Interspeech 2024
1-5 September 2024, Kos, Greece

FluentEditor: Text-based Speech Editing by Considering Acoustic and Prosody Consistency

Rui Liu[1], *Jiatian Xi*[1], *Ziyue Jiang*[2], *Haizhou Li*[3,4]
[1] Inner Mongolia University, Hohhot, China [2] Zhejiang University, China
[3] Shenzhen Research Institute of Big Data, School of Data Science,
The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China
[4] National University of Singapore, Singapore
liurui_imu@163.com, x_jiatian@163.com, ziyuejiang@zju.edu.cn, haizhouli@cuhk.edu.cn

Abstract

The current Text-based Speech Editing (TSE) techniques have focused on reducing the difference between the generated speech segment and the reference target in the editing region, ignoring its local and global fluency in the context and original utterance. To maintain the speech fluency, we propose a fluency speech editing model, termed FluentEditor, by considering fluency-aware training criterion in the TSE training. Specifically, the acoustic consistency constraint aims to smooth the transition between the edited region and its neighboring acoustic segments consistent with the ground truth, while the prosody consistency constraint seeks to ensure that the prosody attributes within the edited regions remain consistent with the overall style of the original utterance. The subjective and objective experimental results on VCTK demonstrate that our FluentEditor outperforms all advanced baselines in terms of naturalness and fluency. The audio samples and code are available at `https://github.com/AI-S2-Lab/FluentEditor`.
Index Terms: Speech Editing, Fluency Modeling, Acoustic Consistency, Prosody Consistency

1. Introduction

Text-based speech editing (TSE) [1] allows for modification of the output audio by editing the transcript rather than the audio itself. With the rapid development of the internet, audio-related media sharing has become a prevalent activity in our daily lives. Note that TSE can bring great convenience to the audio generation process and be applied to a variety of areas with personalized voice needs, including video creation for social media, games, and movie dubbing.

Over the past few years, many attempts adopted text-to-speech (TTS) systems to build neural network-based TSE models. For example, the CampNet [2] conducts mask training on a context-aware neural network based on Transformer to improve the quality of the edited voice. A^3T [3] 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. More recently, the diffusion model has gradually become the backbone of the NN-based TSE with remarkable results. For example, EdiTTS [4] takes the diffusion-based TTS model as the backbone and proposes a score-based TSE methodology for fine-grained pitch and content editing. FluentSpeech [5] proposes a context-aware diffusion model that iteratively refines the modified mel-spectrogram with the guidance of context features.

However, during training, the existing approaches just constrain the Euclidean Distance [6] between the mel-spectrum to be predicted and the ground truth to ensure the naturalness of TSE. Although they consider the use of contextual information to mitigate the over-smoothing problem of edited speech, their objective functions are not designed to ensure fluent output speech [7, 8]. We consider two challenges to be tackled for effective speech fluency modeling. 1) *Acoustic Consistency*: the smoothness of the concatenation between the region to be edited and its neighboring regions should be close to the real concatenation point [9]. 2) *Prosody Consistency*: the prosody style of the synthesized audio in the region to be edited needs to be consistent with the global prosody style of the original utterance [10, 11].

To address the above issues, we propose a novel fluency speech editing scheme, termed FluentEditor, by introducing the acoustic and prosody consistency training criterion to achieve natural and fluent speech editing. Specifically, 1) To achieve the acoustic consistency, we design the *Acoustic Consistency Loss* $\mathcal{L}_{AC}$ to calculate whether the variance at the boundaries is close to the variance at the real concatenation points. 2) To achieve the prosody consistency, we introduce the *Prosody Consistency Loss* $\mathcal{L}_{PC}$ to let the high-level prosody features of the synthesized audio in the region to be edited be close to that of the original utterance. The high-level prosody features are extracted by the pre-trained GST-based prosody extractor [11]. The subjective and objective results on the VCTK [12] dataset show that the acoustic and prosody consistency of the FluentEditor is significantly better than the advanced TSE baselines, while the proposed FluentEditor can ensure a high degree of fluency like real speech. The main contributions of this work can be summarized as follows:

- We propose a novel fluency speech editing scheme, termed FluentEditor;
- We adopt the diffusion model as the backbone and introduce *Acoustic and Prosody Consistency Losses* to conduct the fluency modeling for TSE;
- The proposed model outperforms all advanced TSE baselines in terms of naturalness and fluency.

