Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00064

PHÂN LOẠI NHẠC THEO THỂ LOẠI DÙNG PHÉP BIẾN ĐỔI WAVELET RỜI RẠC

Phan Anh Cang [1], Phan Thượng Cang[2]

[1] Khoa Công Nghệ Thông Tin, Trường Đại học Sư Phạm Kỹ Thuật Vĩnh Long
[2] Khoa Công Nghệ Thông Tin, Trường Đại học Cần Thơ

cangpa@vlute.edu.vn, ptcang@cit.ctu.edu.vn

TÓM TẮT— *Cùng với sự bùng nổ về công nghệ thông tin và sự gia tăng nhu cầu sưu tập nhạc số của mỗi cá nhân hay tổ chức, việc phân loại các bản nhạc để dễ dàng quản lý là một nhu cầu tất yếu. Tuy nhiên, do việc sưu tầm từ nhiều nguồn khác nhau nên việc phân loại chỉ dựa trên thông tin ghi trên tập tin lưu trữ còn gặp nhiều hạn chế. Với một số lượng đồ sộ các bản nhạc thì việc phân loại các bản nhạc là một thách thức đối với người nghe nhạc và các hệ thống lưu trữ âm nhạc. Điều này làm cho nhu cầu xây dựng hệ thống phân loại nhạc tự động trở nên cần thiết. Trong bài báo này, chúng tôi đề xuất một hệ thống phân loại nhạc theo thể loại sử dụng các phương pháp rút trích tập các đặc trưng của tín hiệu audio bao gồm âm sắc (timbral texture), nhịp điệu (rhythmic content) và cao độ (pitch) phục vụ cho việc phân loại nhạc tự động theo thể loại. Trong đó, phép biến đổi wavelet rời rạc để phân tích tín hiệu audio thành các băng tần con dùng cho việc xác định các đặc trưng về nhịp điệu. Nghiên cứu của chúng tôi thực hiện minh hoạ trên bốn thể loại Classical, Rock, Jazz và Pop. Nghiên cứu này có thể áp dụng mở rộng đối với các thể loại nhạc khác hoặc xây dựng các hệ thống truy vấn thông tin nhạc dựa vào nội dung, kiểm tra việc sao chép bản quyền nhạc,....*

Từ khóa— *Phân loại nhạc, wavelet rời rạc, tín hiệu âm nhạc, rút trích đặc trưng tín hiệu audio.*

I. GIỚI THIỆU

Trong những năm gần đây, cùng với sự phát triển của công nghệ thông tin, số lượng bản nhạc dưới hình thức dữ liệu audio trong các kho dữ liệu lớn, trên Internet, đang ngày càng gia tăng nhanh chóng. Điều này làm cho việc sở hữu những bản nhạc trở nên dễ dàng hơn bao giờ hết, kéo theo đó là sự gia tăng nhu cầu sưu tập nhạc số ở mỗi cá nhân hay tổ chức. Hiện nay, hầu hết các hệ thống lưu trữ nhạc số sắp xếp các bản nhạc theo tên nhạc sĩ hoặc theo tên bài hát trong khi người nghe nhạc chỉ quan tâm đến các thể loại nhạc. Điều này đã nảy sinh nhu cầu phân loại nhạc tự động theo thể loại trong các hệ thống lưu trữ nhạc số để cho phép người nghe nhạc có thể tìm kiếm bản nhạc theo yêu cầu. Tuy nhiên, với số lượng lớn nhạc số sưu tầm được, việc phân loại chúng để dễ dàng quản lý trở thành một thách thức đối với các hệ thống phân loại nhạc tự động. Điều này là do việc sưu tầm nhạc thực hiện từ nhiều nguồn nên nó có thể có nhiều thông tin khác nhau cho từng bản nhạc tải về. Bên cạnh đó, người sưu tầm có thể tự nghe lại từng bản nhạc rồi tự phân loại chúng thay vì chỉ dựa vào các thông tin có sẵn được lưu trữ trên tập tin nhạc. Theo cách này, độ chính xác về phân loại đối với các bản nhạc sẽ tùy thuộc vào khả năng hiểu biết về âm nhạc của người phân loại. Điều này cho thấy, việc phân loại các bản nhạc với các phương pháp truyền thống trên còn nhiều hạn chế về độ chính xác và không khả thi với một số lượng lớn các bản nhạc số. Do đó, các hệ thống phân loại nhạc tự động là rất cần thiết đối với các hệ thống lưu trữ nhạc số, phát hiện sao chép bản quyền, tìm kiếm thông tin nhạc trên Internet,... bởi vì chúng cung cấp cơ sở khoa học cho việc phân tích các tín hiệu nhạc dựa vào nội dung.

Nhiều nghiên cứu đã đưa ra các ý tưởng phát triển các hệ thống phân loại nhạc tự động trong thời gian gần đây. Anan et al. đề xuất một tiếp cận phân loại nhạc dựa trên độ đo tương đồng và máy học véctơ hỗ trợ (Support vector machines - SVM) [**1**]. Để xác định mức độ tương đồng giữa các tín hiệu audio, phương pháp này biến đổi các file audio dưới định dạng MIDI thành ba tập dữ liệu dạng chuỗi bao gồm cao độ, nhịp điệu, và nốt nhạc (Pitch string, Rhythm string and Note string). Tuy nhiên, phương pháp này là không thực tế vì nó đòi hỏi tất cả các file audio dưới định dạng MIDI và hệ thống phiên âm đa âm là một bài toán khó giải quyết hơn là phân loại. Một số phương pháp khác phân tích dựa trên hình dạng của tín hiệu audio và ảnh phổ. Costa et al. đã đề xuất cách tiếp cận dựa vào ảnh phổ để phân loại nhạc [**2**]. Phương pháp này phân tích tín hiệu audio thành ảnh phổ và sau đó rút trích các đặc trưng từ ảnh này. Tuy nhiên, chúng ta rất khó để nhận biết thể loại nhạc một cách chính xác nếu chỉ dựa trên việc xem ảnh phổ này mà không có sự phân tích dựa trên tiết tấu, cao độ,… của âm thanh. Một cách tiếp cận khác cho việc phân loại nhạc dựa trên việc rút trích và lựa chọn đặc trưng được đề xuất bởi nhiều nghiên cứu được trình bày trong [**3**], [**4**]. Trong đó, Matsui et al. đã sử dụng các đặc trưng hướng được rút trích dựa trên thuật toán SIFT [**4**]. Đặc trưng này cung cấp các thông tin về tần số của tín hiệu nhạc. Các kết quả thực nghiệm cho thấy việc kết hợp các đặc trưng này với phương pháp SVM làm cho phương pháp phân loại nhạc của họ đạt được độ chính xác 80%. McKay et al. đã cải tiến thuật toán phân loại nhạc bằng cách sử dụng đặc trưng lời bài hát [**5**]. Họ sử dụng nhiều đặc trưng kết hợp được rút trích từ nhiều nguồn audio, lời bài hát, biểu tượng, văn hoá liên quan đến thông tin âm nhạc. Các kết quả thực nghiệm cho thấy rằng đặc trưng được rút trích từ lời bài hát là kém hiệu quả hơn so với các đặc trưng khác. Chathuranga et al. đã xây dựng hệ thống phân loại nhạc theo thể loại dựa trên cách tiếp cận máy học [**6**], trong đó phương pháp SVM với hàm nhân đa thức được sử dụng. Hai tập đặc trưng được đề xuất biểu diễn cho miền tần số, miền thời gian, miền Cepstral và thông tin thay đổi về tần số trong các tín hiệu audio. Kết quả cho thấy rằng phương pháp này có độ chính xác phân loại lần lượt là 78% và 81% trên tập dữ liệu GTZAN và ISMIR2004 tương ứng. Rini Wongso, Diaz D. Santika nghiên cứu kết hợp phương pháp TCWT (Tree Complex Wavelet Transform) và SVM [**7**]. Nghiên cứu này tập trung vào việc phân

loại bốn thể loại nhạc: Pop, Classical, Jazz và Rock bằng cách sử dụng các chỉ số thống kê về trung bình, độ lệch chuẩn, phương sai, và entropy của các đặc trưng tín hiệu nhạc.

