Cross-utterance Conditioned Coherent Speech Editing

Cheng Yu[1], *Yang Li*[2], *Weiqin Zu*[1], *Fanglei Sun*[1,*], *Zheng Tian*[1,*], *Jun Wang*[3]

[1]ShanghaiTech University, China
[2]The University of Manchester, United Kingdom
[3]University College London, United Kingdom

{yucheng,zuwq2022,sunfl,tianzheng}@shanghaitech.edu.cn, yang.li-4@manchester.ac.uk, junwang@cs.ucl.ac.uk

Abstract

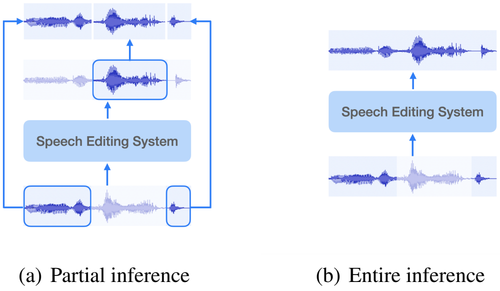
Text-based speech editing systems are developed to enable users to modify speech based on the transcript. Existing state-of-the-art editing systems based on neural networks do partial inferences with no exception, that is, only generate new words that need to be replaced or inserted. This manner usually leads to the prosody of the edited part being inconsistent with the surrounding speech and a failure to handle the alteration of intonation. To address these problems, we propose a cross-utterance conditioned coherent speech editing system, that first does the entire reasoning at the inference time. Our proposed system can generate speech by utilizing speaker information, context, acoustic features, and the mel-spectrogram from the original audio. Experiments conducted on subjective and objective metrics demonstrate that our approach outperforms the baseline on various editing operations regarding naturalness and prosody consistency.

Index Terms: speech editing, variational autoencoder

1. Introduction

Speech editing can be applied to a variety of areas with personalized voice needs and higher demands for speech naturalness, including video creation for social media, games, and movie dubbing. A promising neural-network-based audio editing technology is to synthesize speech according to text transcription and original audio. This system can synthesize speech that matches the tone and timbre of the original audio, according to the aligned transcription altered by content authors. As a result, editors could perhaps lessen their burden by modifying the text transcription rather than editing the original audio. Previous work [1, 2, 3] based on digital signal processing (DSP) has partially overcome the problem of prosody mismatch created by directly concatenating the audio in different scenarios. *Morrison et al.* [4] utilizes the neural network to predict prosodic information and integrates the TD-PSOLA algorithm, denoising, and de-reverberation [5] approaches to realize prosodic modification. Although the above systems support cut, copy, and paste operations, they cannot insert or replace a new word that doesn't exist in the voice data of the same speaker.

More recent research has applied text-to-speech (TTS) systems to synthesize the missing inserted word. VoCo [6] synthesizes the inserted word using a comparable TTS voice, then transforms it using the voice conversion (VC) model to fit the target speaker. EditSpeech [7] proposes the partial inference and bidirectional fusion method to achieve smooth transitions at edit boundaries. CampNet [8] conducts mask-training on a context-aware neural network based on Transformer to improve the quality of the edited voice. *Bai et al.*[9] suggests an alignment-aware acoustic and text pretraining method, which can be directly applied to speech editing by reconstructing masked acoustic signals through text input and acoustic text alignment. What's more, SpeechPainter[10] leverages an auxiliary textual input to fill in gaps of up to one second in speech samples and generalize it to unseen speakers.



(a) Partial inference (b) Entire inference

Figure 1: *Illustrations for different inference ways of the speech editing system.*

However, when applied to speech editing, all the existing methods [6, 7, 8, 9] based on neural networks do partial inference instead of entire inference, as shown in Figure 1(a). Specifically, the input of existing systems is the waveform or mel-spectrogram of the segments that do not need editing. Although the direct output of the editing module is the complete waveform or mel-spectrogram corresponding to the edited transcripts, in order to improve the similarity with the original audio, the existing methods select only the segments that must be modified and then insert them back into the original waveform or mel-spectrogram. Although retaining the original audio as much as possible adheres to our intuition, it will also lead to the following potential problems,

1. Since partial inference artificially inserts the predicted acoustic characteristics of the editing area into the corresponding positions of the original waveform, the discontinuity near the boundary of the editing area is almost inevitable to a certain extent. Meanwhile, the output of the existing speech synthesis system based on partial inference is still the whole audio, including the context. Therefore, it will not spare time or resources compared with the entire inference.