In the rest of this paper, we first introduce the methodology of FluentEditor in Section 2. Afterward, we present the experimental setup in Section 3, which includes the dataset, the baseline, and the implementation details. We also show all the experimental results and conduct in-depth analyses in Section 3. Finally, we conclude this paper and discuss future work in Section 4.

2. FluentEditor: Methodology

We formulate the proposed FluentEditor, a TSE model that ensures speech fluency by considering acoustic and prosody con-

10.21437/Interspeech.2024-2320

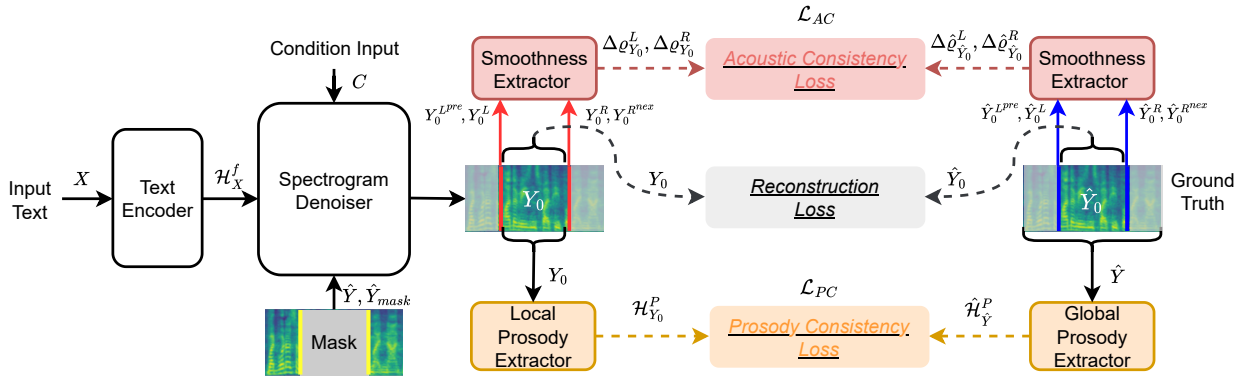


Figure 1: *The overall workflow of FluentEditor. The total loss function includes Reconstruction Loss, and Acoustic and Prosody Consistency Losses.*

sistency. We first introduce the overall workflow, then further elaborate the fluency-aware training criterion and the run-time inference.

2.1. Overall Workflow

As shown in Fig.1, our FluentEditor adopts the mask prediction-based diffusion network as the backbone, which consists of a text encoder, and a spectrogram denoiser. The spectrogram denoiser seeks to adopt the Denoising diffusion probabilistic model (DDPM) to learn a data distribution $p(\cdot)$ by gradually denoising a normally distributed variable through the reverse process of a fixed Markov Chain of length T.

Assume that the phoneme embedding of the input phoneme sequence is $X = (X_1, \ldots, X_{|X|})$ and the acoustic feature sequence for X is $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_{|\hat{Y}|})$. The masked acoustic feature sequence $\hat{Y}_{mask} = Mask(\hat{Y}, \lambda)$ is obtained by replacing the random spans of $\hat{Y}$ with the random vector according to a λ probability. Specifically, the text encoder aims to extract the high-level linguistic feature $\mathcal{H}_X$ for X. The spectrogram denoiser then aggregates the $\mathcal{H}_X$ and the condition input C to guide the reverse process of the diffusion model $\Theta(Y_t|t, C)$ $(t \in T)$, where Y_t is a noisy version of the clean input $\hat{Y}_0$. Similar to [5], the condition input C consists of the frame-level linguistic feature $\mathcal{H}_X^f$, acoustic feature sequence $\hat{Y}$, masked acoustic feature sequence $\hat{Y}_{mask}$, speaker embedding e_{spk} and the pitch embedding e_{pitch}. In the generator-based diffusion models, $p_\theta(Y_0|Y_t)$ is the implicit distribution imposed by the neural network $f_\theta(Y_t, t)$ that outputs Y_0 given Y_t. And then Y_{t-1} is sampled using the posterior distribution $q(Y_{t-1}|Y_t, Y_0)$ given Y_t and the predicted Y_0.