Trong bài báo này, chúng tôi giới thiệu một thuật toán phân loại nhạc tự động theo thể loại dựa trên phương pháp nhận dạng K-NN (K-Nearest Neighbor) và ba tập đặc trưng được rút trích từ tín hiệu nhạc: âm sắc (timbral texture), nhịp điệu (rhythmic content) và cao độ (pitch). Chúng tôi sử dụng phép biến đổi wavelet rời rạc (DWT) để phân tích tín hiệu audio dùng cho việc xác định các đặc trưng về nhịp điệu. Nghiên cứu của chúng tôi thực hiện minh hoạ trên bốn thể loại Classical, Rock, Jazz và Pop. Nó có thể áp dụng mở rộng đối với các thể loại nhạc khác hoặc xây dựng các hệ thống truy vấn thông tin nhạc dựa vào nội dung, kiểm tra việc sao chép bản quyền nhạc,.... Chúng tôi cũng trình bày việc lựa chọn các đặc trưng phù hợp vì chúng ảnh hưởng đáng kể đến độ chính xác phân loại.

II. CÁC CÔNG VIỆC NGHIÊN CỨU LIÊN QUAN

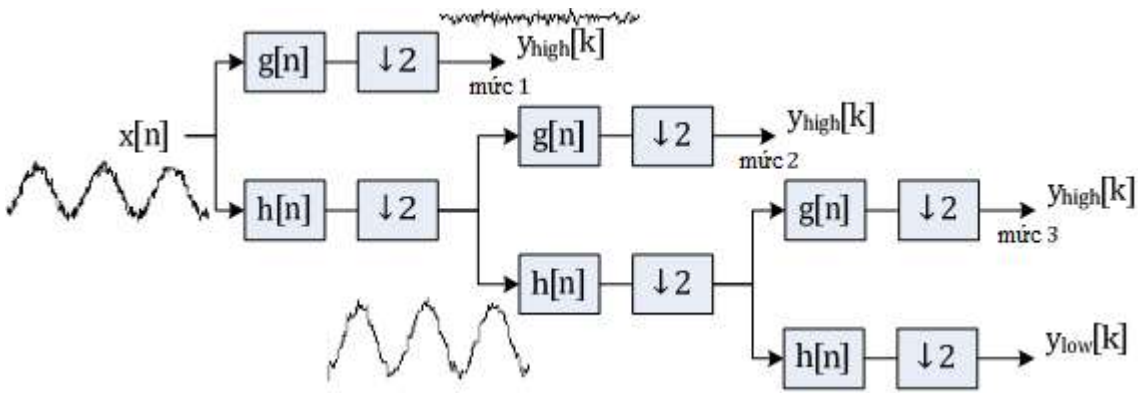
2.1. Phép biến đổi wavelet rời rạc

Phép biến đổi Fourier thường dùng cho phân tích các tín hiệu audio. Tuy nhiên, nó có hạn chế là ta không thể biết được tại một thời điểm sẽ xuất hiện những thành phần tần số nào. Để khắc phục nhược điểm này, các nhà khoa học sử dụng biến đổi STFT (Short time Fourier transform). Theo đó, tín hiệu được chia thành các khoảng nhỏ và được biến đổi Fourier trong từng khoảng đó. Phương pháp này có hạn chế là việc chọn độ rộng của các khoảng tín hiệu phân chia sao cho phù hợp vì nếu độ rộng này càng nhỏ thì độ phân giải thời gian càng tốt nhưng phân giải tần số càng kém và ngược lại. Để khắc phục cả 2 phương pháp trên, biến đổi wavelet ra đời. Biến đổi wavelet (WT) được thực hiện như sau: tín hiệu được nhân với hàm Wavelet (tương tự như nhân với hàm cửa sổ trong biến đổi STFT), sau đó thực hiện phân tích riêng rẽ cho các khoảng tín hiệu khác nhau trong miền thời gian tại các tần số khác nhau.

Phép biến đổi wavelet rời rạc (DWT) là một trường hợp đặc biệt của WT. Nó cung cấp một cách biểu diễn tín hiệu dưới dạng nén trong miền thời gian-tần số giúp cho việc tính toán một cách nhanh chóng và hiệu quả. DWT thực hiện phân tích đa phân giải một tín hiệu audio **x** thành 2 thành phần: thành phần tín hiệu thô A (coarse approximation) tương ứng với thành phần tần số thấp y_{low} và thành phần tín hiệu chi tiết D (detail) tương ứng với thành phần tần số cao y_{high} [**8**]. Sau đó, thành phần tín hiệu thô tiếp tục được phân tích tương tự. Như vậy, một tín hiệu có thể được biểu diễn dưới dạng tổng của thành phần tín hiệu thô và các thành phần tín hiệu chi tiết. Quá trình phân tích này được thực hiện bởi các bộ lọc băng tần cao và thấp đối với tín hiệu **x** như biểu diễn trong Hình 1 và được định nghĩa bởi công thức (1).

$$y_{high}[k] = \sum_n x[n]g[2k-n]$$
$$y_{low}[k] = \sum_n x[n]h[2k-n] \quad (1)$$

Trong đó: $y_{high}[k]$: thành phần tần số cao; $y_{low}[k]$: thành phần tần số thấp; x[n]: tín hiệu audio; g[n]: bộ lọc băng tần cao; h[n]: bộ lọc băng tần thấp.



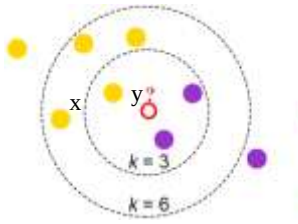
Hình 1. DWT mức 3 đối với tín hiệu x

Tín hiệu x[n] có thể được xác định bằng cách tổng hợp tất cả các hệ số của y_{high} và y_{low} bắt đầu từ mức phân tích cuối cùng. Trong bài báo này, chúng tôi sử dụng DWT trong việc phân tích tín hiệu audio theo miền tần số để rút trích các đặc trưng về nhịp điệu và sử dụng bộ lọc băng tần DAUB4 [**9**] được đề xuất bởi Daubechies.

2.2. Phương pháp phân loại KNN

Có nhiều phương pháp phân lớp như: KNN, Bayes, HMMs, Gaussian,... Trong nghiên cứu này, chúng tôi sử dụng phương pháp K-NN (K-Nearest Neighbor) [**10**] [**11**] vì nó đơn giản và được sử dụng phổ biến trong các bài toán phân lớp. Phương pháp này cho phép bổ sung mẫu huấn luyện mới vào bộ huấn luyện dễ dàng và hiệu quả khi tập huấn luyện lớn. Bên cạnh đó, bộ huấn luyện được huấn luyện từ chính các vectơ đặc trưng rút trích từ tín hiệu audio. Nó xử lý tốt với tập dữ liệu nhiễu do dựa trên khoảng cách giữa các vectơ đặc trưng để quyết định phân lớp, do đó nó phù hợp với hệ thống phân loại nhạc.

Phương pháp K-NN xem các mẫu (vectơ đặc trưng) như là các điểm biểu diễn trong không gian đặc trưng n chiều (Hình 2). Khoảng cách giữa mẫu cần phân loại x và k mẫu láng giềng y là d(x, y) được xác định dựa trên khoảng cách không gian. Thông thường, người ta dùng khoảng cách Euclide để xác định khoảng cách giữa các mẫu trong không gian đặc trưng được xác định bởi công thức (2).



Hình 2. Mô hình phân lớp K-NN

$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (2)$$

Xác suất mẫu x thuộc vào thể loại c_i được xác định bởi công thức (3):

$$p(c_i \mid x) = \frac{\sum_{y \in K, yc = c_i} w_y}{\sum_{y \in K} w_y} \qquad (3)$$

Trong đó: $w_y = (1/d(x,y))$; K là một tập hợp k mẫu láng giềng gần x nhất; yc là thể loại của y; c_i là thể loại thứ i.

