AttentionStitch: How Attention Solves the Speech Editing Problem

Antonios Alexos*
Department of Computer Science
University of California Irvine
aalexos@uci.edu

Pierre Baldi
Department of Computer Science
University of California Irvine
pfbaldi@uci.edu

Abstract

The generation of natural and high-quality speech from text is a challenging problem in the field of natural language processing. In addition to speech generation, speech editing is also a crucial task, which requires the seamless and unnoticeable integration of edited speech into synthesized speech. We propose a novel approach to speech editing by leveraging a pre-trained text-to-speech (TTS) model, such as FastSpeech 2, and incorporating a double attention block network on top of it to automatically merge the synthesized mel-spectrogram with the mel-spectrogram of the edited text. We refer to this model as AttentionStitch, as it harnesses attention to stitch audio samples together. We evaluate the proposed AttentionStitch model against state-of-the-art baselines on both single and multi-speaker datasets, namely LJSpeech and VCTK. We demonstrate its superior performance through an objective and a subjective evaluation test involving 15 human participants. AttentionStitch is capable of producing high-quality speech, even for words not seen during training, while operating automatically without the need for human intervention. Moreover, AttentionStitch is fast during both training and inference and is able to generate human-sounding edited speech.

1 Introduction

TTS synthesis has paved the way for the exploration of various speech tasks like speech editing. In speech editing, the edited part is synthesized and then combined with the rest of the reference audio to produce a smooth and natural-sounding audio sample. The speech editing task can be formally defined as follows: given a reference audio sample S_R with its corresponding text transcript T_R, the objective is to modify T_R by replacing word(s), resulting in an edited text transcript T_E. T_E is then synthesized into an edited audio sample S_E, with the aim of making S_E sound similar to S_R and for the edited portion of S_E to be indistinguishable from the original S_R to the listener. One method for speech editing integrates segments from the same speaker using pitch and prosody features for natural editing Morrison et al. (2021). Another approach generates audio in a generic voice and converts it to the desired target voice Jin et al. (2017), but has noticeable roughness at edit boundaries. EditSpeech Tan et al. (2021) uses forward and backward decoders for fused mel-spectrograms, while A^{3}T Bai et al. (2022) introduces cross-modal alignment embedding. EdiTTS Tae et al. (2021) refines edited speech with perturbations to Gaussian priors. SpeechPainter Borsos et al. (2022) fills speech gaps using an attention-based model, limited to 1-second gaps. MaskedSpeech Zhang et al. (2022) focuses on Mandarin speech editing with a pretrained FastSpeech2 model. Our novel approach expedites training and achieves smoother audio segment integration by incorporating an auxiliary module into a pretrained TTS model like FastSpeech2 (FS2), enhancing efficiency and naturalness in audio stitching. This paper presents a novel method for speech editing that is rooted in state-of-the-art TTS synthesis with the following contributions: (1) Proposal of AttentionStitch, a speech editing model combining FS2 and a double attention block. (2) Fast and high-quality synthesis with automatic editing due to

*corresponding author
37th Conference on Neural Information Processing Systems (NeurIPS 2023).

attention. (3) Subjective and objective evaluation on single (LJSpeech) and multi-speaker (VCTK) data, showing superiority over state-of-the-art methods and an extensive ablation study.

2 Preliminaries

Double Attention Block. The role of the double attention block is to propagate global features from images and enable the model to access them efficiently. It operates in two steps. In the first step, the double attention block gathers image features through an attention-pooling operation. In the second step, it selects and distributes the features via attention. Let $X \in \mathbb{R}^{c \times d \times h \times w}$ be the input to a 3D convolution layer, where c denotes the number of channels, d the dimension, and h and w the dimensions of the images. In a general format, we define F_{distr} as the operation that distributes the features in the layers, and F_{gather} is the operation that gathers the features from the images which are later distributed. So for every location $i, \ldots, dhw$ of the network with feature u_i we have the output of an operation that gathers features from each location i and distributes them back to each location considering the local feature.