To model speech fluency, we design *acoustic consistency loss* $\mathcal{L}_{AC}$ and *prosody consistency loss* $\mathcal{L}_{PC}$ on the basis of the original *reconstruction loss*, to ensure that the acoustic and prosody performance of speech generated in the editing area is consistent with the context and the original utterance. For reconstruction loss, we follow [5] and employ Mean Absolute Error (MAE) and the Structural Similarity Index (SSIM) [13] losses to calculate the difference between Y_0 and the corresponding ground truth segment $\hat{Y}_0$. *Acoustic consistency loss* $\mathcal{L}_{AC}$ and *prosody consistency loss* $\mathcal{L}_{PC}$ serve the purposes of ensuring the smoothness of connecting points between edited speech regions and neighboring acoustic segments, as well as ensuring that the local prosody of the masked speech aligns with the global prosody of the ground truth. In the following subsection, we will introduce $\mathcal{L}_{AC}$ and $\mathcal{L}_{PC}$ in detail.

2.2. Fluency-Aware Training criterion

2.2.1. Acoustic Consistency Loss

Note that *Target Loss* and *Concatenation Loss* are two important training criterion in unit-selection TTS [9, 14, 15], in which Target Loss is used to compute the proximity of candidate units to the target [15] while Concatenation Loss [9] is used to compute the degree of concatenation smoothness of the speech units to be concatenated. Inspired by the *Concatenation Loss* [9], we propose the acoustic consistency loss to quantify the smooth transition between the editing region and the adjacent context.

The acoustic consistency loss $\mathcal{L}_{AC}$ employs smoothness constraints at both the left and right boundaries for the predicted acoustic feature Y_0. We compare the euclidean distance, $\Delta\varrho_{Y_0}^L$ and $\Delta\varrho_{Y_0}^R$, for the left and right boundaries with $\Delta\hat{\varrho}_{\hat{Y}_0}^{L/R}$ of ground truth speech at the corresponding boundaries to serve as the approximate indicator of the overall smoothness $\mathcal{L}_{AC}$.

Specifically, $\mathcal{L}_{AC}$ consists of $\mathcal{L}_{AC}^L$ and $\mathcal{L}_{AC}^R$, and we use Mean Squared Error (MSE) [16] to measure the proximity between the predicted segment and the ground truth:

$$\begin{aligned}\mathcal{L}_{AC} &= \mathcal{L}_{AC}^L + \mathcal{L}_{AC}^R \\ &= \mathrm{MSE}(\Delta\varrho_{Y_0}^{\mathrm{L}}, \Delta\hat{\varrho}_{\hat{Y}_0}^{\mathrm{L}}) + \mathrm{MSE}(\Delta\varrho_{Y_0}^{\mathrm{R}}, \Delta\hat{\varrho}_{\hat{Y}_0}^{\mathrm{R}})\end{aligned} \tag{1}$$

Note that the Euclidean Distance between two adjacent frames is obtained by the smoothness extractor. Take $\Delta\varrho_{Y_0}^L$ as an example,

$$\Delta\varrho_{Y_0}^L = \varrho_{Y_0{}^L} - \varrho_{Y_0^{L^{pre}}} \tag{2}$$

where $Y_0^{L^{pre}}$ denotes the speech frame preceding the left boundary of the masked region. In other words, the ending frame of the adjacent non-masked region is on the left side. Variance is employed as a statistical measure of data dispersion to depict changes in the values of acoustic features [17]. A smaller variance may suggest a greater similarity between two spectra, while a larger variance may indicate greater dissimilarity. To comprehensively capture the statistical properties of the audio signal, we utilize variance to characterize the feature information of each Mel spectrogram frame, denoted as $\varrho_{Y_0{}^L}$ and $\varrho_{Y_0^{L^{pre}}}$.