2. When the transcript is modified, the tone and prosody could also change accordingly. That is, the audio corresponding to the altered text might not be intended to sound exactly like the original audio. A special example is when a general question sentence can be modified into a declarative sentence, partial inference will be difficult to deal with the mood change.

To address the above-mentioned issues, we propose a cross-

[*]Corresponding authors.

10.21437/Interspeech.2023-2558

utterance conditioned coherent speech editing system. This text-based speech editing system applies the variational autoencoder with masked training to reconstruct the unmodified area of the original waveform with high fidelity. Therefore, the entire inference can replace partial inference, so as to avoid the incoherence of the junctions caused by splicing. Also, compared with the existing partial reasoning editing system, our method does not consume additional resources. This point can be intuitively accepted through Figure 1, where the framework of the entire inference is more concise than partial inference.

Moreover, to ensure that the generated audio conforms to both the original audio features and the context after the edition, the variational autoencoder is conditioned on the semantic information of the context and audio features extracted from the original waveform. The subjective and objective results on a challenging dataset show that our proposed model can ensure a high degree of similarity with real audio, while the coherency of the entire inference is significantly better than that of partial inference.

The rest of this paper is organized as follows. Section 2 illustrates our proposed speech editing system. The experimental setup, results, and conclusion are presented in Sections 3, 4 and 5, respectively.

2. Our System

Our proposed text-based speech editing system aims to synthesize the new audio that is consistent with the original audio rhythm and to truly restore the unmodified part of the audio, by virtue of the reconstruction ability of a variational autoencoder conditioned on context information. Figure 2(a) describes the model architecture, which takes the mel-spectrogram $\boldsymbol{x}_i$ extracted from the original waveform, current utterance $\boldsymbol{u}_i$, and l utterances before and after $\boldsymbol{u}_i$ as the input. Using an additional G2P conversion tool, the utterance $\boldsymbol{u}_i$ is translated into phonemes $\boldsymbol{p}_i$. Following *Li et al.* [11], the $2l+1$ neighboring utterances are paired into $2l$ pairs, i.e. $[(\boldsymbol{u}_{i-l}, \boldsymbol{u}_{i-l+1}), \cdots, (\boldsymbol{u}_{i+l-1}, \boldsymbol{u}_{i+l})]$. Then the pretrained BERT is used to capture the cross-utterance information, yielding $2l$ embeddings $[\boldsymbol{b}_{-l}, \cdots, \boldsymbol{b}_{l-1}]$. Also, the start and end times of each phoneme can be extracted by Montreal forced alignment [12]. The following part details the design of our system and biased training.

2.1. Mask CU-Enhanced CVAE

The mask CU-Enhanced CVAE module, as shown in Figure 2(b), is proposed to overcome the limitation that existing speech editing systems cannot restore the unmodified portion of the audio and must splice the modified portion with the original mel-spectrogram or audio.

2.1.1. Implementation of Text-based Speech Editing Operations

To start with, a text-based speech editing system supports the operations of deletion, insertion, and replacement. Without loss of generality, we can divide the original utterance transcript of the original speech as $[\boldsymbol{u}_a, \boldsymbol{u}_b, \boldsymbol{u}_c]$ and the modified utterance to be $[\boldsymbol{u}_a, \boldsymbol{u}_{b'}, \boldsymbol{u}_c]$, where $\boldsymbol{u}_{b'}$ is the modified segment and $\boldsymbol{u}_a, \boldsymbol{u}_c$ remain the same. Correspondingly, the phonemes translated by G2P can be denoted as $\boldsymbol{p}_i = [\boldsymbol{p}_a, \boldsymbol{p}_b, \boldsymbol{p}_c]$, with original speech's mel-spectrogram denoted as $\boldsymbol{x}_i = [\boldsymbol{x}_a, \boldsymbol{x}_b, \boldsymbol{x}_c]$. For $i \in \{a, b, c\}$, $\boldsymbol{x}_i$ contains a sequence of frame-level mel-spectrogram. Since the replacement operation in editing can be regarded as deletion before addition, we can use two flags instead of three to indicate the place to delete and add the corresponding content, i.e., $Flag_{del}$ and $Flag_{add}$.

