So Sánh Hai Phương Pháp Dùng Đường Bao Phổ Và Ảnh Phổ Trong Tìm Kiếm Âm Thanh

Ninh Khánh Duy, Nguyễn Bình Thiên

Trường Đại học Bách Khoa, Đại học Đà Nẵng
nkduy@dut.udn.vn, tucothien@gmail.com

Tóm tắt. Sự phát triển nhanh chóng của các cơ sở dữ liệu âm thanh lớn đòi hỏi nghiên cứu các phương pháp để tìm kiếm nhanh chóng và chính xác thông tin về bài hát khi người dùng cần đến. Trong trường hợp người dùng không nhớ được tên bài hát, hệ thống cần truy xuất được bài hát gốc chỉ cần dựa trên tín hiệu âm thanh của một đoạn thu âm nằm ở vị trí bất kỳ của bài hát. Bài báo này trình bày hai phương pháp tìm kiếm âm thanh dùng đặc trưng đường bao phổ và ảnh phổ. Thử nghiệm bước đầu trên cơ sở dữ liệu nhỏ gồm 100 bài hát có chất lượng âm thanh cao và đoạn tín hiệu đưa vào tìm kiếm được thu âm bằng điện thoại di động dài từ 5 đến 10 giây cho thấy cả hai phương pháp đều đạt độ chính xác từ 96% đến 100%. Tuy nhiên phương pháp ảnh phổ cho thời gian tìm kiếm trung bình nhanh hơn khoảng 80 lần so với phương pháp đường bao phổ và khả thi hơn để áp dụng với các cơ sở dữ liệu âm thanh có kích thước lớn.

Từ khoá: Tìm Kiếm Âm Thanh, Đường Bao Phổ, Ảnh Phổ, Phân Cụm K-means, Lọc Cực Đại.

1 Đặt vấn đề

Tìm kiếm âm thanh là phương pháp tìm kiếm dựa trên một đoạn tín hiệu âm thanh chứa nội dung của bài nhạc đó. Tìm kiếm âm thanh rất hữu ích trong nhiều trường hợp. Trường hợp người dùng nghe một bài nhạc và muốn biết thông tin của bài nhạc thì họ có thể thu âm lại tín hiệu âm thanh và tìm kiếm bài nhạc đó dựa trên đoạn tín hiệu đã thu âm. Những nhà sản xuất âm nhạc có thể sử dụng phương pháp này để tìm ra những vi phạm về bản quyền âm nhạc, ví dụ như Youtube [1]. Để thực hiện tìm kiếm âm thanh thì dấu vân tay âm thanh (audio fingerprint) được sử dụng rộng rãi [2].

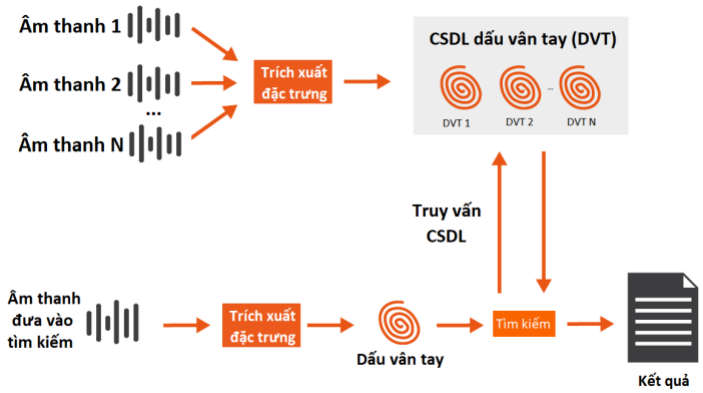
Dấu vân tay âm thanh là một dạng chữ ký số dựa trên tín hiệu âm thanh và được sử dụng làm đặc trưng để nhận dạng một mẫu âm thanh hoặc định vị trí các âm thanh tương tự trong cơ sở dữ liệu (CSDL). Tín hiệu âm thanh có nhiều đặc trưng khác nhau như cao độ, cường độ, trường độ, nhịp điệu và phổ. Bài báo này trình bày hai phương pháp trích đặc trưng dấu vân tay sử dụng đường bao phổ và ảnh phổ do phổ là đặc trưng được sử dụng phổ biến trong các ứng dụng nhận dạng âm thanh và tiếng nói [3]. Vì vậy chúng tôi cho rằng các đặc trưng liên quan đến phổ tín hiệu có thể hữu ích nhất cho bài toán tìm kiếm âm thanh. Bên cạnh đặc trưng phổ, cao độ (pitch) của tín

hiệu cũng được dùng để tìm kiếm âm thanh, nhưng phù hợp hơn khi đữ liệu tìm kiếm là âm thanh của nhạc cụ hoặc giai điệu của bài hát [4].

