End-to-End Speech Synthesis with Explicit Prosody Control and Computational Pipeline Analysis

Arjun Arora (M24CSA003)
Indian Institute of Technology, Jodhpur

Prateek (M24CSA022)
Indian Institute of Technology, Jodhpur

Abstract

Speech synthesis plays a pivotal role in human–computer speech interaction, being an indispensable module in fields such as audiobooks, voice guidance in public service facilities, and intelligent systems like robotics and autonomous driving. Modern advancements in speech synthesis have achieved remarkable quality, with synthesized speech approaching the naturalness of human speech. One of the most prominent models in neural network-based sequence-to-sequence (Seq2Seq) architectures is Tacotron2, which utilizes a modified WaveNet vocoder built upon the Tacotron framework. On the same dataset, the Mean Opinion Score (MOS) value of Tacotron2 reached 4.53, which is very close to human speech at 4.58. Despite these advancements, Tacotron2 and similar autoregressive models rely on spectral sequences for speech synthesis, which introduces cyclic dependencies that result in slow training and prediction times.

In recent years, the VITS (Variational Inference Text-to-Speech) model has emerged as a non-autoregressive alternative, offering improved training efficiency and faster prediction speeds. However, a notable limitation of the VITS model is the inability to control prosody during synthesis. In this work, we aim to address this limitation by introducing an emotional prosody control mechanism. Specifically, we propose the incorporation of a reference audio signal alongside the model's inputs, enabling dynamic prosody modulation for more expressive and natural speech generation.

1. Introduction

Recent advancements in neural text-to-speech (TTS) systems have significantly improved the intelligibility and naturalness of synthesized speech. However, achieving fine-grained and controllable prosody remains a persistent challenge. While current models are capable of generating speech that closely mimics human voice quality, they often lack the ability to modulate prosodic features such as pitch, rhythm, and intonation in a flexible and contextually aware manner.

Prosody is critical in conveying nuances of speech, emphasizing semantic elements, and enhancing expressiveness. Traditional TTS architectures, including sequence-to-sequence models like Tacotron2 [2], typically fail to provide users with mechanisms for explicit prosodic control. These models attempt to learn average representations of prosody across the dataset, leading to bland or uniform outputs. Moreover, due to the complexity and interdependence of prosodic components, it is difficult to isolate and manipulate specific elements without degrading speech quality.

Several methods have been explored to tackle this issue. Techniques based on reference audio, such as Global Style Tokens (GST) [3], enable the extraction of global prosodic characteristics, while Variational Autoencoders (VAE) [5] have shown promise in learning disentangled representations for more structured control. However, most of these approaches are implemented in two-stage TTS pipelines, where the text-to-spectrogram and vocoder components are trained separately. This separation often leads to mismatches between stages and causes a loss of critical information [1].

Furthermore, models that only support single-level granularity of control fail to capture the hierarchical nature of prosody. Expressive speech often requires both global and local variation, such as utterance-level emotion and phoneme-level emphasis. Limiting control to either level may result in monotonous or unnatural outputs [4].

To address these limitations, our work focuses on extending the VITS model, an end-to-end non-autoregressive TTS framework, by introducing a reference audio-based prosody control mechanism. This allows the model to capture and reproduce dynamic prosodic variations, achieving more expressive and context-sensitive speech synthesis.

2. ProsodyVITS

The Variational Inference Text-to-Speech (VITS) model represents a major advancement in end-to-end speech synthesis, combining the strengths of variational autoencoders (VAE), normalizing flows, and generative adversarial networks (GAN) into a unified architecture. Unlike traditional two-stage models that separately handle acoustic feature generation and waveform synthesis, VITS jointly learns all components, allowing for better alignment, expressiveness, and naturalness.

2.1. VITS model

At the core of VITS is a text encoder, comprising modules like Relative Multi-Head Attention and Feed-Forward Networks (FFN), which transforms phoneme or character inputs into contextual embeddings. The text encoder itself contains 6,326,784 parameters, with attention mechanisms (149,952 parameters) and FFN layers (885,696 parameters) forming a major part of the architecture. These embeddings condition the prior distribution in the VAE framework.

The posterior encoder (4,878,720 parameters) estimates the latent distribution from the target mel-spectrogram, whereas the generator (14,337,024 parameters), based on a WaveNet-like structure (1,148,544 parameters), synthesizes the final waveform from the latent representations. Residual Coupling Blocks (4,742,784 parameters), composed of Residual Coupling Layers (1,185,696 parameters), serve as normalizing flows that transform simple latent variables into complex ones, aiding in high-quality synthesis.

