
StyleTTS: A Style-Based Generative Model for Natural and Diverse Text-to-Speech Synthesis

Yinghao Aaron Li
Columbia University
y14579@columbia.edu

Cong Han
Columbia University
ch3212@columbia.edu

Nima Mesgarani
Columbia University
nima@ee.columbia.edu

Abstract

Text-to-Speech (TTS) has recently seen great progress in synthesizing high-quality speech owing to the rapid development of parallel TTS systems, but producing speech with naturalistic prosodic variations, speaking styles and emotional tones remains challenging. Moreover, since duration and speech are generated separately, parallel TTS models still have problems finding the best monotonic alignments that are crucial for naturalistic speech synthesis. Here, we propose StyleTTS, a style-based generative model for parallel TTS that can synthesize diverse speech with natural prosody from a reference speech utterance. With novel Transferable Monotonic Aligner (TMA) and duration-invariant data augmentation schemes, our method significantly outperforms state-of-the-art models on both single and multi-speaker datasets in subjective tests of speech naturalness and speaker similarity. Through self-supervised learning of the speaking styles, our model can synthesize speech with the same prosodic and emotional tone as any given reference speech without the need for explicitly labeling these categories. The generated samples can be found on our demo page at <https://styletts.github.io/>.

1 Introduction

Text-to-speech (TTS), also known as speech synthesis, aims to synthesize natural and intelligible speech from a given text. The recent advances in deep learning have resulted in great progress in TTS technologies to the extent that several recent studies claim to have synthesized speech qualitatively similar to real human speech [1, 2]. However, it remains a challenge to synthesize expressive speech that can accurately capture the extremely rich diversity occurring naturally in prosodic, temporal, and spectral characteristics of speech which together encode the paralinguistic information [3]. For example, the same given text can be spoken in many ways depending on the context, the emotional tone, and dialectic and habitual speaking patterns of a speaker. Hence, TTS is by nature a one-to-many mapping problem that needs to be addressed as such.

Several approaches have been proposed to address such a problem, including methods based on variational inference [1, 4, 5, 6], flow-based modeling [1, 7, 8], controlling pitch, duration and energy [9, 10], and using external prosody encoder [11, 12, 13]. It is worth noting that models such as VITS [1] have used a combination of these techniques to achieve the state-of-the-art performance, however, as we will demonstrate, the synthesized speech from the current models is still perceptually distinguishable from real human speech which warrants further research. In particular, the speaking styles and emotional tones of different speakers remain difficult to model and incorporate adequately.

Many attempts have been made to integrate style information into TTS models [11, 12, 14, 15], but they are mostly based on autoregressive models such as Tacotron. Non-autoregressive parallel TTS models, such as Fastspeech [16] and Glow-TTS [8], have several advantages over autoregressive models. These models fully utilize parallel implementation to enable fast speech synthesis, and they are also more robust to longer and out-of-distribution (OOD) utterances. Moreover, because phoneme

duration, pitch, and energy are predicted independently from speech, models such as FastSpeech2 [10] and FastPitch [17] allow fully controllable speech synthesis.

One limitation of the current models is that the improvements of parallel TTS over autoregressive systems and utilizing styles to enable expressive speech synthesis are done mostly separately. The majority of current TTS models focus on synthesizing speech from a single target speaker [10, 16, 18, 19, 20, 21, 22], and multi-speaker extensions are often done by concatenating speaker embeddings with the encoder output that is given to the synthesizer [1, 8, 23]. Models that explore speech styles also incorporate styles by concatenating style vectors and phoneme embeddings as input to the decoder [11, 12, 14, 15]. This approach for incorporating style information may not be optimal because it cannot fully capture the temporal modulation of acoustic features in the target speech. In the domain of style transfer, styles are introduced through conditional normalization such as adaptive instance normalization (AdaIN) [24]. AdaIN has seen great success in not only neural style transfer [25, 26, 27], but also in generative modeling [28, 29, 30] and neural image editing [31, 32]. Such promising techniques are rarely used in speech synthesis, with only a few exceptions in voice conversion [33, 34, 35] and speaker adaptation [36, 37]. Unlike autoregressive TTS systems, parallel TTS models synthesize the entire speech, eliminating the need for generating every frame of the mel-spectrogram separately. This characteristic of parallel TTS models makes it possible to take advantage of powerful AdaIN modules to integrate generalized styles for diverse speech synthesis.

Recent state-of-the-art models mostly employ the non-autoregressive parallel framework for TTS, but because they do not directly align the input text and speech like autoregressive models do, an external aligner such as the Montreal Forced Aligner [38] pre-trained on a large dataset is usually required to align the text and speech first. Since the external aligner is not trained on the TTS data and objectives, the alignments are not optimally suited for the TTS task. VITS and Glow-TTS [1, 8], on the other hand, use generative flows to search for the monotonic alignment directly. EfficientTTS [39] and Parallel Tacotron 2 [40] also train an internal text aligner. Although training internal aligners solves the generalization problems caused by external aligners, it still suffers from overfitting because the aligners are trained with an only mel-reconstruction loss on a much smaller TTS dataset. Therefore, a text aligner that is pre-trained but transferable for TTS fine-tuning is required to combine the benefits of large-scale pre-training of external aligners and data-specific TTS-oriented internal aligners.

In this study, we address the limitations of the current systems in incorporating diverse speaking styles in synthesis and the difficulty in learning a reliable monotonic aligner. We propose StyleTTS, a style-based generative model for speech synthesis. In our framework, a style encoder extracts style vectors from reference audio, and the style vectors are passed both to the decoder and prosody predictors through adaptive normalization. A pre-trained text aligner is jointly optimized with our Transferable Monotonic Aligner (TMA) training objectives. We apply a novel duration-invariant data augmentation to learn natural prosody independently from phoneme duration estimation. With the help of stylization and a novel training framework, our method produces naturalistic prosodic patterns and emotional tones similar to the reference audio. Using various reference audios, our method synthesizes the same text with diverse speaking styles and enables one-to-many mapping that is still challenging for many TTS systems. We show that our framework significantly outperforms the current state-of-the-art models in terms of naturalness, speaker similarity, and speech diversity.

Our study makes multiple contributions: (i) we propose Transferable Monotonic Aligner (TMA), a novel transfer learning scheme that enables fine-tuning of pre-trained text aligners for TTS tasks, (ii) we introduce novel duration-invariant data augmentation for better prosody prediction, and (iii) we present the first parallel TTS framework that incorporates generalized speech styles for natural and expressive TTS. Together, these contributions significantly advance the state-of-the-art style-based speech synthesis for better TTS technologies that can enhance human-computer interactions.

2 StyleTTS

2.1 Proposed Framework

Given $t \in \mathcal{T}$ the input phonemes and $x \in \mathcal{X}$ an arbitrary reference mel-spectrogram, our goal is to train a system that generates the mel-spectrogram $\tilde{x} \in \mathcal{X}$ that corresponds to the speech of t and reflects the generalized speech styles of x . Generalized speech styles are defined as any characteristics in the reference audio x except the phonetic content [15], including but not limited to prosodic pattern,

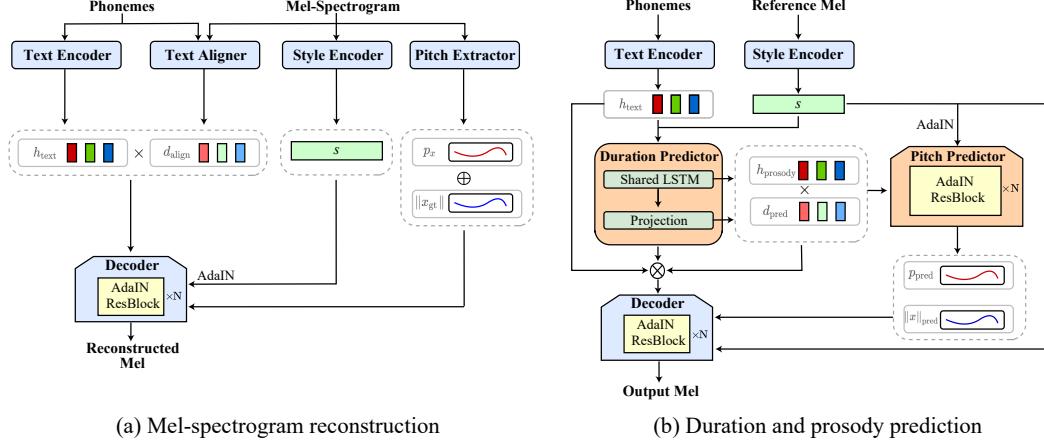


Figure 1: Training and inference schemes of StyleTTS. (a) Stage 1 of our training procedures where the decoder is trained to reconstruct input mel-spectrogram using pitch, energy, phonemes, alignment, and style vectors. (b) Stage 2 of training and inference procedures where pitch, energy, and alignment are predicted based on input text, and a style vector is extracted from a reference mel-spectrogram for synthesis. Modules in blue are fixed during this stage of training while modules in orange are trained.

lexical stress, formants transition, speaking rate, and speaker identity. Our framework consists of eight modules as described below. An overview of our framework is provided in Figure 1.

