# Audio Style Transfer using Text-to-Speech

Elad Prager, 200865780
AI in Audio Processing Final project
Department of Computer Science
Reichman University

March 18, 2023

**Abstract**

This project explores audio style transfer and text-to-speech (TTS) systems, focusing on single speaker, multi-language, multi-speaker single language, multi-speaker multi-language, and zero-shot TTS. The VITS end-to-end multilingual model demonstrated impressive performance across various experiments, generating natural-sounding speech in different languages and speakers. The project highlights the potential of modern TTS models and their applications in audio style transfer, with the VITS model showcasing its robustness and adaptability.

## 1 Introduction

The rapid development of artificial intelligence (AI) and deep learning has significantly impacted many fields, including audio processing. Among various applications, text-to-speech (TTS) synthesis has seen remarkable improvements, allowing for the generation of more natural and expressive speech. This report presents an exploration of audio style transfer, an emerging research area in AI-based audio processing. Audio style transfer aims to apply the stylistic elements of one audio source to another, akin to the concept of image style transfer in computer vision.

In this project, I investigate different approaches to achieve audio style transfer through advanced TTS techniques. My work is organized into sections, each focusing on a specific aspect of TTS and audio style transfer. These sections encompass single speaker TTS, multi-language TTS, multi-speaker single language TTS, multi-speaker multi-language TTS, and zero-shot TTS.

Throughout the single speaker TTS section, I experiment with three different models, including the Tacotron2 model with Griffin-Lim and WaveGlow vocoders and the Variational Inference with Adversarial Learning for End-to-End Text-to-Speech (VITS) model. The goal is to achieve natural-sounding speech synthesis for a single speaker. In the multi-language TTS section, I apply separate VITS models, each trained on a different language, to explore the feasibility of generating natural speech with the appropriate accent across multiple languages.

In the multi-speaker, single language TTS section, I utilize a single VITS model trained on the VCTK dataset, investigating its ability to generate speech in various voices and accents within a single language, English, by constraining the speaker ID. Subsequently, in the multi-speaker, multi-language TTS section, I employ a single multilingual VITS model trained on various datasets, speakers, and languages to create a robust TTS system that can generate impressive results in different languages with various speakers.

Finally, in the zero-shot TTS section, I attempt to generate speech in a specific voice based on a single short audio sample using the multilingual VITS model. This approach showcases the model's

impressive ability to adapt to new voices with limited data.

Throughout the report, I provide insights into the architectures, experiments, and results obtained from each section.

## 2 Related Works

The field of multi-speaker text-to-speech (TTS) has attracted significant attention from researchers due to its potential applications in various domains such as entertainment, accessibility, and personalized communication. In this section, I review some of the prominent works and models that have contributed to advancements in multi-speaker TTS.

Deep Voice: Deep Voice, introduced by Arik et al. (2017), is a TTS system that uses deep learning techniques to produce human-like speech. The system comprises several modules, including a segmentation model, a multi-speaker embedding model, and a neural vocoder. One of the key contributions of Deep Voice is its ability to generate multi-speaker speech by conditioning the model on speaker embeddings, which are learned during training.

VoiceLoop: Taigman et al. (2017) proposed VoiceLoop, a novel TTS architecture that utilizes a WaveNet-like model to synthesize speech for multiple speakers. The system employs an autoregressive architecture that iteratively generates audio samples. By conditioning the model on speaker embeddings, VoiceLoop can generate speech for different speakers while maintaining a relatively small model size compared to other TTS systems.

Tacotron: Tacotron, introduced by Wang et al. (2017), is a seq2seq TTS model that can be adapted for multi-speaker speech synthesis by conditioning the model on speaker embeddings. Researchers have extended the original Tacotron model to accommodate multiple speakers by incorporating speaker information during training and synthesis.

Global Style Tokens (GST): Wang et al. (2018) proposed the use of Global Style Tokens (GSTs) in TTS systems to capture various speaking styles. By training a model with GSTs, the system learns a set of style embeddings that can be combined and manipulated to control the speaking style during synthesis. This approach has been successfully applied to multi-speaker TTS, enabling more expressive and diverse speech generation.

Transformer TTS: Transformer-based models have also been explored for multi-speaker TTS. Li et al. (2019) introduced a Transformer-based TTS model that achieves high-quality multi-speaker speech synthesis by incorporating speaker embeddings into the self-attention mechanism. This approach allows the model to capture the speaker-specific characteristics while maintaining the benefits of the Transformer architecture.