Để giảm kích thước dữ liệu dấu vân tay cần lưu trữ và tăng tốc độ tìm kiếm, chúng tôi đã áp dụng các giải pháp phù hợp. Trong phương pháp sử dụng đặc trưng đường bao phổ, chúng tôi sử dụng thuật toán phân cụm k-means [5] để lượng tử hóa các vector đặc trưng [3], đồng thời sử dụng khoảng cách Euclidean để đánh giá mức độ giống nhau của hai vector. Trong phương pháp sử dụng đặc trưng ảnh phổ, chúng tôi sử dụng kỹ thuật lọc cực đại (maximum filter) của xử lý ảnh để lọc ra những thông tin quan trọng trong ảnh phổ của tín hiệu [6].

Bài báo gồm các phần như sau. Phần 2 mô tả mô hình chung của hệ thống. Phần 3 trình bày chi tiết hai phương pháp đã khảo sát. Phần 4 trình bày thực nghiệm và kết quả. Phần 5 đưa ra kết luận cho bài báo.

2 Mô hình chung của hệ thống



Hình. 1. Mô hình hệ thống tìm kiếm âm thanh.

Hình 1 trình bày mô hình chung của hệ thống tìm kiếm âm thanh. Đầu vào của hệ thống tìm kiếm là một đoạn tín hiệu âm thanh của bài nhạc cần tìm. Hệ thống sẽ trích xuất đặc trưng dấu vân tay từ đoạn tín hiệu đó và tiến hành so khớp với các đặc trưng dấu vân tay của các bài hát trong CSDL. Cuối cùng hệ thống sẽ đưa ra kết quả là thông tin của bài nhạc (ví dụ như tên bài hát, nhạc sỹ sáng tác,…) có đặc trưng dấu vân tay khớp nhất với đoạn nhạc đưa vào tìm kiếm.

3 Các phương pháp đề xuất

3.1 Phương pháp sử dụng đặc trưng đường bao phổ

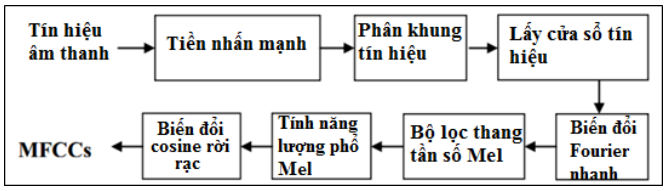
Hệ thống tìm kiếm âm thanh là hệ thống phân biệt các bài nhạc dựa trên âm sắc, giai điệu của chúng, tương tự như hệ thống thính giác của con người. Không phải mọi

thông tin của tín hiệu đều quan trọng trong việc tìm kiếm. Thực tế cho thấy tai người nhận biết âm thanh với tần số thấp tốt hơn tần số cao và nhận biết một nhóm các âm thanh hơn là từng âm thanh với tần số riêng lẻ. Hơn nữa, tai người cũng không thể phân biệt được sự khác nhau giữa hai âm có tần số quá gần nhau. Vì vậy, chúng tôi sử dụng phương pháp trích đặc trưng của bài nhạc dựa trên đường bao phổ được mô tả bởi các hệ số phổ MFCC (Mel-frequency cepstral coefficient). Đây là một phương pháp trích đặc trưng của âm thanh dựa trên sự mô phỏng lại quá trình cảm nhận âm thanh của tai người và được sử dụng phổ biến trong nhận dạng tiếng nói [3].

Để giảm dung lượng lưu trữ của CSDL và tăng tốc độ tìm kiếm dấu vân tay, chúng tôi đã sử dụng thuật toán phân cụm k-means [5] để lượng tử hóa các vector đặc trưng MFCC. Đoạn nhạc sau khi được trích xuất dấu vân tay (là các vector MFCC được lượng tử hóa vào một trong k cụm) sẽ được so khớp với các dấu vân tay của các bài nhạc trong CSDL. Trong thuật toán tìm kiếm, khoảng cách Euclidean được sử dụng để tính khoảng cách giữa các vector đặc trưng dấu vân tay.

Tính đặc trưng đường bao phổ

Đường bao phổ là đường cong đi qua các đỉnh của phổ biên độ của tín hiệu. Kỹ thuật trích đặc trưng MFCC dựa trên việc thực hiện biến đổi để chuyển dữ liệu âm thanh đầu vào về các hệ số trong thang đo Mel để mô phỏng đặc trưng cảm nhận âm thanh phi tuyến của tai người (Hình 2).