Thuật toán K-NN:

1. Xác định giá trị tham số k (số láng giềng gần nhất).

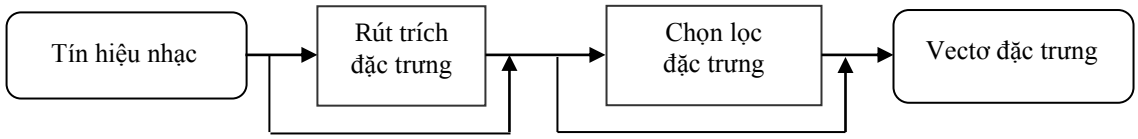
2. Tính khoảng cách giữa mẫu cần phân loại x với các mẫu trong tập huấn luyện (sử dụng công thức (2)).

3. Xác định k láng giềng gần nhất với x và các nhãn thể loại của chúng.

4. Xác định nhãn thể loại của x: x được gán nhãn thể loại c_i khi $p(c_i \mid x)$ là lớn nhất (sử dụng công thức (3)).

III. RÚT TRÍCH ĐẶC TRƯNG TÍN HIỆU AUDIO

Trên thực tế, tất cả các đặc trưng của tín hiệu audio khi đưa trực tiếp vào các mô hình phân loại sẽ làm giảm đi rõ rệt tốc độ huấn luyện và phân loại. Rút trích đặc trưng là một trong những kỹ thuật tiền xử lý tín hiệu nhạc được sử dụng phổ biến trong việc phân loại. Quá trình rút trích sẽ khử nhiễu tín hiệu và chỉ chọn các thông tin cần thiết cho việc phân loại nhạc. Ngoài ra, việc chọn lọc đặc trưng được dùng để tạo ra một tập con đặc trưng từ dữ liệu đầu vào nhằm làm tăng hiệu quả về mặt thời gian trong việc nhận dạng vì nó là tiến trình tự động hoá được dùng để giảm số chiều dữ liệu sao cho dữ liệu đầu vào được chuyển đổi sang dạng đơn giản và nhỏ hơn trước khi đưa vào mô hình phân loại.



Hình 3. Sơ đồ rút trích đặc trưng từ một tín hiệu nhạc

Nhiều nghiên cứu đã đề xuất các đặc trưng của tín hiệu audio để nhận dạng, phân loại trong các hệ thống nhận dạng, phân loại khác nhau. Mỗi nghiên cứu đều đưa ra một số các đặc trưng của tín hiệu audio và phương thức sử dụng để phân loại. Các đặc trưng của tín hiệu audio thường được chia làm hai nhóm chính: các đặc trưng trong miền thời gian – tần số và các đặc trưng cảm thụ âm thanh của con người (nhịp điệu, cao độ) [**6**]. Trong bài báo này, chúng tôi xây dựng hệ thống phân loại nhạc dựa trên ba tập đặc trưng như sau:

- Các đặc trưng về âm sắc (Timbral Texture Features).

- Các đặc trưng về nhịp điệu (Rhythmic Content Features).

- Các đặc trưng về cao độ (Pitch Content Features).

3.1. Đặc trưng về âm sắc

Tập đặc trưng về âm sắc được sử dụng để biểu diễn các đặc trưng của âm nhạc liên quan đến tiết tấu, âm sắc và nhạc cụ. Vectơ đặc trưng về âm sắc được sử dụng trong hệ thống phân loại của chúng tôi bao gồm 19 chiều với các đặc trưng: (Trung bình và độ lệch chuẩn của ***Spectral Centroid, Rolloff, Flux, ZeroCrossing***, ***LowEnergy***, và Trung bình và độ lệch chuẩn của 5 hệ số ***MFCC*** đầu tiên). Trung bình và độ lệch chuẩn của các đặc trưng này được xác định dựa trên STFT với các cửa sổ phân tích chia tín hiệu đầu vào có độ dài 1s thành các đoạn nhỏ khoảng 20ms. Sau đây là các đặc trưng được xác định trên mỗi cửa sổ phân tích:

a) ***Đặc trưng 1: Spectral Centroid***

Spectral Centroid là một độ đo liên quan hình dáng của phổ tần số. Nó xác định điểm cân bằng của phổ tần số. Giá trị Centroid cao tương ứng với phổ có độ sáng chói hơn và chứa nhiều tần số cao. Spectral Centroid được xác định bởi công thức **(**4):

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]} \tag{4}$$

Trong đó: $M_t[n]$ là biên độ của tần số thứ n trong phổ tần số tương ứng với cửa sổ t.

b) ***Đặc trưng 2: Rolloff***

Rolloff cũng là một độ đo liên quan hình dáng của phổ tần số. Điểm Rolloff của phổ tần số (R_t) được định nghĩa như tần số biên mà ở đó 85% phân bố năng lượng được tập trung trong phổ là dưới điểm này. Công thức (5) xác định R_t - điểm Rolloff của phổ tần số.

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 \sum_{n=1}^{N} M_t[n] \tag{5}$$

c) ***Đặc trưng 3: Flux***

Flux được xem là độ biến thiên phổ, cho biết sự thay đổi về biên độ tần số của phân phối quang phổ giữa hai cửa sổ phân tích liên tiếp. Nó được xác định là bình phương hiệu giữa các biên độ chuẩn của tần số trong phổ và được xác định bởi công thức (**6**).

$$F_t = \sum_{n=1}^{N} \left(N_t[n] - N_{t-1}[n]\right)^2 \qquad N_t[n] = \frac{M_t[n]}{\sqrt{\sum_{i=1}^{N} \left(M_t[i]^2\right)}} \tag{6}$$

Với $N_t[n]$ và $N_{t-1}[n]$ là biên độ chuẩn của tần số thứ n trong phổ tần số ở cửa sổ t và t-1 tương ứng.

d) ***Đặc trưng 4: Zero-crossings***

Zero Crossings cho biết mức độ ồn (noisiness) của âm thanh trong tín hiệu. Nó xuất hiện khi các mẫu kề nhau trong tín hiệu khác dấu. Nó được xác định bởi số lần tín hiệu audio vượt qua trục zero trên một đơn vị thời gian và được tính bởi công thức (**7**):

$$Z_t = \frac{1}{2} \sum_{n=1}^{N} |\, sign(x[n]) - sign(x[n-1]) \,| \quad ; \quad sign(x[n]) = \begin{cases} 1 & x[n] > 0 \\ 0 & x[n] \le 0 \end{cases} \tag{7}$$

x[n] là tín hiệu trong miền thời gian đối với cửa sổ t.

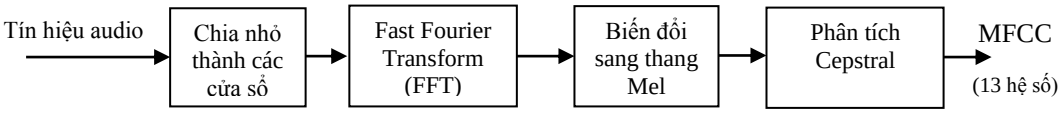
e) ***Đặc trưng 4: Low-Energy***

Khác với các đặc trưng trên, đặc trưng Low-Energy được xác định trên toàn bộ tín hiệu miền thời gian. Nó là tỉ lệ phần trăm của các cửa sổ phân tích có RMS (Root-Mean-Square) năng lượng thấp hơn RMS trung bình năng lượng của các tín hiệu trong các cửa sổ phân tích. Trong đó, RMS năng lượng của tín hiệu ở cửa sổ t được xác định bởi công thức (8):

$$RMS_t = \sqrt{\frac{\sum_{i=1}^{N} (M_t[i]^2)}{N}} \tag{8}$$

f) ***Đặc trưng 6: Các hệ số MFCC (Mel-Frequency Cepstral Coefficients)***

MFCC là một trong các tập đặc trưng được dùng phổ biến trong các hệ thống nhận dạng giọng nói, truy tìm thông tin nhạc,… Nó cung cấp cách biểu diễn nén tín hiệu audio dưới dạng phổ sao cho hầu hết năng lượng của tín hiệu được tập trung vào các hệ số đầu tiên. Hình 4 mô tả các bước thực hiện rút trích đặc trưng MFCC từ tín hiệu audio. Chi tiết về phương pháp rút trích đặc trưng MFCC mô tả trong [**12**].

