Engineering and Technology Journal e-ISSN: 2456-3358

Volume 09 Issue 05 May-2024, Page No.- 3951-3956

DOI: 10.47191/etj/v9i05.10, I.F. – 8.227

Speech Cloning: Text-To-Speech Using VITS

Utkarsh Verma[1], Dr. Padmanaban R[2]

[1,2]Vellore Institute of Technology, Chennai

ABSTRACT: Voice is one of the most common and natural communication methods for humans. Voice is becoming the primary interface for AI voice assistants like Amazon Alexa, as well as in autos and smart home devices. Homes and so on. As human-machine communication becomes more common, researchers are exploring technology that mimics genuine speech. Speech cloning is the practice of copying or mimicking another person's speech, usually utilizing modern technology and artificial intelligence (AI). This entails producing a synthetic or cloned version of someone's voice that sounds very similar to the actual speaker. The objective is to produce speech that is indistinguishable from the genuine person, both in tone and intonation. Instant Voice Cloning (IVC) in text-to-speech (TTS) synthesis refers to the TTS model's capacity to copy the voice of any reference speaker based on a short audio sample, without requiring extra speaker-specific training. This method is usually referred to as zero-shot TTS. IVC provides users with the flexibility to tailor the generated voice, offering significant value across diverse real-world applications. Examples include media content creation, personalized chatbots, and multi-modal interactions between humans and computers or extensive language models.

1 INTRODUCTION

1.1. Speech Cloning

The term "speech cloning" describes the technique of mimicking or reproducing a human voice, usually with the use of cutting-edge technology and artificial intelligence (AI). To do this, a synthetic or cloned voice that can sound amazingly close to the original speaker must be created. The intention is to produce speech that is identical to the genuine speaker in terms of intonation and tone.

Speech cloning can be accomplished through a variety of methods and tools, most frequently deep learning models and neural networks. A popular method entails using a sizable dataset of speech recordings of the target speaker to train a model. This enables the model to pick up on the distinct qualities, nuanced aspects, and speech patterns of the individual. A substantial amount of the target person's speech data is needed to produce a believable clone. Recordings of different words, sentences, and maybe even lengthier discussions are included in this data.

Deep learning models, such recurrent neural networks (RNNs), convolutional neural networks (CNNs), or more sophisticated architectures like transformer models and long short-term memory networks (LSTMs), are frequently used in speech cloning. The model takes out representations, or embeddings, of the voice data that are specific to the speaker's voice characteristics. The audio input is represented concisely and meaningfully by these embeddings. To create fresh, lifelike voice samples, generative models like variational autoencoders (VAEs) or generative adversarial networks (GANs) may be used. This text might come from a variety of sources, including articles, books, websites, or user-generated material. Before converting the text, the TTS system may check it for punctuation, sentence structure, and other linguistic features to determine the appropriate tone and emphasis for speech synthesis. To enhance performance and produce a more accurate representation of the target voice, the model may be fine-tuned utilizing extra data following an initial training phase. Methods for identifying deepfake audio files and artificially generated voices are currently being developed by researchers. The purpose of these detection systems is to find cases of malicious use of cloned voices.

1.2. Text-to-Speech

Text-to-speech (TTS) technology converts written text into spoken language, allowing computers and other devices to "read" it aloud. This technology is crucial in making digital material more accessible to consumers, hence improving the user experience in a wide range of applications. The technique begins with the input of written materials. TTS models use phonetic and prosodic features to produce speech that sounds human. Phonetics is the study of individual sounds, whereas prosody encompasses intonation, rhythm, and stress patterns in speech. TTS systems frequently offer various voices or accents for users to choose from. These voices can be based on recordings of actual human speakers or completely synthetic, created by computers. TTS can work in real time, converting text into voice as the user types it, or in batch mode, converting vast amounts of text into speech in a single operation. Recent advances in artificial intelligence, particularly deep learning, and neural network topologies, have substantially enhanced the naturalness and quality of synthetic speech in TTS systems.