Text Encoder. Given input phonemes t , our text encoder T encodes t into hidden representation $h_{\text{text}} = T(t)$. The text encoder consists of a 3-layer CNN followed by a bidirectional LSTM [41].

Text Aligner. We train a text aligner together with the decoder during the reconstruction stage. Our text aligner A is modeled after the decoder of Tacotron 2 with attention. It is pre-trained for automatic speech recognition (ASR) task on the LibriSpeech corpus [42] and then fine-tuned together with our decoder. The text aligner produces an alignment d_{align} between mel-spectrograms and phonemes.

Style Encoder. Given an input mel-spectrogram x , our encoder extracts the style vector $s = E(x)$. E can produce diverse style representations with different reference audios. This allows our decoder G to synthesize speech that reflects the style s of a reference audio x . Our style encoder consists of four residual blocks [43] followed by an averaging pooling layer across the time axis.

Pitch Extractor. We follow the approach proposed in FastPitch [10] where pitch F0 is extracted directly in Hertz without further processing. This allows easier control and better representation of pitch F0. Unlike FastPitch which estimates the ground truth pitch using acoustic periodicity detection [44], we train our own pitch extractor together with our decoder for better pitch estimation. The pitch extractor F is a JDC network [45] pre-trained on LibriSpeech with ground truth F0 estimated using YIN [46]. It is fine-tuned together with the decoder to predict pitch $p_x = F(x)$ for reconstructing x .

Decoder. Our decoder G is trained to reconstruct the input mel-spectrogram x by $\hat{x} = G(h_{\text{text}} \cdot d_{\text{align}}, s, p_x, \|x\|)$, where $h_{\text{text}} \cdot d_{\text{align}}$ is aligned hidden representation of phonemes, s is the style vector of x , p_x is pitch contour of x and $\|x\|$ is the log norm (energy) of x per frame. Our decoder consists of seven residual blocks with AdaIN, which can be described as

$$\text{AdaIN}(x, s) = L_\sigma(s) \frac{x - \mu(x)}{\sigma(x)} + L_\mu(s) \quad (1)$$

where x is a single channel of the feature maps, s is the style vector, $\mu(\cdot)$ and $\sigma(\cdot)$ denotes the channel mean and standard deviation, and L_σ and L_μ are learned linear projections for computing the adaptive gain and bias using the style vector s . The advantages of AdaIN are discussed in Appendix B.2.

The pitch p_x , energy $\|x\|$, and residual phoneme features $R(h_{\text{text}})$ are concatenated with the output from every residual block as the input to the next residual block (see Appendix D for details) because these features can be diluted through the AdaIN module. We show that concatenating these residual features is helpful for both the naturalness and diversity of synthesized speech in section 4.4.

Discriminator. VITS [1] argues that adversarial training for TTS models greatly improves the sound quality of generated speech. In StyleTTS, we include a discriminator D to facilitate the training of our decoder. The discriminator shares the same architecture as our style encoder.

Duration Predictor. Our duration predictor consists of a 3-layer bidirectional LSTM S with adaptive layer normalization (AdaLN) module followed by a linear projection L , where instance normalization is replaced by layer normalization in equation 1. We use AdaLN because S takes discrete tokens similar to those in NLP applications, where layer normalization [47] is preferred. S is shared with the prosody predictor P through $\mathbf{h}_{\text{prosody}} = S(\mathbf{h}_{\text{text}})$ as input to P .

Prosody Predictor. Our prosody predictor P predicts both the pitch p_{pred} and energy $\|\mathbf{x}\|_{\text{pred}}$ with given text and style vector. The aligned representation $\mathbf{h}_{\text{prosody}} \cdot \mathbf{d}$ is processed through a shared bidirectional LSTM layer and two sets of three residual blocks with AdaIN and a linear projection layer, one for the pitch output and another for the energy output (see Appendix D for details).

2.2 Training Objectives

Our model is trained in two stages so that the duration-invariant prosody data augmentation can be applied. In the first stage, the model is trained to reconstruct the mel-spectrogram from text, pitch, energy, and style. In the second stage, all modules are fixed except the duration and prosody predictors. The predictors are trained to predict the duration, pitch, and energy from given text.

2.2.1 First Stage Objectives

Mel reconstruction. Given a mel-spectrogram $\mathbf{x} \in \mathcal{X}$ and its corresponding text $\mathbf{t} \in \mathcal{T}$, we train our model under the L_1 reconstruction loss

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{\mathbf{x}, \mathbf{t}} [\|\mathbf{x} - G(\mathbf{h}_{\text{text}} \cdot \mathbf{d}_{\text{align}}, s, p_{\mathbf{x}}, \|\mathbf{x}\|)\|_1] \quad (2)$$

where $\mathbf{h}_{\text{text}} = T(\mathbf{t})$ is the encoded phoneme representation, $\mathbf{d}_{\text{align}}$ is the attention alignment from the text aligner, $s = E(\mathbf{x})$ is the style vector of \mathbf{x} , $p_{\mathbf{x}} = F(\mathbf{x})$ is the pitch F0 of \mathbf{x} and $\|\mathbf{x}\|$ is the energy of \mathbf{x} . When training our decoder, 50% of the time we use the raw attention output from A as the alignment $\mathbf{d}_{\text{align}}$ to make the gradient backpropagate through the text aligner, and another 50% of the time we use the monotonic version of $\mathbf{d}_{\text{align}}$ through dynamic programming algorithms [8]. This ensures that the decoder can synthesize intelligible speech from monotonic hard alignment provided during inference. The effectiveness of this 50%-50% training scheme is examined in section 4.4.

TMA objectives. We fine-tune our text aligner with the original sequence-to-sequence ASR loss function \mathcal{L}_{s2s} to ensure that correct attention alignment is kept during the E2E training, where N is the number of phonemes in \mathbf{t} , \mathbf{t}_i is the i -th phoneme token of \mathbf{t} , $\hat{\mathbf{t}}_i$ is the i -th predicted phoneme token, and $\text{CE}(\cdot)$ denotes the cross-entropy loss function. Since this alignment is not necessarily monotonic, we use a simple monotonic loss $\mathcal{L}_{\text{mono}}$ that forces the soft attention alignment to be close to its non-differentiable monotonic version, where \mathbf{d}_{hard} is the monotonic hard alignment obtained through dynamic programming algorithms. A detailed discussion is provided in Appendix B.1.

$$\mathcal{L}_{\text{s2s}} = \mathbb{E}_{\mathbf{x}, \mathbf{t}} \left[\sum_{i=1}^N \text{CE}(\mathbf{t}_i, \hat{\mathbf{t}}_i) \right] \quad (3)$$

$$\mathcal{L}_{\text{mono}} = \mathbb{E}_{\mathbf{x}, \mathbf{t}} [\|\mathbf{d}_{\text{align}} - \mathbf{d}_{\text{hard}}\|_1] \quad (4)$$

Adversarial objectives. Similar to VITS, we employ the following two adversarial loss functions to improve the sound quality of the reconstructed mel-spectrogram: the original cross-entropy loss function \mathcal{L}_{adv} for adversarial training and the additional feature-matching loss [48] \mathcal{L}_{fm} , where $\hat{\mathbf{x}}$ is the reconstructed mel-spectrogram by G , T is the total number of layers in D and D^l denotes the output feature map of l -th layer with N_l number of features.