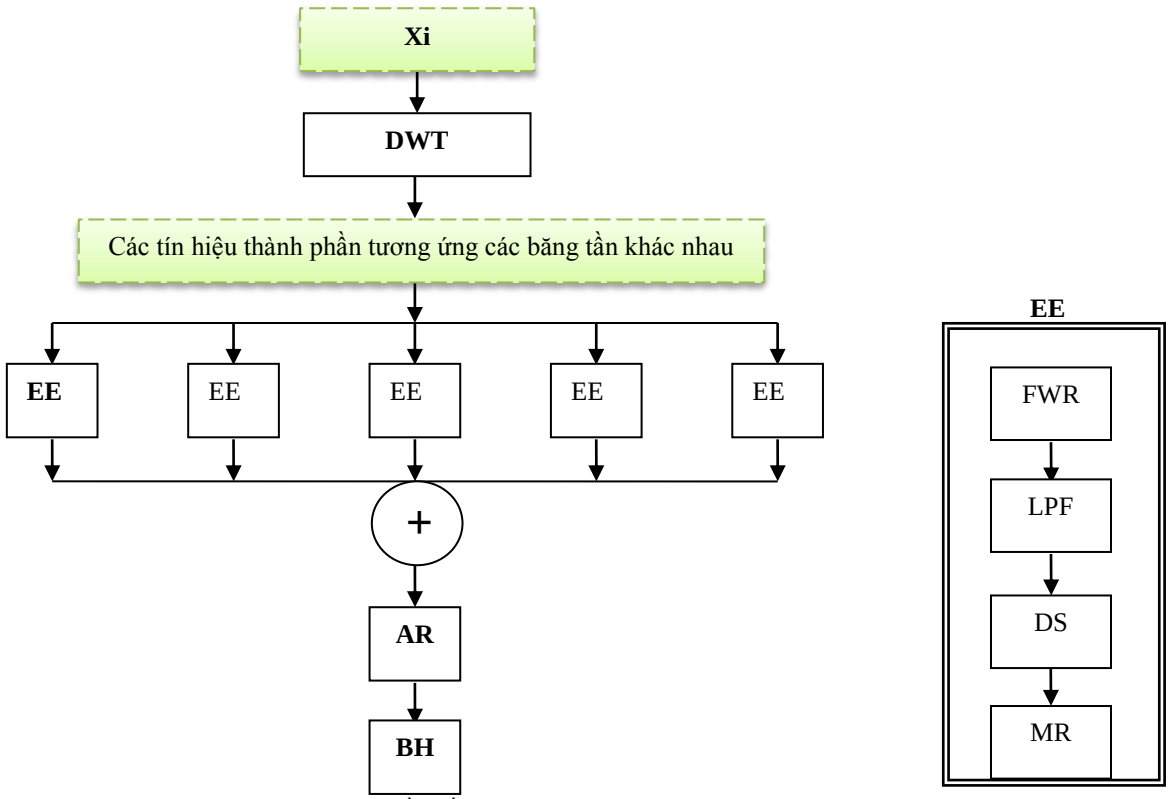


Hình 4. Sơ đồ rút trích đặc trưng MFCC

Kết quả thu được là một tập đặc trưng MFCC gồm 13 hệ số. Tuy nhiên, nhiều nghiên cứu [**13**] [**14**] cho thấy 5 hệ số MFCC đầu tiên cung cấp khá đầy đủ thông tin cho việc phân loại nhạc theo thể loại. Vì vậy, để giảm số chiều cho vectơ đặc trưng, chúng tôi chọn 5 hệ số MFCC đầu tiên cho hệ thống phân loại nhạc theo thể loại của chúng tôi.

3.2. Đặc trưng về nhịp điệu nhạc

Vectơ đặc trưng về nhịp điệu cung cấp rất nhiều thông tin có ích về đặc điểm của các thể loại nhạc. Hầu hết các hệ thống dò tìm nhịp điệu nhạc cung cấp các thuật toán xác định nhịp điệu của bản nhạc và cường độ của chúng. Bên cạnh đó, chúng còn cho biết mối liên hệ giữa các nhịp của bản nhạc. Trong bài báo này, chúng tôi sử dụng phương pháp xác định tập đặc trưng về nhịp điệu nhạc được đề xuất bởi George Tzanetakis et al. [**15**] trong việc phân loại nhạc theo thể loại. Phương pháp này dựa trên việc dò tìm các chu kỳ (đơn vị: bpm - số nhịp/phút) có biên độ lớn nhất của tín hiệu. Tín hiệu audio X được chia nhỏ thành các tín hiệu thành phần Xi bởi cửa sổ phân tích có kích thước 65536 mẫu với tần số lấy mẫu (sampling rate) là 22050 Hz tương ứng xấp xỉ 3s. Sau đó, thuật toán xác định nhịp điệu nhạc được áp dụng đối với mỗi Xi như biểu diễn trong Hình 5.



Hình 5. Sơ đồ khối xác định Histogram nhịp điệu nhạc

Trước tiên, tín hiệu Xi được phân tích thành các tín hiệu thành phần (y_{high} và y_{low}) tương ứng với các băng tần khác nhau dựa vào phép biến đổi Wavelet rời rạc (DWT). Tiếp theo, quá trình phân tích được thực hiện trên mỗi băng tần bằng cách áp dụng các bước trong Envelope Extraction (EE) gồm: Full wave rectification (FWR), low pass filtering (LPF), downsampling (DS) và Mean Removal (MR). Sau đó, chúng được tổng hợp và một hàm tự tương quan (AR) được xác định. Cuối cùng, ba đỉnh cao nhất (có biên độ lớn nhất) của hàm tự tương quan tương ứng với các chu kỳ khác nhau của tín hiệu audio được chọn để đưa vào biểu đồ nhịp điệu (Beat Histogram - BH). Trong đó, trục hoành của BH biểu diễn số nhịp/phút (đơn vị là bpm) và trục tung biểu diễn cường độ của nhịp (Beat strength). Sau đây là các bước phân tích EE trên mỗi băng tần để rút trích đặc trưng nhịp điệu:

1. Full Wave Rectification (FWR): $y[n] = abs(x[n])$ (9)
2. Low Pass Filter (LPF): Bộ lọc với $\alpha = 0.99$: $y[n] = (1- \alpha)\, x[n] - \alpha\, y[n - 1]$ (10)
3. Downsampling (DS) bởi 1 hệ số k (chọn k = 16 trong cài đặt hệ thống này): $y[n] = x[kn]$ (11)
4. Mean Removal (MR) / Normalization: $y[n] = x[n] - E[x[n]]$ (12)
5. Autocorrelation (AR): $$y[k] = \frac{1}{N}\sum_{n} x[n]x[n-k]$$ (13)

Quá trình xác định nhịp điệu nhạc trên tín hiệu audio được áp dụng lặp đi lặp lại trên các tín hiệu thành phần Xi và tích lũy vào trong biểu đồ nhịp điệu BH. Tập các đỉnh cao nhất của hàm tự tương quan tạo nên biểu đồ nhịp điệu nhạc được sử dụng làm cơ sở cho việc xác định các đặc trưng về nhịp điệu. Trong đó, các đỉnh cao nhất trong BH tương ứng với các chu kỳ khác nhau của tín hiệu audio là các nhịp chính của bản nhạc.

Xác định các đặc trưng về nhịp điệu:

Dựa vào BH, các đặc trưng về nhịp điệu: nhịp chính, nhịp phụ, cường độ,… được xác định để cung cấp các thông tin có ích cho việc phân loại nhạc theo thể loại. Gọi Đ1: đỉnh cao nhất và Đ2: đỉnh cao thứ nhì trong BH. Vectơ đặc trưng về nhịp điệu là một vectơ 6 chiều gồm các đặc trưng:

1. **A1, A2**: Đặc trưng này là độ đo sự khác nhau về nhịp so với các nhịp còn lại của tín hiệu. Nó được xác định bởi tỉ số giữa biên độ của lần lượt 2 đỉnh Đ1 và Đ2 với tổng biên độ của tất cả các đỉnh trong BH.