Similarly, we compute the smoothness constraint for the right boundary$\mathcal{L}_{AC}^R$, where $Y_0^{R^{nex}}$ denotes the speech frame succeeding the right boundary of the masked region, in other words, the starting frame of the adjacent non-masked region on the right side.

2.2.2. Prosody Consistency Loss

The prosody consistency loss $\mathcal{L}_{PC}$ is responsible for capturing the prosody feature $\mathcal{H}^P_{Y_0}$ from the predicted masked region Y_0 while also analyzing the overall prosody characteristics $\hat{\mathcal{H}}^P_{\hat{Y}}$ present in the original speech, then employ the MSE loss to conduct the prosody consistency constraints.

$$\mathcal{L}_{PC} = \text{MSE}(\mathcal{H}^P_{Y_0}, \hat{\mathcal{H}}^P_{\hat{Y}}) \tag{3}$$

Note that the prosody features $\mathcal{H}^P_{Y_0}$ and $\hat{\mathcal{H}}^P_{\hat{Y}}$ are obtained by the pre-trained prosody extractor. Specifically, the local prosody extractor and global prosody extractor both employ the reference encoder [11] from the Global Style Token (GST) model [11] for converting Y_0 and $\hat{Y}$ into high-level prosody vectors with fixed lengths, facilitating easy comparison.

$$\mathcal{H}^P_{Y_0} = \text{GST}(Y_0), \quad \hat{\mathcal{H}}^P_{\hat{Y}} = \text{GST}(\hat{Y}) \tag{4}$$

Lastly, following [5], the total loss function is the sum of reconstruction loss and two new loss functions, $\mathcal{L}_{AC}$ and $\mathcal{L}_{PC}$, across all non-contiguous masked regions, since the mask region in a sentence may include multiple non-contiguous segments [5]. In a nutshell, $\mathcal{L}_{AC}$ and $\mathcal{L}_{PC}$ of the FluentEditor are introduced to ensure fluent speech with consistent prosody.

2.3. Run-time Inference

In run-time, given the original text and its speech, the user can edit the speech by editing the text. Note that we can manually define modification operations (i.e., insertion, replacement, and deletion). The corresponding speech segment of the edited word in the given text is treated as the masked regions in Fig. 1. Similar to [5], our FluentEditor reads the edited text and the remaining acoustic feature $\hat{Y} - \hat{Y}_{mask}$ of the original speech to predict the Y_0 for the edited word. At last, the Y_0 and its context $\hat{Y} - \hat{Y}_{mask}$ are concatenated as the final fluent output speech.

3. Experiments and Results

3.1. Dataset

We validate the FluentEditor on the VCTK [12] dataset, which is an English speech corpus uttered by 110 English speakers with various accents. Each recording is sampled at 22050 Hz with 16-bit quantization. The precise forced alignment is achieved through Montreal Forced Aligner (MFA) [18]. We partition the dataset into training, validation, and testing sets, randomly with 98%, 1%, and 1%, respectively.

3.2. Experimental Setup

The configurations of text encoder and spectrogram denoiser are referred to [5]. The diffusion steps T of the FluentEditor system is set to 8. Following GST [11], the prosody extractor comprises a convolutional stack and an RNN. The dimension of the output prosody feature of the GST-based prosody extractor is 256. Following [3], we adopt a random selection strategy, with a fixed masking rate of 80%, for masking specific phoneme spans along with their corresponding speech frames. The pre-trained HiFiGAN [19] vocoder is used to synthesize the speech waveform. We set the batch size is 16. The initial learning rate is set at 2×10^{-4}, and the Adam optimizer [20] is utilized to optimize the network. The FluentEditor model is trained with 2 million training steps on one A100 GPU.