Hình 2. Sơ đồ khối thuật toán trích đặc trưng MFCC [3].

Đầu tiên tín hiệu âm thanh được xử lý tiền nhấn mạnh rồi chia thành các khung với độ dài bằng nhau khoảng 20-40 ms. Các khung nằm chồng lên nhau một đoạn nhỏ hơn độ dài khung. Khoảng cách giữa hai điểm đầu tiên của hai khung liên tiếp gọi là độ dịch khung. Mục đích của việc phân khung là để giảm sự thay đổi của tín hiệu trong một lần trích đặc trưng MFCC. Trên mỗi khung tín hiệu ta tính phổ biên độ sử dụng biến đổi Fourier nhanh. Bằng cách nhân dải các bộ lọc Mel vào phổ biên độ ta tính được một dãy các hệ số. Lấy logarit cơ số 10 của các hệ số này và thực hiện biến đổi cosine rời rạc ta thu được các hệ số MFCC. Đặc trưng MFCC thu được là một chuỗi các vector, mỗi vector đại diện cho một khung tín hiệu. Dựa trên kết quả thực nghiệm ở Phần 4, chúng tôi sử dụng thuật toán trích đặc trưng MFCC với 13 hệ số, độ dài khung là 30 ms và độ dịch khung là 20 ms.

Lượng tử hóa vector đặc trưng

Thuật toán phân cụm k-means được sử dụng để lượng tử hóa vector đặc trưng, nhờ đó giảm dung lượng của CSDL dấu vân tay và tăng tốc độ của thuật toán tìm kiếm.

- Đầu vào: Chuỗi vector MFCC của một đoạn nhạc (hay bài hát) và số cụm k.
- Đầu ra: Chuỗi k vector trung bình của k cụm và một vector chứa nhãn của các vector MFCC đầu vào (một nhãn là một chỉ số cụm nằm từ 1 đến k). Vector nhãn này chính là dấu vân tay của đoạn nhạc (hay bài hát). Hệ thống của chúng tôi sử dụng hệ số k = 64 dựa trên kết quả thực nghiệm ở Phần 4.

Thuật toán tìm kiếm dấu vân tay

Ý tưởng

Đoạn nhạc đưa vào tìm kiếm cũng được phân thành các khung, tính đặc trưng MFCC và lượng tử hóa tương tự như những bài hát trong CSDL. Việc tìm kiếm được thực hiện bằng cách so sánh vector đặc trưng của đoạn nhạc này với vector đặc trưng của từng đoạn nhạc có cùng độ dài trong bài hát CSDL. Mỗi lần so sánh dịch chuyển đoạn nhạc đi một khung so với bài hát cho đến khi kết thúc bài hát. Khoảng cách bé nhất tìm được chính là khoảng cách từ đoạn nhạc đến bài hát đó. Thực hiện tương tự với tất cả các bài hát còn lại trong CSDL ta thu được một mảng các khoảng cách. Bài hát có khoảng cách bé nhất với đoạn nhạc chính là bài hát cần tìm.

Các chi tiết

Với mỗi bài hát gốc, chuỗi vector MFCC tương ứng sau khi lượng tử hóa được lưu vào CSDL dấu vân tay dưới dạng (việc này thực hiện offline một lần duy nhất):

- k vector trung bình: $m_o[1], \dots , \ m_o[\mathrm{k}]$.
- Một vector nhãn v_o có M chiều với M là số khung tín hiệu của bài hát gốc: $v_o = \{v_o[1], \dots, v_o[M]\}$ với $v_o[i] \in \{1,2,..,\mathrm{k}\}$.

Các vector MFCC của đoạn nhạc đưa vào tìm kiếm cũng được lượng tử hóa dưới dạng (thực hiện online mỗi lần tìm kiếm):

- k vector trung bình: $\mathrm{m_f}[1] \dots \mathrm{m_f}[\mathrm{k}]$.
- Một vector nhãn v_f có N chiều với N là số khung tín hiệu của đoạn nhạc: $v_f = \{v_f[1] \dots v_f[N]\}$ với $v_f[i] \in \{1,2,..,\mathrm{k}\}$.

  Bước 1. Tính bảng khoảng cách giữa các vector trung bình của đoạn nhạc và các vector trung bình của bài hát trong CSDL:

$$d\big(m_o[i], m_f[j]\big) = EuclideanDistance\big(m_o[i], m_f[j]\big).$$

Bước 2. Tính khoảng cách từ đoạn nhạc đến bài hát dựa vào bảng khoảng cách các vector trung bình.

