

B**Ộ THÔNG TIN VÀ TRUYỀN THÔNG**

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN 1**



**BÁO CÁO BÀI LỚN**

**SỬ DỤNG NGÔN NGỮ PYTHON ĐỂ XÂY DỰNG CSDL ÂM THANH**

**Giảng viên hướng dẫn:** Nguyễn Đình Hóa

**Môn:** Cơ sở dữ liệu đa phương tiện

**Nhóm:** 02

**Danh sách thành viên:**

Lê Công Tuấn B15DCCN607

Nguyễn Văn Bảy B15DCCN057

Tạ Khắc Tùng B15DCCN629

***Hà Nội 2019***

MỤC LỤC

[LỜI NÓI ĐẦU 1](#_Toc9342361)

[CHƯƠNG I. KỸ THUẬT TRÍCH CHỌN ĐẶC TRƯNG TRONG NHẬN DẠNG ÂM THANH. 2](#_Toc9342362)

[1.1. Trích chọn đặc trưng sử dụng mô hình MFCC. 2](#_Toc9342363)

[1.1.1. Tổng quan về mã hóa âm thanh. 2](#_Toc9342364)

[1.1.2. Các bước trích rút đặc trưng âm thanh. 3](#_Toc9342365)

[1.2. Mô hình Markov ẩn (HMM). 8](#_Toc9342366)

[1.3. Các thuộc tính âm thanh được sử dụng để nhận dạng. 8](#_Toc9342367)

[CHƯƠNG II. BỘ CƠ SỞ DỮ LIỆU. 9](#_Toc9342368)

[2.1. Đặc điểm của cơ sở dữ liệu. 9](#_Toc9342369)

[2.2. Xây dựng cơ sở dữ liệu. 9](#_Toc9342370)

# **LỜI NÓI ĐẦU**

Hãy dụng ngôn ngữ lập trình Python để xây dựng hệ CSDL âm thanh nói chung.

1. Tìm hiểu tài liệu và trình bày về các kỹ thuật nhận dạng âm thanh. Trình bày các thuộc tính âm thanh thường được sử dụng để nhận dạng các đoạn âm thanh khác nhau.

2. Hãy sưu tầm 100 đoạn âm thanh của nguồn âm khác nhau, các đoạn âm thanh đều có cùng độ dài.

3. Xây dựng chương trình nhận dạng âm thanh với đầu vào là một mẫu âm thanh mới, đầu ra là đoạn âm thanh giống nhất với mẫu đầu vào từ 100 đoạn âm thanh đã có ở mục 2. Các thuộc tính được sử dụng để so sánh các đoạn âm thanh được trình bày ở phần 1.

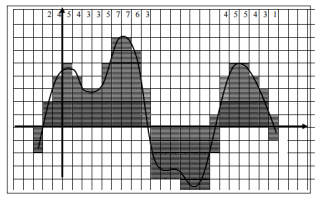
4. Đánh giá kết quả đã đạt được.

# **CHƯƠNG I. KỸ THUẬT TRÍCH CHỌN ĐẶC TRƯNG TRONG NHẬN DẠNG ÂM THANH.**

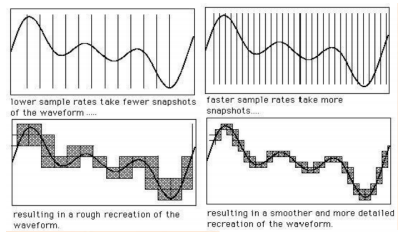
## **1.1. Trích chọn đặc trưng sử dụng mô hình MFCC.**