3.3. Evaluation Metric

For subjective evaluation, We conduct a Mean Opinion Score (MOS) [21] listening evaluation in terms of speech fluency, termed *FMOS*. Note that FMOS allows the listener to feel whether the edited segments of the edited speech are fluent compared to the context. We keep the text content and text modifications consistent among different models to exclude other interference factors, only examining speech fluency. Furthermore, Comparative FMOS (C-FMOS) [21] is also used to conduct the ablation study. For objective evaluation, we utilize MCD [22], STOI [23], and PESQ [24] to measure the overall quality of the edited speech.

3.4. Comparative Study

We develop four neural TSE systems for a comparative study, that includes: 1) **CampNet** [2] propose a context-aware mask prediction network to simulate the process of text-based speech editing; 2) $\boldsymbol{A^3T}$ [3] propose the alignment-aware acoustic-text pre-training that takes both phonemes and partially-masked spectrograms as inputs; 3) **FluentSpeech** [5] takes the diffusion model as backbone and predict the masked feature with the help of context speech; and 4) **FluentEditor (Ours)** designs the acoustic and prosody consistency losses. We also add the **Ground Truth** speech for comparison. Note that two ablation systems, that are "$\boldsymbol{w/o\ \mathcal{L}_{AC}}$" and "$\boldsymbol{w/o\ \mathcal{L}_{PC}}$", are built to validate the two new losses.

3.5. Main Results

In this section, we present the main results of the study, focusing on the comprehensive assessment of reconstructed and edited speech, including a comprehensive analysis of objective metrics and subjective assessments.

3.5.1. Evaluation of Reconstructed Speech

Objective results: We select 400 test samples from the test set randomly and report the objective results in the second to fourth columns of Table 1. Note that we follow [5] and just measure the objective metrics of the masked region using the reconstructed speech. We observe that our FluentEditor achieves the best performance in terms of overall speech quality. For example, the MCD and STOI values of FluentEditor obtain optimal results and PESQ achieves suboptimal results among all systems. It suggests that the FluentEditor performs proper acoustic feature prediction for the speech region to be edited. Note that objective metrics do not fully reflect the human perception [25], we further conduct subjective listening experiments.

3.5.2. Evaluation of Edited Speech

Subjective results: For FMOS evaluation, we selected 50 audio samples from the test set and invited 20 listeners to evaluate speech fluency. Following [26], we test the insertion and replacement operations and present the FMOS results in the last two columns of Table 1. We find that FluentEditor consistently achieves superior fluency-related perceptual scores. For example, FluentEditor obtains the top FMOS value of 4.25 for insertion and 4.26 for replacement, that very close to that of ground truth. This demonstrates the effectiveness of the fluency-aware training criterion. By considering the acoustic and prosody consistency constraints, our FluentEditor allows for weakening the editing traces and improving the prosody performance of the edited speech.

Table 1: *Objective and subjective evaluation results of comparative study. * means the value achieves suboptimal.*

Method	Objective Evaluation			Subjective Evaluation (FMOS)	
	MCD (↓)	STOI (↑)	PESQ (↑)	Insertion	Replacement
Ground Truth	NA	NA	NA	4.37 ± 0.05	4.42 ± 0.01
CampNet	3.85	0.53	1.38	3.89 ± 0.01	3.94 ± 0.03
A^3T	3.79	0.76	1.59	3.82 ± 0.03	3.83 ± 0.02
FluentSpeech	3.57	0.72	1.50	4.02 ± 0.04	4.04 ± 0.01
FluentEditor (Ours)	**3.59**	**0.80**	**1.85**	$\mathbf{4.25 \pm 0.03}$	$\mathbf{4.26 \pm 0.01}$