Khoảng cách từ đoạn nhạc đến bài hát tại khung thứ i:

$$d[i] = \sum_{l=1}^{N} d_e\big[v_o[l+i-1]\big]\big[v_f[l]\big]$$

với *l* là chỉ số khung của đoạn nhạc, *i* là độ dịch khung của đoạn nhạc so với bài hát ($i = \overline{1, M-N+1}$).

Khoảng cách từ đoạn nhạc đến bài hát được xác định là khoảng cách bé nhất trong các khoảng cách $d[i]$: $d = mi\,n(d[i])$ $(i = \overline{1, M-N+1})$.

Bước 3. Tính khoảng cách từ đoạn nhạc đến tất cả các bài hát còn lại trong CSDL.

Bước 4. Ra quyết định bài hát cần tìm là bài hát có khoảng cách *d* bé nhất trong số các bài hát gốc.

3.2 Phương pháp sử dụng đặc trưng ảnh phổ

Ảnh phổ (spectrogram) là biểu diễn trực quan biên độ của các tần số biến thiên theo thời gian của một tín hiệu cụ thể. Ảnh phổ là một đồ thị 3 chiều: chiều thứ nhất là thời gian, chiều thứ hai là tần số, chiều thứ ba được miêu tả bởi độ đậm nhạt của màu sắc và đại diện cho độ lớn của một tần số tại một thời điểm nào đó (màu càng đậm thì biên độ của tần số càng lớn). Tín hiệu âm thanh được phân khung, lấy cửa sổ và tính biến đổi Fourier nhanh (tương tự như Hình 2) để nhận được phổ biên độ.

Trích xuất đặc trưng dấu vân tay từ ảnh phổ

Trong thuật toán tìm kiếm âm thanh sử dụng ảnh phổ, đặc trưng dấu vân tay của một đoạn nhạc là các điểm trên ảnh phổ có biên độ tần số đạt cực đại cục bộ [7]. Để tìm được các điểm này ta làm giãn nở các điểm ảnh trên ảnh phổ ban đầu (Hình 3a) bằng cách áp dụng lọc cực đại [6], các điểm lân cận nhỏ hơn kích thước của cửa sổ lọc được hợp nhất (Hình 3b). Các điểm trên ảnh phổ gốc có biên độ bằng với biên độ của điểm đó trên ảnh phổ được giãn nở sẽ là điểm cực đại cục bộ (Hình 3c). Số lượng điểm cực đại cục bộ (Hình 3d) phụ thuộc vào độ rộng của cửa sổ lọc: cửa sổ lọc càng lớn thì số điểm đặc trưng thu được càng ít và ngược lại.

Thuật toán tìm kiếm dấu vân tay

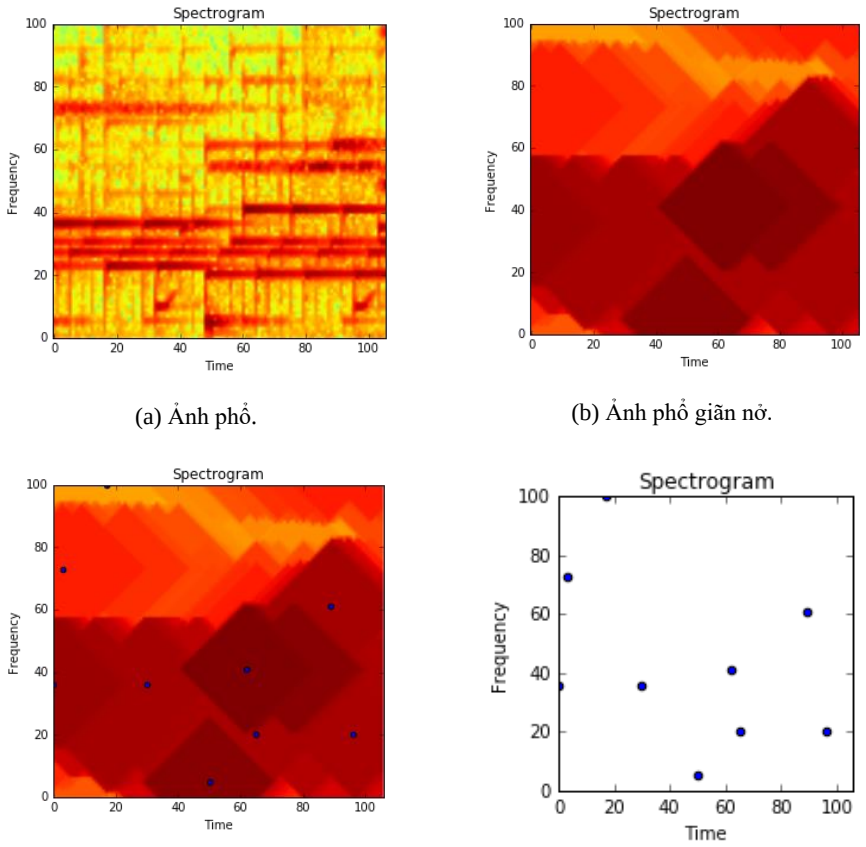
- *Bước 1:* Trích xuất đặc trưng ảnh phổ của các bài hát gốc và lưu vào CSDL dấu vân tay cùng với thông tin của bài hát đó (thực hiện offline một lần duy nhất).
- *Bước 2:* Trích xuất đặc trưng ảnh phổ của đoạn nhạc thu âm cần tìm kiếm (thực hiện online mỗi lần tìm kiếm).
- *Bước 3*: Tìm các điểm trong đặc trưng của các bài hát gốc có cùng tần số với các điểm trong đặc trưng của đoạn nhạc thu âm ta thu được k cặp điểm. Đối với mỗi cặp điểm như vậy, ta tính được:

