Expressive Voice Conversion: From FreeVC to Bi-Referenced Transformation

情感與語音轉換： 從FreeVC到雙參考轉換

組別：B456、指導教授：李祈均 、組員姓名 ： 趙育樵 & 趙翊程、專題領域 ： 資工領域

Abstract

This research advances expressive voice conversion by merging FreeVC's [1] text-free adaptability with Emotional VITS's [2] emotional depth [3]. The proposed hybrid model integrates speaker and emotion embeddings through cross-embedding normalization and employs ContentVec [4] in place of WavLM to achieve balanced and natural speech synthesis. Evaluation results highlight significant improvements in emotional expressiveness and speaker identity preservation, particularly for explicit emotions like "Sad" and "Happy." However, challenges also persist in synthesizing nuanced emotions and optimizing the model for real-time and cross-language applications. Overall, this work establishes a foundation for innovative AI-driven speech technologies that enhances human-computer interaction.

Motivation

AI-driven speech synthesis has transformed human-computer interaction, yet integrating emotional expressiveness into text-free voice conversion remains a challenge. FreeVC excels in flexibility but lacks emotional depth, while Emotional VITS enriches emotion but relies on annotated text. This highlights the need for a hybrid model combining their strengths.

Background

FreeVC:

A text-free voice conversion model using pre-trained speech features to separate and recombine content and speaker characteristics. It excels in flexibility and low-data scenarios but struggles to preserve emotional nuances.

Emotional VITS:

Emotional VITS extends VITS by adding emotion embeddings from annotated text for expressive speech synthesis. Though powerful for generating emotion-rich speech, it depends on extensive annotations and high computational resources, limiting scalability.

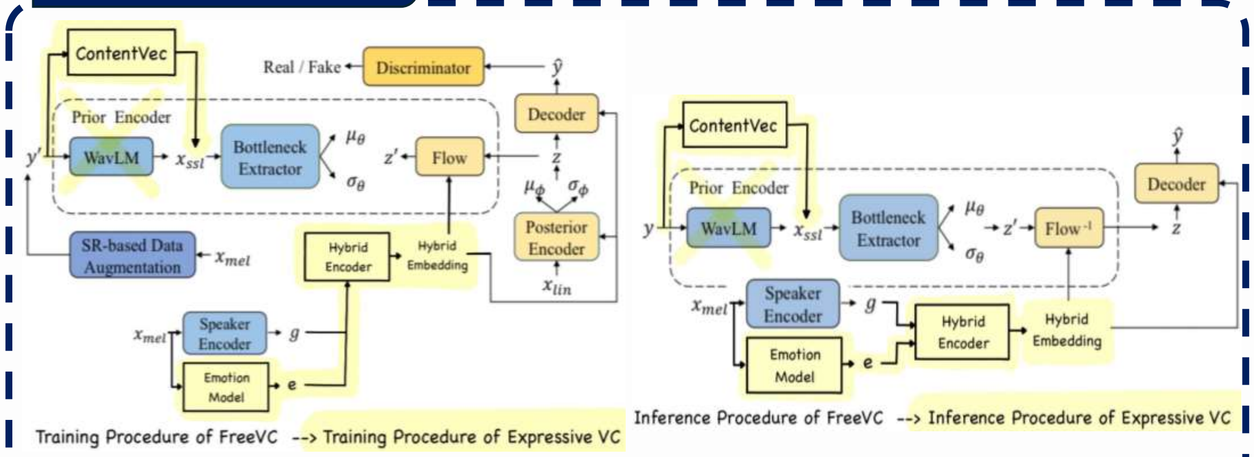
WavLM:

A pre-trained model that captures linguistic and acoustic features from audio. In FreeVC, it enables effective content and speaker separation for text-free conversion.

ContentVec:

ContentVec reduces redundancy in content embeddings, focusing on speaker and emotional features to improve voice conversion accuracy and expressiveness.

Design



The block diagrams illustrate the training and inference procedures of the expressive VC model, highlighting our key innovations: the hybrid encoder (lower half of both graphs) and the integration of ContentVec (upper half), with other components mostly inherited from existing models (still some details such as weight controller layers aren't shown).

