MÔ HÌNH SINH GIỌNG HÁT VỚI CÁC THUẬT TOÁN CHỈNH SỬA CAO ĐỘ

Đặng Phước Sang - 21521377
Ngô Cao Lộc - 21521088

GVHD: PGS.TS Lê Đình Duy

Tóm tắt

- Lớp: CS519.O11
- Link Github của nhóm: https://github.com/PhuocSang16/CS519.O11
- Link YouTube video:

Đặng Phước Sang - 21521377

Ngô Cao Lộc - 21521088

Giới thiệu

- Trong ngành công nghiệp âm nhạc hiện đại, công nghệ âm thanh đang trở nên ngày càng quan trọng. Bài toán Chỉnh sửa cao độ (Pitch Correction) là một phần của việc làm đẹp giọng hát.
- Bài toán chỉnh sửa cao độ được định nghĩa như sau:

  • Đầu vào: Một đoạn âm thanh chứa giọng hát một người.

  • Đầu ra: Đoạn âm thanh chứa giọng hát đã được sửa lỗi cao độ

- Thách thức lớn là làm sao để sửa lỗi cao độ mà vẫn giữ nguyên tính tự nhiên của giọng hát.

Mục tiêu

- Đề xuất thuật toán Shape-Aware Dynamic Time Warping (SADTW) cho bài toán Pitch Correction, so sánh với các thuật toán DTW truyền thống.
- Đề xuất framework NSVB giải quyết bài toán Singing Voice Beautifying.
- Chứng minh hiệu quả của NSVB trên các các chỉ số đánh giá mục tiêu và chỉ số đánh giá chủ quan.

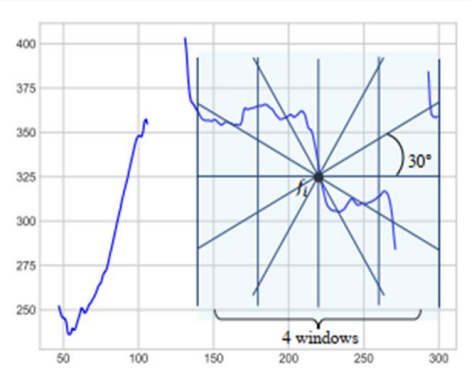
Nội dung và Phương pháp

Tìm hiểu các kiến thức biểu diễn âm thanh, đặc trưng giọng hát, các thuật toán ước lượng tần số cơ bản, cùng các thuật toán căn chỉnh cao độ.

Tìm hiểu kiến trúc mô hình CVAE và các mô hình liên quan như ASR, Conformer, WaveNet, ... được sử dụng như các encoder giải mã các đặc trưng giọng hát được đưa vào mô hình CVAE.

Nội dung và Phương pháp

Xây dựng thuật toán Shape-Aware Dynamic Time Warping (SADTW), cải tiến từ DTW truyền thống bằng cách sử dụng thông tin về hình dạng của các biến đổi cao độ, tức thay thế khoảng cách Euclid của DTW bằng khoảng cách mô tả ngữ cảnh hình dạng (shape context descriptor distance).

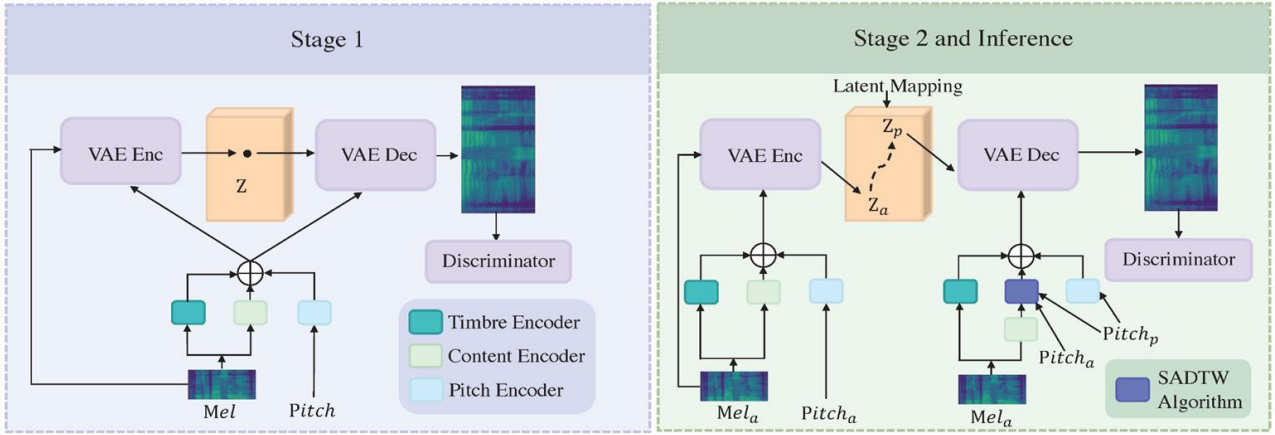


$$h_i(k) = |\{f_j \neq f_i, f_j \in bin(k)\}|,$$

$$C(a,p) = \frac{1}{2}\sum_{k=1}^{m*n} \frac{[h_a(k) - h_p(k)]^2}{h_a(k) + h_p(k)},$$

Nội dung và Phương pháp

Xây dựng và huấn luyện framework NSVB với cốt lõi là mô hình CVAE, kết hợp thuật toán SADTW và tối ưu hóa thuật toán ánh xạ tiềm ẩn (latent-mapping algorithm).



Nội dung và Phương pháp

Thử nghiệm và đánh giá thuật toán Pitch Correction trên thang đo PAA và F0-RMSE, đánh giá chất lượng âm thanh sinh ra dựa trên chỉ số mục tiêu và các chỉ số chủ quan.

Viết báo cáo kết quả nghiên cứu, xây dựng trang web minh họa, cho phép người dùng đưa vào file âm thanh giọng hát của bản thân vào mô hình, sinh ra file giọng hát đã được làm đẹp.

Kết quả dự kiến

- Một báo cáo chi tiết về các kiến thức tìm hiểu, kết quả thử nghiệm, đánh giá và so sánh với các phương pháp khác.
- Một chương trình minh họa cho phép sinh ra file âm thanh giọng hát chất lượng đã được căn chỉnh cao độ.

Tài liệu tham khảo

[1] Jinglin Liu, Chengxi Li, Yi Ren, Zhiying Zhu, Zhou Zhao: Learning the Beauty in Songs: Neural Singing Voice Beautifier. CoRR abs/2202.13277 (2022)
[2] Matthias Mauch, Simon Dixon: PYIN: A fundamental frequency estimator using probabilistic threshold distributions. ICASSP 2014: 659-663
[3] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang: Conformer: Convolution-augmented Transformer for Speech Recognition. CoRR abs/2005.08100 (2020)
[4] Artidoro Pagnoni, Kevin Liu, Shangyan Li: Conditional Variational Autoencoder for Neural Machine Translation. CoRR abs/1812.04405 (2018)
[5] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, Koray Kavukcuoglu: WaveNet: A Generative Model for Raw Audio. CoRR abs/1609.03499 (2016)