$$d_t = t_s - t_f$$

với d_t là độ lệch thời gian giữa hai điểm cực đại cục bộ của bài hát gốc (tại thời điểm t_s) và cực đại cục bộ của đoạn nhạc thu âm (tại thời điểm t_f) có cùng tần số.

Với mỗi bài hát, ta sẽ tìm được một mảng $d_t[i]$ các độ lệch thời gian của các cặp điểm với $i \in \{1,2,..,\mathrm{k}\}$.

- *Bước 4*: Với các d_t từ mỗi mảng $d_t[i]$, ta tính số lượng cặp điểm có cùng giá trị d_t trong mảng $d_t[i]$, từ đó tìm được N_{max} là số lượng cặp điểm lớn nhất có cùng d_t hay là số lượng điểm trùng nhau lớn nhất của bài hát thu âm tại vị trí d_t trong bài hát gốc. Bài hát nào có N_{max} lớn nhất sẽ là bài hát cần tìm.



(a) Ảnh phổ.

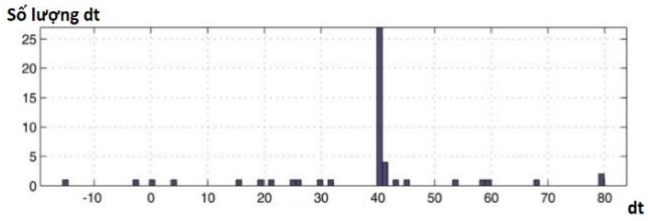
(b) Ảnh phổ giãn nở.

(c) Các điểm cực đại cục bộ.

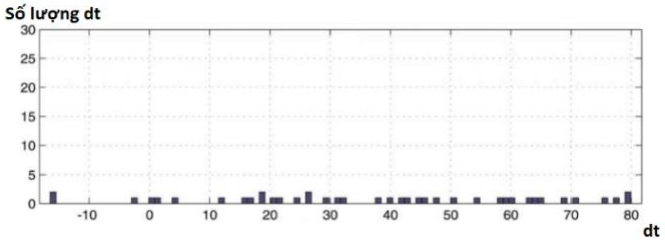
(d) Các điểm đặc trưng của đoạn nhạc.

Hình 3. Trích xuất đặc trưng dấu vân tay từ ảnh phổ.

Hình 4 và Hình 5 lần lượt thể hiện sự biến thiên của số lượng điểm trùng nhau theo các vị trí khác nhau trong bài hát gốc trong hai trường hợp: đoạn nhạc có nội dung nằm trong và không nằm trong với bài hát gốc. Có thể thấy trường hợp trước có giá trị N_{max}=26 (tại vị trí d_t=40 và giá trị đỉnh nổi trội rất rõ), cao hơn nhiều so với trường hợp sau có giá trị N_{max}=3 (giá trị đỉnh không nổi trội rõ ràng).



Hình 4. Số điểm trùng nhau trong trường hợp đoạn nhạc có nội dung nằm trong bài hát.



Hình 5. Số điểm trùng nhau trong trường hợp đoạn nhạc có nội dung không nằm trong bài hát.

4 Thực nghiệm và kết quả

4.1 Chuẩn bị dữ liệu

CSDL bài hát gốc: gồm 100 bài hát tiếng Việt được chọn ngẫu nhiên từ Internet.

Dữ liệu đưa vào tìm kiếm: 10 bài hát thu âm bằng điện thoại di động trong khi mở loa phát 10 bài hát gốc trong CSDL. Tần số lấy mẫu của các file âm thanh là 16000 Hz. Các bài hát được thu âm bằng điện thoại di động trong môi trường yên tĩnh. 10 bài hát thu âm được chọn lựa bao gồm nhiều thể loại nhạc và tiết tấu nhanh chậm khác nhau.