FastSpeech 2. The core model utilized in our approach is a pre-trained FS2 model, which is a fast-training non-autoregressive TTS synthesis model. FS2 follows a two-step synthesis process, where it initially predicts a mel-spectrogram and then transforms it into an audio waveform. To enhance the naturalness and controllability of the synthesized speech, FS2 incorporates prosody features such as energy, pitch, and duration. These features are predicted individually through dedicated predictors during training, while during inference, they are predicted by the model itself. The model follows an Encoder-Decoder architecture, taking phonemes as input and generating a mel-spectrogram as output. The mel-spectrogram is subsequently processed by a vocoder, which generates the corresponding audio waveform. The Encoder module receives the phoneme embedding and positional embedding as input and produces a hidden sequence. A module called the variance adaptor enriches this hidden sequence with informative features such as energy, pitch, and duration. Finally, the Decoder module parallelly converts the enriched hidden sequence into a mel-spectrogram.

3 Proposed Method

AttentionStitch comprises a pre-trained FS2 model and a double attention block. We chose FS2 since it is renowned as one of the top-performing models in the TTS community, offering fast training and inference speeds with high-quality speech synthesis. The FS2 model is already pre-trained as part of AttentionStitch, enabling us to save time during training. We only need to train the remaining components of the model, which is more manageable due to its smaller size. The double attention block employed is intuitive in its operation and serves the purpose of gathering features from the synthesized mel-spectrogram to fill the masked regions of the reference mel-spectrogram. The application of this model variation in the context of audio data makes it a suitable candidate for achieving the desired combination of features from the synthesized and reference mel-spectrograms. This choice is motivated by the need to combine specific parts of two mel-spectrograms effectively and automatically, which aligns with the objective of synthesizing high-quality speech output.

FS2 takes the phonemes of the edited text as input and generates a mel-spectrogram as output. During the training phase, we randomly mask 10% of the reference mel-spectrogram near its center, although the procedure works if we mask in the beginning or in the end instead. The mask consists of zeros, and we provide further details on the masking strategy in Section 4.4. After masking the reference mel-spectrogram, we concatenate it with the synthesized mel-spectrogram and feed it into the double attention block which redistributes the features of the synthesized mel-spectrogram within the masked region of the reference mel-spectrogram. Additionally, we employ a Postnet module to refine the mel-spectrogram further. Finally, a Hifi-GAN Kong et al. (2020) vocoder, transforms the final mel-spectrogram into an audio waveform. We incorporate skip connections between the output of the double attention block and the Postnet, as they have proven to be beneficial components in speech synthesis Tu & Zhang (2017); Shi et al. (2018). During inference we mask the phoneme sequence based on the word(s) that need to be edited, as we know their boundaries. As in training, we perform the masking operation by replacing the corresponding part of the mel-spectrogram with zeros. Additionally, we modify the reference text T_R by replacing the word(s) with the target word(s). To ensure the reference mel-spectrogram and the synthesized mel-spectrogram have the same length, we leverage the duration predictor of FS2 and resize the mask accordingly. The speech editing operation

takes place within the double attention block. The proposed AttentionStitch model is depicted in Figure 1.

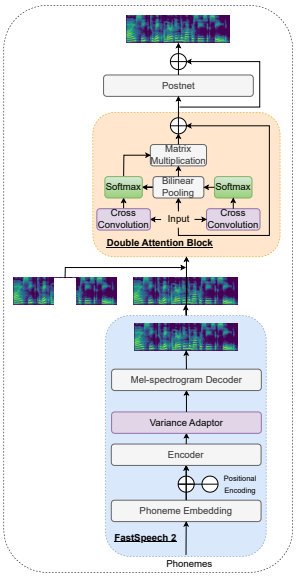

Figure 1: Overview of our proposed AttentionStitch model. AttentionStitch consists of a pre-trained FS2 model and a Double Attention Block.

Method	MOS (↑)
Complete synthesis	2.56±0.33
FeatSwitch	2.73±0.35
AttentionStitch	**3.86±0.28**
Reference	4.82±0.12

Figure 2: LJSpeech

Method	MOS (↑)	MCD (↓)
EditSpeech	3.28±0.33	7.54
A^3T	3.3±0.35	7.97
AttentionStitch	**3.51±0.23**	**6.5**
Reference	4.43±0.2	-

Figure 3: VCTK

Figure 4: MOS (↑) and MCD (↓) scores for AttentionStitch, the compared methods, and the reference samples with 95% confidence intervals for LJSpeech and VCTK. AttentionStitch outperforms the compared methods in both metrics.