Evaluation

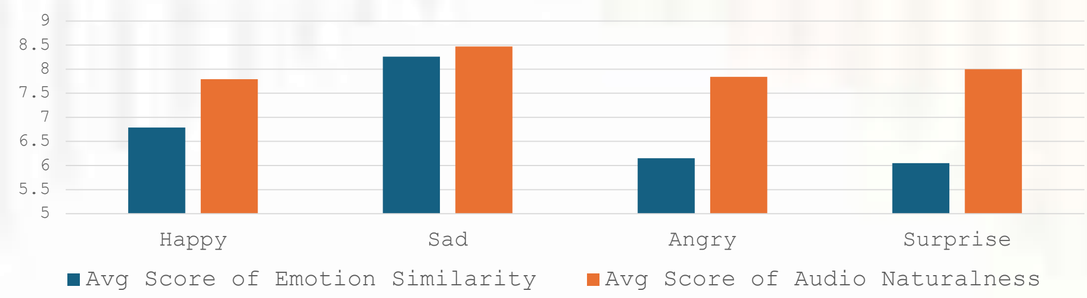
First, we prepared three experiment sets to examine the performance of conversion with respect to emotion, voice, and both. Next, we conducted inferences using randomly selected input audio clips under required conditions.

Finally, since the model involves multiple embeddings and emotion perception varies among individuals, we resorted to real-world user feedback through Google Form as evaluation.

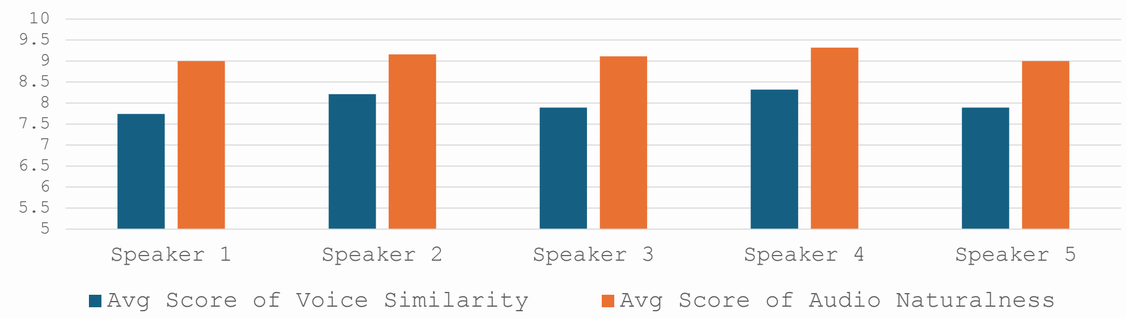
* See Google form at: https://reurl.cc/EgzeAg

Result

Part1 (Fix content and speaker ref, change emotion ref):



Part2 (Fix content and emotion ref, change speaker ref):



Compare to FreeVC: Better in all five cases

Part3 (Change both emotion and speaker ref):

Emotion: 7.95/ **Voice:** 7.95/ **Naturalness:** 9.05

*** Note that the scores take value from 1 (worst) to 10 (best), and that y-axis on charts above start from 5 for clearer view**

Conclusion

Expressive voice conversion model bridges the gap between FreeVC's flexibility and Emotional VITS's emotional richness, achieving text-free, one-shot voice conversion with enhanced emotional expressiveness. Our experimental results demonstrate significant improvements in emotional conversion and speaker similarity compared to existing models. Despite some room for improvement in emotional tones like "Surprise" due to their nuanced and context-dependent nature, the overall achievements still underscorde the potential of hybrid architectures that leverage the strengths of multiple frameworks.

Reference

[1] Jingyi li, Weiping tu, Li xiao. (2022, Oct 27) FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion. https://github.com/OlaWod/FreeVC

[2] emotional-vits. https://github.com/innnky/emotional-vits

[3] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, Björn W. Schuller. (2023, Sep 11).Dawn of the transformer era in speech emotion recognition: closing the valence gap

[4] Qian, K., Zhang, Y., Gao, H., Ni, J., Lai, C., Cox, D., Hasegawa-Johnson, M. & Chang, S. (2022). ContentVec: An Improved Self-Supervised Speech Representation by Disentangling Speakers. Proceedings of the 39th International Conference on Machine Learning, in Proceedings of Machine Learning Research 162:18003-18017 Available from https://proceedings.mlr.press/v162/qian22b.html.