Cách thức thực nghiệm: Mỗi bài hát thu âm được cắt ra lần lượt thành 10 đoạn dài 5 giây và 10 đoạn dài 10 giây ở những vị trí khác nhau. Tổng cộng ta có 100 đoạn 5 giây và 100 đoạn 10 giây. Thực hiện tìm kiếm trên những đoạn này. Tổng số lần tìm kiếm là 200 lần.

4.2 Tìm các tham số tối ưu của hai phương pháp

Phương pháp sử dụng đặc trưng đường bao phổ

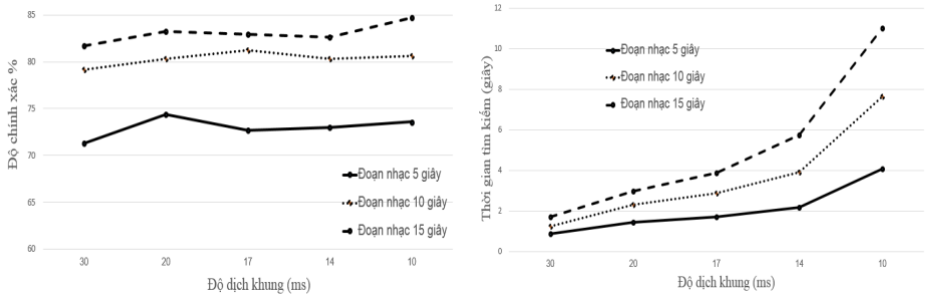
Khảo sát độ dịch khung trong thuật toán trích đặc trưng MFCC

Độ dịch khung trong thuật toán trích đặc trưng MFCC là một tham số quan trọng ảnh hưởng đến độ chính xác và thời gian tìm kiếm của hệ thống. Giảm độ dịch khung có thể làm giảm sự mất mát thông tin của tín hiệu, từ đó tăng độ chính xác của thuật toán, tuy nhiên lại làm cho số lượng vector đặc trưng và thời gian tìm kiếm tăng lên. Chúng tôi tiến hành khảo sát độ chính xác và thời gian tìm kiếm của thuật toán khi thay đổi độ dịch khung nhằm chọn ra được độ dịch khung tối ưu.

Phương pháp thực nghiệm: Tìm một đoạn nhạc được thu âm từ một bài hát trong CSDL gồm 10 bài hát. Độ dài của đoạn nhạc lần lượt là 5, 10, 15 giây, được thu tại các vị trí khác nhau từ đầu đến cuối bài hát. Mỗi khung có độ dài 30 ms. Thay đổi độ dịch khung lần lượt là 30 ms, 20 ms, 17 ms, 14 ms, và 10 ms

Kết quả của thực nghiệm được thể hiện trên Hình 6. Nếu độ dịch khung là 30 ms thì thời gian tìm kiếm chỉ dưới 2 giây, tuy nhiên độ chính xác lại thấp hơn nhiều so với các độ dịch khung khác. Khi thay đổi độ dịch khung từ 20 ms đến 10 ms thì độ

chính xác không thay đổi nhiều, tuy nhiên thời gian tìm kiếm lại tăng khá nhanh. Từ đó, chúng tôi chọn được độ dịch khung tối ưu là 20 ms.



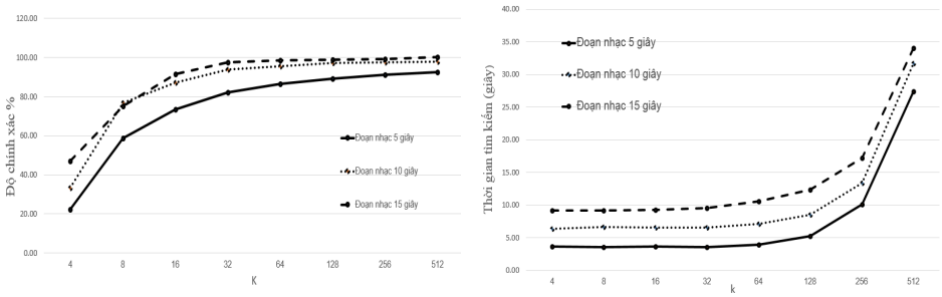
Hình 6. Khảo sát độ chính xác và thời gian tìm kiếm theo độ dịch khung.