### 1.1.1. Tổng quan về mã hóa âm thanh.

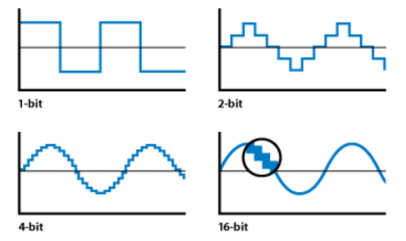
Có nhiều phương pháp mã hoá âm thanh. Cách đơn giản nhất là mã hoá bằng cách xấp xỉ dao động sóng âm bằng một chuỗi các byte thể hiện biên độ dao dộng tương ứng theo từng khoảng thời gian bằng nhau. Các đơn vị thời gian này cần phải đủ nhỏ để không làm “nghèo” âm thanh. Đơn vị thời gian này gọi là tần số lấy mẫu (sample rate). Giá trị tại mỗi lần lấy mẫu được biểu diễn trong một miền giá trị xác định được gọi là độ sâu số (bit depth). Khi phát, một mạch điện sẽ khôi phục lại sóng âm với một sai lệch chấp nhận được.



*Hình 1. Số hóa tin hiệu âm thanh.*

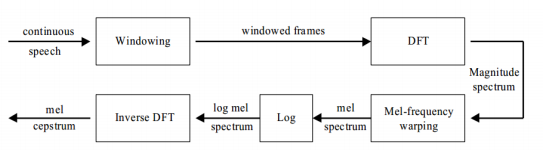


*Hình 2 : Lấy mẫu với tần số khác nhau.*



*Hình 3. Lấy mẫu với độ sâu khác nhau.*

### 1.1.2. Các bước trích rút đặc trưng âm thanh.



*Hình 4. Bước trích rút đặc trưng âm thanh.*

Bước 1. windowing, chia tín hiệu âm thanh ban đầu thành các frame liên tiếp nhau. Mỗi frame này sẽ được đưa vào và rút trích đặc trưng MFCC tương ứng.

Bước 2. biến đổi fourier rời rạc. Bước này nhằm chuyển đổi tín hiệu ban đầu thành tổ hợp của các sinusoid tương ứng với từng tần số khác nhau.

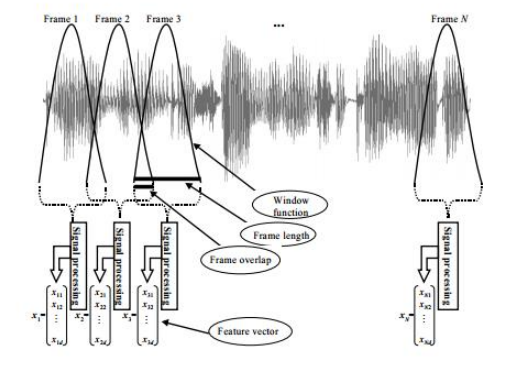
Bước 3. Chuyển tín hiệu ở dạng tần số thu được ở bước hai sang một vùng tần số theo cảm nhận của tai người.

Bước 4. Lấy log để tách tín hiệu tần số thấp và tần số cao thành 2 vùng khác nhau.

Bước 5. thực hiện phép biến đổi fourier đảo, ta thu được đặc trưng MFCC.

**Chi tiết hoạt động của từng bước.**

Bước 1. Windowing.

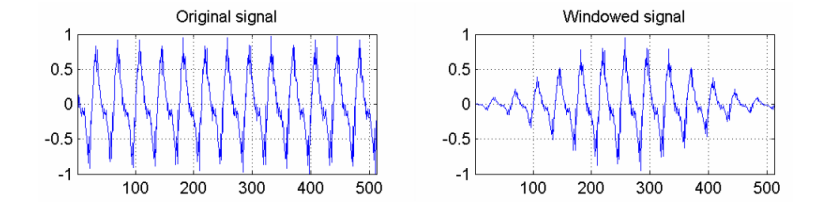


*Hình 5. Quá trình framming.*

Tín hiệu nói ban đầu được chuyển thành các frame có kích thước cố định (20-30 ms), mỗi window sẽ có một phần chồng lên nhau (30 – 50%) với các frame cạnh nó nhằm tránh mất mát thông tin. Để tránh biến đổi đột ngội ở cuối frame, mỗi frame thường được nhân với một hàm window (window function), mà phổ biến nhất là hamming window function:

