THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo: https://www.youtube.com/watch?v=MU_ak3xlAsY
- Link slides: https://github.com/PhuocSang16/CS519.O11/blob/main/Slide.pdf

• Họ và Tên: Đặng Phước Sang
• MSSV: 21521377
 
	• Lớp: CS519.O11
	• Tự đánh giá (điểm tổng kết môn): 8.5/10
	• Số buổi vắng: 1
	• Số câu hỏi QT cá nhân: 14
	• Số câu hỏi QT của cả nhóm: 2
	• Link Github: https://github.com/PhuocSang16/CS519.O11
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Lên ý tưởng cho đề cương
	  ○ Viết đề cương
	  ○ Làm Powerpoint phần tóm tắt, giới thiệu, nội dung và phương pháp.
	  ○ Làm video YouTube

• Họ và Tên: Ngô Cao Lộc • MSSV: 21521088 	• Lớp: CS519.O11 • Tự đánh giá (điểm tổng kết môn): 8/10 • Số buổi vắng: 2 • Số câu hỏi QT cá nhân:13 • Số câu hỏi QT của cả nhóm:2 • Link Github: https://github.com/PhuocSang16/CS519.O11 • Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:

	○ Viết đề cương ○ Làm Powerpoint phần mục tiêu và kết quả mong đợi ○ Làm poster ○ Làm video Youtube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

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

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

GENERATIVE SINGING VOICE MODELING WITH PITCH CORRECTION ALGORITHMS

TÓM TẮT

Đề tài liên quan đến lĩnh vực Xử lý âm thanh và tiếng nói, tập trung vào bài toán mới là Làm đẹp giọng hát (Singing Voice Beautifying). Với giọng hát của một ca sĩ nghiệp dư, mục tiêu của SVB là cải thiện cao độ của giọng hát, trong khi giữ nguyên nội dung và âm sắc giọng hát. Trong đề tài này, chúng tôi giới thiệu một framework Neural Singing Voice Beautifier (NSVB), sử dụng mô hình sinh Conditional Variational Autoencoder (CVAE) làm cốt lõi và học các biểu diễn tiềm ẩn giọng hát. Chúng tôi đề xuất một thuật toán Chỉnh sửa cao độ (Pitch Correction) mới: Shape-Aware Dynamic Time Warping (SADTW) cùng một thuật toán ánh xạ tiềm ẩn (latent-mapping algorithm) để chuyển giọng điệu và màu sắc giọng hát nghiệp dư sang giọng điệu và màu sắc giọng hát chuyên nghiệp. Cuối cùng, chúng tôi thử nghiệm trên bộ dữ liệu PopBuTFy gồm cả bài hát tiếng Trung và tiếng Anh, chứng minh hiệu quả của NSVB trên các chỉ số đánh giá mục tiêu và chủ quan (objective and subjective metrics).

GIỚI THIỆU

Trong ngành công nghiệp âm nhạc hiện đại, công nghệ âm thanh đang ngày càng trở nên quan trọng trong việc sản xuất và tạo ra nội dung âm thanh chất lượng. Bài toán Chỉnh sửa cao độ (Pitch Correction) là một trong những bài toán phổ biến và là một phần trong bài toán Làm đẹp giọng hát (SVB).

Bài toán chỉnh sửa cao độ (Pitch Correction) 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

Một trong những thách thức lớn là làm thế nào để tạo ra giọng hát đúng cao độ mà vẫn giữ

nguyên tính tự nhiên của giọng hát. Điều này đặt ra một thách thức đối với cả các nghệ sĩ và kỹ sư âm thanh, đặc biệt là khi muốn tạo ra sản phẩm âm nhạc hoặc giọng hát có chất lượng cao. Có nhiều công cụ chỉnh sửa cao độ giọng hát như Autotune, Melodyne, nhưng cần phải sử dụng chúng một cách cẩn thận để tránh làm mất đi tính tự nhiên và cảm xúc của giọng hát. Trong khi đó, các thuật toán Pitch Correction truyền thống như DTW thường không đạt được sự ổn định. Với sự phát triển của các mô hình học sâu, đặc biệt là các mô hình sinh, chúng tôi đề xuất framework NSVB, sử dụng mô hình CVAE làm cốt lõi để học cách biểu diễn các đặc trưng phức tạp trong giọng hát để sinh ra giọng hát chất lượng, kết hợp với thuật toán Shape-Aware DTW cải thiện sự ổn định so với các thuật toán trước đó.

