VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY

HO CHI MINH UNIVERSITY OF TECHNOLOGY

FACULTY OF COMPUTER SCIENCE AND ENGINEERING

Modern Speech Processing (CO5257)

Assignment Report

Enhancing Vocoder Designs
with Emotion Embeddings and Emotional Cues

Mentor: Nguyễn Đức Dũng
Student: Trần Hà Tuấn Kiệt – 2011493

Ho Chi Minh City, 1/2025

Table of Contents

Abstract

This article explores the integration of emotion embeddings and emotional cues into vocoder architectures to enhance expressiveness in speech synthesis. Modern vocoders, while capable of producing high-quality audio, often fail to effectively capture the nuanced prosody and variability required for emotionally rich speech. By conditioning vocoders with emotion-specific embeddings and leveraging emotional cues such as pitch, rhythm, and intensity, this study aims to bridge the gap between technical accuracy and human-like expressiveness. Moreover, the study proposes a universal emotion-aware conditioning framework applicable to neural, statistical, and hybrid vocoder designs. Through a combination of theoretical exploration and empirical validation, the methodology demonstrates improvements in emotional fidelity and naturalness across synthesized speech outputs.

1 Introduction

In recent years, synthetic speech technology has seen a marked expansion within human-computer interaction. This trend reflects the inherent complexity of human speech, which encodes not only linguistic content but also para-linguistic and non-linguistic cues. Linguistic information can be derived directly from written text or contextual inferences, whereas para-linguistic features are deliberately introduced by the speaker to enhance or modify the message. By contrast, non-linguistic aspects—such as emotional state—are often beyond the speaker's voluntary control, presenting additional challenges in achieving truly expressive synthesized speech [1]. These complexities have motivated extensive research in text-to-speech (TTS) synthesis aiming to generate stylish, expressive, or emotional speech [2, 3, 4, 5, 6, 7, 8, 9]. Ultimately, bridging these complexities is crucial for creating speech synthesis that feels natural and truly conveys the richness of human speech.

To tackle these challenges and more accurately model the multifaceted nature of human speech, researchers have turned to advanced neural network architectures. *Neural vocoders* represent a cutting-edge technology in speech synthesis, specializing in transforming intermediate acoustic representations, such as spectrograms, into lifelike audio waveforms [10, 11, 12]. Renowned for their capability to produce high-fidelity, natural-sounding speech, these models have become a pivotal component of contemporary speech synthesis frameworks [13, 14]. Moreover, emotion embeddings have emerged as a key strategy to capture and encode subtle emotional cues—including intonation, prosody, and timbre—in a learned representation. When integrated into modern TTS pipelines, such embeddings enrich the expressive range of synthesized speech, making it more engaging and human-like [15]. However, accurately modeling the nuances of emotion remains an ongoing challenge, driven by the inherent complexity of emotional expression and the diverse contexts in which it appears[16]. This paper aims to integrate emotion embeddings directly into the waveform generation process through a novel neural vocoder architecture. By leveraging these representations, the system aspires to produce speech that is both high-fidelity and emotionally expressive, unlocking applications in empathetic virtual agents, personalized audiobooks, and emotion-driven interfaces that enhance user engagement and interactivity.

The current state of research in emotional speech synthesis and text-to-speech (TTS) systems highlights the use of specialized datasets and intermediate modules, such as prosody-driven conditioning, to achieve expressive outputs. Frameworks like Tacotron [17], paired with neural vocoders such as WaveNet [10] and HiFi-GAN [18], have become foundational in this domain. Many systems incorporate emotion or style tokens [5], which allow modulation of the synthesized speech to reflect various emotional states. While these approaches have successfully synthesized clear emotional expressions, challenges persist in generating nuanced and diverse emotions, especially in handling complex states or transitions. Additionally, ensuring robustness remains an issue, as small variations in embeddings can lead to significant changes in the quality or nature of the output [19]. Emotional Voice Conversion (EVC) and our project, "Enhancing Vocoder with Emotion Embeddings & Cues," share some conceptual similarities but differ in scope and focus. EVC is a broader approach that seeks to transform a neutral or source utterance into a target emotional style without altering speaker identity, often involving a comprehensive pipeline that includes acoustic feature extraction, prosody modeling, and vocoder synthesis. These systems frequently use parallel data or disentanglement techniques to separate speaker traits from emotional features [20]. By contrast, our work focuses specifically on the vocoder stage, where we integrate learned emotional embeddings or cues directly into the waveform generation process. This enables finer control over intonation, timbre, and microprosody, allowing for more precise and natural emotional rendering. Unlike EVC, which modifies the entire pipeline to achieve emotional transformation, our method emphasizes refining the vocoder to achieve stable, contextually appropriate emotional outputs. This targeted approach enhances the vocoder's ability to express subtle emotional nuances, providing greater flexibility and control over the synthesized speech's emotional quality [21].