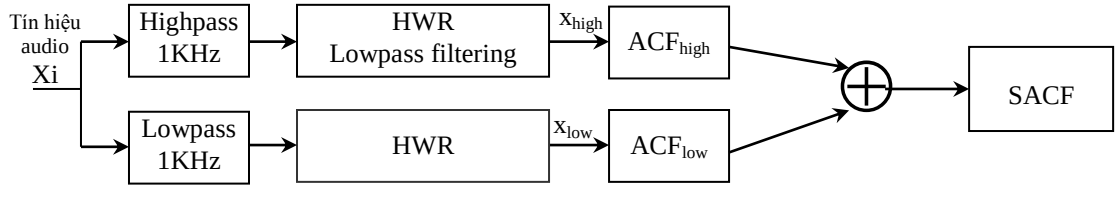
2. **RA**: là tỷ số giữa biên độ của đỉnh Đ2 với biên độ của đỉnh Đ1. Đặc trưng này biểu diễn mối quan hệ giữa nhịp chính và nhịp phụ đầu tiên.

3. **P1, P2**: Chu kỳ của đỉnh Đ1 và Đ2 được tính bằng số nhịp trong 1 phút (đơn vị tính: bpm).

4. **SUM**: Tổng biên độ của các đỉnh trong BH. Đặc trưng này cho biết độ mạnh của nhịp nhạc.

3.3. Đặc trưng về cao độ

Cao độ (pitch) là đại lượng tỉ lệ nghịch với tần số cơ bản của tín hiệu audio và liên quan đến đặc trưng về cảm thụ âm thanh của con người. Mặc dù việc phân loại nhạc theo thể loại không thể dựa hoàn toàn vào đặc trưng liên quan đến cao độ, nhưng nó cung cấp thông tin rất có ích cho việc phân loại. Chẳng hạn, nhạc Jazz hoặc Classical thường có mức độ thay đổi cao độ nhiều hơn so với nhạc Rock hoặc Pop. Ngược lại, biểu đồ về cao độ của nhạc Pop hoặc Rock sẽ có số đỉnh trội (có biên độ lớn) ít hơn nhưng các đỉnh này sẽ cao hơn so với biểu đồ về cao độ của nhạc Jazz hoặc Classical. Hiện nay, nhiều nghiên cứu đưa ra thuật toán và phương thức ước lượng cao độ. Các thuật toán ước lượng này hầu hết dựa vào phương pháp tự tương quan hoặc biến thể của nó. Trong nghiên cứu của chúng tôi, tập đặc trưng về cao độ được rút trích từ tín hiệu audio dựa trên thuật toán dò tìm cao độ đề xuất bởi Tolonen và Karjalainen [**16**]. Để xác định các đặc trưng về cao độ, tín hiệu audio X được chia nhỏ thành các tín hiệu thành phần Xi bởi cửa sổ phân tích có kích thước 512 mẫu với tần số lấy mẫu là 22050 Hz (xấp xỉ 23ms). Các bước rút trích đặc trưng về cao độ áp dụng đối với mỗi Xi được biểu diễn trong Hình 6.



Hình 6. Sơ đồ rút trích đặc trưng về cao độ

Trong thuật toán này, tín hiệu audio Xi được phân tích thành 2 băng tần trên và dưới 1000 Hz kèm theo biên độ được rút trích đối với mỗi băng tần. Việc xử lý các tín hiệu thành phần được thực hiện bởi HWR (Half wave rectification) và lọc Lowpass đối với băng tần cao. Sau đó, hàm tự tương quan tương ứng với 2 tín hiệu thành phần (x_{high} và x_{low}) được xác định tương tự với phương pháp dò tìm nhịp điệu. Kết quả hai hàm tự tương quan ACF_{high} và ACF_{low} được tạo ra. Hai hàm này được tổng hợp lại tạo thành hàm tự tương quan tổng hợp SACF (Summary autocorrelation function). Ba đỉnh cao nhất (có biên độ lớn nhất) của hàm SACF được chọn để đưa vào biểu đồ cao độ (Pitch Histogram - PH). Quá trình này được áp dụng lặp đi lặp lại trên các tín hiệu thành phần Xi và tích lũy vào trong biểu đồ cao độ PH. Tập hợp ba đỉnh cao nhất của mỗi SACF tạo thành PH đối với tín hiệu audio. Trong đó, các đỉnh cao nhất của mỗi SACF tương ứng với các cao độ chính đối với đoạn âm thanh đó. Từ biểu đồ cao độ PH, hai kiểu biểu đồ cao độ được tạo ra: UPH (Unfolded Pitch Histogram) chứa các thông tin về pitch range và FPH (Folded Pitch Histogram) chứa các thông tin về các pitch class hoặc hoà âm của bản nhạc. Chi tiết về phương pháp tạo UPH và FPH được mô tả trong [**16**].

Xác định các đặc trưng về cao độ:

Dựa vào UPH và FPH, các đặc trưng về cao độ được xác định để cung cấp các thông tin có ích cho việc phân loại nhạc theo thể loại. Gọi Đ1_U, Đ2_U: đỉnh cao nhất và nhì tương ứng trong UPH; Đ1_F, Đ2_F: đỉnh cao nhất và nhì tương ứng trong FPH. Vectơ đặc trưng về cao độ là một vectơ 5 chiều gồm các đặc trưng:

1. **FA0**: Biên độ của Đ1_F tương ứng với pitch class cao nhất của bản nhạc (tương ứng với âm chủ).
2. **UP0**: Chu kỳ của Đ1_U. Đặc trưng này tương ứng với vùng bát độ của pitch cao nhất trong bản nhạc.
3. **FP0**: Chu kỳ của Đ1_F. Đặc trưng này tương ứng với pitch class chính của bản nhạc.
4. **IPO1**: Quãng cao độ giữa 2 đỉnh Đ1_F và Đ2_F.
5. **FAVG**: Trung bình biên độ của các pitch trong FPH. Đặc trưng này là biên độ trung bình của các pitch.

IV. XÂY DỰNG HỆ THỐNG PHÂN LOẠI NHẠC THEO THỂ LOẠI

4.1. Tập dữ liệu dùng cho huấn luyện và kiểm tra

Trong nghiên cứu này, chúng tôi sử dụng nguồn dữ liệu cho huấn luyện và kiểm tra là bộ sưu tập nhạc GTZAN [**15**] gồm 10 thể loại nhạc vì nó được sử dụng phổ biến như một bộ sưu tập nhạc tham khảo chuẩn cho các nghiên cứu về phân loại nhạc theo thể loại. Dựa trên cấu trúc phân loại âm thanh của bộ sưu tập nhạc GTZAN, bốn thể loại nhạc Classical, Rock, Jazz, Pop được chọn ngẫu nhiên để minh hoạ cho hệ thống phân loại nhạc theo thể loại.