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\mathbf{x}, \mathbf{t}} [\log D(\mathbf{x}) + \log (1 - D(\hat{\mathbf{x}}))] \quad (5)$$

$$\mathcal{L}_{\text{fm}} = \mathbb{E}_{\mathbf{x}, \mathbf{t}} \left[\sum_{l=1}^T \frac{1}{N_l} \|D^l(\mathbf{x}) - D^l(\hat{\mathbf{x}})\|_1 \right] \quad (6)$$

First stage full objectives. Our full objective functions in the first stage can be summarized as follows with hyperparameters λ_{s2s} , λ_{mono} , λ_{adv} and λ_{fm} :

$$\min_{G,A,E,F,T} \max_D \mathcal{L}_{rec} + \lambda_{s2s}\mathcal{L}_{s2s} + \lambda_{mono}\mathcal{L}_{mono} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{fm}\mathcal{L}_{fm} \quad (7)$$

2.2.2 Second Stage Objectives

Duration prediction. We employ the L_1 loss to train our duration predictor

$$\mathcal{L}_{dur} = \mathbb{E}_a [\|\mathbf{a} - \mathbf{a}_{pred}\|_1] \quad (8)$$

where \mathbf{a} is the ground truth duration obtained by summing \mathbf{d}_{align} along the mel frame axis. $\mathbf{a}_{pred} = L(S(\mathbf{h}_{text}, s))$ is the predicted duration under the style vector s .

Prosody prediction. We train our prosody predictor via a novel data augmentation scheme. Instead of using the ground truth alignment and prosody of the original mel-spectrogram, we first apply a 1-D interpolation to stretch or compress the mel-spectrogram in time so the speech becomes either slower or faster. The alignment, pitch, and energy are then extracted through this modified mel-spectrogram. Even though the speed of the speech has been changed, the pitch and energy of the original speech stay the same. This introduces prediction invariance of pitch and energy independent of the speed of speech. Since the predicted duration has no direct interaction with the generated mel-spectrogram during training, predicted prosody is susceptible to incorrect duration prediction. By introducing invariance of predicted prosody over phoneme duration, we can alleviate the problems of unnatural prosody when the predicted duration is wrong. Specifically, the prosody predictor is trained with:

$$\mathcal{L}_{f0} = \mathbb{E}_{\tilde{p}} [\|\tilde{p} - P_p(S(\mathbf{h}_{text}, s) \cdot \tilde{\mathbf{d}}_{align})\|_1] \quad (9)$$

$$\mathcal{L}_n = \mathbb{E}_{\tilde{x}} [\|\|\tilde{x}\| - P_n(S(\mathbf{h}_{text}, s) \cdot \tilde{\mathbf{d}}_{align})\|_1] \quad (10)$$

where \tilde{p} , $\|\tilde{x}\|$ and $\tilde{\mathbf{d}}_{align}$ are the ground truth pitch, energy and alignment of $\tilde{x} \in \tilde{\mathcal{X}}$ the augmented dataset. P_p denotes the pitch output from the prosody predictor and P_n denotes the energy output.

Decoder reconstruction. Lastly, we want to ensure that the predicted pitch and energy can be utilized by the decoder. Since the mel-spectrogram is stretched or compressed, using them as the ground truth may lead to unwanted artifacts in the predicted prosody. Instead, we train the prosody predictor to produce pitch and energy usable to reconstruct the decoder output

$$\mathcal{L}_{de} = \mathbb{E}_{\tilde{x}, t} [\|\hat{x} - G(\mathbf{h}_{text} \cdot \tilde{\mathbf{d}}_{align}, s, \hat{p}, \hat{n})\|_1] \quad (11)$$

where $\hat{x} = G(\mathbf{h}_{text} \cdot \tilde{\mathbf{d}}_{align}, s, \tilde{p}, \|\tilde{x}\|)$ is the reconstruction of augmented sample $\tilde{x} \in \tilde{\mathcal{X}}$, $p = P_p(S(\mathbf{h}_{text}, s) \cdot \tilde{\mathbf{d}}_{align})$ the predicted pitch and $n = P_n(S(\mathbf{h}_{text}, s) \cdot \tilde{\mathbf{d}}_{align})$ the predicted energy.

Second stage full objectives. Our full objective functions in the second stage can be summarized as follows with hyperparameters λ_{dur} , λ_{f0} , and λ_n :

$$\min_{S,L,P} \mathcal{L}_{de} + \lambda_{dur}\mathcal{L}_{dur} + \lambda_{f0}\mathcal{L}_{f0} + \lambda_n\mathcal{L}_n \quad (12)$$

3 Experiments

3.1 Datasets

We conducted experiments on two datasets. We trained a single-speaker model on the LJSpeech dataset [49]. The LJSpeech dataset consists of 13,100 short audio clips with a total duration of approximately 24 hours. We used the same split as VITS where the training set contains 12,500 samples, the validation set 100 samples and the test set 500 samples. We also trained a multi-speaker

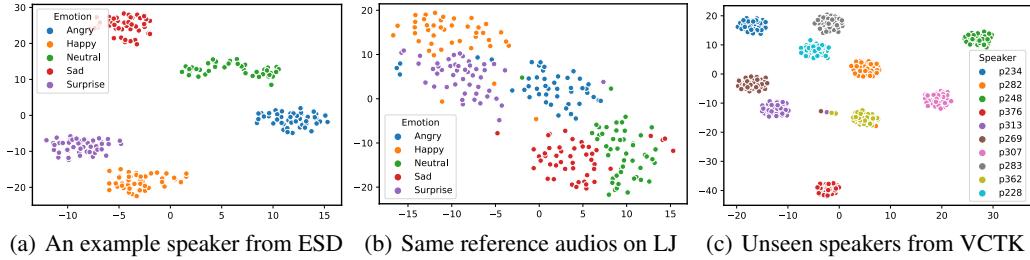


Figure 2: t-SNE visualization of style vectors. All styles are learned without explicit emotion or speaker labels. (a) Style vectors of reference audios in five different emotions of the speaker 0017 in ESD, computed by the multi-speaker model trained on ESD. (b) Style vectors of the same reference audios as in Fig. 2a, computed by the single-speaker model trained on the LJSpeech dataset. (c) Style vectors from the model trained on the LibriTTS data of 10 unseen speakers in the VCTK dataset.

model on the LibriTTS dataset [50]. The LibriTTS train-clean-460 subset consists of approximately 245 hours of audio from 1,151 speakers. We removed utterances with a duration longer than 30 seconds and shorter than one second. We randomly split the combined train-clean-460 subset into a training (98%), a validation (1%), and a test (1%) set and use the test set for evaluation following [37].

In addition, we trained a multi-speaker model on the emotional speech dataset (ESD) [51] to demonstrate the capacity of synthesizing speech with diverse prosodic patterns. ESD consists of 10 Chinese and 10 English speakers reading the same 400 short sentences in five different emotions. We trained our model on 10 English speakers with all five emotions. We also used the VCTK [52] dataset to show that our model is capable of zero-shot speaker adaptation. We upsampled samples from the LJSpeech dataset and ESD into 24 kHz to match those in the LibriTTS dataset. We converted the text sequences into phoneme sequences using an open-source tool¹. We extracted mel-spectrograms with a FFT size of 2048, hop size of 300, and window length of 1200 in 80 mel bins using TorchAudio [53].

3.2 Training

For both stages, we trained all models for 200 epochs using the AdamW optimizer [54] with $\beta_1 = 0, \beta_2 = 0.99$, weight decay $\lambda = 10^{-4}$, learning rate $\gamma = 10^{-4}$ and batch size of 64 samples. We set $\lambda_{s2s} = 0.2, \lambda_{adv} = 1, \lambda_{mono} = 5, \lambda_{fm} = 0.2, \lambda_{dur} = 1, \lambda_{f0} = 0.1$, and $\lambda_n = 1$. This setting of hyperparameters makes sure that all loss values are in the same scale and the training is not sensitive to these hyperparameters. We randomly divided the mel-spectrograms into segments of the shortest length in the batch. The training was conducted on a single NVIDIA A40 GPU.

3.3 Evaluations

We performed two subjective evaluations: mean opinion score of naturalness (MOS-N) to measure the naturalness of synthesized speech, and mean opinion score of similarity (MOS-P) to evaluate the similarity between synthesized speech and reference for the multi-speaker model. We recruited native English speakers located in the U.S. to participate in the evaluations on Amazon Mechanical Turk. In every experiment, we randomly selected 100 text samples from the test set. For each text, we synthesized speech using our model and the baseline models and included the ground truth for comparison. The baseline models include Tacotron 2 [20], FastSpeech 2 [10], and VITS [1] (see Appendix C). The generated mel-spectrograms were converted into waveforms using the Hifi-GAN

Table 1: Comparison of evaluated MOS with 95% confidence intervals (CI) on the LJSpeech dataset.