2 Background

Denoising diffusion probabilistic models (DDPMs) are a class of generative models that learn a data distribution by iteratively denoising a noisy signal. These models operate through two main processes: the *forward process* and the *reverse process*.

2.1 Forward Process

The forward process $q(\cdot)$ progressively adds Gaussian noise to the data $\mathbf{x}_0$ over T timesteps, resulting in a latent variable $\mathbf{x}_T$ that approximates an isotropic Gaussian distribution. This process is parameterized as a Markov chain:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where $\beta_t \in \{\beta_1, \dots, \beta_T\}$ represents the predefined noise schedule.

By repeatedly applying this transition, we can express $\mathbf{x}_t$ in terms of $\mathbf{x}_0$:

$$\mathbf{x}_t = \sqrt{\gamma_t}\mathbf{x}_0 + \sqrt{1-\gamma_t}\varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where $\gamma_t = \prod_{i=1}^{t}(1-\beta_i)$.

2.2 Reverse Process

The reverse process aims to reconstruct the original data $\mathbf{x}_0$ from the noisy sample $\mathbf{x}_T$ by inverting the forward process. This is achieved by approximating the conditional distribution $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ using a neural network:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}), \tag{3}$$

where σ_t^2 is derived from the noise schedule, and the mean μ_θ is defined as:

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{1-\beta_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\gamma_t}}\varepsilon_\theta(\mathbf{x}_t, t)\right). \tag{4}$$

2.3 Training Objective

The neural network ε_θ is trained to predict the added noise ε, minimizing the following loss function:

$$\mathscr{L}_{\text{simple}} = \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_t, \varepsilon} \left[\|\varepsilon - \varepsilon_\theta(\mathbf{x}_t, t)\|^2 \right]. \tag{5}$$

Extensions like PriorGrad modify this framework by introducing a prior distribution to incorporate domain-specific knowledge, such as frame-level energy in mel-spectrograms, leading to an adjusted loss function:

$$\mathscr{L}_{\text{diff}} = \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_t, \varepsilon} \left[\|\varepsilon - \varepsilon_\theta(\mathbf{x}_t, t, \mathbf{X})\|_{\Sigma}^2 \right], \tag{6}$$

where Σ is a diagonal matrix representing normalized frame-level energy of the mel-spectrogram $\mathbf{X}$.

3 Experiments

3.1 Training Setup

We conducted experiments using the LJSpeech dataset [22], a publicly available 24-hour speech dataset comprising 13,100 audio clips from a single female speaker. The dataset was preprocessed to generate 80-band mel-spectrograms computed with a 1024-point FFT, a hop length of 256, and a frequency range of 80Hz to 7,600Hz.

For training, we used 13,000 clips, with 5 clips reserved for validation and the remaining 95 clips for testing. A publicly available implementation [23] was employed, utilizing a model with 2.62M parameters optimized using the Adam optimizer [24] with a learning rate of 2×10^{-4}. Training was performed over 1M iterations, requiring approximately 7 days on a single NVIDIA A40 GPU.

The diffusion process was configured with $T = 50$ steps and a linear beta schedule ranging from 1×10^{-4} to 5×10^{-2}, consistent with the default settings of the DiffWave$_{\text{BASE}}$ model [23]. For inference, a faster noise schedule with $T_{\text{infer}} = 6$ steps was utilized, as specified in the same implementation. To ensure a fair comparison, all models, including WaveGrad [25] and PriorGrad [26], were trained and evaluated under identical conditions.

