# TÁCH GIỌNG HÁT TỪ BÀI HÁT SỬ DỤNG MẠNG NƠ-RON TÍCH CHẬP

## EXTRACT SINGING VOICE FROM MUSIC USING CONVOLUTIONAL NEURAL NETWORKS

***SVTH: Dương Huỳnh Sơn***

*Lớp 15TCLC1, Khoa Công Nghệ Thông Tin, Trường Đại học Bách Khoa, Đại học Đà Nẵng; Email: sonduong305@gmail.com*

***GVHD: Nguyễn Tấn Khôi***

*Khoa Công Nghệ Thông Tin, Trường Đại học Bách Khoa, Đại học Đà Nẵng; Email: ntkhoi@dut.udn.vn*

**Tóm tắt –** Hiện nay, cùng với sự phát triển nhanh chóng của nền công nghiệp âm nhạc giải trí và karaoke đã dẫn đến nhu cầu ngày càng tăng của không chỉ những nghệ sĩ cần đoạn âm thanh giọng hát của ca sĩ để tạo ra những sản phẩm đặc sắc dựa trên bài hát gốc mà rộng hơn là còn nhu cầu giải trí karaoke của đại đa số người yêu âm nhạc cần đoạn âm thanh riêng biệt của các nhạc cụ. Bài báo này trình bày phương pháp tách riêng đoạn âm thanh chứa giọng hát ra khỏi đoạn âm thanh gốc sử dụng mạng nơ ron tích chập. Thử nghiệp bước đầu trên CSDL MUSDB18 [1] và một số bài hát Việt Nam cho kết quả khả quan với thời gian thực hiện nhanh, có thể thực hiện trong thời gian thực khi bài hát đang được phát.

**Từ khóa –** Âm nhạc; tách giọng hát; tách karaoke; ảnh phổ; mạng nơ ron tích chập.

**Abstract –** Now adays, along with the rapid growth of current music industry and karaoke industry, not only artists need singing voice audio for making their own product but the music lovers also need the instruments audio for their karaoke entertainment. This article presents the method for extracting singing voice audio and instruments audio from music audio using Convolutional Neural Networks. Initial testing on MUSDB18 datasets and some Vietnamese songs shows the remarkable result with the fast processing time, the possibility of real-time processing while playing the music.

**Key words –** Music; Singing voice extraction; karaoke audio extraction; spectrogram; Convolutional Neural Networks.

### 1. Đặt vấn đề

Có rất nhiều bài hát khi được đưa ra thị trường nhưng không được công khai những tệp âm thanh của giọng hát và karaoke riêng biệt, khiến cho nhiều nghệ sĩ và người yêu âm nhạc không thể thực hiện nhu cầu giải trí ca hát của họ. Với đoạn âm thanh giọng hát có thể được kết hợp cùng với bản phối khí của nghệ sĩ mới để tạo ra sự mới mẻ cho bài hát, cũng như việc tách đoạn âm thanh này có thể được dùng để chuyển đổi từ âm thanh sang văn bản lời bài hát và được dùng cho tìm kiếm và lưu trữ, cũng như gắn thẻ nội dung dựa trên những đoạn âm thanh giọng hát này. Còn với đoạn âm thanh karaoke, người yêu nhạc có thể tự hát theo giai điệu của bài hát mà không bị nhiễu bởi giọng của ca sĩ gốc, từ đó có thể tạo ra những sản phẩm văn hóa văn nghệ và có tính giải trí cho người hát lẫn cộng đồng.

Trong bài báo này, chúng tôi nghiên cứu ứng dụng mạng nơ ron tích chập được huấn luyện trên nền cơ sở dữ liệu MUSDB18 được cung cấp bởi SigSep [2] với kết quả được đánh giá trong chiến dịch được nêu ở phần kết quả. Với phương pháp này, tất cả các dữ liệu âm thanh đều được chuyển về ảnh phổ [3] dựa trên phương pháp biến đổi Fourier thời gian ngắn (Short-time Fourier Transform) [4], chuyển về ma trận 2 chiều cho mục đích huấn luyện.