Deletion The deletion procedure enables the user to eliminate a segment of the speech waveform which is associated with a set of certain words. The target utterance to be synthesized after deletion is $[\boldsymbol{u}_a, \boldsymbol{u}_c]$, where $\boldsymbol{u}_b$ is the part to be deleted. By comparing the utterance before and after editing, we can get the corresponding deletion indicator, which is further used to instruct the editing of mel-spectrogram

$$Flag_{del} = [\mathbf{0}_a, \mathbf{1}_b, \mathbf{0}_c].$$

Insertion and Replacement Different from the deletion operation, the target synthesized speech after insertion or replacement is based on the edited utterance $[\boldsymbol{u}_a, \boldsymbol{u}_{b'}, \boldsymbol{u}_c]$, where $\boldsymbol{u}_{b'}$ is the content to replace $\boldsymbol{u}_b$. Noted that the insertion process can be considered as the special case where $\boldsymbol{u}_b = \boldsymbol{p}_b = \boldsymbol{x}_b = \varnothing$. Correspond to the deletion operation, we have the addition indicator

$$Flag_{add} = [\mathbf{0}_a, \mathbf{1}_{b'}, \mathbf{0}_c].$$

Based on $Flag_{del}$, the reference mel-spectrogram $[\boldsymbol{x}_a, \boldsymbol{x}_c]$ is sent into the Mask CU-Enhanced CVAE module since $\boldsymbol{x}_{b'}$ is to be generated. The mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$ are learned from two one-dimensional convolutions. Referring to $Flag_{add}$, 0 and 1s are added to the corresponding position of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, that is, $\hat{\boldsymbol{\mu}} = [\boldsymbol{\mu}_a, \mathbf{0}_{b'}, \boldsymbol{\mu}_c]$ and $\hat{\boldsymbol{\sigma}} = [\boldsymbol{\sigma}_a, \mathbf{1}_{b'}, \boldsymbol{\sigma}_c]$. This allows the speech generated by the editing area to be sampled from the utterance-specific prior, while the audio of the area that has no modification is sampled from the real audio and the utterance-specific prior. During the training process, the edited real audio is unavailable, so we can only mask specific audio segments and restore the same content to simulate the editing scenario, that is $b' = b$.

2.1.2. Enhancement of Coherence and Prosody

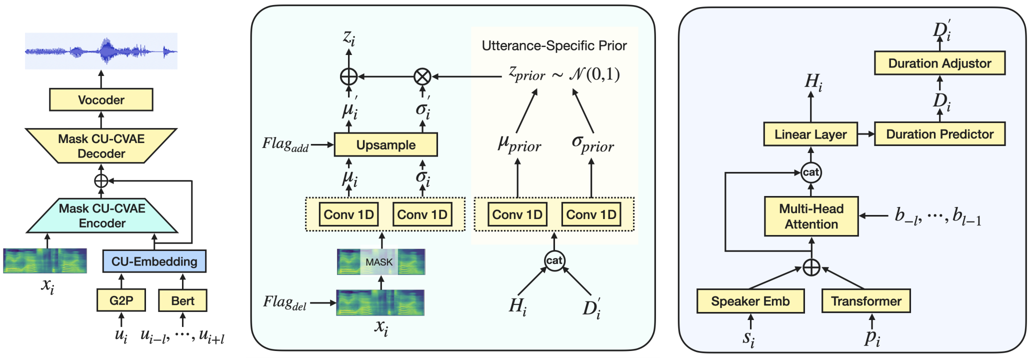
We have introduced more mechanisms to ensure that the output of the mask CU-CVAE module can further synthesize coherent and contextual audio. In order to make the editing boundary more fluent, $\boldsymbol{\mu}'$ and $\boldsymbol{\sigma}'$ are further generated through one-dimensional convolution from $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$. At this time, the module can sample from the estimated prior and can be re-parameterized as

$$\boldsymbol{z} = \boldsymbol{\mu}' \oplus \boldsymbol{\sigma}' \otimes \boldsymbol{z}_{prior}$$

where $\oplus, \otimes$ are element-wise addition and multiplication operations, and $\boldsymbol{z}_{prior}$ is sampled from the learned utterance-specific prior, corresponding to *Li et al.* [11]. The re-parameterization is as follows,

$$\boldsymbol{z}_{prior} = \boldsymbol{\mu}_{prior} \oplus \boldsymbol{\sigma}_{prior} \otimes \boldsymbol{\epsilon}$$

