gamma_{mix}(H_c) = \lambda\gamma(H_c) + (1-\lambda)\gamma(\tilde{H}_c), \tag{4}$$

$$\beta_{mix}(H_c) = \lambda\beta(H_c) + (1-\lambda)\beta(\tilde{H}_c), \tag{5}$$

$$MCLN(x, H_c) = \gamma_{mix}(H_c)\frac{x-\mu}{\sigma} + \beta_{mix}(H_c), \tag{6}$$

where $\tilde{H}_c$ denotes the shuffled context vector, x denotes the input feature vector, μ and σ represent the mean and variance, respectively, and λ is sampled from the beta distribution [28]. The MCLN can dynamically adjust input features based on a given context, thereby enhancing the naturalness and expressiveness of the synthesized speech.

3.3. Flow-based post-net with context

The effectiveness of normalizing flows [29, 30] has been proven to enhance the quality of the mel spectrograms. Although generative models based on the VAE tend to produce blurry outputs in images, flow-based models address the smoothing problem, thereby enabling the generation of images that are closer to reality. To generate a mel spectrogram with complex distributions, we design a flow-based post-net conditioned on the output of a context module using Glow [31]. During training, the post-net transforms the mel spectrogram into a latent prior Gaussian distribution and calculates the log-likelihood. In the inference, a latent variable is sampled from the latent prior distribution and conditioned on the context embedding H_c, and the inverse process of the post-net is carried out to generate a high-quality mel spectrogram.

4. Experiment

4.1. Setup

For our experiments, we utilized the publicly available English corpus, DailyTalk [17]. The DailyTalk dataset contains 2,541 dialogues, each featuring one male speaker and one female speaker. For convenience in training, these dialogues were segmented into 23,773 audio clips by dialogue turns, totaling 20 hours of audio. We excluded dialogues with fewer than 5 turns from our dataset because of the difficulty in capturing the contextual meanings in short conversational contexts. The dataset was divided into 128 dialogues for the validation set, 15 dialogues for the test set, and others for the training set. To process the data, the raw waveforms were converted into 80-dimensional mel spectrograms, setting the frame and hop sizes to 1024 and 256, respectively, at a sampling rate of 22,050 Hz.

For experimental comparison, we prepared the four TTS models: Tacotron2 [7], FastSpeech2 [18], DailyTalk [17], and the proposed model. All models, including the proposed one, were trained with a batch size of 8 and a learning rate of 1×10^{-9}, using the Adam optimizer [32] on two NVIDIA RTX 2080TI GPUs. We employed a HiFi-GAN [33] for the vocoder, which was trained with a batch size of 16 on two NVIDIA RTX 2080Ti GPUs.

We utilized the Sentence BERT pretrained model[3] from sentence transformers to extract semantic features from the text. To extract acoustic features from the audio, we utilized the Wav2vec 2.0 pretrained model[4] from Speechbrain, which was trained on the IEMOCAP emotion dataset.

4.2. Evaluation metrics

To evaluate the speech quality, we performed both objective and subjective evaluations of the synthesized speech and the ground truth (GT) speech. For the subjective evaluation, we used the mean opinion score (MOS), which is a method for evaluating satisfaction with sound quality, using a scale from 1 (worst) to 5 (best), as rated by humans. All scores are reported with a 95% confidence interval.

For an objective evaluation, we employed mel-cepstral distortion (MCD) and log-F0 root mean square error (F0 RMSE). The MCD was employed to quantify the difference between the GT and the generated samples in voice synthesis. It calculates the error between the mel-cepstral coefficients of the GT and the generated samples, with lower MCD values indicating better synthesis quality. The F0 RMSE is a measure used to evaluate the accuracy of pitch prediction in speech synthesis. It calculates the square root of the average squared differences between the logarithmic fundamental frequencies (log-F0) of the original and synthesized speech, with lower values indicating better pitch prediction.

Table 1: *Comparison with different acoustic models for subjective and objective metrics.*

Model	MOS (↑)	MCD (↓)	F0 RMSE (↓)
GT	4.53 ± 0.06	-	-
Tacotron2	3.72 ± 0.06	5.8043	0.3137
FastSpeech2	3.86 ± 0.08	5.7055	0.3025
DailyTalk	3.95 ± 0.07	5.5038	0.3070
Proposed model	**4.09 ± 0.07**	**5.4624**	**0.2978**

5. Results

5.1. Naturalness

To compare the speech quality of the proposed model, we selected four TTS models and a GT. The four TTS models were Tacotron2, FastSpeech2, DailyTalk, and the proposed model [5], with HiFi-GAN as the vocoder. We performed evaluations using a test set comprising 15 dialogues with 20 individuals participating in the assessment. The evaluation metrics include MOS, F0 RMSE, and MCD. The evaluation process was designed to closely mimic real-life conversational scenarios. Participants were initially presented with four preceding utterances from a dialogue, denoted by $U_{i-4:i-1}$, to immerse them in the context. They were then asked to evaluate the subsequent i-th utterance U_i focusing on its naturalness and how well it fits within the given conversational context. As shown in Table 1, the proposed model achieved a MOS rating of 4.09, marking it as significantly more natural-sounding than the other TTS models in comparison. Moreover, the F0 RMSE and MCD also demonstrated better performance compared to other TTS models. In conclusion, the evaluation process demonstrated the superior capability of the proposed model to deliver high-quality, natural, and contextually appropriate speech.