2. METHODOLOGY

In this part, we will discuss our suggested technique and its architecture. The first three subsections primarily cover the proposed method: a conditional VAE formulation, alignment estimation based on variational inference, and adversarial training to improve synthesis quality. This section concludes with a description of the general architecture. Figures 1a and 1b depict the training and inference procedures for our technique, respectively. From now on, we will refer to our technique as Variational Inference with Adversarial Learning for End-to-End Text-to-Speech (VITS). The term "Variational Inference with Adversarial Learning for End-to-End Text-to-Speech" implies a hybrid of variational inference, adversarial learning, and an end-to-end method in the context of Text-to-Speech (TTS) systems. Let us break down the components:

2.1. VITS

2.1.1. Variational Inference

Variational Inference (VI): - Variational Inference is a technique used in Bayesian statistics to estimate complicated probability distributions. VI might be used in TTS to simulate the uncertainty involved in speech generation. It can assist capture the distribution of possible voice outputs given a text input, resulting in more robust and expressive synthesis.

2.1.2. Adversarial Learning

Adversarial Learning: Training a model using a game-like dynamic between two neural networks (generator and discriminator). In the case of TTS, this may include training a generator (speech synthesis model) to create realistic speech and an adversarial network to discriminate between actual and synthetic speech. The adversarial training method is intended to increase the overall quality and naturalness of synthesized speech.

2.1.3. End to End Text-to-Speech

An end-to-end method in TTS refers to a system that accepts raw text as input and creates synthesized voice without explicitly separating intermediary language or acoustic characteristics. End-to-end models seek to streamline the TTS pipeline, frequently employing deep learning techniques to learn complicated mappings straight from text to speech.

Combining these features in "Variational Inference with Adversarial Learning for End-to-End Text-to-Speech" suggests a more advanced method to TTS synthesis. The use of variational inference aids in modeling uncertainties and capturing the natural variability in voice production. Adversarial learning creates a competitive training environment that encourages the model to produce more realistic and natural-sounding speech. The end-to-end method streamlines the design and may enable more seamless integration of different components.

This technique is most likely intended to solve difficulties such as enhancing the naturalness, expressiveness, and robustness of synthesized speech, all of which are critical for developing high-quality and human-like TTS systems. It also suggests a desire to use powerful machine learning techniques to push the limits of what is possible in text-to-speech synthesis. The base speaker TTS model offers a wide range of options. The VITS model may be extended to include style and language embedding in its text encoder and duration predictor. Other options, such as InstructTTS, can also handle style prompts. It is also Microsoft TTS, which supports speech synthesis markup language (SSML) for emotion, pauses, and articulation, is a commercially accessible and cost-effective option. Users can read the text in their preferred style and language without using the base speaker TTS model. VITS is a conditional VAE that aims to maximize the variational lower bound (ELBO) for intractable problems. The marginal log-likelihood of data is

$$\log p_\theta(x|c) \geq \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p_\theta(z|c)}\right]$$

2.2. Tone Colour Converter

"Tone color" in speech may refer to the emotive or expressive elements inherent in a person's voice, such as intonation, pitch changes, and other features that contribute to the overall tone of their speech.

In the context of speech cloning, a "Tone Color Converter" might theoretically be a tool or algorithm that allows for the conversion or adaption of the emotional tone or expressiveness of one speaker's voice to another. This might entail collecting and recreating the subtle emotional aspects of the original speaker's voice throughout the cloning procedure.

The tone color converter uses an encoder-decoder structure with an invertible normalizing flow in the center. The encoder is a 1D convolutional neural network that uses the short-time Fourier converted spectrum of X(LI, SI, and CI) as input. All convolutions are single-layered. The feature maps produced by the encoder are designated as Y(LI, SI, CI).