where $\boldsymbol{\mu}_{prior}, \boldsymbol{\sigma}_{prior}$ are learned from the utterance-specific prior, and $\boldsymbol{\epsilon}$ is sampled from the standard Gaussian distribution $\mathcal{N}(0, 1)$. $\boldsymbol{H}_i$ is the output of CU-Embedding, as shown in Figur 2(c). The CU-embedding module is proposed to create phoneme-level embeddings from nearby utterances to enhance prosody modeling. It uses a pretended BERT to capture contextual information from $2l$ utterances surrounding the current one. The phoneme sequence is encoded using a transformer encoder, and contextual data is collected using a multi-head attention layer. Also, an additional duration predictor takes $\boldsymbol{H}_i$ as inputs and predicts the duration of each phoneme. In addition, in order to effectively utilize the duration information extracted from the original audio, similar to the method in [7, 9], we further adjust the phoneme duration of the edited area by multiplying it with



(a) Our system

(b) Mask CU-CVAE encoder

(c) CU-embedding module

Figure 2: *The overall architecture of our system, consisting of the mask cross-utterance enhanced CVAE and cross-utterance embedding module.* $\oplus$ *and* $\otimes$ *are element-wise addition and multiplication operations.* ⓒ *is the concatenate operation.*

the ratio of the original audio and the predicted audio duration of the unedited area to get D_i'. The estimated duration is rounded after the duration predictor and the adjustor.

3. Experimental Setup

3.1. Dataset

We conducted experiments on a multi-speaker dataset, LibriTTS[13]. Both the train-clean-100 and train-clean-360 subsets were used, containing 245 hours of English audiobooks from 1151 speakers (553 female speakers and 598 male speakers). We randomly select 90%,5%,5% data from datasets for train, valid and test set, respectively.

3.2. Configuration Detail

The proposed Mask CU-CVAE TTS system was based on the framework of FastSpeech 2. In the CU-embedding module, we used the "BERT_BASE" configuration. Additionally, information about different speakers learned from a 256-dim lookup table was added to the Transformer output.

Four 1D-convolutional layers with 1 kernel size were utilized in the Mask CU-enhanced CVAE module to predict the mean and variance of 2-dim latent features. Meanwhile, an additional upsampling layer was applied to make the predicted sequence length consistent with the phoneme sequence length after editing, and also promote the naturalness of synthesized audio. We randomly select the part to be masked by taking a word instead of a phoneme as a unit to faithfully recreate the actual editing scenario. In addition, to balance the system's ability to learn and predict audio information, we set the shielding rate to 50%, which has been proven effective in *Bai et al.* [9]. The decoder, optimizer, and other hyperparameters are the same as that used in FastSpeech2.

The length regulator in FastSpeech 2 was modified to accommodate the outputs of the CU-embedding module and the vocoder HifiGAN [14] was finetuned for 1200 steps on an open-sourced, pre-trained version of "UNIVERSAL_V1" to synthesize a waveform from the predicted mel-spectrogram.

3.3. Evaluation Metrics

Both subjective and objective tests were conducted in order to measure the performance of our proposed method. First of all, using a 5-scale mean opinion score (MOS) evaluation, 20 volunteers participated in a subjective listening test over 15 synthesized audios in which they were asked to assess the level of naturalness and similarity of speech samples. 95% confidence intervals and p-value were provided with the MOS results.

For the objective evaluation, FFE [15] and MCD [16] were utilized to test the reconstruction performance of different VAEs and different settings of loss weights. FFE was used to assess the accuracy of the F0 track reconstruction. Besides, MCD estimated timbral distortion from the first 13 MFCCs.

Moreover, WER from an automatic speech recognition model was also reported. Complementary to naturalness, the WER metric demonstrated the degree of intelligibility and consistency between synthetic and real speech. The attention-based encoder-decoder model utilized in this study was trained on Librispeech 960-hour data.

4. Results

This section presents a series of experiments for our proposed speech editing system. First, the naturalness and similarity of synthesized audio generated by EditSpeech [7] and our system via both partial and entire inference were evaluated. Next, an ablation study was performed to progressively show the influence of restrictions on context information in our system, based on MOS and reconstruction performance. At last, the effect of the degree of biased training on reconstruction performance was also investigated. Our audio examples are available on the demo page [1].