Model	MOS-N (CI)
Ground Truth	4.32 (± 0.04)
Tacotron 2 + HiFi-GAN	3.01 (± 0.06)
FastSpeech 2 + HiFi-GAN	2.97 (± 0.06)
VITS	3.78 (± 0.06)
StyleTTS + HiFi-GAN	4.01 (± 0.05)

¹<https://github.com/Kyubyong/g2p>

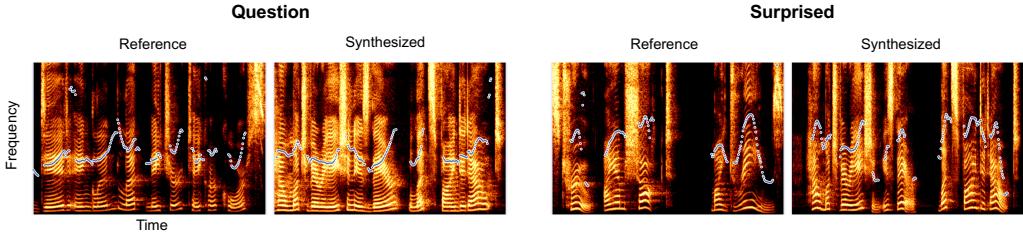


Figure 3: Spectrograms of example reference audios and their corresponding generated speech reading “How much variation is there? Let’s find it out.” from the single-speaker model trained on LJSpeech. The estimated pitch contour is shown as white dots. **Left:** Reference audio is a question, “Did England let nature take her course?”. Note the pitch is mostly going up at the end of each word. The same pattern of pitch rising at the end of the words is present in synthesized speech. **Right:** Reference audio is surprised speech saying “It’s true! I am shocked! My dreams!”. Note the pitch goes up first and then down for each word. Synthesized speech has the same pattern of the pitch going up and down for most of the words. Same patterns of pauses between words are also present.

vocoder [55] for all models. Each set of speech was rated by 10 raters on a scale from 1 to 5 with 0.5 point increments. For a fair comparison, we downsampled our synthesized audio into 22 kHz to match those from baseline models. We used random references when synthesizing speech with our single-speaker models. When evaluating each set, we randomly permuted the order of the models and instructed the subjects to listen and rate them without telling them the model labels. It is similar to multiple stimuli with hidden reference and anchor (MUSHRA), allowing the subjects to compare subtle differences among models. We used the ground truth as hidden attention checkers: raters were dropped from analysis if the MOS of the ground truth was not ranked top two among all the models.

4 Results

4.1 Model Performance

The results of human subjective evaluation on the LJSpeech and LibriTTS dataset are shown in Tables 1 and 2. StyleTTS significantly outperforms other models in the LJSpeech dataset (Table 1). It can be seen that our multi-speaker model also outperforms other models in both naturalness (MOS-N) and similarity (MOS-P) on the LibriTTS dataset (Table 2). Our models are more robust compared to other models (Table 4), especially for long input texts. Since we do not use generative flows that require inverse Jacobian computation, our model is also faster than VITS for inference (Table 5).

4.2 Visualization of Style Vectors

To verify that our model can learn meaningful style representations, we projected the style vectors extracted from reference audios into a 2-D plane for visualization using t-SNE [56]. We selected 50 samples of each emotion from a single speaker in ESD and projected the style vectors of each audio into the 2-D space. It can be seen in Fig. 2(a) that our style vector distinctively encodes the emotional tones of reference sentences even though the training does not use emotion labels. We also computed the style vectors using speech samples from the same speaker with our single-speaker model. This model is only trained on the LJSpeech dataset and therefore has never seen the selected speaker from ESD during training. Nevertheless, in Fig. 2(b), we see that our model can still clearly capture the emotional tones of the sentences, indicating that even when the reference audio is from a

Table 2: Comparison of evaluated MOS with 95% confidence intervals (CI) on the LibriTTS dataset.

Model	MOS-N (CI)	MOS-S (CI)
Ground Truth	4.35 (\pm 0.04)	3.90 (\pm 0.07)
FastSpeech 2 + HiFi-GAN	3.00 (\pm 0.06)	3.51 (\pm 0.07)
VITS	3.62 (\pm 0.06)	3.70 (\pm 0.07)
StyleTTS + HiFi-GAN	4.03 (\pm 0.05)	3.79 (\pm 0.07)

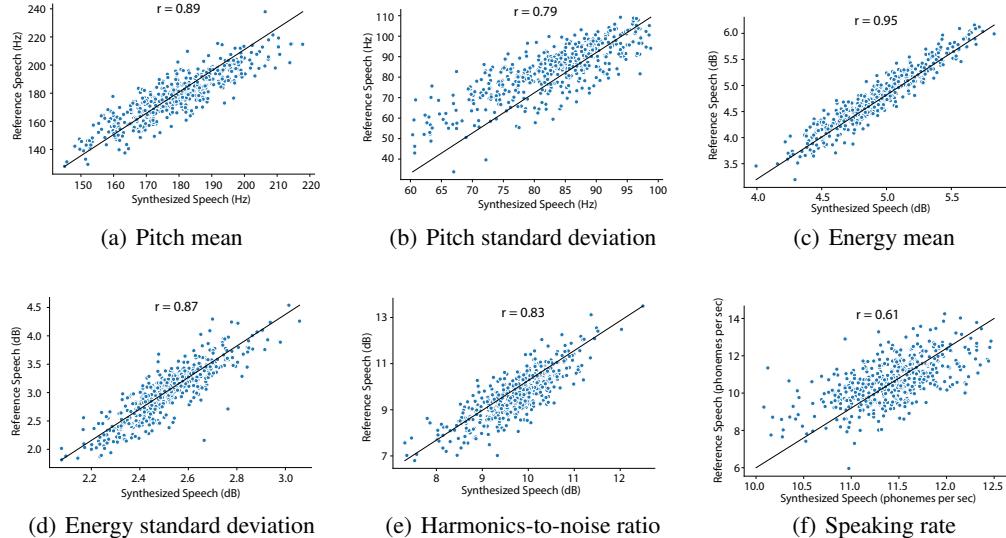


Figure 4: Pearson correlation coefficients of six acoustic features associated with emotions between reference and synthesized speech on LJ Speech dataset.

speaker different from the single speaker seen during training, it still can synthesize speech with the correct emotional tones. This shows that our model can implicitly extract emotions from an unlabeled dataset in a self-supervised manner. Lastly, we show projected style vectors from 10 unseen VCTK speakers each with 50 samples in Fig 2(c). Different speakers are perfectly separated from each other in the 2-D projection. This indicates that our model can learn speaker identities without explicit speaker labels and hence perform zero-shot speaker adaptation (see Appendix E for more details).

4.3 Style-Enabled Diverse Speech Synthesis

To show that the learned style vectors indeed enable diverse speech synthesis, we provide an example of synthesized speech with two different reference audios using our single-speaker model trained on the LJSpeech dataset in Figure 3. It can be seen clearly that the synthesized speech captures various aspects of the reference speech, including the pitch, prosody, pauses, and formant transitions.

To systematically quantify this effect, we drew six scatter plots showing the correlations between synthesized and reference speech in acoustic features traditionally used for speech emotion recognition (Figure 4). The six features are pitch mean, pitch standard deviation, energy mean, energy standard deviation, harmonics-to-noise ratio, and speaking rate [57]. All six features demonstrate a significant correlation between the synthesized and reference speech ($p < 0.001$) with the correlation coefficients all above 0.6. The results indicate that multiple aspects of the synthesized speech are matched to the reference, allowing flexible control over synthesized speech simply by selecting appropriate reference audios. Since our models also allow fully controllable pitch, energy, and duration, our approach is among the most flexible models in terms of controllability for speech synthesis. Our model also outperforms other models on multi-speaker datasets in acoustic feature correlations (Table 4).

4.4 Ablation Study

We further conduct an ablation study to verify the effectiveness of each component in our model with experiments of subjective human evaluation. We instructed the subjects to compare our single-speaker model to the models with one component ablated. We converted the ratings into comparative mean opinion scores (CMOS) by taking the difference between the MOS of the baseline model and component-ablated models. The results are shown in table 3 with more details in Appendix B.

The left-most table shows the results related to our Transferable Monotonic Aligner (TMA) training. We see that when training consists of 100% hard alignment so that no gradient is back-propagated to the parameters of the aligner (equivalent to using an external aligner such as in FastSpeech 2),

Table 3: Ablation study for verifying the effectiveness of each proposed component.