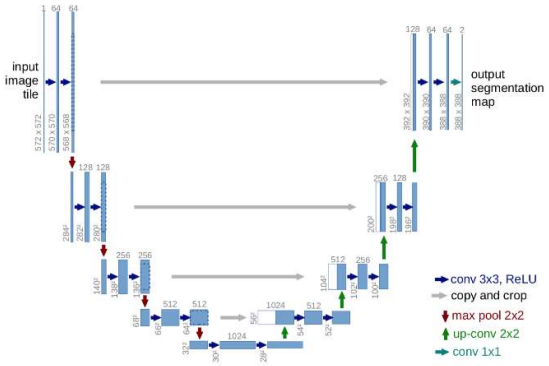
Bố cục của bài báo gồm các nội dung chính sau: Phần 2 trình bày về nhận diện vật thể. Phần 3 trình bày một số kết quả ban đầu triển khai thực nghiệm hệ thống. Kết luận và hướng nghiên cứu trình bày trong phần cuối.

### 2. Mạng nơ-ron tích chập

Phương pháp chính chúng tôi đề cập ở đây sẽ dùng kiến trúc của mạng nơ ron tích chập [5] để phân tích và thực hiện bài toán, với đầu vào là các ảnh phổ được biến đổi từ dạng sóng của âm thanh dựa trên phương pháp biến đổi thời gian ngắn, đầu ra là ảnh phổ tương ứng với các thông tin tần số và thời gian của giọng hát.

Mục tiêu của mô hình mạng nơ ron là để dự đoán ảnh phổ của riêng giọng hát khi nhận đầu vào là ảnh phổ của đoạn nhạc tương ứng, kích thước của ma trận đầu vào và ma trận đầu ra là bằng nhau.

Kiến trúc U-Net [6] đầu tiên được dùng trong công việc trích xuất hình ảnh những tế bào sinh học, dùng để tăng độ chính xác trong việc định vị các điểm ảnh của các tế bào vi sinh vật trong hình ảnh y tế, cho đầu ra là hình ảnh gồm 2 giá trị: 1 – tế bào sinh học, 0 – không phải tế bào sinh học.




*Hình 1: Kiến trúc mạng U-Net trong bài toán phân tích hình ảnh sinh học*

Kiến trúc U-Net được xây dựng dựa trên mô hình encoder-decoder. Phần encoder bao gồm những lớp tích chập, mục đích làm giảm số chiều của ma trận dữ liệu đầu vào trong khi dữ lại và tạo ra những thông tin hữu ích cho việc tính toán output. Phần decoder có trách nhiệm nhận input là output của phần encoder và tái tạo lại thông tin dựa trên input đó.

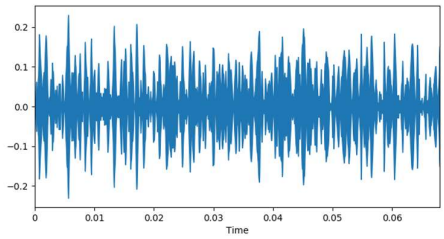
Nếu xem các ảnh phổ đã được xử lý từ đoạn âm thanh ở phần trước như một hình ảnh thì ta có thể thấy thay vì cung cấp đầu ra của mạng nơ ron gồm 2 giá trị 0 và 1 thì

trong bài toán trích xuất giọng hát, chúng ta cần phải dự đoán giá trị thực cho từng điểm ảnh trong ảnh phổ tương ứng với biên độ của tần số ở thời điểm xác định. Và trong bài toán trích xuất hình ảnh, một vài giá trị đầu ra không chính xác vẫn không ảnh hưởng nhiều đến kết quả chúng ta cần tìm, nhưng trong bài toán này, một điểm ảnh sai có thể dẫn đến sự sai lệch về tần số trong cả toàn bộ bài hát. Do đó, buộc chúng ta phải xây dựng và thử nghiệm một mô hình có độ chính xác rất cao. Trong kiến trúc của mạng U-Net có những cấu trúc lược bỏ kết nối giữa các lớp trong mạng nơ ron (skip connection), thông tin của một lớp trong phần encoder sẽ được đưa thẳng sang và nối tiếp với những lớp phía sau có cùng kích thước trong phần decoder, do đó thông tin từ những lớp ban đầu sẽ được giữ lại và tạo ra kết quả tốt hơn.

## 3. Giải phát đề xuất

### *3.1. Trích xuất ảnh phổ từ tệp âm thanh bài hát*

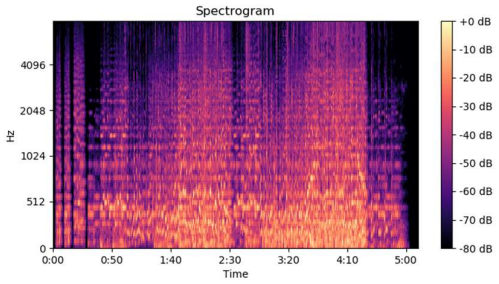
Thông thường, chúng ta sẽ lưu tệp âm thanh dưới dạng sóng, như hình bên dưới.