2.3. Tone Colour Extractor

In speech cloning, a "Tone Color Extractor" may be used to recognize and extract distinct emotional indicators from the original speaker's voice. This might include pitch modulation, intonation patterns, and other factors that influence the emotional tone of the speech. The program may transform the retrieved emotional characteristics into a format suitable for use as input in a speech cloning model. This might entail producing a depiction that captures the emotional subtleties of the voice. Once the emotional tone has been collected and represented, it may be used in a speech cloning system to increase the expressiveness of the cloned speech. The objective might be to make the synthetic speech seem more natural and emotionally engaging, similar to the real speaker. The collected tone color information may also be utilized to supplement training data for speech cloning models, allowing them to better capture and mimic the emotional traits of various speakers.

The tone color extractor is a 2D convolutional neural network that uses the input voice's mel-spectrogram to produce a single feature vector containing tone color information. We first apply it to X(LI, SI, CI) to get vector v(CI), then to X(LO, SO, CO) to get vector v(CO). The normalizing flow layers use Y(LI, SI, CI) and v(CI) as input and produce a feature representation Z(LI, SI) that removes tone color information while retaining the remaining style attributes.

The characteristic Z(LI,SI) aligns with the International Phonetic Alphabet (IPA) in the time dimension. The next part will discuss the process of learning feature representations. The normalizing flow layers are then applied in the inverse direction, with Z(LI,SI) and v(CO) as input and Y(LI,SI,CO) as output. This phase incorporates the tone color CO from the reference speaker into the feature maps. HiFi-Gan decodes Y(LI,SI,CO) into raw waveforms X(LI,SI,CO) via a stack of transposed 1D convolutions.

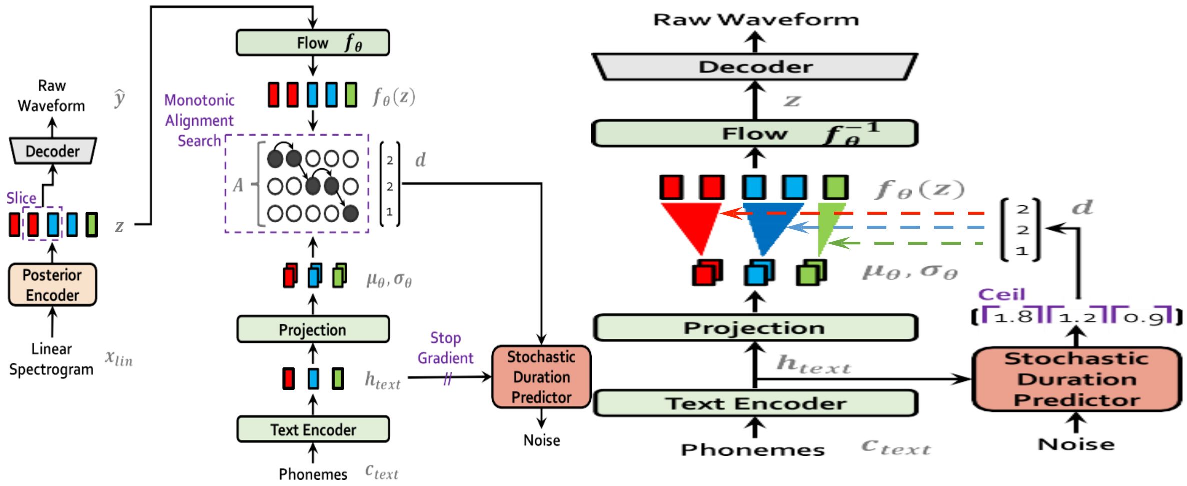

Figure 1. System diagram depicting (a) training procedure and (b) inference procedure. The proposed model can be viewed as a conditional VAE; a posterior encoder, decoder, and conditional prior (green blocks: a normalizing flow, linear projection layer, and text encoder) with a flow-based stochastic duration predictor.