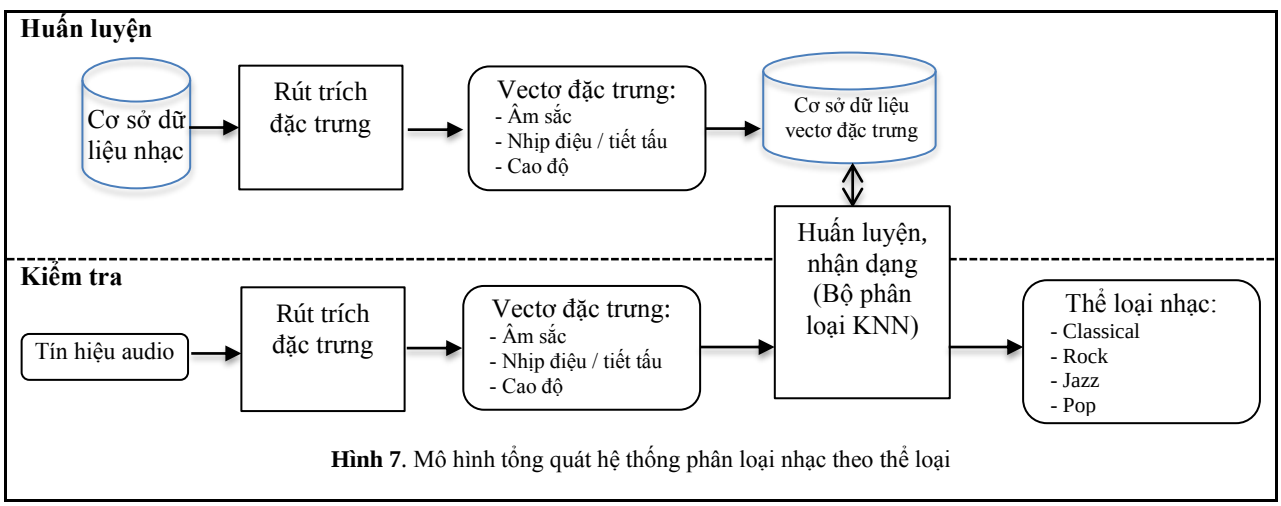
Trong phương pháp của chúng tôi, nguồn dữ liệu được chia thành 2 tập dữ liệu: huấn luyện và kiểm tra. Tập file audio huấn luyện được sử dụng để huấn luyện cho bộ phân loại KNN để đưa ra các quyết định cho hệ thống phân loại nhạc trong khi tập file audio kiểm tra sẽ được sử dụng để đánh giá hiệu quả của phương pháp đề xuất. Các file này được chọn ngẫu nhiên từ bộ sưu tập GTZAN. Mỗi file audio có độ dài 30s với tần số 22050 Hz Mono 16-bit ở định dạng .wav. Số file audio sử dụng trong tập huấn luyện và kiểm tra tương ứng từng thể loại được trình bày trong Bảng 1.

Bảng 1. Số lượng tập tin audio dùng cho huấn luyện và kiểm tra

STT	Tên thể loại	Số lượng tập tin audio	
		Huấn luyện	Kiểm tra
1	Classical	150	197
2	Rock	58	74
3	Jazz	81	104
4	Pop	90	93
CỘNG		**379**	**468**

4.2. Mô hình tổng quát hệ thống phân loại nhạc theo thể loại

Chúng tôi đề xuất hệ thống phân loại nhạc theo thể loại gồm 2 pha: rút trích đặc trưng và huấn luyện hoặc phân loại. Kết quả sau khi rút trích đặc trưng của tín hiệu audio là một tập gồm các đặc trưng về âm sắc, nhịp điệu, cao độ. Chi tiết việc rút trích đặc trưng được trình bày trong phần III. Chúng tôi sử dụng phương pháp biến đổi wavelet rời rạc (DWT) để rút trích đặc trưng về nhịp điệu. Phương pháp phân loại KNN được sử dụng để nhận dạng các thể loại nhạc (trình bày chi tiết trong phần 2.2). Quá trình huấn luyện bao gồm việc sử dụng các vectơ đặc trưng đã được gán nhãn thể loại để huấn luyện cho bộ phân loại KNN. Từ đó, bộ phân loại sẽ gán nhãn thể loại cho các vectơ đặc trưng mới một cách tự động. Mô hình tổng quát hệ thống phân loại nhạc theo thể loại được minh hoạ trong Hình 7.



Hình 7. Mô hình tổng quát hệ thống phân loại nhạc theo thể loại

Tập các đặc trưng sử dụng cho hệ thống phân loại nhạc trong nghiên cứu này bao gồm các đặc trưng sau đây:

- ***Các đặc trưng về âm sắc:*** Gồm 19 đặc trưng: Trung bình và phương sai của Centroid, Rolloff, Flux, ZeroCrossing (8), LowEnergy (1); Trung bình và phương sai của 5 hệ số MFC đầu tiên (10).
- ***Các đặc trưng về nhịp điệu / tiết tấu:*** Gồm 6 đặc trưng: A1, A2, RA, P1, P2, SUM được xác định từ biểu đồ nhịp điệu.
- ***Các đặc trưng về cao độ:*** Gồm 5 đặc trưng: FA0, UP0, FP0, IPO1, FAVG được xác định từ biểu đồ cao độ.

4.3. Ma trận đánh giá độ chính xác phân loại

Việc đánh giá phương pháp đề xuất được thực hiện bởi các file audio trong tập dữ liệu kiểm tra. Kết quả phân loại của hệ thống sẽ được trình bày trong ma trận đánh giá độ chính xác phân loại như Bảng 2.

Bảng 2. Ma trận đánh giá độ chính xác phân loại

	Thể loại	**Thể loại tiên đoán** (Kết quả tiên đoán từ hệ thống đề xuất)				**Tổng cộng**
		Classical	**Rock**	**Jazz**	**Pop**	
Thể loại thực tế	**Classical**	**C**	C1	C2	C3	197
	Rock	R1	**R**	R2	R3	74
	Jazz	J1	J2	**J**	J3	104
	Pop	P1	P2	P3	**P**	93

Trong ma trận này, các giá trị trong ma trận là số lượng tập tin audio trong tập dữ liệu kiểm tra. Các phần tử trong ma trận được giải thích như sau:

- C, R, J, P: số tiên đoán đúng đối với các file nhạc có nhãn thể loại Classical, Rock, Jazz, Pop tương ứng.
- Ci, Ri, Ji, Pi (i = 1,..,4): số tiên đoán sai đối với các file nhạc được gán nhãn thể loại Classical, Rock, Jazz, Pop tương ứng.

Như vậy, dòng tương ứng với thể loại thật sự của các file nhạc và cột tương ứng với thể loại tiên đoán của các file nhạc sau khi hệ thống đề xuất thực hiện phân loại. Số tập tin nhạc được gán nhãn thể loại đúng nằm trên đường chéo của ma trận (các giá trị in đậm: C, R, J, P). Để đánh giá hiệu quả của phương pháp đề xuất, độ chính xác phân loại A (Accuracy) được sử dụng và được xác định bởi công thức (14):

$$A(\%) = \frac{C + R + J + P}{(C + R + J + P) + \sum_{i=1}^{4} Ci + Ri + Ji + Pi} x100\% \quad (14)$$

V. KẾT QUẢ

Phương pháp của chúng tôi được thực hiện trong môi trường Visual C++ trên máy tính PC 2.27GHz CPU Core i5 với 3GB Ram để thực hiện cài đặt hệ thống phân loại nhạc theo thể loại. Các kết quả trong nghiên cứu này sẽ sử dụng hai tập dữ liệu huấn luyện (379 files audio) và kiểm tra (468 files audio) tương ứng với 4 thể loại Classical, Rock, Jazz, Pop. Các file này được chọn ngẫu nhiên từ bộ sưu tập GTZAN. Để đánh giá độ chính xác phân loại của phương pháp đề xuất, chúng tôi sử dụng ma trận đánh giá độ chính xác phân loại như Bảng 2.

Việc phân loại nhạc được thực hiện chủ yếu dựa vào 3 tập đặc trưng được rút trích từ tín hiệu audio như sau:

- Tập đặc trưng 1 (ĐT1): 19 đặc trưng về âm sắc.
- Tập đặc trưng 2 (ĐT2): 6 đặc trưng về nhịp điệu.
- Tập đặc trưng 3 (ĐT3): 5 đặc trưng về cao độ.

Chúng tôi kiểm tra trên hệ thống với việc rút trích đặc trưng dựa trên một trong các tập đặc trưng trên hoặc kết hợp chúng với nhau và sau đó tìm giá trị tham số k (số láng giềng gần nhất) sao cho hệ thống đạt hiệu quả về độ chính xác phân loại cao nhất. Sau đây là các kết quả thực nghiệm trên hệ thống trong một số trường hợp:

5.1. Phân loại dựa trên 1 hoặc 2 tập đặc trưng

Chúng tôi thực nghiệm trên hệ thống với việc phân loại dựa trên chỉ một hoặc 2 tập đặc trưng. Chúng tôi cũng kiểm tra trên 1 số giá trị tham số k (k = 3,…, 7). Trong đó, với giá trị k = 4, hệ thống cho kết quả phân loại tốt nhất. Vì vậy, chúng tôi chọn trình bày trong trường hợp này. Sau đây là kết quả đánh giá độ chính xác của việc phân loại.