𝑤(𝑡) = 0.54 − 0.64𝑐𝑜𝑠 ()

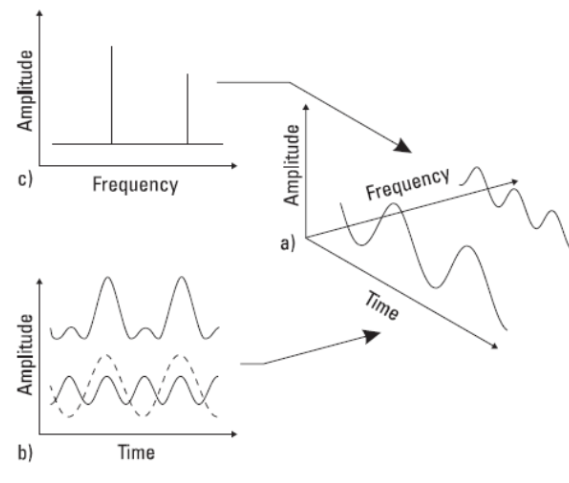
Với N là kích thước của frame. Kết quả thu được sẽ được lần lượt đưa và quá trình rút trích đặc trưng.



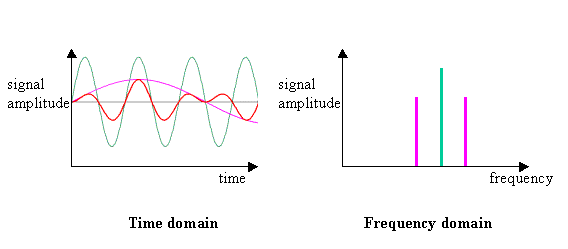
*Hình 6. Tín hiệu trước và sau khi sử dụng hàm window.*

Bước 2. Biến đổi fourier rời rạc.

Theo lý thuyết biến đổi Fourier bất kỳ một hàm nào đó biến đổi theo thời gian cũng có thể phân tích được thành các hàm sin đơn giản ở các tần số khác nhau và ngược lại.



*Hình 7. Sự tương quan qua lại giữ thời gian và tần số.*



*Hình 8. Kết quả quá trình biến đổi từ miền thời gian sang miền tần số.*

Mỗi frame thu được sau quá trình xử lý sẽ được đưa vào phép biến đổi fourier rời rạc để chuyển từ miền thời gian sang miền tần số.

Một biến đổi Fourier nhanh (FFT) là một thuật toán hiệu quả để tính biến đổi Fourier rời rạc (DFT) và biến đổi ngược. Khi cài đặt thực tế, ta sử dụng phép FFT này lên các frame, kết quả sẽ được chuyển qua bước tiếp theo, đó là lọc Mel-frequency.

Sau khi thực hiện biến đổi này, ta thu được một biểu diễn khác của âm thanh được gọi là cepstrum.

Bước 3. Chuyển tín hiệu ở dạng tần số thu được ở bước hai sang một vùng tần số theo cảm nhận của tai người.

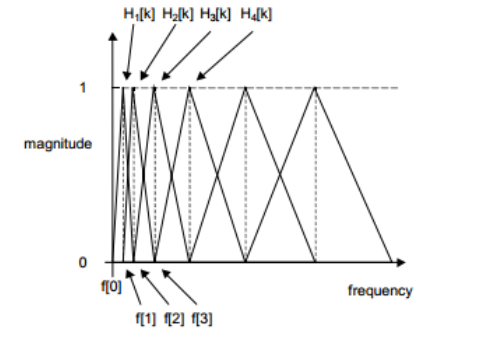
Sau bước biến đổi Fourier rời rạc(DFT), ta thu được thông tin về tín hiệu dưới dạng tần số và cường độ. Tuy nhiên, tai người cảm nhận âm thanh không giống với những biến đổi vật lý của âm thanh. Do đó, người ta sử dụng một thang đo tần số khác được gọi là tần số mel, được đo theo cảm nhận của tai con người. Bảng chuyển đổi tần số vật lý sang thang đo mel như sau, người ta lấy 1000 Hz làm mốc chuyển đổi giữa hai thang đo:



Người ta xây dựng nhiều công thức để chuyển từ Hz sang mel, trong đó phổ biến nhất là công thức của Lindsay và Loman:

M = 2410.