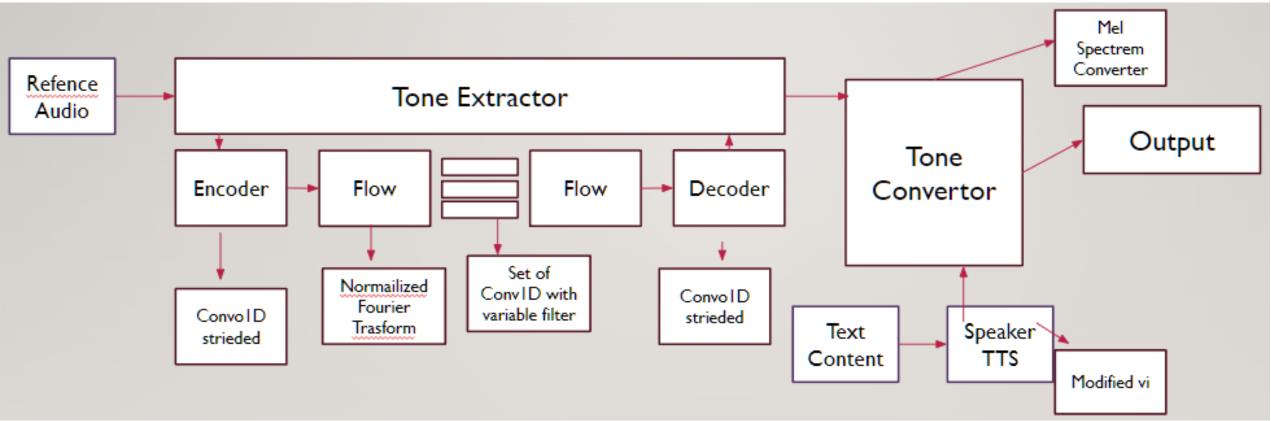

Figure 2. **Workflow Architecture**

3. TRAINING LOSS

3.1. GAN Loss Function

In the context of Generative Adversarial Networks (GANs), the loss function is an important component that directs training. A GAN is made up of two neural networks, a generator, and a discriminator, which are trained concurrently in a competitive way. The loss functions for the generator and discriminator are essential to this adversarial training method. The following are the typically used loss functions for GANs:

-Generator Loss: The generator aims to create data that is indistinguishable from genuine data. The generator loss indicates how successfully the generator fools the discriminator. The minimax loss is a typical loss function for generators.

-Loss of Discriminator: The discriminator's ability to discern between produced and genuine data is measured by the discriminator loss.

-Generator Regularization: To promote a variety of realistic outputs, a regularization term may occasionally be introduced to the generator loss. This can involve concepts like gradient penalty or feature matching loss.

We refer to MSD and MPD as a single discriminator. The training objectives for the generator and discriminator are based on LSGAN (Mao et al., 2017), replacing the binary cross-entropy terms from the original GAN. Goodfellow et al. (2014) used least squares loss functions for non-vanishing gradient flows. The discriminator is taught to categorize ground truth samples as 1 and generated samples as 0. The generator is trained to mimic the discriminator by adjusting sample quality to almost equal 1. The GAN losses for generator G and discriminator D are defined as:

$$\mathcal{L}_{Adv}(D;G) = \mathbb{E}_{(x,s)}\Bigg[(D(x)-1)^2 + (D(G(s)))^2\Bigg]$$

$$\mathcal{L}_{Adv}(G;D) = \mathbb{E}_{s}\Bigg[(D(G(s))-1)^2\Bigg]$$

3.2. Mel Spectrogram Loss Function

A loss function called the Mel-Spectrogram Loss is used to calculate how much two Mel-Spectrograms vary from one another. When the objective is to create or alter audio signals while preserving their spectrogram properties, this loss is frequently used. In essence, the Mel-Spectrogram Loss measures how different the created and original Mel-Spectrograms are from one another. Depending on the demands and nature of the activity, this loss may be expressed in many ways. A popular option is the Mean Squared Error (MSE) loss, which calculates the average squared difference between the two Mel-Spectrograms' corresponding elements. To enhance the quality of the produced audio and the generator's training effectiveness, we use mel-spectrogram loss in addition to the GAN objective. According to earlier research, adding a reconstruction loss to the GAN model aids in producing results that are more realistic. Similarly, Yamamoto et al. (2020) optimize both the adversarial loss functions and the multi-resolution spectrogram in order to successfully capture the time-frequency distribution. Because of the properties of the human auditory system, we employed mel-spectrogram loss in accordance with the input circumstances, which may also be anticipated to have the impact of concentrating more on enhancing the perceptual quality. The L1 difference between the mel-spectrogram of a waveform that the generator produced and that of a ground truth waveform is known as the mel-spectrogram loss. It is described as

$$\mathcal{L}_{Mel}(G) = \mathbb{E}_{(x,s)}\Bigg[||\phi(x) - \phi(G(s))||_1\Bigg]$$