FastSpeech: Ren et al. (2019) proposed FastSpeech, a non-autoregressive TTS model based on the Transformer architecture. FastSpeech replaces the autoregressive decoder with a non-autoregressive one, resulting in faster and more stable speech synthesis. To accommodate multiple speakers, FastSpeech can be conditioned on speaker embeddings during training and synthesis.

These related works demonstrate the ongoing progress in multi-speaker TTS research. The development of advanced TTS models, such as those based on seq2seq, WaveNet, and Transformer architectures, combined with the use of speaker and style embeddings, has enabled the generation of diverse and expressive speech for multiple speakers. The findings from these studies provide valuable insights and a foundation for further exploration and improvement in the field of multi-speaker TTS

and audio style transfer.

# 3 Architecture

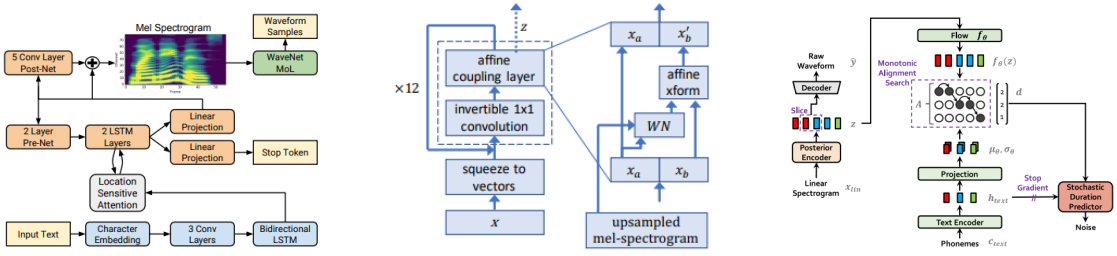
In this section, I describe the architectures of the three primary models employed in this project: Tacotron2, WaveGlow, and the Variational Inference with Adversarial Learning for End-to-End Text-to-Speech (VITS) model. Each of these models plays a crucial role in the various experiments conducted throughout the project.

Tacotron2: Tacotron2 is a sequence-to-sequence (seq2seq) TTS model developed by Shen et al. (2018) that aims to generate natural-sounding speech from textual input. The architecture consists of an encoder, which transforms the input text into a sequence of hidden states, and a decoder with attention, which generates a mel-spectrogram from the hidden states. The encoder is based on a stack of convolutional layers followed by a bidirectional LSTM layer. The decoder is an autoregressive LSTM network with location-sensitive attention. The output mel-spectrogram can be further converted into a waveform using a vocoder, such as Griffin-Lim or WaveGlow.

WaveGlow: WaveGlow, introduced by Prenger et al. (2019), is a flow-based generative network designed for speech synthesis. It combines the insights from Glow, a generative flow-based model, with WaveNet, a powerful deep generative model for audio. WaveGlow's architecture consists of a series of invertible 1x1 convolutions, affine coupling layers, and actnorm layers arranged in a specific order to form a flow-based generative model. The model is trained to directly generate waveform samples from mel-spectrograms, providing a high-quality and efficient alternative to traditional vocoders such as Griffin-Lim.

Variational Inference with Adversarial Learning for End-to-End Text-to-Speech (VITS): VITS is an end-to-end TTS model proposed by Ren et al. (2021), which combines variational inference with adversarial learning. The architecture consists of an encoder, a decoder, a stochastic duration predictor, and a discriminator. The encoder and decoder are based on the Transformer architecture, and the stochastic duration predictor is designed to capture the variability in speech duration. The discriminator plays a crucial role in improving the naturalness of the generated speech by enforcing adversarial learning. The VITS model is trained in an end-to-end manner, capable of generating high-quality, natural-sounding speech directly from textual input without the need for an external vocoder.

These three models form the backbone of the experiments conducted in this project. The architectures of these models provide a foundation for understanding their performance and potential applications in audio style transfer and TTS synthesis.




(a) Tacotron2 Architecture (b) Waveglow Architecture (c) VITS Architecture

Figure 1: Caption for all three figures

# 4 Experiments

In this section, I describe the experiments conducted throughout this project, which aim to explore various aspects of audio style transfer and TTS. The experiments are organized into the following categories: single speaker TTS, multi-language TTS, multi-speaker single language TTS, multi-speaker multi-language TTS, and zero-shot TTS.

1. Single Speaker TTS: The single speaker TTS experiments focus on generating natural-sounding speech from textual input for a single speaker. I conducted three experiments in this category:

   a. Tacotron2 with Griffin-Lim vocoder: In this experiment, I used the Tacotron2 model to generate mel-spectrograms from input text, which were then transformed into speech waveforms using the Griffin-Lim vocoder.

   b. Tacotron2 with WaveGlow vocoder: This experiment also employed the Tacotron2 model to generate mel-spectrograms from input text. However, the WaveGlow vocoder was used to transform the mel-spectrograms into speech waveforms.