Khảo sát hệ số k trong thuật toán phân cụm k-means

Hệ số k đóng vai trò quan trọng trong quá trình lượng tử hóa vector. Nếu k càng lớn thì số lượng vector trung bình sau khi lượng tử hóa càng tăng, điều này làm chậm việc tính toán khoảng cách giữa hai vector. Tuy nhiên, nếu hệ số k quá nhỏ thì việc lượng tử hóa vector làm mất quá nhiều thông tin dữ liệu ban đầu, dẫn đến giảm độ chính xác tìm kiếm. Vì vậy chúng tôi thực hiện một thực nghiệm khác để khảo sát độ chính xác và thời gian tìm kiếm của hệ thống khi thay đổi k nhằm tìm giá trị k tối ưu.

Phương pháp thực nghiệm: Tìm một đoạn nhạc được thu âm từ một bài hát trong CSDL gồm 10 bài hát. Độ dài của đoạn nhạc lần lượt là 5, 10, 15 giây, được thu tại các vị trí khác nhau từ đầu đến cuối bài hát. Mỗi khung tín hiệu có độ dài 30 ms. Thay đổi hệ số k lần lượt bằng 4, 8, 16, 32, 64, 128, 256, 512.

Kết quả của thực nghiệm được thể hiện trên Hình 7. Có thể thấy khi tăng hệ số k thì độ chính xác của thuật toán tăng, đồng thời cũng làm cho thời gian tìm kiếm tăng. Khi tăng k từ 4 đến 64 thì độ chính xác tăng nhanh nhưng thời gian tìm kiếm tăng không đáng kể. Khi thay đổi k từ 64 đến 512 thì độ chính xác gần như đã bão hòa, tuy nhiên thời gian tìm kiếm lại tăng khá nhanh. Từ đó, chúng tôi chọn được hệ số k tối ưu là 64.



Hình 7. Khảo sát độ chính xác và thời gian tìm kiếm theo hệ số k.

Phương pháp sử dụng đặc trưng ảnh phổ

Khảo sát hệ số giãn nở của bộ lọc cực đại

Hệ số giãn nở là số điểm lân cận xung quanh điểm cần giãn nở trong ảnh phổ. Hệ số giãn nở lớn nghĩa là nhiều điểm xung quanh điểm giãn nở được hợp nhất làm giảm số lượng cục bộ cực đại, dẫn đến tăng tốc độ tìm kiếm nhưng có thể làm giảm độ chính xác tìm kiếm dấu vân tay.

Phương pháp thực nghiệm: Tìm 100 đoạn nhạc được thu âm từ 10 bài hát trong CSDL. Độ dài của đoạn nhạc là 5 giây, được thu tại các vị trí khác nhau từ đầu đến cuối bài hát.

Kết quả thực nghiệm được thể hiện trên Bảng 1. Từ đó chúng tôi chọn được hệ số giãn nở tối ưu là 20.

Bảng 1. Thống kê độ chính xác và thời gian tìm kiếm theo hệ số giãn nở.

Hệ số giãn nở	Độ chính xác (%)	Thời gian tìm kiếm (giây)
30	92	1,09
20	100	0,57
10	100	0,55

4.3 Kết quả thực nghiệm

Chúng tôi đã thử nghiệm hai phương pháp trên với CSDL gồm 100 bài hát tiếng Việt được chọn ngẫu nhiên từ Internet và dữ liệu tìm kiếm là đoạn nhạc được thu âm bằng điện thoại di động từ 10 bài hát có giai điệu khác nhau trong 100 bài hát đó. Tần số lấy mẫu của các bài hát gốc trong CSDL là 44100 Hz và của đoạn thu âm bằng điện thoại di động để đưa vào tìm kiếm là 16000 Hz.

Đối với phương pháp đường bao phổ, các tham số được sử dụng trong thực nghiệm là: tần số lấy mẫu của bài hát gốc được giảm xuống 16000 Hz, trích đặc trưng MFCC với 13 hệ số, độ dài khung tín hiệu là 30 ms, độ dịch khung là 20 ms, số điểm FFT là 1200, số cụm của thuật toán k-means là 64.

Đối với phương pháp ảnh phổ, các tham số được sử dụng trong thực nghiệm là: tần số lấy mẫu của các đoạn thu âm được tăng lên lên 44100 Hz, độ dài cửa sổ tín hiệu để tính FFT là 4096 mẫu tương đương 92,87 ms, độ xếp chồng của hai cửa sổ tín hiệu liên tiếp là 50%, hệ số giãn nở của bộ lọc cực đại là 20.

