THÔNG TIN CHUNG

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/6FoznYZu4_s*
- Link slides: https://github.com/Limdim1604/CS519.Q11.KHTN/blob/main/VIETNAMESE-TO-ENGLISH_MUSIC_CONVERSION_SYSTEM.pdf
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

• Họ và Tên: Nguyễn Thiên Bảo
• MSSV: 23520127
 
	• Lớp: CS519.Q11.KHTN
	• Tự đánh giá (điểm tổng kết môn): 9.5/10
	• Số buổi vắng: 0
	• Số câu hỏi QT cá nhân: 9
	• Số câu hỏi QT của cả nhóm: 9
	• Link Github: https://github.com/Limdim1604/CS519.Q11.KHTN
	• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:
	  ○ Tự tìm ý tưởng về đề tài đồ án
	  ○ Làm hết Slide, Poster, Docs
	  ○ Làm hết video YouTube

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

HỆ THỐNG CHUYỂN ĐỔI BÀI NHẠC TIẾNG VIỆT SANG PHIÊN BẢN TIẾNG ANH

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

VIETNAMESE-TO-ENGLISH MUSIC CONVERSION SYSTEM

TÓM TẮT *(Tối đa 400 từ)*

Trong bối cảnh hội nhập và nhu cầu học ngoại ngữ ngày càng cao, việc thưởng thức âm nhạc bằng tiếng Anh trên nền giai điệu và nội dung quen thuộc của các bài hát tiếng Việt là một phương pháp giải trí kết hợp học tập hiệu quả. Đề tài này đề xuất xây dựng một hệ thống tự động chuyển đổi bài hát tiếng Việt sang phiên bản tiếng Anh (Singing Voice Translation), giải quyết bài toán khó về việc dịch thuật đảm bảo tính "hát được" (singability). Hệ thống được thiết kế theo mô hình phân tầng (cascaded), bao gồm ba module chính: (1) Hệ thống tiền xử lý để tách nguồn âm thanh, nhận dạng lời hát và trích xuất giai điệu; (2) Mô hình dịch lời bài hát có ràng buộc âm nhạc để đảm bảo số lượng âm tiết, nhịp điệu và vần điệu; (3) Tổng hợp giọng hát tiếng Anh (Singing Voice Synthesis) dựa trên giai điệu gốc. Kết quả đầu ra là một file âm thanh bài hát tiếng Anh hoàn chỉnh, giữ nguyên cảm xúc và giai điệu của bản gốc, hỗ trợ người dùng có trải nghiệm âm nhạc mới mẻ và hữu ích.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Âm nhạc là cầu nối văn hóa và là công cụ đắc lực trong việc học ngôn ngữ. Hiện nay, nhiều người học tiếng Anh có nhu cầu nghe lại các bài hát tiếng Việt yêu thích dưới phiên bản tiếng Anh để vừa giải trí, vừa học từ vựng và ngữ điệu, cũng như nhà sáng tạo nội dung muốn cover bài hát nước ngoài nhưng gặp rào cản ngôn ngữ... Tuy nhiên, việc chuyển ngữ bài hát là một thách thức lớn, đặc biệt là với một ngôn ngữ đơn âm và có thanh điệu (Vietnamese - Tonal) sang một ngôn ngữ đa âm, không có

thanh điệu nhưng có trọng âm (English - Non-tonal/Stress-timed) đòi hỏi không chỉ khả năng dịch thuật chính xác về ngữ nghĩa mà còn phải tuân thủ nghiêm ngặt các ràng buộc về âm nhạc như giai điệu, nhịp phách và sự luyến láy.

Các công cụ dịch thuật hiện nay (như Google Translate) chỉ tập trung vào văn bản thuần túy mà bỏ qua cấu trúc âm nhạc, dẫn đến kết quả dịch không thể hát được trên nền nhạc gốc. Trong khi đó, các nghiên cứu về Chuyển đổi giọng hát (Singing Voice Conversion) thường chỉ tập trung vào thay đổi chất giọng (timbre) mà không thay đổi ngôn ngữ. Hơn nữa, tiếng việt cũng là một ngôn ngữ không quá phổ biến trên cộng đồng quốc tế, dẫn đến việc ngôn ngữ này không được chú trọng nghiên cứu nhiều bằng những ngôn ngữ khác, đặc biệt là đối với dịch thuật âm nhạc.

Từ đó, đề tài "Hệ thống chuyển đổi bài nhạc tiếng Việt sang phiên bản tiếng Anh" gồm 3 module chính được đề xuất nhằm khắc phục những thiếu sót trên. Hệ thống hướng tới việc tự động hóa quy trình phức tạp từ việc tách nhạc, dịch lời ca cho đến tổng hợp giọng hát, tạo ra một công cụ hỗ trợ đắc lực cho người sáng tạo nội dung và người học ngoại ngữ.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

1. Xây dựng hệ thống xử lý đầu vào tự động: Phát triển module có khả năng nhận đầu vào là file âm thanh (hoặc tùy chọn kèm theo lời bài hát/sheet nhạc), tự động tách giọng hát (vocal) & nhạc nền (Instrumental track), và trích xuất chính xác thông tin về giai điệu (melody) và lời hát (lyrics).
2. Đề xuất mô hình dịch Lyrics có ràng buộc âm nhạc: Xây dựng thuật toán dịch thuật lời bài hát đảm bảo giữ được nội dung ngữ nghĩa, đồng thời tuân thủ cấu trúc bài hát bao gồm: tương đồng về thời lượng câu, khớp nhịp tiết tấu, và đảm bảo vần điệu/nhấn âm phù hợp với tiếng Anh.
3. Tạo giọng hát tiếng Anh theo giai điệu gốc: Xây dựng mô hình tổng hợp giọng hát (Singing Voice Synthesis) có khả năng sinh ra giọng hát tiếng Anh tự nhiên,

khớp với melody gốc và hòa trộn hoàn chỉnh với nhạc nền (instrumental).