4 Experiments and Evaluation

4.1 Experimental Setup

We utilized the experimental setup described in the corresponding papers for the pre-trained FS2 model and double attention network Chien (2021); Vo (2019). Our results were obtained for single and multi-speaker data on the LJSpeech dataset Ito & Johnson (2017) and the VCTK dataset Yamagishi et al. (2019) respectively. We pre-trained FS2 for 200,000 steps on both LJSpeech and VCTK. In an effort to improve the performance of the model for speech editing task, we also attempted training the FS2 for 900,000 steps, but observed that the model overfit and performed poorly for this task. We then froze the entire FS2 model and in the second training phase, we trained the double attention block along with a postnet for 200,000 steps. The double attention block was utilized for speech editing, while the postnet was used for better refinement of the mel-spectrogram. The postnet consisted of 5 1-D convolutions with 512 channels and a kernel size of 5. We also observed that high-quality audio samples could be produced with limited training steps. We use Mean Average Error (MAE) as the loss for the double attention block and the postnet loss, while the rest of the model is frozen.

We first evaluate AttentionStitch using a single speaker dataset, LJSpeech, against two baselines derived from the pre-trained FS2 model denoted as “FeatSwitch" and “Complete synthesis and swap" which do not require any additional tuning or modules for speech editing. Subsequently, we extend our evaluation to VCTK, a more challenging multi-speaker dataset, where we compare to state-of-the-art methods.

Complete synthesis and swap. We first use a pretrained FS2 model to synthesize speech corresponding to the edited text T_e. We predict the durations from a pretrained FS2 and use them to find the word boundaries. Then we replace the source word(s) in the reference with the synthesized word(s).

FeatSwitch. Within FeatSwitch, we perform mid-inference prosody feature switching in FS2 for speech editing. We extract phoneme-level energy, pitch, and duration features from the reference audio R and predict the same features from edited text T_e using FS2. We then replace ground truth features in all non-target phonemes in the synthesized audio.

We deliberately chose not to compare with fully resynthesized edited text, since the edited part may be indistinguishable, but the final audio sample still differs from the reference audio. For VCTK, we compared AttentionStitch to more competitive and state-of-the-art methods, namely A^3T and EditSpeech. Although other modern methods exist like SpeechPainter, EdiTTS, and MaskedSpeech, we chose not to compare with these based on the nature of these works. SpeechPainter fills limited gaps of audio samples with the same text; EdiTTS is evaluated only on single-speaker data; and MaskedSpeech works on sentence level (not word level) and is evaluated only in Mandarin.

4.2 Evaluation Setup

We evaluated AttentionStitch using the objective metric Mel-cepstral distance (MCD) Kubichek (1993), and the subjective metric Mean Opinion Scores (MOS) Sector & Itu (2017). MCD measures the difference between two mel-spectrograms, and MOS on the other hand, is a score given by human listeners based on audio quality. It is worth noting that a higher MOS score and a lower MCD score indicate better quality of speech. 15 subjects participated in our subjective evaluation study and they could listen to the audio samples as many times as they want. For both datasets we used 2 tests to obtain the MOS scores, where the first one is that the subjects have to give their quality opinion score between 1 and 5, about audio samples of different methods, and in the second they have to give their opinion about audio samples of AttentionStitch which we edited with unseen words. For the first part we chose 3 samples for each method for LJSpeech and 4 samples for each method for VCTK. For the second part we chose 12 samples of AttentionStitch for each dataset.

4.3 Results on single speaker data

In the single-speaker part of the evaluation, the subjects scored the samples based on the audio quality and general preference. The results of this part can be found in Table 2, which presents the MOS scores with 95% confidence intervals for each method and the reference audio samples. We observe that AttentionStitch achieved a high MOS score, indicating a high quality of edited synthesized speech. It significantly outperformed the two baselines, although it had a wide range of reported results. Specifically, FeatSwitch achieved a MOS score of 2.73, Complete Synthesis achieved 2.56, AttentionStitch achieved 3.87, and the reference audio samples achieved 4.82. Although the gap between AttentionStitch and the reference seems large, the subjective evaluation results in speech synthesis can vary widely and are highly dependent on individual preferences.