Thông qua một bộ lọc, người ta có thể tính toán lại tần số và biên độ ở thang đo Hz sang thang đo mel, khi đó ta thu được một vector tần số và biên độ mới.



*Hình 9. Kết quả biến thiên biên độ theo tần số*

Bước 4. Lấy log để tách tín hiệu tần số thấp và tần số cao thành 2 vùng khác nhau.

Tín hiệu tiếng nói của con người có thể được biểu diễn bởi hai thành phần là những thành phần biến đổi nhanh và vùng biến đổi chậm. Các đỉnh ở phổ âm thanh cùng với Có thể biểu diễn sự tương quan của hai thông tin “nhanh” và “chậm” này như sau:

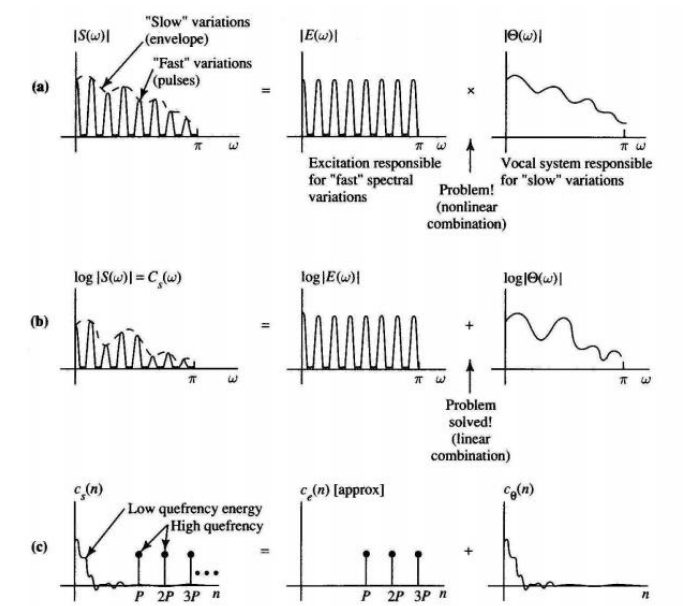
|S(x)| = |E(x).H(x)|

Trong đó E(x) là thành phần có tần số cao, H(x) là thành phần có tần số thấp, S(x) là tín hiệu gốc.

Khi thực hiện logarit trên biểu thức trên, ta có thể chuyển tổng thành tích như sau: log(|S(x)|) = log(|E(x)|) + log(|H(x)|)

Để thực hiện phân tách, người ta thực hiện một mẹo, đó là sử dụng phép biến đổi fourier trên chính log(|S(x)|) và phép biến đổi này được gọi là phép biến đổi fourier đảo.

Từ kết quả của phép biến đổi này, ta có thể lọc ra hai vùng có tần số cao và thấp, vùng cần lấy là vùng có tần số thấp. Biểu diễn trực quan của cách làm này như sau:



*Hình 10. Kết quả thu được sau toàn bộ quá trình này là đặc trưng Mel-frequency Cepstral Coefficients. Mỗi frame sẽ thu được một vector đặc trưng và các vector này sẽ được đưa vào quá trình mô hình hóa và nhận diện người nói.*

## **1.2. Mô hình Markov ẩn (HMM).**

HMM được bắt đầu xây dựng và công bố từ những năm 1960, đây là mô hình toán học về thống kê.

Do đạt được độ chính xác cao và có khả năng thay đổi cấu trúc dễ dàng nên mô hình này ngày càng được sử dụng rộng rãi trong nhiều lĩnh vực, đặc biệt là trong lĩnh vực nhận dạng tiếng nói.