Table 2: *Audio prosody and quality comparisons for ablation study. RC denotes the recent chat; MCLN denotes the mixed context layer normalization; FPNC denotes the flow-based post-net with context.*

Model	MOS (↑)	MCD (↓)	F0 RMSE (↓)
Proposed model	**4.09 ± 0.07**	**5.4624**	**0.2978**
w/o RC	3.98 ± 0.06	5.7926	0.3067
w/o MCLN	4.05 ± 0.06	5.5385	0.2982
w/o FPNC	4.02 ± 0.05	5.4931	0.3016

5.2. Ablation study

We evaluated the proposed model by removing modules and following the evaluation method described in Section 5.1, and conducted a total of three experiments: without *recent chat* (RC), without mixed context layer normalization (MCLN), and without flow-based post-net with context (FPNC). As shown in Table 2, the removal of the RC-related module and training solely on *chat history* (w/o RC) decreased the performance compared to the proposed model by 0.11. Second, the experiment without applying MCLN to the mel decoder (w/o MCLN) also exhibits a performance decrease compared to the proposed model by 0.04. Finally, removing the flow-based post-net and training the model (w/o FPNC) resulted in a performance decrease compared to the proposed model by 0.07. Additionally, the performances of F0 RMSE and MCD also decreased compared to the proposed model. These results indicate that the three methods presented in this paper improved the performance and, above all, modeling the context module with *recent chat* had a significant impact on the performance.

6. Conclusion

In this paper, we proposed conversational TTS by leveraging multi-modal historical context, addressing the issue in existing conversational TTS where the inability to distinguish between recent and past contexts leads to the generation of voices that may seem inappropriate or unnatural for the current situation. Our approach includes a context module for nuanced conversational flow, a context-adapting mel decoder with MCLN, and a context-incorporating flow-based post-net. Consequently, the proposed model generates speech that is both natural and appropriately aligned with the given context. In future work, we aim to broaden the scope of our conversational TTS model by exploring methods that utilize diverse data types and languages, further advancing its applicability and versatility.

[3] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

[4] https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP

[5] Audio samples: `https://seongdonghyun.github.io/H4C-TTS-DEMO/`

7. Acknowledgements

This work was partly supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.RS-2023-00302424) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University)).

8. References

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[2] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *Proc. International Symposium on Chinese Spoken Language Processing* (*ISCSLP*), 2021, pp. 1–5.

[3] Y. Lei, S. Yang, X. Wang, and L. Xie, "MsEmoTTS: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 853–864, 2022.

[4] 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 *Proc. International Conference on Machine Learning* (*ICML*). PMLR, 2018, pp. 5180–5189.

[5] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *Proc. International Conference on Machine Learning* (*ICML*). PMLR, 2021, pp. 7748–7759.

[6] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational end-to-end tts for voice agents," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 403–409.

[7] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2018, pp. 4779–4783.

[8] Y. Nishimura, Y. Saito, S. Takamichi, K. Tachibana, and H. Saruwatari, "Acoustic modeling for end-to-end empathetic dialogue speech synthesis using linguistic and prosodic contexts of dialogue history," in *Proc. Interspeech*, 2022, pp. 3373–3377.

[9] J. Xue, Y. Deng, F. Wang, Y. Li, Y. Gao, J. Tao, J. Sun, and J. Liang, "M2-CTTS: End-to-end multi-scale multi-modal conversational text-to-speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2023, pp. 1–5.

[10] J. Li, Y. Meng, C. Li, Z. Wu, H. Meng, C. Weng, and D. Su, "Enhancing speaking styles in conversational text-to-speech synthesis with graph-based multi-modal context modeling," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2022, pp. 7917–7921.

[11] J. Li, Y. Meng, X. Wu, Z. Wu, J. Jia, H. Meng, Q. Tian, Y. Wang, and Y. Wang, "Inferring speaking styles from multi-modal conversational context by multi-scale relational graph convolutional networks," in *ACM Multimedia*, 2022, pp. 5811–5820.

[12] K. Ito and L. Johnson, "The LJ speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[13] V. Christophe, Y. Junichi, and M. Kirsten, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit," https://datashare.ed.ac.uk/handle/10283/2651, 2017.

[14] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A corpus derived from librispeech for text-to-speech," https://openslr.org/60/, 2019.

[15] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," https://sail.usc.edu/iemocap/iemocap_release.htm, 2008.

[16] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," *arXiv preprint arXiv:1710.03957*, 2017.

[17] L. Keon, P. Kyumin, and K. Daeyoung, "DailyTalk: Spoken dialogue dataset for conversational text-to-speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2023, pp. 1–5.

[18] 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," in *Proc. International Conference on Learning Representations* (*ICLR*), 2020.

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

[20] R. Badlani, A. Łańcucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, "One tts alignment to rule them all," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2022, pp. 6092–6096.

[21] J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, "Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling," *arXiv preprint arXiv:2010.04301*, 2020.

[22] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *Proc. Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 30, 2017, pp. 6309–6318.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 33, 2020, pp. 12 449–12 460.

[24] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 30, 2017, pp. 6000–6010.

[26] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit *et al.*, "Mlp-mixer: An all-mlp architecture for vision," in *Proc. Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 34, 2021, pp. 24 261–24 272.

[27] R. Huang, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "GenerSpeech: Towards style transfer for generalizable out-of-domain text-to-speech synthesis," *arXiv preprint arXiv:2205.07211*, 2022.

[28] J. Mauldon, "A generalization of the beta-distribution," *The Annals of Mathematical Statistics*, pp. 509–520, 1959.

[29] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, "Flow-TTS: A non-autoregressive network for text to speech based on flow," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, pp. 7209–7213.

[30] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," in *Proc. Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 33, 2020, pp. 8067–8077.

[31] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Proc. Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 31, 2018.

[32] P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations* (*ICLR*), 2015.

[33] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Advances in Neural Information Processing Systems* (*NeurIPS*), vol. 33, 2020, pp. 17 022–17 033.