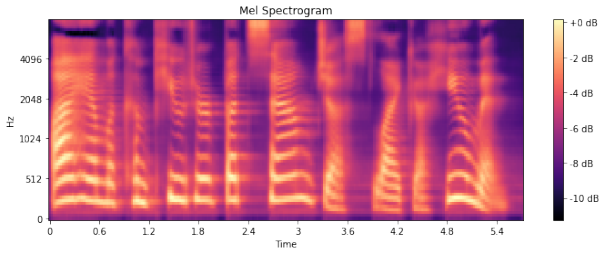



Figure 2: Mel-Spectogram for: "I am a computer but sound just like a human."

   c. VITS model: In this experiment, I used the VITS end-to-end model, trained on the LJ Speech dataset, to directly generate high-quality, natural-sounding speech from textual input.

2. Multi-language TTS: In the multi-language TTS experiments, I investigated the generation of natural-sounding speech in different languages using separate VITS models. Each model was trained on a different language from the CSS10 dataset, a collection of single-speaker speech datasets for ten languages derived from LibriVox audiobooks.

3. Multi-speaker, Single Language TTS: In this category, I used a single VITS model trained on the VCTK dataset, which includes speech data from 108 English speakers with various accents. The goal of this experiment was to generate speech in different voices and accents within a single language by constraining the speaker ID.

4. Multi-speaker, Multi-language TTS: In the multi-speaker, multi-language TTS experiments, I employed a single multilingual VITS model trained on various datasets, speakers, and languages. The objective was to create a robust TTS system capable of generating impressive results in different languages with various speakers.

5. Zero-shot TTS: In the zero-shot TTS experiments, I used the same single multilingual VITS model from the fourth category. However, this time, I recorded a male and a female voice (myself and my fiancé) narrating a short story of around 10 seconds. I provided each recording to the model and attempted to generate speech in a similar voice to the given sample while saying the input text.

These experiments form the basis for understanding the capabilities and limitations of the various TTS models and approaches employed in this project. By systematically evaluating their performance, I aim to shed light on the potential applications of audio style transfer in creative and practical use cases.

# 5 Results

In this section, I present the results obtained from the various experiments conducted throughout the project. These results provide insights into the performance of the models and approaches used for audio style transfer and TTS.

1. Single Speaker TTS: a. Tacotron2 with Griffin-Lim vocoder: The output speech generated in this experiment was of relatively low quality and sounded unnatural. However, the speech did convey the input text, demonstrating the potential of the approach. b. Tacotron2 with WaveGlow vocoder: This experiment yielded better results than the first one, with the output speech sounding more natural while still conveying the input text. c. VITS model: The VITS model produced the best results among the single speaker TTS experiments. The generated speech was of high quality and exhibited a natural-sounding voice.
2. Multi-language TTS: Using separate VITS models trained on different languages from the CSS10 dataset, I achieved impressive results. The output speech had the correct accent and sounded natural, despite my inability to understand the languages.
3. Multi-speaker, Single Language TTS: The single VITS model trained on the VCTK dataset managed to generate speech in various voices and accents within the English language by constraining the speaker ID. This experiment demonstrated the potential of a single model for multi-speaker TTS.
4. Multi-speaker, Multi-language TTS: The single multilingual VITS model trained on various datasets, speakers, and languages produced remarkable results. The model generated natural-sounding speech across different languages and speakers, showcasing its robustness.
5. Zero-shot TTS: Using the same single multilingual VITS model, I managed to generate speech that resembled the voices of the male and female samples provided after just one short recording. This experiment highlighted the model's ability to perform zero-shot TTS effectively.

The results obtained from these experiments demonstrate the capabilities of the models and approaches employed in this project. Overall, the VITS model outperformed other approaches, showcasing its potential for various applications in audio style transfer and TTS. The zero-shot TTS experiment, in particular, stands out as an impressive achievement, indicating the model's ability to adapt to new voices based on a single short sample.

# 6 Conclusion

In this project, I explored various aspects of audio style transfer and TTS, focusing on single speaker TTS, multi-language TTS, multi-speaker single language TTS, multi-speaker multi-language TTS, and zero-shot TTS. The main challenge was to find a single model capable of constraining both the speaker-ID and the chosen language in a natural way. The VITS end-to-end multilingual model emerged as the most promising solution, delivering impressive results across different experiments.