4.1. Partial vs. Entire Inference

To investigate the performance of partial inference versus entire inference, experiments were conducted on the following systems: 1) *GT*, the ground truth audio; 2) *GT (Mel+HifiGAN)*, first convert the ground truth audio to the ground truth mel-spectrogram, and then convert it back to audio using HifiGAN vocoder; 3) *Wave_cut*, manually cut the modified region from the generated waveform, and insert it back into the original waveform; 4) *EditSpeech [7]*, using partial inference and bidirectional fusion to improve the prosody near boundaries; 5) *Our system (Mel_cut)*, cut the modified region from the generated mel-spectrogram,

[1] `http://bitly.ws/uMFd`

Method	Insert		Replace		Delete		Reconstruct	
	Nat.	Sim.	Nat.	Sim.	Nat.	Sim.	Nat.	Sim.
Our system vs. Mel_cut	0.0662	0.793	0.0294	0.771	0.0168	0.298	0.0525	0.691
Our system vs. Wave_cut	0.0219	0.135	0.0163	0.287	0.369	0.310	0.0564	0.143

Table 1: *The significance analysis of our system using entire inference vs. "Mel_cut" and "Wave_cut" on naturalness and similarity MOS scores.*

and insert it back to the original mel-spectrogram with a forced aligner; 6) *Our system*, regenerate a complete mel-spectrogram from the whole sentence to be edited, and then use HifiGAN vocoder to generate the complete waveform;

Method	Insert		Replace		Delete	
	Nat.	Sim.	Nat.	Sim.	Nat.	Sim.
Wave_cut	2.93	3.76	2.82	3.50	3.25	3.82
EditSpeech (Mel_cut)	2.35	3.21	2.47	3.36	2.82	**3.81**
Our system (Mel_cut)	3.11	**3.57**	2.97	3.41	2.82	**3.81**
Our system	**3.37**	3.56	**3.39**	**3.43**	**3.37**	3.67

Table 2: *Subjective naturalness and similarity results on EditSpeech and our system using partial and entire inference. Note that since the deletion operation of EditSpeech, which can only do partial inference, is to combine segments of the real mel-spectrogram, there is no difference in the results of different editing systems using partial inference.*

According to the MOS scores on naturalness shown in Table 2, our model with entire inference achieved the highest score on all the editing operations. The gap in replacement was noticeable, as the speech editing models based on partial reasoning have difficulty dealing with intonation conversion. The score of "Mel_cut" in deletion was relatively low since "Mel_Cut" is highly dependent on the accuracy of MFA. Especially when short words were deleted, its performance could be worse than manually careful deletion based on waveform. "Wave_cut" had a relatively lower naturalness MOS score in insertion and replacement since it involves the insertion of new words, and there is disharmony between the original audio and the generated audio.

MOS scores on similarity suggested that the performance of our system based on entire inference was close to partial inference "Mel_cut" and surpassed EditSpeech in insertion and replacement. It was also close to "Wave_cut", which served as an upper bound indicator of similarity, with the maximum difference around 0.2.

Method	Nat.	Sim.	**FFE**	**MCD**	**WER**
GT	4.56	-	-	-	3.124
GT (Mel+HifiGAN)	4.39	4.68	0.170	4.651	3.887
EditSpeech	3.14	3.80	0.372	6.345	6.702
Our system (Mel_cut)	3.66	**3.91**	**0.326**	**5.957**	**5.174**
Our system	**3.90**	3.83	0.327	6.657	5.377

Table 3: *EditSpeech and our system's reconstruction performance using partial and entire inference. Lower objective results imply better similarity to the original audio.*

The p-value in Table 1 presented that the naturalness of our model using entire inference was obviously superior to "Mel_cut" and "Wave_cut", while there was no significant difference in similarity between entire and the two partial inference methods. The only exception was in the case of deletion, where the naturalness of our model using entire inference was not significantly different from that of the "Wave_cut". Table 3 also demonstrated the capability of our mask CU-enhanced CVAE module to reconstruct the mel-spectrogram. Since partial inference directly copied the real mel-spectrogram of the unedited area, it is reasonable that partial inference had better reconstruction performance on similarity and MCD(Mel-cepstral distortion). Nevertheless, our system using entire inference still surpassed EditSpeech on FFE and WER.

4.2. Ablation Study

In this section, we investigate the performance impact of using different VAEs in our system. We compare the reconstruction performance and MOS scores of the synthesized audio among the following systems: 1-2) same as the above experimental settings; 3) *Baseline1*, use a fine-grained VAE instead of CU-CVAE; 4) *Baseline2*, use a CVAE without the context embeddings, i.e.$l = 0$; 5) *Baseline2*, use CU-CVAE with 2 neighbouring utterances, i.e.$l = 2$; 6) *Our system*, use CU-CVAE with 5 neighbouring utterances, i.e.$l = 5$.