4. EXPERIMENTATION RESULTS

We used crowd-sourced MOS exams to assess the quality. Raters listened to randomly chosen audio clips and scored their naturalness on a 5-point scale from 1 to 5. Raters were permitted to evaluate each audio sample once, and we normalized all the audio clips to eliminate the impact of amplitude variances on the score. All the quality assessments in this study were carried out in this manner. The evaluation findings are presented in Table 1. VITS outperforms other TTS systems and achieves a comparable MOS as ground truth. The VITS (DDP), which uses the same deterministic duration predictor architecture as Glow-TTS rather than the stochastic duration predictor, ranks second among TTS systems in the MOS evaluation. These findings suggest that 1) the stochastic duration predictor generates more realistic phoneme duration than the deterministic duration predictor, and 2) our end-to-end training method is an effective way to produce better samples than other TTS models while maintaining the same duration predictor architecture.

Table 4.1: MOS and CMOS comparisons between NaturalSpeech and previous TTS systems.

System	MOS	CMOS
FastSpeech 2 (ren2021fastspeech) + HiFiGAN (kong2020hifi)	4.32±0.15	−0.33
Glow-TTS (kim2020glow) + HiFiGAN (kong2020hifi)	4.34±0.13	−0.26
Grad-TTS (popov2021grad) + HiFiGAN (kong2020hifi)	4.37±0.13	−0.24
VITS (kim2021conditional)	4.43±0.13	−0.20
NaturalSpeech	4.56±0.13	0

We performed ablation research to illustrate the effectiveness of our approaches, which included normalized flow in the previous encoder and linear-scale spectrogram posterior input. All models in the ablation research were trained up to 300,000 steps. The findings are presented in Table 2. Removing the normalizing flow in the previous encoder leads to a 1.52 MOS drop from the baseline, indicating that the prior distribution's flexibility has a considerable impact on synthesis quality. Replacing the linear-scale spectrogram for posterior input with the mel-spectrogram causes a quality decrease (-0.19 MOS), demonstrating that high-resolution information is useful for VITS in increasing synthesis quality.

Table 4.2: The CMOS of each component to its upper bound. Negative CMOS means this component setting is worse than its upper bound.

Component	Setting	Upper Bound	CMOS
Vocoder	GTMel→Vocoder	Human Recordings	−0.04
Mel Decoder	GT Pitch/Duration→Mel Decoder	GT Mel	−0.15
Variance Adaptor	Predicted Pitch/Duration	GT Pitch/Duration	−0.14
Phoneme	Phoneme Encoder	Phoneme Encoder + Pre training	−0.12

5. RELATED WORK

The Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech by Jaehyeon Kim, Jungil Kong, Juhee Son where they proposed a two staged model. The first stage is to produce intermediate speech representations such as melspectrograms or linguistic features from the preprocessed text, and the second stage is to generate raw waveforms conditioned on the intermediate representations. They proposed the method as Variational Inference with adversarial learning for end-to-end Text-to-Speech(VITS) which produces a more natural sounding audio.

Another model we look into is YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone by Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Golge, Moacir Antonelli Ponti. They built upon VITS and added several modifications for multi level multi speaker training.

XTTS: Taking TTS to the Next Level by Eren Gölge and Kelly Davis develops on XTTS model which although has very expressive outputs, better voice cloning, and delivers all the enhanced Coqui Studio features. It is however very resource expensive and cannot provide instant voice cloning.

The work Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search by J. Kim, S. Kim, J. Kong, and S. Yoon. Which is a flow-based generative model for parallel TTS that does not require any external aligner. By combining the properties of flows and dynamic programming, the proposed model searches for the most probable monotonic alignment between text and the latent representation of speech on its own.