Đây là một mô hình thống kê, thành phần của mô hình bao gồm tập N trạng thái {Si}, các trạng dịch chuyển qua lại với nhau với một xác suất nhất định, tập xác suất di chuyển này được gọi là ma trận dịch chuyển trạng thái A=[]. Mô hình hoạt động khi cho chuỗi dữ liệu đầu vào O=[] gọi là chuỗi quan sát, đây là dữ liệu trích rút từ tiếng nói cần nhận dạng trong ứng dụng nhận dạng tiếng nói.

Mô hình Markov ẩn (HMM) là một mô hình thống kê, thích hợp ứng dụng trong việc nhận dạng mẫu: tiếng nói, hình ảnh và chữ viết…

HMM được ứng dụng rộng rãi trong những năm gần đây vì hai lý do.

- Mô hình có độ chính xác cao trong nhiều ứng dụng.

- Cấu trúc mô hình có thể thay đổi dễ dàng cho phù hợp với từng ứng dụng cụ thể.

HMM có nhiều tham số, vì vậy việc lựa chọn tham số sao cho tốt nhất cũng được thực hiện trong đề tài. Việc lựa chọn này rất quan trọng, nó phải đạt được sự cân bằng giữa tốc độ xử lý và độ chính xác. Hệ thống nhận dạng này được cài đặt để nhận dạng các từ đơn, số lượng từ trong bộ từ vựng có thể thay đổi nhờ khả năng có thể huấn luyện của HMM.

Thích hợp ứng dụng trong giao tiếp người-máy, robot, điều khiển bằng tiếng nói hay hỗ trợ người khuyết tật…

## **1.3. Các thuộc tính âm thanh được sử dụng để nhận dạng.**

**- Độ cao:**

Là mức độ cao thấp của âm, phụ thuộc vào sự chấn động nhanh hay chậm của không khí trong một khoảng thời gian nhất định, được gọi là tần số dao động. Tần số dao động càng lớn thì âm thanh càng cao.

**- Độ mạnh:**

Thường được gọi là cường độ, do biên độ dao động quyết định. Trong ngôn ngữ, phụ âm thường mạnh hơn nguyên âm, đây chính là một trong những đặc điểm góp phần nhận diện sự khác biệt giữa phụ âm và nguyên âm trong âm thanh tiếng nói.

**- Độ dài:**

Là trường độ của âm, phụ thuộc vào sự chấn động lâu hay mau của các phần tử không khí. Độ dài được sử dụng để phân biệt các nguyên âm dài và ngắn, như phân biệt “a” với “ă”, “ơ” với “â” trong tiếng Việt.

**- Âm sắc:**

Là sắc thái riêng của một âm do các cá thể khác nhau tạo ra. Âm sắc là nguyên nhân gây ra sự khác biệt giữa giọng nói của người này với người khác. Âm sắc có được là do hiện tượng cộng hưởng.

**- Tiếng ồn và tiếng thanh:**

Tiếng ồn là do sự chuyển động không nhịp nhàng (không có chu kỳ ổn định) của các phần tử không khí gây ra. Tiếng thanh là do sự chuyển động nhịp nhàng (có chu kỳ ổn định) của các phần tử không khí gây ra.