Bảng 3. Độ chính xác phân loại dựa trên 1 tập đặc trưng với giá trị tham số k = 4

	Tập các đặc trưng		
	ĐT1 (âm sắc)	ĐT2 (nhịp điệu)	ĐT3 (cao độ)
A (%)	75,2%	63%	59,6%

Bảng 4. Độ chính xác phân loại dựa trên 2 tập đặc trưng với giá trị tham số k = 4

	Tập các đặc trưng		
	ĐT1 và ĐT2	ĐT1 và ĐT3	ĐT2 và ĐT3
A (%)	79,7%	79,5%	63,5%

Từ kết quả trình bày trong Bảng 3 và Bảng 4, chúng tôi nhận xét: Nếu chúng tôi chỉ sử dụng 1 hoặc 2 tập đặc trưng thì việc phân loại nhạc theo thể loại từ tín hiệu audio đạt độ chính xác thấp. Kết quả này khó có thể chấp nhận được. Vì vậy, chúng tôi tiếp tục kiểm tra trên hệ thống mà trong đó sử dụng cả 3 tập đặc trưng 1, 2 và 3 trong việc phân loại nhạc (tạo nên một vectơ đặc trưng 30 chiều) với mong muốn làm tăng độ chính xác của việc phân loại.

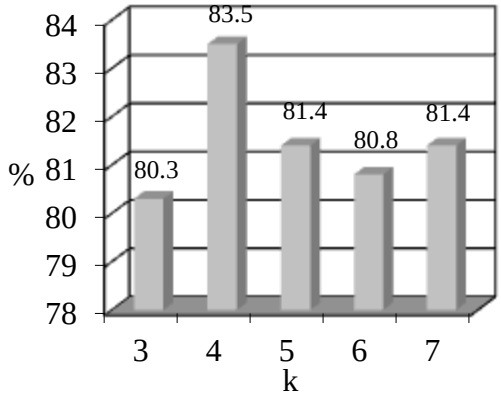
5.2. Phân loại dựa trên cả 3 tập đặc trưng

Chúng tôi thực nghiệm trên hệ thống với việc phân loại dựa trên cả 3 tập đặc trưng: âm sắc, nhịp điệu và cao độ. Từ Bảng 5, chúng tôi nhận thấy kết quả phân loại của hệ thống dựa trên cả 3 tập đặc trưng với giá trị tham số k = 4 như sau: tổng số file audio kiểm tra: 468, số trường hợp hệ thống tiên đoán đúng: 391 (A=83,5%), số trường hợp tiên đoán sai: 77 (16,5%).

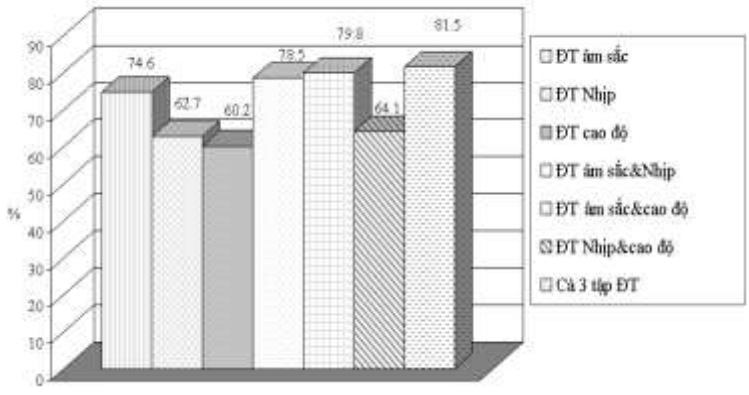
Bảng 5. Ma trận đánh giá độ chính xác phân loại dùng cả 3 tập đặc trưng với k = 4

	Classical	Rock	Jazz	Pop
Classical	**195**	2	0	0
Rock	6	**58**	9	1
Jazz	23	10	**59**	12
Pop	1	4	9	**79**

Chúng tôi cũng kiểm tra trên 1 số giá trị k khác nhau (k=3,5,6,7). Hình 8 biểu diễn độ chính xác phân loại sử dụng cả 3 tập đặc trưng với các giá trị k khác nhau, trong đó hệ thống cho kết quả phân loại tốt nhất với k = 4. Nguyên nhân là với k = 4 hệ thống phân loại nhạc theo thể loại đề xuất đạt độ chính xác là: 83,5%. Với các giá trị khác của k, kết quả độ chính xác phân loại thấp hơn. Chẳng hạn: k = 3 độ chính xác chỉ đạt 80,3%.



Hình 8. Đồ thị biểu diễn độ chính xác phân loại sử dụng kết hợp cả 3 tập đặc trưng



Hình 9. Đồ thị biểu diễn độ chính xác trung bình phân loại nhạc dựa vào các tập đặc trưng

Từ các kết quả thực nghiệm trên tập dữ liệu kiểm tra biểu diễn trong Hình 9 cho thấy: nếu hệ thống chỉ sử dụng một trong 3 tập đặc trưng về âm sắc, nhịp điệu hoặc cao độ, thì việc phân loại nhạc theo thể loại từ tín hiệu audio được thực hiện nhanh hơn (thời gian thực hiện trung bình là 30,7 giây) do số chiều của vectơ đặc trưng nhỏ hơn, nhưng độ chính xác của việc phân loại sẽ thấp hơn (đạt khoảng 60,2% - 74,6%) so với trường hợp phân loại nhạc dựa trên cả 3 tập đặc trưng này. Tương tự, nếu hệ thống phân loại chỉ dựa trên việc kết hợp 2 tập đặc trưng: âm sắc và nhịp điệu; âm sắc và cao độ; nhịp điệu và cao độ thì cũng cho kết quả thời gian thực hiện nhanh hơn (trung bình là 63,1giây), trong khi kết quả độ chính xác của việc phân loại thấp hơn (đạt 64,1% - 79,8%) so với kết quả phân loại trong trường hợp hệ thống sử dụng kết hợp cả 3 tập đặc trưng với độ chính xác phân loại trung bình 81,5% và thời gian thực hiện trung bình là 97,4 giây. Vì vậy, việc sử dụng kết hợp cả 3 tập đặc trưng âm sắc, nhịp điệu, cao độ là rất cần thiết đối với hệ thống phân loại nhạc theo thể loại vì nó cho kết quả phân loại khá chính xác.

Như vậy, phương pháp đề xuất của chúng tôi là kết hợp cả 3 tập đặc trưng âm sắc, nhịp điệu và cao độ trong việc phân loại nhạc theo thể loại bởi vì hệ thống đưa ra kết quả phân loại với độ chính xác cao (trung bình 81,5%). Kết quả của phương pháp đề xuất này là tốt hơn so với một số phương pháp đã nghiên cứu trước đây với độ chính xác trung bình dưới 80%. Chẳng hạn, trong nghiên cứu của Marco Grimaldi et al. [**17**] sử dụng 182 file nhạc với 7 thể loại khác nhau để kiểm tra hệ thống cho kết quả độ chính xác phân loại là 52,75% trong trường hợp dùng phép biến đổi wavelet rời rạc và kỹ thuật phân loại KNN. Phương pháp đề xuất cũng có kết quả tốt hơn so với phương pháp của Panagakis et al. [**18**] vì độ chính xác phân loại đạt 78,2%, 77,9% và 75,01% khi rút trích các tập đặc trưng khác nhau từ cùng bộ sưu tập nhạc GTZAN với bộ phân loại SVM. Mặt khác, kết quả nghiên cứu của chúng tôi cũng có độ chính xác phân loại cao hơn so với phương pháp đề xuất bởi Tao et al. [**19**] vì độ chính xác phân loại đạt 78,6% thực hiện trên cùng tập dữ liệu GTZAN và máy học SVM. Bên cạnh đó, Chathuranga et al. [**6**] đã đề xuất phương pháp phân loại nhạc theo thể loại với tiếp cận máy học SVM. Kết quả phân loại đạt độ chính xác thấp hơn phương pháp đề xuất vì nó chỉ đạt 78% khi thực hiện trên tập dữ liệu GTZAN.