MỤC TIÊU

1. Đề xuất thuật toán Shape-Aware Dynamic Time Warping (SADTW) cho bài toán Pitch Correction, so sánh với các thuật toán DTW truyền thống.
2. Xây dựng framework NSVB giải quyết bài toán Singing Voice Beautifying.
3. 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

Nội dung

- Tìm hiểu các kiến thức về kỹ thuật xử lý âm thanh, các thuật toán và mô hình có liên quan
- Xây dựng thuật toán Shape-Aware Dynamic Time Warping (SADTW), cải tiến từ DTW truyền thống.
- Xây dựng và huấn luyện framework NSVB, kết hợp với thuật toán SADTW. Thử nghiệm trên bộ dữ liệu PopBuTFy và chứng minh hiệu quả của NSVB.

Phương pháp

- Tìm hiểu các cách biểu diễn âm thanh (spectrogram, mel-spetrogram, MFCC…), cách biểu diễn các đặc trưng giọng hát (pitch, content, timbre, …) cùng với các kỹ thuật tăng cường dữ liệu có thể áp dụng cho mô hình (time shift, pitch shift, add noise, …)
- Tìm hiểu mô hình CVAE và các mô hình liên quan như ASR, Conformer, WaveNet, …

- Tìm hiểu các thuật toán ước lượng tần số cơ bản cùng các thuật toán căn chỉnh cao độ theo đường cong mẫu, đề xuất thuật toán SADTW bằng cách sử dụng thông tin về hình dạng của các biến đổi cao độ.
- Xây dựng framework NSVB dựa trên mô hình cốt lõi là CVAE, Sử dụng SADTW để khớp các biến đổi cao độ của giọng hát nghiệp dư với các biến đổi cao độ của giọng hát chuyên nghiệp.
- Tối ưu hóa thuật toán ánh xạ tiềm ẩn (latent-mapping algorithm) để chuyển giọng điệu và âm sắc giọng hát nghiệp dư sang giọng điệu và âm sắc giọng hát chuyên nghiệp.
- Thử nghiệm trên bộ dữ liệu PopBuTFy, triển khai các kỹ thuật xử lý âm thanh, tăng cường dữ liệu và tiến hành huấn luyện cho NSVB.
- 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 Mean Cepstral Distortion (MCD) và các chỉ số chủ quan như Mean Opinion Score (MOS), Comparison Mean Opinion Score (CMOS).
- Viết báo cáo kết quả nghiên cứu và 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 MONG ĐỢI

1. 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.
2. Một chương trình minh họa cho phép sinh ra file âm thanh giọng hát chất lượng đã được căn chỉnh cao độ.

Ý NGHĨA NGHIÊN CỨU

SVB là một lĩnh vực nghiên cứu mới mẻ trong lĩnh vực xử lý âm thanh và học máy, đóng góp vào việc phát triển các phương pháp và công nghệ mới để cải thiện chất lượng giọng hát. Nghiên cứu này có thể giúp sản xuất các sản phẩm âm nhạc chất lượng cao hơn bằng cách cải thiện giọng hát của ca sĩ, từ đó tạo ra các bản ghi âm và các buổi biểu diễn âm nhạc hấp dẫn. Ngoài ra còn có thể ứng dụng trong giáo dụng âm nhạc và các ứng dụng giải trí khác.

KẾ HOẠCH THỜI GIAN

Dự án nghiên cứu dự kiến sẽ hoàn thành trong khoảng sáu tháng, với các cột mốc quan trọng như sau:

Tháng 1: Tìm hiểu kiến thức về xử lý âm thanh, các thuật toán và các mô hình liên quan

Tháng 2-3: Xây dựng thuật toán SADTW, khởi chạy và đánh giá kết quả của thuật toán, so sánh với các thuật toán DTW khác. Xây dựng mô hình NSVB

Tháng 4-6: Huấn luyện mô hình NSVB, đánh giá kết quả trên các chỉ số mục tiêu và chỉ số chủ quan. Xây dựng trang web minh họa tương tác với người dùng.

KẾT LUẬN

Dự án nghiên cứu về Mô hình sinh giọng hát với các thuật toán chỉnh sửa cao độ cho phép tổng hợp và chia sẻ kiến thức âm nhạc, xử lý âm thanh, các thuật toán chỉnh sửa cao độ và các mô hình sinh. Nghiên cứu mang lại tiềm năng ứng dụng không chỉ trong lĩnh vực sản xuất âm nhạc, mà còn trong các lĩnh vực giáo dục âm nhạc và các ứng dụng giải trí khác.

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)