*Hình 2: Biểu diễn âm thanh dưới dạng sóng*

Như hình 1, chỉ có thông tin về thời gian và biên độ của sóng âm, rất khó để chúng ta có thể xác định được các tần số theo thời gian, dẫn đến sự không chính xác trong việc phân tích và trích xuất.

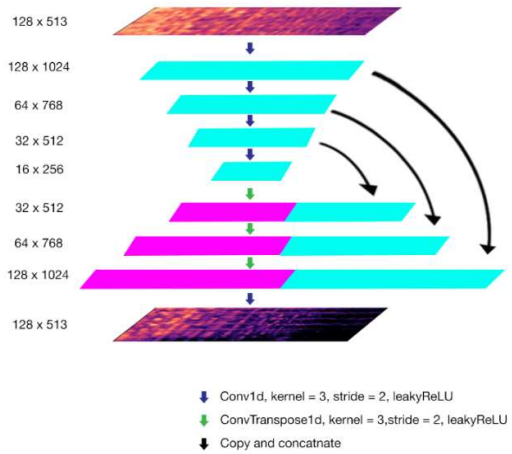
Ảnh phổ (Spectrogram) là cách biểu diễn trực quan hơn biên độ của các tần số biến thiên theo thời gian của một tín hiệu. Ảnh phổ là một đồ thị 3 chiều: thời gian, tần số và độ lớn của tần số tại thời gian tương ứng.




*Hình 3: Biểu diễn âm thanh dưới dạng ảnh phổ*

### *3.2. Kiến trúc mạng nơron cho bài toán tách giọng hát*

Bài báo đề xuất sử dụng 4 lớp tích chập có kích thước *m * 3* với *m * n* là kích thước của input cho phần encoder, sau đó là 3 lớp tích chập đảo (*ConvTranspose*), đồng thời sao chép thông tin và nối trực tiếp các lớp tương ứng ở phần encoder vào những lớp này. Cuối cùng là một lớp tích chập để cho ra kết quả cuối cùng. Kiến trúc của mạng được biểu diễn như hình bên dưới.




*Hình 4: Kiến trúc mạng nơ ron dùng để tách giọng hát*

Ở đây chúng tôi sử dụng tích chập 1 chiều vì sau khi thử nghiệp với bộ tích chập 2 chiều cho kết quả không khả quan và thời gian xử lý input lâu hơn bộ tích chập 1 chiều.

#### *3.2.1. Tái tạo lại tín hiệu từ ảnh phổ*

Mạng nơ ron sử dụng biên độ tần số của ảnh phổ để làm input cho việc xử lý, cung cấp cho chúng ta đầu ra cũng là một ma trận biên độ tần số với kích thước tương ứng. Giá trị pha của tín hiệu cần tìm được lấy trực tiếp từ pha của tín hiệu gốc. Từ những thông tin về biên độ, pha của tần số theo thời gian, chúng ta sử dụng biến đổi Fourier thời gian ngắn ngược (*Inverse Short-time Fourier Transform*) để tái tạo lại tín hiệu về miền thời gian theo công thức sau:

$$y = istft(s \,.\exp\left(i \,.\, angle\left(stft(x)\right)\right))$$

Trong đó y là tín hiệu trên miền thời gian cần tái tạo lại, *istft* là phép biến đổi Fourier thời gian ngắn ngược, *s* là giá trị biên độ được cung cấp bởi output của mạng nơ ron, *i* là đơn vị số ảo = $\sqrt{-1}$ , angle được lấy trực tiếp tự phép biến đổi Fourier của tín hiệu đầu vào, và *x* là tín hiệu đầu vào tương ứng trên miền thời gian. Từ đó, ta áp dụng cho ma trận ảnh phổ để tái tạo lại tín hiệu:

$$Y = istft(S \odot \exp\left(i \,.\, angle\left(stft(X)\right)\right))$$

với *X* và *Y* lần lượt là các vector tín hiệu đầu vào và tín hiệu cần tái tạo, *S* là ma trận ảnh phổ được lấy từ tín hiệu *X*.

Sau khi đã có được dữ liệu của tín hiệu giọng hát, chúng ta tìm dữ liệu tín hiệu của đoạn nhạc không có giọng hát (karaoke):

$$Y_{Karaoke} = X - Y_{giọng\ hát}$$

với *X* là dữ liệu của bài hát gồm nhạc và giọng hát.