# **CHƯƠNG II. BỘ CƠ SỞ DỮ LIỆU.**

## **2.1. Đặc điểm của cơ sở dữ liệu.**

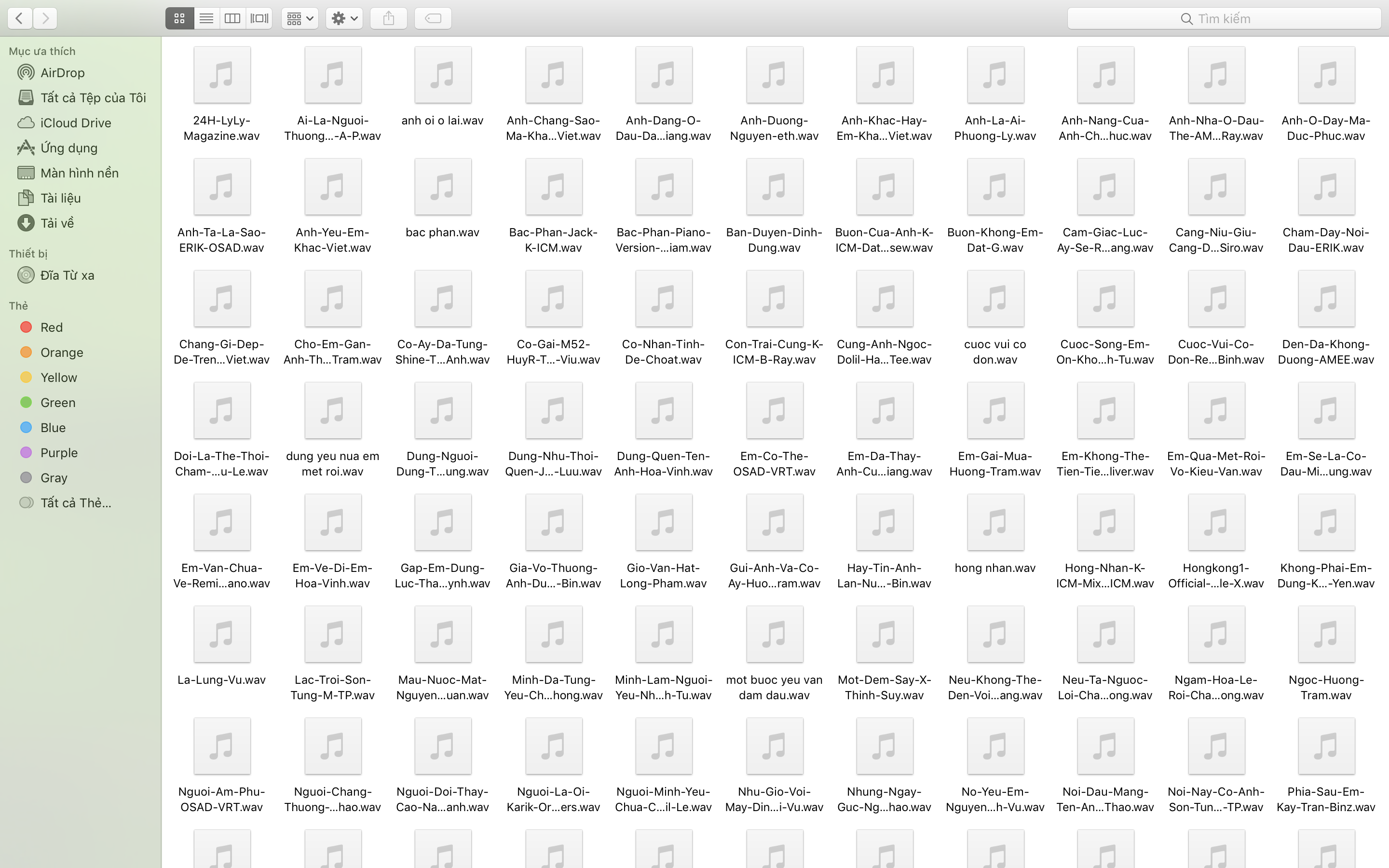
- Bộ cơ sở dữ liệu gồm 100 đoạn âm thanh với độ dài là 30 giây.

- Mỗi đoạn âm thanh là một đoạn âm nhạc các bài hát được cắt ra và lưu dưới dạng \*.wav

- Các đoạn âm thanh gồm có âm thanh điệu nhạc khác nhau và lời hát của các ca sỹ giống nhau.

- Có thể một số bài hát được hát bởi cùng 1 ca sỹ.

## **2.2. Xây dựng cơ sở dữ liệu.**



*Hình 11. Tất cả bản ghi của CSDL.*