Table 2: *Objective and subjective results of ablation study.*

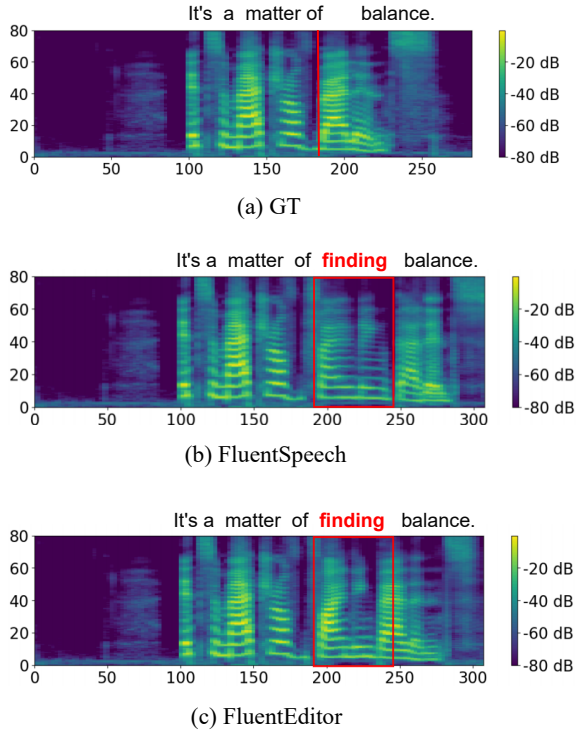
Method	C-FMOS	MCD (↓)
FluentEditor	**0.00**	**3.47**
w/o $\mathcal{L}_{AC}$	-0.16	3.48
w/o $\mathcal{L}_{PC}$	-0.21	3.51

3.6. Ablation Study

To further validate the contribution of our $\mathcal{L}_{AC}$ and $\mathcal{L}_{PC}$ respectively, the subjective and objective ablation results, of insertion and replacement, are reported in Table 2. We follow the previous section to prepare the samples and listeners. It's observed that the C-FMOS and MCD values of these two ablation systems both drop when we remove the $\mathcal{L}_{AC}$ and $\mathcal{L}_{PC}$ respectively, indicating that the acoustic and prosody consistency constraints play a vital role in enhancing both the naturalness and fluency of the edited speech. Notably, the Prosody Consistency exhibits more remarkable prediction results after coding and extracting speech rhythmic features, emphasizing its effectiveness in terms of prosody fluency for speech editing.

3.7. Visualization Analysis

As illustrated in the Fig.2, we visualize the mel-spectrograms produced by FluentEditor and the FluentSpeech baseline[1]. The red boxes indicate the random masked and its reconstructed speech segment of utterance "It's a matter of finding balance.". We can see that FluentEditor can generate mel-spectrograms with richer frequency details compared with the baseline, resulting in natural and expressive sounds, which further demonstrates the effectiveness of acoustic and prosody consistency losses. Nevertheless, we recommend that the reader listen to our speech samples[1] to visualize the advantages.



(a) GT

(b) FluentSpeech

(c) FluentEditor

Figure 2: *Visualizations of the generated mel-spectrograms by GT, FluentEditor and FluentSpeech baseline.*

4. Conclusion

In this paper, we introduce a novel text-based speech editing (TSE) model, termed FluentEditor, that involves two novel fluency-aware training criterions to improve the acoustic and prosody consistency of edited speech. The acoustic consistency loss $\mathcal{L}_{AC}$ to calculate whether the variance at the boundaries is close to the variance at the real concatenation points, while the prosody consistency loss $\mathcal{L}_{PC}$ to let the high-level prosody features of the synthesized audio in the region to be edited be close to that of the original utterance. In this way, our FluentEditor allows for generating encouraging speech editing performance in terms of speech fluency. The objective and subjective experiments on VCTK demonstrate that incorporating $\mathcal{L}_{AC}$ and $\mathcal{L}_{PC}$ yields superior results and ensures fluent speech with consistent prosody. In future work, we will consider the multi-scale consistency and further improve the FluentEditor architecture.

[1]Due to space limits, we just report the FluentSpeech baseline. More visualization results and speech samples are referred to our website: `https://anonymous.4open.science/w/FluentEditor-B684/`.