### *3.3. Dữ liệu huấn luyện, quá trình huấn luyện và đánh giá mô hình*

Musdb18 là tập dữ liệu huấn luyện được công khai trong chiến dịch trích xuất tín hiệu (SigSep), bao gồm 100 bài hát cho phần huấn luyện và 50 bài hát cho phần đánh giá. Mỗi bài hát gồm có 4 phần âm thanh riêng biệt, bao gồm giọng hát, tiếng trống, tiếng bass, và phần còn lại gồm piano, guitar, các nhạc cụ khác.

#### *3.3.1. Huấn luyện mô hình*

Chúng tôi thực hiện huấn luyện mô hình với hàm tổn thất MSE được định nghĩa như sau:

$$L(X,Y,\theta) = \frac{1}{N}(\sum_{i=1}^{N}(Y - f(X,\theta))^2)$$

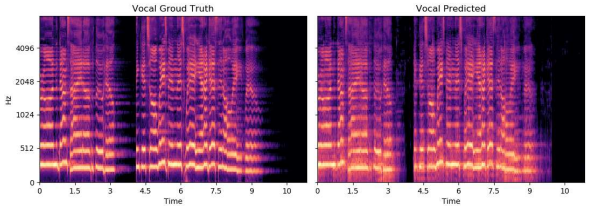
trong đó $f(X,\theta)$ là output của mạng nơ ron với các tham số $\theta$, $Y$ là giá trị đúng của output tương ứng với $X$.

Sau khi huấn luyện 60 epoch bằng phương pháp Cycle learning [7], chúng tôi chuyển từ hàm tổn thất MSE sang MAE để giảm bớt ảnh hưởng của nhiễu đến đầu ra của mạng nơ ron. Hàm mất mát MAE được định nghĩa như sau:

$$L(X,Y,\theta) = \frac{1}{N}(\sum_{i=1}^{N}|Y - f(X,\theta)|)$$

*3.3.2. Đánh giá mô hình*

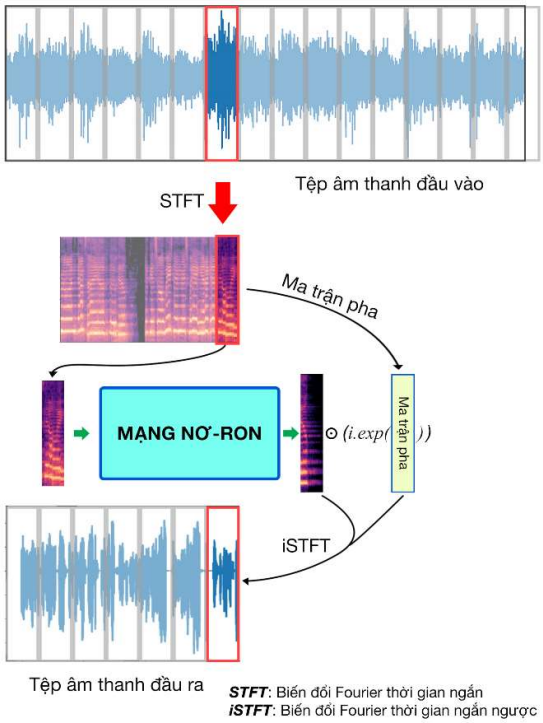
Sau khi huấn luyện tổng cộng 100 epoch, mô hình cho ra được giá trị của hàm mất mát MSE trong tập dữ liệu đánh giá ~0.006, và giá trị của MAE trong tập dữ liệu đánh giá ~0.024.




*Hình 5: Ảnh phổ giọng hát thật và ảnh phổ được dự đoán bởi mô hình*

### *3.4. Sơ đồ tổng thể*

Từ các tệp âm thanh có định dạng **.mp3, *.wav*, chúng tôi sẽ cho ra đầu ra gồm 2 tệp âm thanh của giọng hát riêng và phần karaoke riêng theo sơ đồ như sau:




*Hình 6: Sơ đồ tổng thể của hệ thống*

## 4. Kết quả thực nghiệm và đánh giá

Kết quả thực nghiệm được đánh giá trên bộ dữ liệu đánh giá của MUSDB18.