Model	CMOS	Model	CMOS	Model	CMOS
StyleTTS	0	StyleTTS	0	StyleTTS	0
w/ 100% hard	-0.26	w/o pitch extractor	-0.11	w/o residual	-0.30
w/ 0% hard	-2.98	w/o pre-trained aligner	-0.39	AdaIN → AdaLN	-0.21
w/o \mathcal{L}_{mono}	-0.10	w/o augmentation	-0.39	AdaIN → Concat.	-0.17
w/o \mathcal{L}_{s2s}	-2.48	w/o discriminator	-1.79	AdaIN → IN	-0.03

the rated MOS is decreased by -0.26. This is due to the covariate shift between the training data (LibriSpeech) and testing data (LJ Speech). An example of bad alignment of the pre-trained external aligner is shown in Figure 5. This shows that fine-tuning the aligner is effective in improving the quality of synthesized speech. However, when taken to another extreme of using 0% hard alignment (100% soft attention alignment), the model gets overfitted to reconstructing speech with soft alignment and is unable to produce audible speech using hard alignment during inference (-2.98 CMOS). We also see that both TMA objectives (equation 3 and 4) are important for high-quality speech synthesis.

The table in the middle shows the effects of removing various training techniques and components. Using an external pitch extractor (such as acoustic-based methods in FastPitch) decreases MOS by -0.11. This is likely caused by the acoustic-based pitch extraction method sometimes failing to extract the correct F0 curve, and fine-tuning the pitch extractor along with the decoder makes the model learn better pitch representation (see Appendix B.3). Without a pre-trained text aligner (such as VITS), rated MOS is decreased by -0.39. This indicates that our transfer learning is helpful for training reliable text aligners. Removing our novel duration-invariant data augmentation also lowers the performance. Lastly, training without discriminators significantly affects the perceived sound quality.

The rightmost table shows architecture changes by removing the residual features and replacing the AdaIN components in the decoder and predictor with instance normalization (IN), AdaLN, and simple feature concatenation (Concat). Their effects on diversity are also shown in Table 7. Removing the residual features in the decoder decreases both naturalness and correlations between synthesized and reference speech. Layer normalization is also worse than IN for both metrics. Concatenating styles in place of AdaIN dramatically decreases the correlations and lowers rated naturalness, confirming our hypothesis that all previous methods that use concatenation ([1, 8, 23, 11, 12, 14, 15]) are not as effective as AdaIN due to the lack of temporal modulations (see Appendix B.2). Lastly, we see that replacing AdaIN with IN does not significantly affect the rated naturalness, suggesting that the improved naturalness is not due to the introduction of styles but our novel technical improvements including TMA, data augmentation, use of IN, pitch extractor, and residual features. Nevertheless, styles enable diverse speech synthesis which models without styles cannot do.

5 Conclusions

In this work, we proposed StyleTTS, a parallel TTS model that can synthesize natural and diverse speech from reference audios. We take advantage of parallel TTS systems and introduce style information through AdaIN, which we show is an effective approach to incorporating styles into TTS systems. The style vectors from our model encode a rich set of information in the reference audio, including pitch, energy, speaking rates, formant transitions, and speaker identities. They together form generalized speech styles that are lacking in most TTS systems. Our model allows easy control of the prosodic patterns and emotional tones of the synthesized speech by choosing an appropriate reference style while benefiting the robust and fast speech synthesis of parallel TTS systems. The experimental results show that our method outperforms state-of-the-art TTS models. Our approach allows various applications of which other TTS models without styles are not capable, including movie dubbing, book narration, unsupervised speech emotion recognition, personalized speech generation, and any-to-any voice conversion (see Appendix E for more details).

We also note that although our model benefits from the freedom of choosing any audio as the reference for synthesis, a randomly chosen reference is not always the best for speech with specific contexts such as in book narrations. Future work includes discovering a better way of selecting references

and predicting the most suitable style from the text. We will release our source code and pre-trained models to further research in the future directions.

6 Acknowledgments

We thank Gavin Mischler and Vinay Raghavan for their efforts in listening to and rating the naturalness of the synthesized speech and providing feedback for the quality of models during the development stage of this work. This work was funded by the national institute of health (NIHNIDCD) and a grant from Marie-Josée and Henry R. Kravis.

References

- [1] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540, 18–24 Jul 2021.
- [2] Ye Jia, Heiga Zen, Jonathan Shen, Yu Zhang, and Yonghui Wu. Png bert: augmented bert on phonemes and graphemes for neural tts. *arXiv preprint arXiv:2103.15060*, 2021.
- [3] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*, 2021.
- [4] Wei-Ning Hsu, Yu Zhang, Ron Weiss, Heiga Zen, Yonghui Wu, Yuan Cao, and Yuxuan Wang. Hierarchical generative modeling for controllable speech synthesis. In *International Conference on Learning Representations*, 2019.
- [5] Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE, 2019.
- [6] Yoonhyung Lee, Joongbo Shin, and Kyomin Jung. Bidirectional variational inference for non-autoregressive text-to-speech. In *International Conference on Learning Representations*, 2020.
- [7] Rafael Valle, Kevin J Shih, Ryan Prenger, and Bryan Catanzaro. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis. In *International Conference on Learning Representations*, 2020.
- [8] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33, 2020.
- [9] Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6189–6193. IEEE, 2020.
- [10] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*, 2021.
- [11] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR, 2018.
- [12] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, pages 5180–5189. PMLR, 2018.
- [13] Liping Chen, Yan Deng, Xi Wang, Frank K Soong, and Lei He. Speech bert embedding for improving prosody in neural tts. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6563–6567. IEEE, 2021.

- [14] Guangzhi Sun, Yu Zhang, Ron J Weiss, Yuan Cao, Heiga Zen, and Yonghui Wu. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6264–6268. IEEE, 2020.
- [15] Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li. Expressive tts training with frame and style reconstruction loss. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.
- [16] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: fast, robust and controllable text to speech. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3171–3180, 2019.
- [17] Adrian Łaniccki. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592. IEEE, 2021.
- [18] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.
- [19] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech*, page pages 4006–4010, 2017.
- [20] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerry-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.
- [21] Yi Ren, Jinglin Liu, and Zhou Zhao. Portaspeech: Portable and high-quality generative text-to-speech. *Advances in Neural Information Processing Systems*, 34, 2021.
- [22] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. *arXiv preprint arXiv:2105.06337*, 2021.
- [23] Mingjian Chen, Xu Tan, Yi Ren, Jin Xu, Hao Sun, Sheng Zhao, Tao Qin, and Tie-Yan Liu. Multispeech: Multi-speaker text to speech with transformer. In *Proc. Interspeech*, page pages 4024–4028, 2020.
- [24] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.
- [25] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.
- [26] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10551–10560, 2019.
- [27] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.
- [28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
- [29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.
- [30] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021.

- [31] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020.
- [32] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020.
- [33] Ju Chieh Chou and Hung-Yi Lee. One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization. In *Proc. Interspeech 2019*, pages 664–668, 2019.
- [34] Yen-Hao Chen, Da-Yi Wu, Tsung-Han Wu, and Hung-yi Lee. Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5954–5958. IEEE, 2021.
- [35] Yinghao Aaron Li, Ali Zare, and Nima Mesgarani. StarGANv2-VC: A Diverse, Unsupervised, Non-Parallel Framework for Natural-Sounding Voice Conversion. In *Proc. Interspeech 2021*, pages 1349–1353, 2021.
- [36] Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Sheng Zhao, and Tie-Yan Liu. Adaspeech: Adaptive text to speech for custom voice. *arXiv preprint arXiv:2103.00993*, 2021.
- [37] Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. Meta-stylespeech: Multi-speaker adaptive text-to-speech generation. *arXiv preprint arXiv:2106.03153*, 2021.
- [38] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502, 2017.
- [39] Chenfeng Miao, Liang Shuang, Zhengchen Liu, Chen Minchuan, Jun Ma, Shaojun Wang, and Jing Xiao. Efficienttts: An efficient and high-quality text-to-speech architecture. In *International Conference on Machine Learning*, pages 7700–7709. PMLR, 2021.
- [40] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, RJ Skerry-Ryan, and Yonghui Wu. Parallel tacotron 2: A non-autoregressive neural tts model with differentiable duration modeling. *arXiv preprint arXiv:2103.14574*, 2021.
- [41] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.
- [42] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.
- [43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [44] Paul Boersma et al. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences*, volume 17, pages 97–110. Citeseer, 1993.
- [45] Sangeun Kum and Juhan Nam. Joint detection and classification of singing voice melody using convolutional recurrent neural networks. *Applied Sciences*, 9(7):1324, 2019.
- [46] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.
- [47] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [48] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016.
- [49] Keith Ito and Linda Johnson. The lj speech dataset. <https://keithito.com/LJ-Speech-Dataset/>, 2017.