The duration predictor (345,857 parameters) plays a crucial role in aligning text and speech via the Monotonic Alignment Search (MAS) algorithm, which computes latent alignment and duration with variational augmentation. Discriminators are key to the adversarial training regime: Discriminator$_\text{S}$ (5,641,362 parameters), Discriminator$_\text{P}$ (8,221,154 parameters), and the Multi-Period Discriminator (46,747,132 parameters) collectively distinguish real and generated waveforms while guiding the generator to produce perceptually realistic outputs.

VITS is trained to maximize the Evidence Lower Bound (ELBO) on the marginal likelihood of waveform data, treating the speech generation process as a conditioned variational inference task:

$$\log p(x|c) \geq \mathbb{E}_{q(z|x)}[\log p(x|z)] - \mathrm{KL}(q(z|x)\|p(z|c))$$

Here, $p(z|c)$ is the prior conditioned on text, $q(z|x)$ is the approximate posterior from mel-spectrogram, and $p(x|z)$ is the decoder likelihood. The loss is composed of a reconstruction loss:

$$\mathcal{L}_{\text{recon}} = \|x_{\text{mel}} - \hat{x}_{\text{mel}}\|_1$$

Table 1. Parameter Counts for Core Components of VITS

Module	Parameter Count
Relative Multi-Head Attention	149,952
Feed-Forward Network (FFN)	885,696
Encoder	6,218,496
Text Encoder	6,326,784
WaveNet Decoder	1,148,544
Posterior Encoder	4,878,720
Residual Coupling Layer	1,185,696
Residual Coupling Block	4,742,784
Generator	14,337,024
Duration Predictor	345,857
Discriminator$_\text{S}$	5,641,362
Discriminator$_\text{P}$	8,221,154
Multi-Period Discriminator	46,747,132

and a KL divergence term:

$$\mathcal{L}_{\text{kl}} = \mathrm{KL}(q(z|x_{\text{lin}})\|p(z|c, A))$$

To model phoneme durations explicitly, VITS includes a variational loss over durations:

$$\mathcal{L}_{\text{dur}} = -\log p(d|c) + \mathrm{KL}(q(u, v|d, c)\|p(u, v|c))$$

where u and v are auxiliary latent variables for dequantization and augmentation, respectively.

For realism, adversarial losses are introduced via discriminators:

$$\mathcal{L}_{\text{adv}}(G) = \mathbb{E}_z[(D(G(z)) - 1)^2]$$

$$\mathcal{L}_{\text{fm}}(G) = \sum_{l=1}^{T} \frac{1}{N_l}\|D_l(y) - D_l(G(z))\|_1$$

Combining all, the final loss for VITS training becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{kl}} + \mathcal{L}_{\text{dur}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{fm}}$$

The VITS model however is unble to generate prosody controlled outputs.

2.2. Prosody Conditioning Mechanism

To enable expressive and contextually appropriate speech synthesis, **ProsodyVITS** extends the baseline VITS architecture by incorporating a prosody conditioning mechanism. This mechanism introduces a dedicated prosody encoder network that extracts and conditions prosodic information from a reference speech signal at both global (utterance-level) and local (frame-level) resolutions.

The raw audio reference is first passed through a pretrained wav2vec 2.0 model, fine-tuned on a prosody-rich dataset. This model acts as a robust feature extractor, converting the waveform into a sequence of high-dimensional, content-independent representations. These representations are then fed into two parallel branches designed to capture prosody at different granularities:

- **Global Prosody Encoder:** This branch is responsible for capturing the overall prosodic style of the entire utterance, such as average pitch, intonation contour, and speaking rate. The process involves:
  1. A linear projection layer followed by a ReLU activation, which reduces the dimensionality of the wav2vec features while introducing nonlinearity.
  2. A single-layer Long Short-Term Memory (LSTM) network that models temporal dependencies across the entire sequence.
  3. A masked average pooling operation (MaskAvg) that aggregates the LSTM output into a fixed-length vector by averaging only over valid time steps. The result is a 192-dimensional global prosody embedding vector.
- **Local Prosody Encoder:** This branch retains the temporal resolution of the prosodic features and captures fine-grained variations in prosody that occur across phonemes or syllables. It consists of:
  1. A linear transformation layer that projects the wav2vec features to a lower-dimensional space.
  2. A temporal average pooling operation that smooths short-term variations and expands the receptive field to capture local context.
  3. Another linear layer that refines the smoothed features. The resulting output preserves the time dimension and yields a sequence of local prosody embeddings.

These two encoders extract complementary aspects of prosody: the global encoder captures the overall prosodic style of the reference utterance, while the local encoder preserves time-dependent variations in pitch, energy, and rhythm. These two sets of embeddings are later combined in the prosody fusion module to generate a unified prosodic representation that conditions the text encoder during both training and inference.