6. CONCLUSION AND FUTURE WORK

Speech cloning, also known as voice cloning or voice synthesis, has advanced significantly in recent years, with several uses. The technology entails copying or imitating a person's voice, enabling the development of synthetic voices that closely resemble the real speaker. Speech cloning provides promise and possibility in a number of domains. Our work introduces VITS, a parallel TTS system capable of learning and generating from start to finish. To convey different speech rhythms, we proposed the stochastic duration predictor. The method generates natural-sounding voice waveforms directly from text, bypassing specified intermediary representations. Our experimental findings demonstrate that our technique outperforms two-stage TTS systems and reaches near-human quality. We hope that the suggested technique would Two-stage

TTS systems are commonly employed in speech synthesis to increase performance and simplify training procedures. Our solution unifies two different generating processes in TTS systems, although there is still a text preparation difficulty. Self-supervised language learning might potentially eliminate the need for text preparation. We will make our source code and pre-trained models available for further research.

Our method improves the expressive capacity of generative modeling by integrating variational inference, normalizing flows, and adversarial training. We provide a stochastic duration predictor that generates speech with various rhythms from input text. To reflect the natural one-to-many link in which a text input can be spoken in a variety of pitches and rhythms, we use uncertainty modeling over latent variables and a stochastic duration predictor. The future scope of speech cloning holds promises and potential in several areas. The Variational Inference with Adversarial Learning for End-to-End Text-to-Speech (VITS) model has an intricate and sophisticated design. To improve its performance in the context of speech cloning, use speech-specific data augmentation approaches. Pitch, tempo, and background noise changes can all help to make a model more resilient. Include speaker embeddings in the model. This can assist capture distinctive speech features and increase voice cloning accuracy. Train the model on a dataset that includes numerous speakers. This allows the model to generalize more accurately across diverse speakers during speech cloning. Improving existing models of speech cloning with text-to-speech (TTS) approaches entails fine-tuning numerous system components to improve the quality, naturalness, and adaptability of synthetic voices. Improving the training dataset by include a wide variety of speakers, accents, and linguistic variances. A more representative dataset can result in greater generalization.

REFERENCES

1. Kim, J., Kong, J., & Son, J. (2021). Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. Proceedings of the 38th International Conference on Machine Learning, in Proceedings of Machine Learning Research, 139, 5530-5540
2. Qin, Z., Zhao, W., Yu, X., & Sun, X. (2023). OpenVoice: Versatile Instant Voice Cloning. arXiv preprint arXiv:23
   12.01479.
3. Kim, J., Kim, S., Kong, J., & Yoon, S. (2020). Glow-tts: A generative flow for text-to-speech via monotonic alignment search. Advances in Neural Information Processing Systems, 33, 8067-8077.
4. Kong, J., Kim, J., & Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems, 33, 17022-17033.
5. Binkowski, M., Donahue, J., Dieleman, S., Clark, A., Elsen, E., Casagrande, N., Cobo, L. C., and Simonyan, K. High Fidelity Speech Synthesis with Adversarial Networks. In International Conference on Learning Representations, 2019.
6. M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. arXiv preprint arXiv:2306.15687, 2023.
7. J. Li, W. Tu, and L. Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE, 2023.
8. A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux. Speech resynthesis from discrete disentangled self-supervised representations. arXiv preprint arXiv:2104.00355, 2021.
9. D. Rezende and S. Mohamed. Variational inference with normalizing flows. In International conference on machine learning, pages 1530–1538. PMLR, 2015.
10. B. van Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper. A comparison of discrete and soft speech units for improved voice conversion. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6562–6566. IEEE, 2022.
11. 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. In Interspeech, pages 1526–1530, 2019.
12. Chen, J., Lu, C., Chenli, B., Zhu, J., and Tian, T. Vflow: More expressive generative flows with variational data augmentation. In International Conference on Machine Learning, pp. 1660–1669. PMLR, 2020.
13. R. J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. 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 4700–4709, 2018