- [50] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.
- [51] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924. IEEE, 2021.
- [52] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). 2019.
- [53] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsich, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi. Torchaudio: Building blocks for audio and speech processing. *arXiv preprint arXiv:2110.15018*, 2021.
- [54] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam, 2018.
- [55] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33, 2020.
- [56] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [57] Carlos Busso, Murtaza Bulut, Shrikanth Narayanan, J Gratch, and S Marsella. Toward effective automatic recognition systems of emotion in speech. *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds*, pages 110–127, 2013.
- [58] John Garofolo, David Graff, Doug Paul, and David Pallett. Csr-i (wsj0) complete ldc93s6a. *Web Download. Philadelphia: Linguistic Data Consortium*, 83, 1993.
- [59] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211, 2018.
- [60] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.
- [61] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.
- [62] Seung-won Park, Doo-young Kim, and Myun-chul Joe. Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data. *arXiv preprint arXiv:2005.03295*, 2020.
- [63] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 862–871, 2021.
- [64] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.
- [65] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29:901–909, 2016.

Appendix A Objective Evaluation

We evaluated the robustness of the models to different lengths of text input. We created four test sets with text length $L < 10$, $10 \leq L < 50$, $50 \leq L < 100$, and $100 \leq L$, respectively. Each set contains 100 texts sampled from the WSJ0 dataset [58]. We calculated the word error rate of the synthesized speech from both single-speaker and multi-speaker models using a pre-trained ASR model² from ESPnet [59]. The results are shown in Table 4. Our model is more robust than other models in most cases except one case where the input number of words is less than 10 in multi-speaker settings.

Table 4: Robustness evaluation on the LJSpeech and LibriTTS dataset. Word error rates (%) are reported for different lengths of text (L).

Model	WER (\downarrow)			
	$L < 10$	$10 < L < 50$	$50 < L < 100$	$L > 100$
<i>Single-speaker models (on LJSpeech)</i>				
Tacotron 2 ++ HiFi-GAN	17.22	12.61	16.95	46.33
FastSpeech 2 + HiFi-GAN	15.37	11.02	14.42	23.04
VITS	14.35	10.69	12.59	32.39
StyleTTS + HiFi-GAN	9.42	7.44	11.97	22.24
<i>Multi-speaker models (on LibriTTS)</i>				
FastSpeech 2 + HiFi-GAN	12.73	8.90	17.20	17.48
VITS	20.97	15.67	20.95	21.05
StyleTTS + HiFi-GAN	17.35	8.26	14.58	15.83

We also measured the inference speed with the real-time factor (RTF), which denotes the time (in seconds) needed for the model to synthesize a one-second waveform. RTF was measured on a server with one NVIDIA 2080Ti GPU and a batch size of 1. Our model is faster than the state-of-the-art model, VITS [1], even though our model was not trained end-to-end like VITS (Table 5). We believe it is possible to make the inference time shorter if we train in an end-to-end manner in future work.

Table 5: Real time factor (RTF) in second.

Model	RTF (s)
Tacotron 2 + HiFi-GAN	0.0868
VITS	0.0428
StyleTTS + HiFi-GAN	0.0388

In addition, we conducted the same analysis on the correlations of acoustic features associated with emotions between reference and synthesized speech using four multi-speaker models. Since there is no style in FastSpeech 2 and VITS, we used a pre-trained X-vector model [60] from Kaldi [61] to extract the speaker embedding as the reference vector. The results are given in Table 6. We can see that our model obtains higher correlation coefficients of every acoustic feature for the multi-speaker datasets.

Appendix B Ablation Study Details

In this section, we describe the detailed settings of each condition in Table 3 and provide more discussions of the results in Table 3 and Table 7.

B.1 TMA-related

There are three Transferable Monotonic Aligner (TMA) related innovations in this work: the decoder is trained with hard monotonic alignment and soft attention in a 50%-50% manner and two TMA

²kamo-naoyuki/librispeech_asr_train_asr_conformer5_raw_bpe5000_scheduler_confwar_mup_steps25000_batch_bins140000000_optim_conf1r0.0015_initnone_accum_grad2_sp_valid

Table 6: Comparison of Pearson correlation coefficients of acoustic features associated with emotions between reference and synthesized speech in multi-speaker experiments. Fastspeech 2 and VITS employ the X-vector as the reference.

Model	Pitch mean	Pitch standard deviation	Energy mean	Energy standard deviation	Harmonics-to-noise ratio	Shimmer	Jitter
FastSpeech 2	0.95	0.43	0.23	0.51	0.81	0.81	0.58
VITS	0.97	0.32	0.14	0.5	0.84	0.81	0.54
StyleTTS	0.99	0.51	0.91	0.52	0.9	0.87	0.65

Table 7: Comparison of Pearson correlation coefficients of acoustic features associated with emotions between reference and synthesized speech in ablation study.

Model	Pitch mean	Pitch standard deviation	Energy mean	Energy standard deviation	Harmonics-to-noise ratio	Shimmer	Jitter
Baseline	0.90	0.53	0.77	0.15	0.79	0.66	0.64
AdaIN → AadLN	0.89	0.52	0.67	0.19	0.76	0.53	0.66
AdaIN → Concat.	0.36	0.16	0.19	-0.07	0.58	0.36	0.40
w/o residual	0.88	0.51	0.68	0.11	0.79	0.64	0.60

objectives functions. The 50%-50% training is motivated by the fact that the monotonic alignment search proposed in [8] is not differentiable, and the soft attention alignment does not necessarily provide correct alignments for duration prediction in the second stage of training. This 50%-50% split is arbitrary and can be changed to anything from 10%-90% to 90%-10%, depending on the dataset and the application. When the ratio is 100%-0%, it becomes the case where the external aligners are not fine-tuned like in most parallel TTS systems such as FastSpeech [16], while when the ratio is 0%-100%, it becomes the case we fine-tune the aligner with only soft attention such as in Cotatron [62] for voice conversion applications. We find that training with external aligners (100% hard, no fine-tuning) decreases the naturalness of the synthesized speech because bad alignments can happen due to covariate shifts between the training dataset (LibriSpeech) and testing dataset (LJSpeech) as in the case of Montreal Forced Aligner [38]. One example is given in the leftmost figure in Figure 5. On the other hand, if we only fine-tune the decoder with soft alignment, the decoder will overfit on the soft alignment and be unable to synthesize audible speech from hard alignment because the soft alignments are not either 0 or 1 and the precise numerical values of alignments are used by the decoder to generate speech.

Another notable case is when we do not use a pre-trained text aligner such as in the case of VITS. This case makes MOS even lower than the case of no fine-tuning, suggesting that overfitting on a smaller dataset can be more detrimental than failure in generalization on the TTS dataset for some samples. The figure in the middle in Fig. 5 shows an alignment with gaps and no background noises. This indicates overfitting of the text aligner to the smaller dataset for the mel-spectrogram reconstruction objective. However, since our goal is to synthesize the speech from predicted alignment, overfitting to speech reconstruction can be harmful to natural speech synthesis during inference.

In addition to the 50%-50% training, we also introduced two TMA objectives \mathcal{L}_{s2s} and \mathcal{L}_{mono} . This is motivated by the fact that \mathcal{L}_{s2s} learns correct alignments for S2S-ASR but not necessarily monotonic while non-differentiable monotonic alignments obtained through dynamic programming algorithms proposed in [8] do not necessarily produce correct alignments. By combining \mathcal{L}_{s2s} and \mathcal{L}_{mono} , we can learn an aligner that produces both correct and monotonic alignments.

B.2 AdaIN, AdaLN, and Concatenation

As shown in Table 3 and Table 7, AdaIN outperforms AdaLN and simple concatenation for both naturalness and style reflection. Here we describe our intuitions behind these results.