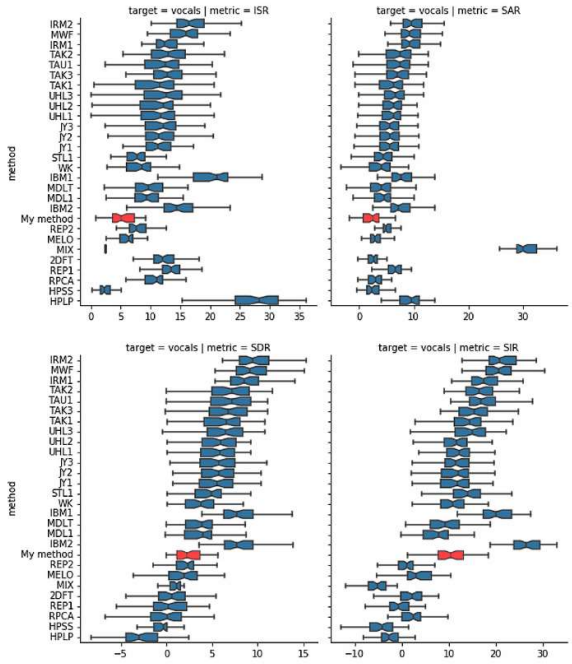
### *4.1. Dữ liệu đánh giá*

Dữ liệu đánh giá có độ dài xấp xỉ *3.5* giờ nhạc, được phân chia theo thể loại như sau:

| **Thể loại** | **Số lượng bài hát** | **Tỉ lệ** |
|---|---|---|
| Electronic | 4 | 8% |
| Heavy Metal | 4 | 8% |
| Pop/Rock | 36 | 72% |
| Rap | 3 | 6% |
| Reggae | 2 | 4% |
| Rock | 1 | 2% |
| **Tổng cộng** | **50** | **100%** |

### *4.2. Đánh giá*

Bên dưới là kết quả của mô hình so sánh với các phương pháp hiện tại trên thế giới:




*Hình 7: Kết quả đánh giá bởi chiến dịch SigSep, so sánh với các phương pháp khác*

Ở đây, các phương pháp khác đa phần được huấn luyện trên nhiều bộ dữ liệu khác nhau, kích thước của bộ dữ liệu cũng rất lớn, nên các phương pháp đó đạt kết quả rất cao.

Khi nghe trực tiếp kết quả cho ra từ mô hình với các bài hát tiếng Anh với chất lượng tốt, tuy vẫn còn một số tạp âm của các âm thanh khác nhưng không đáng kể. Với những bài hát tiếng Việt, tạp âm nhiều hơn, ảnh hưởng đến chất lượng đầu ra.

Về thời gian xử lý, mô hình có thời gian xử lý trung bình đối với một đoạn nhạc 180s là 40s trên vi xử lý Intel® Xeon® Processor E3-1231v3. Đối với những thiết bị có hỗ trợ GPU, mô hình có thời gian xử lý nhanh hơn đáng kể - thời gian xử lý đoạn nhạc 180s là 20s trên GPU GeForce GTX 1050Ti.

## 5. Kết luận và hướng phát triển

Trong bài báo này, chúng tôi trình bày hướng nghiên

cứu xây dựng một hệ thống tách âm thanh giọng hát và karaoke sử dụng mạng nơ ron tích chập. Kết quả thực nghiệm cho thấy độ chính xác ổn với thời gian thực thi nhanh.

Hướng phát triển tiếp theo sẽ triển khai thu thập dữ liệu nhiều hơn để hệ thống có thể xử lý tốt hơn. Đồng thời sẽ triển khai cài đặt hệ thống xử lý theo thời gian thực ngay khi đang phát đoạn âm thanh để ứng dụng vào thực tiễn.

## Tài liệu tham khảo


[1] https://sigsep.github.io/datasets/musdb.html.

[2] Fabian-Robert Stoter, Antoine Liutkus, Nobutaka Ito. The 2018 Signal Separation Evaluation Campaign .

[3] Ahmad Zuri bin Sha'ameri and Tan Jo Lynn, "Spectrogram time-frequency analysis and classification of digital modulation signals," 2007 IEEE International Conference on Telecommunications and Malaysia International Conference on Communications, Penang, 2007.

[4] Tatsuro Baba. Time-Frequency Analysis Using Short Time Fourier Transform

[5] Andreas Jansson1, 2 , Eric Humphrey2 , Nicola Montecchio2 , Rachel Bittner2 , Aparna Kumar2 , Tillman Weyde. SINGING VOICE SEPARATION WITH DEEP U-NET CONVOLUTIONAL NETWORKS.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation.

[7] Smith, Leslie N. Cyclical Learning Rates for Training Neural Networks.