2.3. Prosody Feature Fusion

To integrate global and local prosodic information into a unified representation, we employ an attention-based fusion mechanism. The fused prosody embedding is computed as:

$$\mathbf{h}_{\text{fused}} = \lambda \mathbf{h}_{\text{global}} + (1 - \lambda)\mathbf{h}_{\text{local}}, \tag{1}$$

where $\mathbf{h}_{\text{global}}$ and $\mathbf{h}_{\text{local}}$ represent the global and local prosody feature vectors, respectively. The scalar fusion weight $\lambda \in [0, 1]$ is computed dynamically through a prosody attention network that processes both inputs.

Specifically, the fusion network consists of two submodules:

- **Global Attention Module:** Applies convolutional layers, followed by batch normalization, ReLU activation, and global average pooling to emphasize utterance-level prosodic saliency.
- **Local Attention Module:** Similar to the global module but without global pooling, allowing it to capture temporal nuances in local prosody.

The outputs of both attention modules are combined and passed through a sigmoid activation to yield the final fusion weight λ.

The fused prosody embedding $\mathbf{h}_{\text{fused}}$ is then concatenated with phoneme embeddings at the input of the text encoder. This conditioning ensures that prosodic information influences the alignment, acoustic modeling, and ultimately the waveform generation stages in ProsodyVITS.

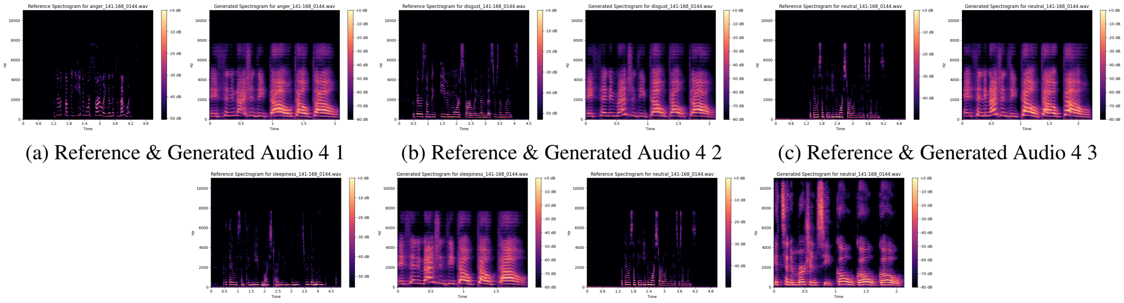
3. Results

We took the pretrained VITS model, which was originally trained on the LJSpeech dataset, and added the prosody network. We then trained this extended model on the SLR dataset, which consists of speech samples with different emotional tones. The model was trained for approximately 15,000 iterations, where the classes were non-IID. Four reference audio samples were used during training. The generated and reference audio samples' mel spectrograms were plotted for comparison.

Figures 1 show the spectrograms of the reference audios (1, 2, 3, 4) and their corresponding generated audio samples. Additionally, the spectrogram generated by the original VITS model is shown in Figure 1e. These spectrograms provide a visual comparison between the reference and generated audios, showing how the prosody network influences the speech synthesis.

Moreover, to evaluate the training progress, we plotted the spectrograms generated at different iterations, ranging from 1000 to 15000. These spectrograms, shown in Figure 2, illustrate the evolution of the model as it trained.