Bảng 2 và 3 lần lượt thể hiện độ chính xác và thời gian tìm kiếm trung bình của hai phương pháp. Có thể thấy cả hai phương pháp đều đạt độ chính xác từ 96% trở lên. Tuy nhiên phương pháp ảnh phổ cho thời gian tìm kiếm trung bình nhanh hơn khoảng 80 lần so với phương pháp đường bao phổ (mặc dù phải xử lý tín hiệu với tần số lấy mẫu cao hơn nhiều) và khả thi để áp dụng với các CSDL âm thanh có kích thước lớn.

Bảng 2. Thống kê độ chính xác tìm kiếm của hệ thống.

Độ dài đoạn nhạc tìm kiếm	Phương pháp đường bao phổ	Phương pháp ảnh phổ
5 giây	99%	96%
10 giây	100%	100%

Bảng 3. Thống kê thời gian tìm kiếm trung bình của hệ thống.

Độ dài đoạn nhạc tìm kiếm	Phương pháp đường bao phổ	Phương pháp ảnh phổ
5 giây	85 giây	1,12 giây
10 giây	176 giây	2,18 giây

5 Kết luận và hướng phát triển

Chúng tôi đã đánh giá so sánh hai phương pháp tìm kiếm âm thanh sử dụng đặc trưng đường bao phổ và ảnh phổ. Kết quả thực nghiệm cho thấy với CSDL gồm 100 bài hát, độ chính xác của cả hai phương pháp đều từ 96% trở lên đối với đoạn nhạc cần tìm kiếm dài 5 giây và 100% đối với đoạn nhạc cần tìm kiếm dài 10 giây. Thời gian thực hiện tìm kiếm đối với phương pháp sử dụng đường bao phổ lần lượt là 85 giây và 176 giây (tương ứng với độ dài đoạn nhạc đưa vào tìm kiếm là 5 giây và 10 giây), cao hơn nhiều so với thời gian tìm kiếm của phương pháp sử dụng ảnh phổ lần lượt là 1,12 giây và 2,18 giây. Điều đó cho thấy đặc trưng ảnh phổ khả thi hơn trong việc cài đặt các hệ thống tìm kiếm âm thanh với CSDL lớn.

Trong tương lai chúng tôi sẽ tiếp tục nghiên cứu và cải tiến để giảm thời gian tìm kiếm và tăng độ chính xác của các thuật toán. Việc xây dựng một hệ thống có khả năng xử lý CSDL âm thanh lớn hơn và cho kết quả chính xác ngay cả với những đoạn nhạc thu âm trong môi trường có nhiều nhiễu là các hướng nghiên cứu thách thức cần triển khai tiếp theo.

Tài liệu tham khảo

1. M. Maksimović, P. Aichroth and L. Cuccovillo: Detection and Localization of Partial Audio Matches, 2018 International Conference on Content-Based Multimedia Indexing (CBMI), La Rochelle, 2018.
2. C. Ouali, P. Dumouchel and V. Gupta: A robust audio fingerprinting method for content-based copy detection, 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI), Klagenfurt, 2014.
3. L. R. Rabiner: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, Vol. 77, No. 2, 1989.
4. T. T. H. Phung, X. N. Doan, T. N. Phung: So sánh hai phương pháp trích chọn đặc trưng âm thanh: đường bao phổ (MFCC) và cao độ Pitch trong việc tìm kiếm âm nhạc theo nội dung, Tạp chí Khoa học và Công nghệ - Đại học Thái Nguyên, Tập 112, số 12/2, trang 33 – 38, 2013.
5. https://en.wikipedia.org/wiki/K-means_clustering.
6. P. W. Verbeek, H. A. Vrooman, L. J. Van Vliet: Low-level image processing by max-min filters, Signal Processing, Vol. 15, No. 3, 1988.
7. A. Li-Chun Wang: An Industrial-Strength Audio Search Algorithm, Proceedings of the 4 th International Conference on Music Information Retrieval, 2003.

Comparison On Spectral Envelope And Spectrogram Based Methods In Audio Retrieval

Abstract. The rapid development of large audio databases requires advanced methods to quickly and accurately find song information when needed. In case the user cannot remember the song, the system needs to retrieve the original song just based on the audio signal of a recording segment located in any position of the song. This paper presents two methods of audio retrieval using spectral envelope and spectrogram based features. Initial testing on a small database of 100 high-quality songs and input signal segments for searching recorded by mobile phone lasting from 5 to 10 seconds showed that both methods achieved the accuracy of from 96% to 100%. The spectrogram based method, however, provides an 80 times faster average searching time than the spectral envelope based method, and thus is more feasible for use with large audio databases.

Keywords: Audio Retrieval, Spectral Envelope, Spectrogram, K-Means Clustering, Maximum Filter.