4.4 Results on multi-speaker data

In the multi-speaker evaluation, following the same methodology as in Section 4.3 and in Section 4.2, we collected MOS scores from the participants to assess their overall preference and audio quality for each method and reference audio sample. The MOS scores with a 95% confidence interval for each method and reference audio sample are presented in Figure 3. AttentionStitch achieved the highest MOS score, indicating a superior quality of the synthesized speech. Specifically, AttentionStitch obtained a MOS score of 3.51, while EditSpeech and A^3T achieved scores of 3.28 and 3.3, respectively. In contrast, the reference audio samples received a MOS score of 4.43. It is worth noting that the VCTK dataset is more challenging than the LJSpeech dataset due to the variety of accents present, resulting in lower MOS scores for both AttentionStitch and the reference audio samples. We also present our findings with a short objective evaluation with MCD scores which show that our method performs better than A^3T and EditSpeech.

As part of an ablation study, we subjected AttentionStitch to the task of synthesizing edited audio containing unseen words; words not encountered during the model's training phase. Remarkably, the resulting MOS obtained from this exercise stood at $\mathbf{3.66 \pm 0.17}$, a value closely aligned with the original MOS reported in Figure 3. Furthermore, we explored the feasibility of altering multiple words within a sentence simultaneously. This endeavor presented challenges, as the model's training involved applying the mask to only one part at a time. Sequential word changes led to a decline in audio quality due to the emergence of electronic artifacts. Nonetheless, AttentionStitch retains the capability to replace a single word with multiple words, showcasing its versatility in handling certain editing tasks.

5 Conclusions and Discussion

AttentionStitch is a novel method for speech editing that utilizes a pre-trained FS2 model and incorporates the unique double attention block which effectively gathers the features of the edited part and distributes them within the masked area of the reference mel-spectrogram. It is a fast approach to speech editing for researchers with limited resources, something that the community has not addressed yet.

References

Bai, H., Zheng, R., Chen, J., Ma, M., Li, X., and Huang, L. A3t: Alignment-aware acoustic and text pretraining for speech synthesis and editing. In *International Conference on Machine Learning*, pp. 1399–1411. PMLR, 2022.

Borsos, Z., Sharifi, M., and Tagliasacchi, M. Speechpainter: Text-conditioned speech inpainting. *arXiv preprint arXiv:2202.07273*, 2022.

Chien, C.-M. Fastspeech 2 - pytorch implementation. `https://github.com/ming024/FastSpeech2`, 2021.

Ito, K. and Johnson, L. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

Jin, Z., Mysore, G. J., Diverdi, S., Lu, J., and Finkelstein, A. Voco: Text-based insertion and replacement in audio narration. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017.

Kong, J., Kim, J., and Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

Kubichek, R. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, volume 1, pp. 125–128. IEEE, 1993.

Morrison, M., Rencker, L., Jin, Z., Bryan, N. J., Caceres, J.-P., and Pardo, B. Context-aware prosody correction for text-based speech editing. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7038–7042. IEEE, 2021.

Sector, S. and Itu, O. Recommendation itu-t p. 10/g. 100, vocabulary for performance, quality of service and quality of experience, 2017.

Shi, Y., Rong, W., and Zheng, N. Speech enhancement using convolutional neural network with skip connections. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 6–10. IEEE, 2018.

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

Tan, D., Deng, L., Yeung, Y. T., Jiang, X., Chen, X., and Lee, T. Editspeech: A text based speech editing system using partial inference and bidirectional fusion. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 626–633. IEEE, 2021.

Tu, M. and Zhang, X. Speech enhancement based on deep neural networks with skip connections. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5565–5569. IEEE, 2017.

Vo, N. Pytorch implementation of a2-nets: Double attention networksfastspeech 2 - pytorch implementation. `https://github.com/nguyenvo09/Double-Attention-Network`, 2019.

Yamagishi, J., Arai, T., Dong, M., King, S., Chávez, S. R., King, S. E., O'Shea, J. D., Oura, K., Kawai, H., and Zhang, J. CSTR VCTK Corpus: English multi-speaker corpus for Cstr voice cloning toolkit (version 0.92). In *Proc. of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 3636–3642, Portorož, Slovenia, 2019.

Zhang, Y.-J., Song, W., Yue, Y., Zhang, Z., Wu, Y., and He, X. Maskedspeech: Context-aware speech synthesis with masking strategy. *arXiv preprint arXiv:2211.06170*, 2022.