5. Acknowledgement

The research by Rui Liu was funded by the Young Scientists Fund of the National Natural Science Foundation of China (No. 62206136), Guangdong Provincial Key Laboratory of Human Digital Twin (No. 2022B121201 0004), and the "Inner Mongolia Science and Technology Achievement Transfer and Transformation Demonstration Zone, University Collaborative Innovation Base, and University Entrepreneurship Training Base" Construction Project (Supercomputing Power Project) (No.21300-231510). The research by Haizhou Li was partly supported by the Internal Project Fund from Shenzhen Research Institute of Big Data (Grant No. T00120220002), and the Shenzhen Science and Technology Research Fund (Fundamental Research Key Project Grant No. JCYJ20220818103001002).

6. References

[1] Z. Jin, G. J. Mysore, S. Diverdi, J. Lu, and A. Finkelstein, "Voco: Text-based insertion and replacement in audio narration," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–13, 2017.

[2] T. Wang, J. Yi, R. Fu, J. Tao, and Z. Wen, "Campnet: Context-aware mask prediction for end-to-end text-based speech editing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2241–2254, 2022.

[3] H. Bai, R. Zheng, J. Chen, M. Ma, X. Li, and L. Huang, "A^3T: Alignment-aware acoustic and text pretraining for speech synthesis and editing," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1399–1411.

[4] J. Tae, H. Kim, and T. Kim, "Editts: Score-based editing for controllable text-to-speech," *arXiv preprint arXiv:2110.02584*, 2021.

[5] Z. Jiang, Q. Yang, J. Zuo, Z. Ye, R. Huang, Y. Ren, and Z. Zhao, "FluentSpeech: Stutter-oriented automatic speech editing with context-aware diffusion models," in *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 11 655–11 671. [Online]. Available: https://aclanthology.org/2023.findings-acl.741

[6] H. Reyes, S. Subramaniam, N. Kaabouch, and W. C. Hu, "A spectrum sensing technique based on autocorrelation and euclidean distance and its comparison with energy detection for cognitive radio networks," *Computers & Electrical Engineering*, vol. 52, pp. 319–327, 2016.

[7] C.-y. Tseng, S.-h. Pin, Y. Lee, H.-m. Wang, and Y.-c. Chen, "Fluent speech prosody: Framework and modeling," *Speech communication*, vol. 46, no. 3-4, pp. 284–309, 2005.

[8] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive tts training with frame and style reconstruction loss," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1806–1818, 2021.

[9] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, vol. 1. IEEE, 1996, pp. 373–376.

[10] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," *Proc. Interspeech 2020*, pp. 4432–4436, 2020.

[11] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 5180–5189.

[12] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, vol. 6, p. 15, 2017.

[13] Y. Ren, X. Tan, T. Qin, Z. Zhao, and T.-Y. Liu, "Revisiting oversmoothness in text to speech," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8197–8213.

[14] M. Dong, K.-T. Lua, and H. Li, "A unit selection-based speech synthesis approach for mandarin chinese." *J. Chin. Lang. Comput.*, vol. 16, no. 3, pp. 135–144, 2006.

[15] R. Fu, J. Tao, Y. Zheng, and Z. Wen, "Deep metric learning for the target cost in unit-selection speech synthesizer." in *INTERSPEECH*, 2018, pp. 2514–2518.

[16] P. S. Mandhare, V. R. Borkar, and M. R. Kumbhakarna, "The generalized approximation of an arbitrary function using its mean and quadratic variance," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 12, pp. 2096–2100, 2019.

[17] N. N. W. N. Hashim, M. A.-E. A. Ezzi, and M. D. Wilkes, "Mobile microphone robust acoustic feature identification using coefficient of variance," *International Journal of Speech Technology*, vol. 24, no. 4, pp. 1089–1100, 2021.

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

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

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] P. C. Loizou, "Speech quality assessment," in *Multimedia analysis, processing and communications*. Springer, 2011, pp. 623–654.

[22] 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. IEEE, 1993, pp. 125–128.

[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[25] 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 *International Conference on Learning Representations*, 2020.

[26] 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 *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 626–633.