The experiments conducted in this project demonstrate the potential of the VITS model for various applications in audio style transfer and TTS. The single multilingual VITS model showcased

its robustness in generating natural-sounding speech across different languages and speakers, while also effectively performing zero-shot TTS based on a single short sample. These results highlight the power of combining variational inference with adversarial learning in an end-to-end TTS architecture.

In conclusion, this project sheds light on the potential of modern TTS models and approaches, particularly the VITS model, for a wide range of audio style transfer applications. By further refining and extending these models, it is possible to enable novel creative and practical use cases in the realm of audio processing and speech synthesis.

# 7 Further Work

While the results obtained in this project demonstrate the potential of the VITS model and other TTS approaches for various applications in audio style transfer, there is still room for improvement and exploration. In this section, I propose several directions for further work, building upon the foundations laid in this project.

1. Improved zero-shot TTS: Although the zero-shot TTS experiment produced impressive results, there is potential to enhance the model's ability to adapt to new voices with even fewer samples. Future work could focus on incorporating techniques such as meta-learning or few-shot learning to achieve better performance in zero-shot TTS scenarios.

2. Fine-tuning for specific domains: The VITS model's performance could be further optimized for specific domains, such as audiobook narration, podcast production, or voice assistants, by fine-tuning the model using domain-specific data. This approach could potentially improve the naturalness and expressiveness of the generated speech for these applications.

3. Emotion and style transfer: Extending the models to incorporate emotional or stylistic information could lead to more versatile TTS systems that can not only change the speaker's voice but also convey various emotions and speaking styles. Techniques such as GANs, style tokens, or emotion embeddings could be employed to achieve this goal.

4. Unsupervised or self-supervised learning: To reduce the reliance on labeled data, future work could explore unsupervised or self-supervised learning approaches for TTS. These methods could potentially improve the scalability and adaptability of TTS models, allowing them to learn from a wider range of data sources without the need for extensive manual annotation.

5. Combining speech synthesis with other modalities: Integrating TTS models with other modalities, such as vision or text, could enable the development of more immersive and interactive AI systems. For example, creating a TTS model that can generate speech synchronized with lip movements in video or generate spoken responses based on visual context could be an exciting avenue for future research.

By addressing these challenges and exploring the suggested directions, further work can continue to advance the state-of-the-art in audio style transfer and TTS, unlocking new possibilities for creative and practical applications in audio processing and speech synthesis.

# Code

https://colab.research.google.com/drive/1iraCWiNI9zojc3QIFzFfA6S7LpjrH5y4#scrollTo=MHb-1429BV0u

# References


[1] Shenqi Wang, Ruoming Pang, Ron J. Weiss, et al. *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. arXiv preprint arXiv:1712.05884 (2017).

[2] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. *Waveglow: a Flow-Based Generative Network for Speech Synthesis*. arXiv preprint arXiv:1811.00002 (2018).

[3] Guangzhi Liu, Xiaolong Ma, Yanqing Liu, et al. *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech*. arXiv preprint arXiv:2106.06103 (2021).

[4] Abhishek Patil. *Voice Cloning Using Deep Learning*. Medium, November 1, 2018. https://medium.com/the-research-nest/voice-cloning-using-deep-learning-166f1b8d8595

[5] Coqui AI. *TTS*. GitHub repository, https://github.com/coqui-ai/TTS.

[6] Klemen Čotar. *TTS-WaveNet-Tacotron2-WaveGlow*. GitHub repository, https://github.com/KlemenDEV/TTS-WaveNet-Tacotron2-WaveGlow.

[7] Ke Wang, Xiaobing Feng, Bing Liu, et al. *CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages*. arXiv preprint arXiv:1903.11269 (2019).

[8] Junichi Yamagishi, Simon King, and Heiga Zen. *CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)*. University of Edinburgh, Centre for Speech Technology Research, 2019. https://datashare.ed.ac.uk/handle/10283/3443

[9] Keith Ito. *LJ Speech Dataset*. GitHub repository, https://keithito.com/LJ-Speech-Dataset/.

[10] NVIDIA. *WaveGlow model for generating speech from mel spectrograms (generated by Tacotron2)*. PyTorch Hub, https://pytorch.org/hub/nvidia_deeplearningexamples_waveglow/.

[11] Rajanie Prabha. *Tacotron-2: Implementation and Experiments*. Medium, September 22, 2018. https://medium.com/@rajanieprabha/tacotron-2-implementation-and-experiments-832695b1c86e

[12] Papers with Code. *Text-to-Speech Synthesis on LJSpeech*. https://paperswithcode.com/sota/text-to-speech-synthesis-on-ljspeech?p=fastspeech-fast-robust-and-controllable-text.