4 Evaluations

5 Conclusion

References

[1] G. An, L. Wu, J. Xie, M. Huang, and X. Xie, "Emotional text-to-speech synthesis: A review of methods and techniques," *IEEE Access*, vol. 9, pp. 10001–10019, 2021.

[2] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for hmm-based expressive speech synthesis," *IEICE Transactions on Information and Systems*, vol. 90, pp. 1406–1413, 2007.

[3] G. E. Henter, J. Lorenzo-Trueba, X. Wang, and J. Yamagishi, "Principles for learning controllable tts from annotated and latent variation," in *Proc. Interspeech*, pp. 3956–3960, 2017.

[4] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis," *Speech Communication*, 2018.

[5] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *arXiv preprint arXiv:1803.09017*, 2018.

[6] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *arXiv preprint arXiv:1803.09047*, 2018.

[7] X. Wu, L. Sun, S. Kang, S. Liu, Z. Wu, X. Liu, and H. Meng, "Feature based adaptation for speaking style synthesis." Manuscript, 2018.

[8] X. Wu, Y. Cao, M. Wang, S. Liu, S. Kang, Z. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Rapid style adaptation using residual error embedding for expressive speech synthesis," in *Proc. Interspeech*, pp. 3072–3076, 2018.

[9] S. Liu, Y. Cao, and H. Meng, "Multi-target emotional voice conversion with neural vocoders," *arXiv preprint arXiv:2004.03782*, 2020.

[10] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *Proceedings of the 9th ISCA Speech Synthesis Workshop*, p. 125, ISCA, 2016.

[11] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," in *Proceedings of Interspeech 2021*, pp. 2207–2211, ISCA, 2021.

[12] H. Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis," *arXiv preprint arXiv:2201.08337*, 2022.

[13] E. A. AlBadawy, A. Gibiansky, Q. He, J. Wu, M.-C. Chang, and S. Lyu, "Vocbench: A neural vocoder benchmark for speech synthesis," *arXiv preprint arXiv:2112.03099*, 2021.

[14] P. Pérez Zarazaga, Z. Malisz, G. E. Henter, and L. Juvela, "Speaker-independent neural formant synthesis," *arXiv preprint arXiv:2203.17032*, 2022.

[15] Y. Lei, S. Yang, X. Wang, and L. Xie, "Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 853–864, 2022.

[16] S. Raptis, S. Karabetsos, A. Chalamandaris, and P. Tsiakoulis, "Towards expressive speech synthesis: Analysis and modeling of expressive speech," in *Proceedings of the 2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom)*, pp. 461–465, 2014.

[17] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[18] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, pp. 17087–17096, 2020.

[19] X. Zhang, M. Xu, J. Cui, *et al.*, "Deep learning based emotional speech synthesis: A review," *arXiv preprint arXiv:2106.08395*, 2021.

[20] K. Zhou, J. Yamagishi, and C. Veaux, “Transforming spectrum and prosody for emotional voice conversion with non-parallel training data,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7754–7758, IEEE, 2020.

[21] D. Kim, D. Ahn, H. Kim, and N. S. Kim, “Conditional variational autoencoder with attention for emotional speech synthesis,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 533–540, IEEE, 2021.

[22] K. Ito, “The lj speech dataset.” `https://keithito.com/LJ-Speech-Dataset/`, 2017. Available at `https://keithito.com/LJ-Speech-Dataset/`.

[23] Z. Kong, W.-H. Kim, and J. Forth, “Diffwave: A versatile diffusion model for audio synthesis,” in *International Conference on Learning Representations (ICLR)*, 2021.

[24] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[25] N. Chen, Y. Zhang, and H. Zen, “Wavegrad: Estimating gradients for waveform generation,” *arXiv preprint arXiv:2009.00713*, 2020.

[26] M. Chen, X. Zhang, *et al.*, “Priorgrad: Incorporating prior knowledge in diffusion probabilistic models,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.