Method	**FFE**	**MCD**	**WER**
GT	-	-	3.124
GT (Mel+HifiGAN)	0.170	4.651	3.887
Baseline1	0.371	6.919	7.404
Baseline2	0.333	6.750	5.503
Baseline3	0.332	6.697	5.392
Our system	**0.327**	**6.657**	**5.377**

Table 4: *Objective metrics of the reconstruction performance of our system with different VAEs.*

The reconstruction metrics in Table 4 suggest that using more cross-utterances can improve the reconstruction capability. These results indicated that the CU-embedding and mask CU-CVAE module played a crucial role in generating more coherent audio.

5. Conclusion

In this paper, we propose a cross-utterance conditioned coherent speech editing system, which is the first text-based speech editing system that can entirely generate audio corresponding to the edited transcript. A variational autoencoder conditioned on speaker information, context, and audio prior is integrated into a high-quality text-to-speech model to ensure both the restoration and generation quality of audio. Experiments show that our proposed system has the ability to reconstruct the acoustic characteristics of original audio with high fidelity and that the prosody of the synthesized speech conforms to the context of the edited transcript.

6. Limitations

The experiments in this paper focused on seen speakers. When it is transferred to unseen speakers, it will need additional components, such as GST [17], to better extract speaker features.

7. Acknowledgements

Yang Li is supported by the China Scholarship Council (CSC) Scholarship.

8. References

[1] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5-6, pp. 453–467, 1990. [Online]. Available: https://doi.org/10.1016/0167-6393(90)90021-Z

[2] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99-D, no. 7, pp. 1877–1884, 2016. [Online]. Available: https://doi.org/10.1587/transinf.2015EDP7457

[3] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 349–353, 2006.

[4] M. Morrison, L. Rencker, Z. Jin, N. J. Bryan, J. P. Cáceres, and B. Pardo, "Context-aware prosody correction for text-based speech editing," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 7038–7042. [Online]. Available: https://doi.org/10.1109/ICASSP39728.2021.9414633

[5] J. Su, Z. Jin, and A. Finkelstein, "Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 4506–4510. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-2143

[6] Z. Jin, G. J. Mysore, S. DiVerdi, J. Lu, and A. Finkelstein, "Voco: text-based insertion and replacement in audio narration," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 96:1–96:13, 2017. [Online]. Available: https://doi.org/10.1145/3072959.3073702

[7] D. Tan, L. Deng, Y. T. Yeung, X. Jiang, X. Chen, and T. Lee, "Editspeech: A text based speech editing system using partial inference and bidirectional fusion," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*. IEEE, 2021, pp. 626–633. [Online]. Available: https://doi.org/10.1109/ASRU51503.2021.9688051

[8] T. Wang, J. Yi, R. Fu, J. Tao, and Z. Wen, "Campnet: Context-aware mask prediction for end-to-end text-based speech editing," *CoRR*, vol. abs/2202.09950, 2022. [Online]. Available: https://arxiv.org/abs/2202.09950

[9] H. Bai, R. Zheng, J. Chen, M. Ma, X. Li, and L. Huang, "A^{3}t: Alignment-aware acoustic and text pretraining for speech synthesis and editing," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 1399–1411. [Online]. Available: https://proceedings.mlr.press/v162/bai22d.html

[10] Z. Borsos, M. Sharifi, and M. Tagliasacchi, "Speechpainter: Text-conditioned speech inpainting," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 431–435. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-194

[11] Y. Li, C. Yu, G. Sun, H. Jiang, F. Sun, W. Zu, Y. Wen, Y. Yang, and J. Wang, "Cross-utterance conditioned VAE for non-autoregressive text-to-speech," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Association for Computational Linguistics, 2022, pp. 391–400. [Online]. Available: https://doi.org/10.18653/v1/2022.acl-long.30

[12] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, F. Lacerda, Ed. ISCA, 2017, pp. 498–502. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1386.html

[13] 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," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 1526–1530. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-2441

[14] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html

[15] W. Chu and A. Alwan, "Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*. IEEE, 2009, pp. 3969–3972. [Online]. Available: https://doi.org/10.1109/ICASSP.2009.4960497

[16] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, 1993, pp. 125–128 vol.1.

[17] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-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 *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 5167–5176. [Online]. Available: http://proceedings.mlr.press/v80/wang18h.html