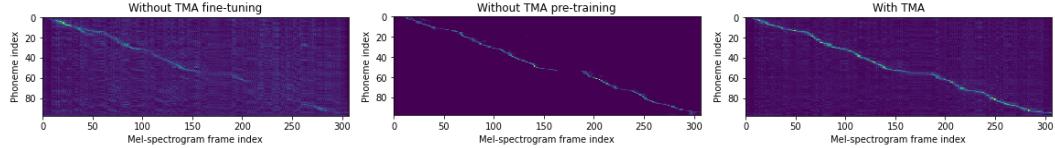


Figure 5: An example showing text alignments under different conditions. **Left:** No TMA fine-tuning (100% hard alignment such as FastSpeech). This is an example of a failed alignment. **Middle:** No pre-trained text aligner (such as VITS). Note the gaps between alignments and clean attention (with no background noise), indicating some degrees of overfitting to the TTS speech dataset. **Right:** Full TMA fine-tuning. Note that TMA learns an alignment that is both continuous and monotonic compared to without fine-tuning and pre-training.

Concatenation vs. AdaIN. When we concatenate the style vector to each frame of the encoded phonetic representations, we create a representation $\mathbf{h}_{\text{style}} = \begin{bmatrix} \mathbf{h}_{\text{text}} \\ \mathbf{s} \end{bmatrix}$. When the $\mathbf{h}_{\text{style}}$ is passed to the next convolution layer whose parameter is W , we get

$$\mathbf{h}_{\text{style}} \cdot W = \begin{bmatrix} \mathbf{h}_{\text{text}} \\ \mathbf{s} \end{bmatrix} \cdot [W_{\text{text}} | W_{\text{style}}] = \mathbf{h}_{\text{text}} \cdot W_{\text{text}} + \mathbf{s} \cdot W_{\text{style}} = \mathbf{h}_{\text{text}} \cdot W_{\text{text}} + \text{Concat}(\mathbf{h}_{\text{text}}, \mathbf{s}) \quad (13)$$

where W_{text} and W_{style} are block matrix notation of the corresponding weights for $\mathbf{h}_{\text{style}}$ and \mathbf{s} and $\text{Concat}(\mathbf{h}_{\text{text}}, \mathbf{s}) = \mathbf{s} \cdot W_{\text{style}}$ denotes the concatenation operation as a function of input \mathbf{h}_{text} and style vector \mathbf{s} . This $\text{Concat}(\mathbf{x}, \mathbf{s})$ function is almost like AdaIN in equation 1 where $L_\mu(s) = W_{\text{style}}$ except we do not have the temporal modulation term $L_\sigma(s)$. The modulation term is very important in style transfer, and some works argue that modulation alone is enough for diverse style representations [29, 63]. In contrast, concatenation only provides the addition term (L_μ) but no modulation term (L_σ). Intuitively, the modulation term can determine the variance of the pitch and energy, for example, and therefore without such a term, correlations for pitch and energy standard deviation are much lower than AdaIN and AdaLN as shown in Table 7.

AdaLN vs. AdaIN. Generative models for speech synthesis learn to generate mel-spectrograms, which is essentially a 1-D feature map with 80 channels. Each channel in the mel-spectrogram represents a single frequency range. When we apply AdaIN, we learn a distribution with a style-specific mean and variance for *each channel*, compared to AdaLN, where a single mean and variance are learned for the *entire feature map*. This inherent difference between feature distributions makes AdaIN more expressive in terms of style reflection than AdaLN.

B.3 Pitch Extractor

Acoustic methods for pitch estimation sometimes fail because of the presence of non-stationary speech intervals and sensitivity of hyper-parameters as discussed in the original papers that propose these methods [44, 46]. A neural network trained with ground truth from these methods, however, can leverage the problems of failed pitch estimation because the failed pitch estimation can be regarded as noises in the training set, so it does not affect the generalization of the pitch extractor network. Moreover, since the pitch extractor is fine-tuned along with the decoder, there is no ground truth for the pitch beside the sole objective that the decoder needs to use extracted pitch information to reconstruct the speech. This fine-tuning allows better pitch representations beyond the original F0 in Hertz, but it also allows flexible pitch control as we can still recognize the pitch curves and edit them later when needed during inference.

Appendix C Subjective Evaluation Details

We used the publicly available pre-trained models as baselines for comparison. For the single-speaker experiment on the LJSpeech dataset, we used pre-trained Tacotron2³, Fastspeech2⁴, HiFiGAN⁵ from ESPnet, and VITS⁶ from the official implementation. We randomly selected 100 text samples from the test set to synthesize the speech. Since audios from our model were synthesized using Hifi-GAN trained with audios sampled at 24 kHz, for a fair comparison, we resampled all the audios into 22 kHz and then normalized their amplitude. For multi-speaker models, because our training did not require speaker labels, for a fair comparison with other models that use explicit speaker embeddings during training, we averaged the style vectors computed using all samples in the training set from the same speaker as the reference style. We used the pre-trained model for Min et. al. [37]⁷ from a public repository in GitHub for comparison of zero-shot speaker adaptation in Appendix E. We did not use the official implementation because the vocoder used was MelGAN sampled at 16 kHz while the implementation we employed uses Hifi-GAN sampled at 22 kHz, which is comparable to other models.

To reduce the listening fatigue, we randomly divided these 100 sets of audios into 5 batches⁸ with each batch containing 20 sets of audios for comparison. We launched the 5 batches sequentially on Amazon Mechanical Turk (AMT)⁹. We required participating subjects to be native English speakers located in the United States. For each batch, we made sure that we had collected completed responses from at least 10 self-reported native speakers whose IP addresses were within the United States and residential (i.e., not VPN or proxies). We used the average score that a subject rated on ground truth audios to check whether this subject carefully finished the survey as the subjects did not know which audio was the ground truth. We then disqualified and dropped all ratings from the subjects whose average ground truth score was not ranked top two among all the models. Finally, 46 out of 50 subjects were qualified for this experiment.

In the multi-speaker experiments, we used pre-trained Fastspeech2¹⁰, VITS¹¹, and HiFiGAN¹² from ESPnet. We used pre-trained VITS from ESPnet instead of the official repository because we need the model to be trained on the LibriTTS dataset; however, the official models were trained on the LJSpeech or VCTK dataset.

Similar to the single-talker experiment, we launched 5 batches¹³ on AMT when we tested the multi-talker models on the LibriTTS dataset. 48 out of 58 subjects were qualified. We launched 3 batches¹⁴ with batch sizes 33, 33, 34, respectively, when we tested the multi-talker models on the VCTK dataset. 28 out of 30 subjects were qualified.

Appendix D Detailed Model Architectures

In this section, we provide detailed model architectures of StyleTTS, which consists of eight modules. Since we use the same text encoder as in Tacotron 2 [20], very similar architecture to the decoder of Tacotron 2 for text aligner and the same architecture as the JDC network [45] for pitch extractor, we leave the readers to the above references for detailed descriptions of these modules. Here, we only provide detailed architectures for the other five modules. All activation functions used are leaky ReLU (LReLU) with a negative slope of 0.2. We apply spectral normalization [64] to all trainable

³The model was kan-bayashi/ljspeech_tacotron2 from ESPNet

⁴The model was kan-bayashi/ljspeech_fastspeech2 from ESPNet

⁵The model was parallel_wavegan/ljspeech_hifigan.v1 from ESPNet

⁶The implementation can be found at <https://github.com/jaywalnut310/vits>

⁷The implementation can be found at <https://github.com/keonlee9420/StyleSpeech>

⁸The survey (batch 1) can be found at <https://survey.alchemer.com/s3/6696322/LJ100-B1>

⁹<https://www.mturk.com/>

¹⁰The model was kan-bayashi/libritts_xvector_conformer_fastspeech2 from ESPNet

¹¹The model was kan-bayashi/libritts_xvector_vits from ESPNet

¹²The model was parallel_wavegan/libritts_hifigan.v1 from ESPNet

¹³The survey (batch 1) can be found at <https://survey.alchemer.com/s3/6705095/LibriTTS-seen100-B1>

¹⁴The survey (batch 1) can be found at <https://survey.alchemer.com/s3/6706053/zero-shot-B1>

Table 8: Decoder architecture. T represents the input length of the mel-spectrogram, p is the input F0, n is the input energy, and s is the style code. \tilde{n} and \tilde{p} are the processed pitch and energy, and h_{res} is the output of the phoneme residual sub-module.