NỘI DUNG VÀ PHƯƠNG PHÁP

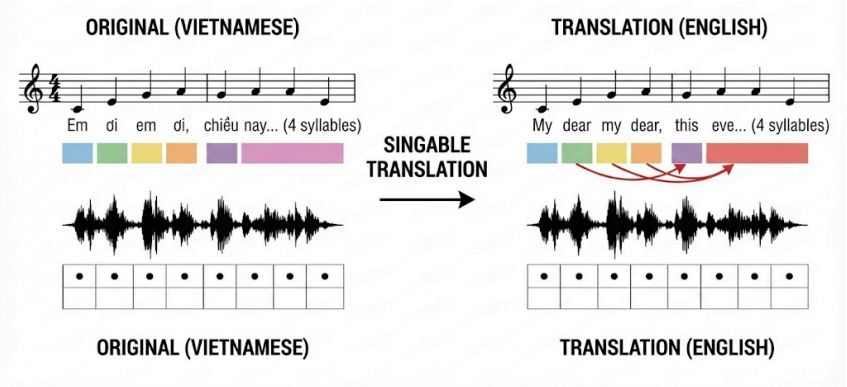
Hệ thống được xây dựng dựa trên quy trình xử lý gồm 4 bước chính:

Bước 1: Tiền xử lý dữ liệu đầu vào (Pre-processing)

- Sử dụng mô hình tách nguồn âm thanh SOTA (ví dụ: Demucs) để tách riêng phần Vocal (lời hát) và Instrumental (nhạc nền) từ file audio gốc.
- Áp dụng mô hình nhận dạng giọng nói (ví dụ: Whisper) để chuyển đổi Vocal thành văn bản (Lyrics) kèm thông tin thời gian.
- Sử dụng thuật toán trích xuất giai điệu (ví dụ: Omnizart hoặc CREPE) để lấy thông tin cao độ (pitch) và trường độ (duration).
- *Phương pháp lai (Hybrid):* Cho phép người dùng tải lên Music Sheet hoặc Lyrics chuẩn, thông qua các kỹ thuật đề xuất cải thiện (như Forced Alignment, Score-Audio Synchronization,...) để tăng độ chính xác của dữ liệu đầu vào.

Bước 2: Dịch thuật lời bài hát có ràng buộc (Music-Constrained Lyric Translation)

- Có thể sử dụng Mô hình Ngôn ngữ Lớn (LLM) để sinh ra nhiều ứng viên dịch thuật cho từng câu hát.
- Áp dụng thuật toán tối ưu hóa để lựa chọn phương án dịch tốt nhất dựa trên các hàm mục tiêu (Reward Functions) về:
    - Syllable count: Số âm tiết tiếng Anh khớp với số nốt nhạc.
    - Rhyme & Rhythm: Gieo vần và trọng âm phù hợp.
    - Meaning: Độ tương đồng ngữ nghĩa.



Bước 3: Tổng hợp giọng hát (Singing Voice Synthesis - SVS)

- Xây dựng mô hình Acoustic Model (ví dụ: dựa trên kiến trúc Diffusion hoặc Transformer) để dự đoán phổ âm thanh (Mel-spectrogram) từ lời bài hát tiếng Anh và thông tin giai điệu gốc.
- Sử dụng Vocoder (ví dụ: HifiGAN) để chuyển đổi Mel-spectrogram thành dạng sóng âm thanh (Waveform).

Bước 4: Hậu xử lý & Tích hợp

- Ghép giọng hát tiếng Anh vừa tạo với phần nhạc nền (Instrumental) đã tách ở tiền xử lý để tạo ra sản phẩm cuối cùng. Tích hợp các bước thành một hệ thống hoàn chỉnh.

KẾT QUẢ MONG ĐỢI

1. Bộ lời bài hát tiếng Anh hoàn chỉnh: Văn bản lời bài hát tiếng Anh đã được dịch và căn chỉnh, phù hợp để hát trên nền nhạc gốc (Singable Lyrics).
2. Sản phẩm âm thanh (Audio English Version): File âm thanh bài hát tiếng Anh hoàn chỉnh, giữ nguyên giai điệu và cảm xúc của bản gốc.
3. Ứng dụng Web (Demo): Một giao diện web đơn giản cho phép người dùng tải lên bài hát tiếng Việt và nhận lại kết quả là bài hát tiếng Anh.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Linan Ou, Xiaojuan Ma, Min-Yen Kan, Ye Wang: Songs Across Borders: Singable and Controllable Neural Lyric Translation. ACL (1) 2023: 447-467

[2]. Jinglin Liu, Chengxi Li, Yi Ren, Feilong Chen, Zhou Zhao: DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism. AAAI 2022: 11020-11028

[3]. Simon Rouard, Francisco Massa, Alexandre Défossez: Hybrid Transformers for Music Source Separation. ICASSP 2023: 1-5

[4]. Yu-Te Wu, Berlin Chen, Li Su: Multi-Instrument Automatic Music Transcription with Self-Attention-Based Instance Segmentation. IEEE/ACM Trans. Audio Speech Lang. Process. 29: 2798-2811 (2021)

[5]. Chengxi Li, Kai Fan, Jiajun Bu, Boxing Chen, Zhongqiang Huang, Zhi Yu: Translate the Beauty in Songs: Jointly Learning to Align Melody and Translate Lyrics. Findings of the Association for Computational Linguistics: EMNLP 2023: 27-39