VI. KẾT LUẬN

Một phương pháp phân loại nhạc theo thể loại nhanh và chính xác là rất cần thiết đối với các hệ thống quản lý một số lượng lớn nhạc số. Tuy nhiên, đây là một công việc không đơn giản vì các thể loại nhạc vẫn còn là một khái niệm mở, tùy thuộc vào ý kiến chủ quan của con người. Trong nghiên cứu thực nghiệm này, chúng tôi đề xuất sử dụng các tập đặc trưng được rút trích bởi các công cụ STFT, DWT và bộ phân loại KNN. DWT là một kỹ thuật phân tích tín hiệu, cung cấp một cách biểu diễn tín hiệu trong miền thời gian và tần số dưới dạng nén làm cho việc tính toán nhanh và hiệu quả. Nghiên cứu này tập trung vào việc phân loại 4 thể loại nhạc: Classical, Rock, Jazz và Pop bằng cách sử

dụng kết hợp cả 3 tập đặc trưng về âm sắc, nhịp điệu và cao độ tạo nên một vectơ đặc trưng 30 chiều. Tập dữ liệu được sử dụng trong nghiên cứu này lấy từ bộ sưu tập nhạc GTZAN. Dựa trên các kết quả thực nghiệm, phương pháp đề xuất của chúng tôi đạt độ chính xác trung bình 81,5%. Kết quả nghiên cứu này cho độ chính xác phân loại cao hơn một số nghiên cứu trước đó mà chỉ đạt độ chính xác dưới 80%.

Việc phân loại nhạc theo thể loại được thực hiện một cách tự động bằng máy tính và cho kết quả khá chính xác là hoàn toàn có thể. Nghiên cứu này cung cấp cơ sở khoa học cho phát triển các hệ thống: truy vấn thông tin nhạc dựa vào nội dung, phát hiện sao chép bản quyền nhạc, tìm các bản nhạc có các đặc trưng gần giống với các đặc trưng mà người sử dụng mong muốn, phân tích nhạc và lời bài hát, phân loại bản nhạc theo ca sĩ - nhạc sĩ, chú thích tự động các tập tin nhạc với những mô tả,... Phương pháp đề xuất có thể áp dụng mở rộng trên các thể loại nhạc khác: Opera, Rap, Blues, Country, Hip Hop,… Ngoài ra, nó có thể áp dụng cho việc phân loại nhạc truyền thống của Việt Nam như: dân ca Bắc bộ, dân ca Nam bộ, Chèo, Bội, Cải lương. Hệ thống đề xuất cũng có thể áp dụng với các bộ phân loại kết hợp khác như: SVM, Gaussian, mạng Neural,… Chúng tôi dự định thực nghiệm hệ thống đề xuất trên một tập dữ liệu lớn (Big Data); nghiên cứu và sử dụng các đặc trưng mới để có thể trích xuất các thông tin âm nhạc có ý nghĩa từ các tín hiệu âm thanh. Đó chính là những hướng nghiên cứu của chúng tôi trong thời gian sắp tới.

VII. TÀI LIỆU THAM KHẢO

[1] Anan, Yoko, Hatano, Kohei, Bannai, Hideo, and Takeda, Masayuki, "Music Genre Classification Using Similarity Functions", Proceedings of the 12th International Society for Music Information Retrieval Conference (Miami (Florida), USA, pp. 693-698, 2011.

[2] Costa, Y. M. G., Oliveira, L. S., Koericb, A. L., and Gouyon, F., "Music genre recognition using spectrograms", 18th International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 1-4, 2011.

[3] Jang, Dalwon, Jin, Minho, and Yoo, Chang Dong, "Music genre classification using novel features and a weighted voting method", Proceedings of International Conference on Multimedia and Expo, Hannover, Germany, pp. 1377-1380, 2008.

[4] Matsui, Tomoko, Goto, Masataka, Vert, Jean-Philippe, and Uchiyama, Yuji, "Gradient-based musical feature extraction based on scale-invariant feature transform". *EUSIPCO*, IEEE, pp. 724-728, 2011.

[5] Mckay, Cory, Burgoyne, John Ashley, Hockman, Jason, Smith, Jordan B.L., Vigliensoni, Gabriel, and Fujinaga, Ichiro, "Evaluating the Genre Classification Performance of Lyrical Features Relative to Audio, Symbolic and Cultural Features", Proceedings of the 11th International Society for Music Information Retrieval Conference (Utrecht, The Netherlands August 9-13 2010), pp. 213-218, 2010.

[6] Chathuranga, Dhanith and Jayaratne, Lakshman, "Automatic Music Genre Classification of Audio Signals with Machine Learning Approaches", *GSTF Journal on Computing (JoC)*, Vol. 3, No. 2, pp. 1-12, 2013.

[7] Rini Wongso, Diaz D. Santika, "Automatic music genre classification using dual tree complex wavelet transform and support vector machine", Journal of Theoretical and Applied Information Technology, Vol. 63, No. 1, pp. 61-68, 2014.

[8] Mallat, S. G., "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 11, pp. 674-693, 1989.

[9] Daubechies, Ingrid, "Orthonormal bases of compactly supported wavelets", Journal of Communications on Pure and Applied Mathematics, Vol. 41, No. 7, pp. 909-996, 1988.

[10] Altman, N. S., "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression", the American Statistician, Vol. 46, No. 3, pp. 175-185, 1992.

[11] Theodoridis, Sergios and Koutroumbas, Konstantinos, "Pattern Recognition", Third Edition, Academic Press, Inc., Orlando, FL, USA, 2006.

[12] Logan, Beth. "Mel Frequency Cepstral Coefficients for Music Modeling", Proceedings of the 1st International Conference on Music Information Retrieval (Plymouth (Massachusetts), USA October 23, 2000.

[13] Li, Tao and Tzanetakis, G. , "Factors in automatic musical genre classification of audio signals", Applications of Signal Processing to Audio and Acoustics, IEEE Workshop, pp. 143-146, 2003.

[14] Cataltepe, Zehra, Yaslan, Yusuf, and Sonmez, Abdullah, "Music Genre Classification Using MIDI and Audio Features", EURASIP Journal on Advances in Signal Processing, Vol. 1, pp. 1-8, 2007.

[15] Tzanetakis, George, Essl, Georg, and Cook, Perry, "Automatic Musical Genre Classification of Audio Signals", Proceedings of the 2nd Annual International Symposium on Music Information Retrieval (Bloomington (Indiana), USA, pp. 205-210, 2001.

[16] Tolonen, Tero and Karjalainen, Matti. "A computationally efficient multipitch analysis model", IEEE Trans. Speech and Audio Processing, Vol. 8, No. 6, pp. 708-716, 2000.

[17] Grimaldi, Marco, Kokaram, Anil, and Cunningham, Pádraig, "Classifying music by genre using a discrete wavelet transform and a round-robin ensemble", Computer Science Dept, Trinity College Dublin, Ireland, 2003.

[18] Panagakis, Ioannis, Benetos, Emmanouil, and Kotropoulos, Constantine, "Music Genre Classification: A Multilinear Approach", Proceedings of the 9th International Conference on Music Information Retrieval, pp. 583-588, 2008.

[19] Ran Tao, Zhenyang Li, Ye Ji, "Music genre classification using temporal information and support vector machine", ASCI Conference, Vol. 77, 2010.

MUSIC CLASSIFICATION BY GENRE USING DISCRETE WAVELET TRANSFORM

Phan Anh Cang, Phan Thượng Cang

ABSTRACT*— As the demand for multimedia grows, the development of musical genre classification systems including information about musical genre is of increasing concern. However, due to the collection of audio files from various sources, the musical genre classification based on information recorded on file may encounter limitations. To automate searching, organizing and classifying a huge number of audio files based on their genre is a challenging task. In this paper, we present an approach to identifying musical genres based on their content including three feature sets for representing timbral texture, rhythmic content and pitch content. We apply the discrete wavelet transform for decomposing audio signals to determine their rhythmical features. Our method is applied to identify four musical genres including Classical, Rock, Jazz and Pop. It can be extended to applications related to the different musical genres or the music information retrieval systems, the music copyright matching systems,… We present experimental results that show that our approach is effective in identifying the musical genre of the audio file with acceptable level of confidence.*