Submodule	External Input	Layer	Norm	Output Shape
F0 processing	p	Input F0 p	-	$1 \times T$
	-	ResBlk	-	$64 \times T$
	-	Conv 1×1	IN	$1 \times T$
Energy processing	n	Input energy n	-	$1 \times T$
	-	ResBlk	-	$64 \times T$
	-	Conv 1×1	IN	$1 \times T$
Phoneme residual	h_{text}	Input h_{text}	-	$512 \times T$
	-	Conv 1×1	IN	$64 \times T$
IN ResBlks	$\tilde{p}, \tilde{n}, h_{\text{text}}$	Concat	-	$(512 + 2) \times T$
	-	ResBlk	IN	$1024 \times T$
	-	ResBlk	IN	$1024 \times T$
AdaIN ResBlks	$\tilde{p}, \tilde{n}, h_{\text{res}}$	Concat	-	$(1024 + 2 + 64) \times T$
	s	ResBlk	AdaIN	$1024 \times T$
	$\tilde{p}, \tilde{n}, h_{\text{res}}$	Concat	-	$(1024 + 2 + 64) \times T$
	s	ResBlk	AdaIN	$1024 \times T$
	$\tilde{p}, \tilde{n}, h_{\text{res}}$	Concat	-	$(1024 + 2 + 64) \times T$
	s	ResBlk	AdaIN	$512 \times T$
	s	ResBlk	AdaIN	$512 \times T$
	s	ResBlk	AdaIN	$512 \times T$
	-	Conv 1×1	-	$80 \times T$

parameters in style encoder and discriminator and weight normalization [65] to those in decoder because they are shown to be beneficial for adversarial training.

Decoder (Table 8). Our decoder takes four inputs: the aligned phoneme representation, the pitch F0, the energy, and the style code. It consists of seven 1-D residual blocks (ResBlk) along with three sub-modules for processing the input F0, energy, and residual of the phoneme representation. The normalization consists of both instance normalization (IN) and adaptive instance normalization (AdaIN). We concatenate (Concat) the processed F0, energy, and residual of phonemes with the output from each residual block as the input to the next block for the first three blocks.

Table 9: Style encoder and discriminator architectures. T represents the input length of the mel-spectrogram, and D is the output dimension. For style encoder, $D = 128$. For discriminator, $D = 1$.

Layer	Pooling	Norm	Output Shape
Mel x	-	-	$1 \times 80 \times T$
Conv 1×1	-	-	$64 \times 80 \times T$
ResBlk	Dilated Conv	-	$128 \times 40 \times T/2$
ResBlk	Dilated Conv	-	$256 \times 20 \times T/4$
ResBlk	Dilated Conv	-	$512 \times 10 \times T/8$
ResBlk	Dilated Conv	-	$512 \times 5 \times T/16$
LReLU	-	-	$512 \times 5 \times T/16$
Conv 5×5	-	-	$512 \times 1 \times T/80$
LReLU	-	-	$512 \times 1 \times T/80$
-	AdaAvg	-	512×1
Linear	-	-	$D \times 1$

Style Encoder and Discriminator (Table 9). Our style encoder and discriminator share the same architecture, which consists of four 2-D residual blocks (ResBlk). The dimension of the style vector is set to 128. We use learned weights for pooling through a dilated convolution (Dilated Conv) layer with a kernel size of 3×3 . We apply an adaptive average pooling (AdaAvg) along the time axis of the feature map to make the output independent of the size of the input mel-spectrogram.

Table 10: Duration and prosody predictor architectures. N represents the number of input phonemes and T represents the length of the alignment. h_{text} is the hidden phoneme representation from the text encoder, d_{align} is the monotonic alignment, s is the style code, a_{pred} is the predicted duration, p_{pred} is the predicted pitch and $\|x\|_{\text{pred}}$ is the predicted energy. h_{prosody} and h_{aprosody} are intermediate outputs from submodules.

Submodule	External Input	Layer	Norm	Output Shape	Submodule Output
Prosody Encoder	h_{text}, s	Concat	-	$(512 + 128) \times N$	
	s	bi-LSTM	AdaLN	$512 \times N$	
	s	Concat	-	$(512 + 128) \times N$	
	s	bi-LSTM	AdaLN	$512 \times N$	h_{prosody}
	s	Concat	-	$(512 + 128) \times N$	
Duration Projection	s	bi-LSTM	AdaLN	$512 \times N$	
	h_{prosody}	bi-LSTM	-	$512 \times N$	
	-	Linear	-	$1 \times N$	a_{pred}
Shared LSTM	$h_{\text{prosody}}, d_{\text{align}}$	Dot	-	$512 \times T$	
	s	Concat	-	$(512 + 128) \times T$	h_{aprosody}
	-	bi-LSTM	-	$512 \times T$	
Pitch Predictor	h_{aprosody}, s	ResBlk	AdaIN	$512 \times T$	
	s	ResBlk	AdaIN	$256 \times T$	
	s	ResBlk	AdaIN	$256 \times T$	p_{pred}
	-	Linear	-	$1 \times T$	
Energy Predictor	h_{aprosody}, s	ResBlk	AdaIN	$512 \times T$	
	s	ResBlk	AdaIN	$256 \times T$	
	s	ResBlk	AdaIN	$256 \times T$	$\ x\ _{\text{pred}}$
	-	Linear	-	$1 \times T$	

Duration and Prosody Predictors (Table 11). The duration predictor and prosody predictors are trained together in the second stage of training. There is a shared 3-layer bidirectional LSTM (bi-LSTM) S between the duration predictor and prosody predictor named text feature encoder, each followed by an adaptive layer normalization (AdaLN). AdaLN is similar to AdaIN where the gain and bias are predicted from the style vector s . However, unlike AdaIN which normalizes each channel independently, AdaLN normalizes the entire feature map. The style vector s is also concatenated (Concat) with the output to every token from each LSTM layer as the input to the next LSTM layer. Lastly, we have a final bidirectional LSTM and a linear projection L that maps h_{prosody} into the predicted duration.

The hidden representation h_{prosody} is dotted with the alignment d_{align} and sent to the prosody decoder. The prosody encoder consists of one bidirectional LSTM and two sets of three residual blocks (ResBlk) with AdaIN followed by a linear projection, one for predicting the F0 and another for predicting the energy, respectively.

Appendix E Zero-Shot Speaker Adaptation and Voice Conversion

We note that our style encoder is speaker-independent and therefore can perform zero-shot speaker adaptation similar to Min et. al. [37]. We compared our models for zero-shot speaker adaptation with an evaluation of naturalness and similarity on the VCTK dataset. The results are shown in table 11.

Table 11: Comparison of evaluated MOS with 95% confidence intervals (CI) on the VCTK dataset for unseen speaker adaptation.

Model	MOS-N (CI)	MOS-S (CI)
Ground Truth	4.39 (± 0.05)	4.28 (± 0.06)
Min et. al.	2.13 (± 0.06)	2.43 (± 0.07)
Proposed	3.55 (± 0.06)	3.57 (± 0.07)

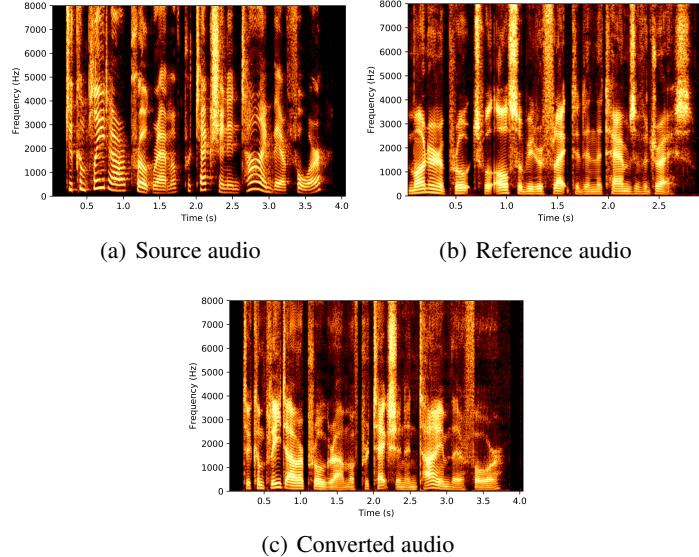


Figure 6: An example of any-to-any voice conversion. The source audio is from the LJSpeech dataset and the reference audio is from the VCTK dataset, both unseen during training.

Moreover, since our text encoder, text aligner, pitch extractor, and decoder are trained in a speaker-agnostic manner, our decoder can reconstruct speech from any aligned phonemes, pitch, energy, and reference speakers. Therefore, our model can perform any-to-any voice conversion by extracting the alignment, pitch, and energy from an input mel-spectrogram and generating speech using a style vector of reference audio from an arbitrary target speaker. Our voice conversion scheme is transcription-guided, similar to Mellotron [9] and Cotatron [62]. We provide one example in Figure 6 with both source and target speaker unseen from the LJSpeech and VCTK datasets. We refer our readers to our demo page for more examples.