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, 12/2024

Table of Contents

1 Introduction **3**

2 Methodology **4**

3 Experiments **5**

4 Evaluations **6**

5 Conclusion **7**

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

Neural vocoders represent a cutting-edge technology in speech synthesis, specializing in transforming intermediate acoustic representations, such as spectrograms, into lifelike audio waveforms [1, 2, 3]. Renowned for their capability to produce high-fidelity, natural-sounding speech, these models have become a pivotal component of contemporary speech synthesis frameworks [4, 5]. Despite their significant advances, a notable gap remains in achieving nuanced emotional expressiveness, which is crucial for applications like virtual assistants, dubbing, and human-computer interaction.

One of the primary shortcomings of existing vocoders lies in their inability to accurately model the subtle variations in pitch, rhythm, intensity, and spectral quality that characterize emotional speech [3]. Emotions such as happiness, sadness, and anger often involve intricate acoustic modulations, yet current models struggle to replicate these nuances. This limitation stems partly from the design priorities of most vocoders, which focus on naturalness and intelligibility rather than expressiveness.

Another challenge is the lack of generalization across speakers. Emotional expressiveness varies significantly between individuals due to differences in vocal characteristics and delivery styles. While vocoders can achieve high performance when tailored to a specific speaker, their performance tends to degrade when applied to other voices. This limitation arises because emotional cues are often entangled with speaker-specific features, making it difficult for models to adapt without extensive retraining [5].

The scarcity of high-quality, emotion-annotated datasets further exacerbates the issue. Emotional speech synthesis requires large datasets that encompass a wide range of emotions, speakers, and languages. However, collecting such datasets is resource-intensive, and the subjective nature of emotion labeling introduces inconsistencies. As a result, many vocoders are trained on datasets dominated by neutral or mildly expressive speech, leaving them ill-equipped to handle more dramatic or subtle emotional variations.

2 Methodology

3 Experiments

4 Evaluations

5 Conclusion

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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