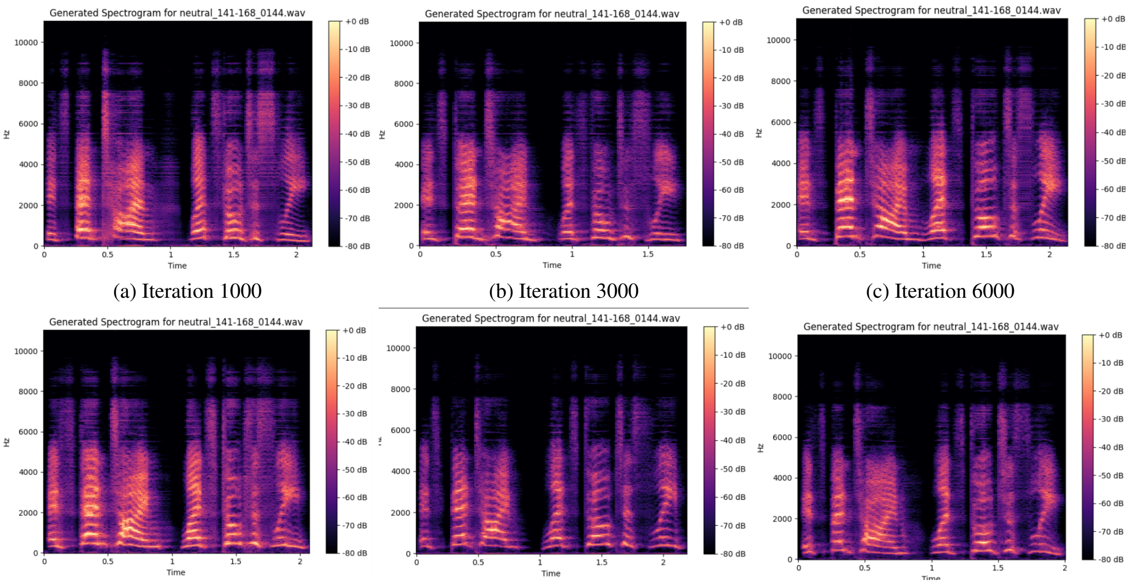
As observed over the course of training, the spectrograms at different iterations show a clear progression in terms of quality. At early iterations, the generated audio exhibits significant noise and lacks clarity. However, as the training progresses, this noise diminishes, and the spectrograms become more refined. This reduction in noise and improvement in synthesis quality is reflected in the decreasing loss during training. The loss function typically penalizes discrepancies between the generated and reference



(a) Reference & Generated Audio 4 1 (b) Reference & Generated Audio 4 2 (c) Reference & Generated Audio 4 3

(d) Reference & Generated Audio 4 (e) Audio from Original VITS Model

Figure 1. Comparison of reference audio and generated audio spectrograms. Figures 1-4 show the reference and corresponding generated audios, while Figure 5 shows the generated audio from the original VITS model.



(a) Iteration 1000 (b) Iteration 3000 (c) Iteration 6000

(d) Iteration 9000 (e) Iteration 12000 (f) Iteration 15000

Figure 2. Spectrograms at different training iterations. The plots show how the generated audio improves with training, from iteration 1000 to 15000.

audio, guiding the model towards reducing errors in frequency, pitch, and prosody. As the model learns more about the distribution of prosodic features and speech patterns, the resulting spectrograms progressively approach the reference samples with higher accuracy. We observe that the ProsodyVITS model generates audio samples with noticeable variation given a reference audio sample. This variation in the output speech is crucial for synthesizing diverse prosody, such as different emotional tones or speech rates. When the dataset is independent and identically distributed (IID), the generated samples are more distinguishable from each other. This is because the model has learned to generalize the prosody features effectively, leading to clear differences between samples. When using a non-IID dataset, the prosodic features extracted from reference audios are more similar to one another, making the generated samples less distinguishable from one another. However, by using a non-IID dataset, we are essentially incorporating more diverse data, which improves the model's ability to generalize across varied prosody patterns. While the variations in the generated audio samples become subtler in the non-IID case, they are still meaningful and contribute to stronger, more robust prosody synthesis. This is because the model learns to integrate a broader range of prosodic cues, enhancing its capacity to generate high-quality, diverse speech despite the overlapping features in the reference samples.

4. Conclusion

In this work, we explored how the VITS model functions and the role each component plays in the synthesis of speech. We also examined the impact of adding prosody embeddings to VITS, which allows for greater flexibility in controlling the tone of the generated audio. While the inclusion of prosody embeddings enhances the ability to vary the tone, it also introduces additional noise in the generated audio, especially in the early stages of training. As demonstrated in the results section, the performance of the model improves with more training iterations. Over time, the noise in the generated spectrograms decreases, and the quality of the synthesized speech improves. This suggests that continued training can help refine the prosody and overall audio quality, making it more natural and accurate. Adding prosody embeddings to VITS provides a promising approach for controlling the emotional tone and variation in synthesized speech. While noise remains a challenge, further training can help mitigate this issue, leading to more realistic and diverse speech synthesis.

References

[1] Atsushi Ando, Yusuke Igarashi, and Junichi Yamagishi. Analysis of information loss in text-to-speech systems based on two-stage models. *Interspeech*, 2019. 1

[2] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *ICASSP*, 2018. 1

[3] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, and Samy Bengio. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *ICML*, 2018. 1

[4] Yusuke Yasuda, Xin Wang, and Junichi Yamagishi. Investigation of enhanced tacotron text-to-speech synthesis systems with self-attention for pitch accent language. *ICASSP*, 2019. 1

[5] Yi Zhao, Xin Wang, and Junichi Yamagishi. Towards semi-supervised and unsupervised methods for neural speech synthesis with variational autoencoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1394–1405, 2020. 1