A Query-by-example Framework to Retrieve Music Documents by Singer

Wei-Ho Tsai and Hsin-Min Wang

Institute of Information Science, Academia Sinica, Taiwan, Republic of China

{wesley,whm}@iis.sinica.edu.tw

Abstract

In this paper, we present a framework for music document retrieval that allows users to retrieve a specified singer’s music recordings from an unlabeled database by submitting a fragment of music as a query to the system. Such a framework can be of great use for those wishing to know more about a particular singer but having no idea about what the name of this singer is. In order for the searched documents to be relevant to the query, methods are proposed to compare the similarity between document and query based on automatic extraction of singer voice characteristics from a music recording.

1. Introduction

Music is more than just a collection of notes, beats, chords or rhythmic words. It is an information rich medium in itself, capable of conveying various aspects ranging from concrete and abstract, such as compositions, themes, artistic styles, moods, sentiments, or even ethnic culture, which cannot be exhaustively described and attached textually to every music recording. As music materials encoded in digital formats rapidly grow in size and number, finding the desired information from innumerable available options can be, however, no easy task. This problem has consequently motivated research towards content-based retrieval of music information. Many techniques have been developed for automatically extracting information from music residing in audio data, such as melodies [1], instruments [2], genres [3], singers [4-6], and so on. This study resembles this research target and further extends its scope to retrieving music documents based on performing artist or singer.

Imagine the following scenario.¹ When you walk on the street or drive your car and suddenly you hear a song that fascinates you pretty much. You may have heard this song many times, but for some reasons you still have no idea about what the title of this song is and who performed this song. Out of curiosity and amusement, you record a piece of this music via, for example, a Digital Voice Recorder or a Personal Digital Assistant, and hope to get more information about this song via accessing some websites after back home. Under this circumstance, the most efficient way to satisfy your desire is through the so-called query-by-example, which uses an excerpt from a musical performance as a query to retrieve the relevant information or objects. Since the singer(s) involving in this song might be unfamiliar or totally fresh to you, you probably want to query, “Find me all the songs performed by the singer of this attached recording.” This study aims to make the above scenario a reality. We present a singer-based music document retrieval framework, which extracts the singer voice characteristics from a music recording, compares the characteristic similarity between an exemplar recording and each of the music documents to be searched, and determines the relevance of each music document to the submitted query.

In addition to the above scenario, retrieving music documents by singer can be trivially applied to many problems. For instance, many rock music artists such as Phil Collins, Sting, Ozzy Osbourne, or even Michael Jackson, are known to have joined a band prior to becoming famous for solo work. Since the vast majority of rock music data is only labeled by band name, the proposed method may be useful for those wishing to locate the full works of those artists. Moreover, retrieving music documents by singer can easily distinguish between an original song and a cover-band, compensating for the shortage of the title-based or melody-based music retrieval.

2. Task definition and method overview

Let \( X_1, X_2, \ldots, X_N \) denote \( N \) music documents in a database to be searched, \( Y \) be an exemplar music query submitted by a user, and \( S(\cdot) \) represent the underlying singer of a music recording. The singers associated to both the documents and the query are assumed unknown. Our aim is to choose, as many as possible, the music documents satisfying \( S(X_i) = S(Y) \). Following the general information retrieval framework, retrieving music documents by singer can be converted into a task of ranking the possibilities of music documents performed by the singer appearing in a given music query in descending order. The documents ranked higher will be regarded as relevant to the query. In order for this framework to be effective, methods are proposed to compare the similarity between document and query in terms of singer voice characteristics; i.e., the similarity measurement is built upon the basis of extraction and modeling of singer’s solo vocal signal from accompanied voices. Let \( \mathcal{L}(X, Y) \) denote the similarity measurement taken over the music document \( X \) and the query \( Y \), in which a large value of \( \mathcal{L}(X, Y) \) indicates high similarity, and \( \mathcal{R} \{ \mathcal{L}(X_i, Y) \} \) denote the rank of \( \mathcal{L}(X, Y) \) among \( \mathcal{L}(X_1, Y), \mathcal{L}(X_2, Y), \ldots, \mathcal{L}(X_N, Y) \) in

¹ Part of the idea is similar to the one in [7]
descending order. A music document \( \mathbf{X} \) is hypothesized as relevant to the query \( \mathbf{Y} \) if
\[
R(\mathcal{L}(\mathbf{X}, \mathbf{Y})) < \theta,
\]
where \( \theta \) controls the number of documents that will be presented to the user.

It is known that the above rank-based approach may be improved by applying relevance feedback (RF) techniques [8] that refine the query via exploiting the information from the documents deemed relevant to the query. However, during the initial design stage, such a strategy is not adopted explicitly in this study, since there are many potential methods [9,10] for realizing the RF, and most of them could be trivially carried out here without specific tailoring. Instead methods [9,10] for realizing the RF, and most of them could be trivially carried out here without specific tailoring. Instead, the scope of this study is restricted to dealing with the music recordings (both document and query) that refine the query via exploiting the information from the documents deemed relevant to the query. However, during training and testing, the proposed framework remains applicable if the recordings of a music document
\[
\text{music document } \mathbf{X} = \{x_1, x_2, ..., x_T\}
\]
are then stochastically represented as a parametric model
\[
\pi(x | \lambda_i, \lambda_b) = \frac{1}{W_i} \sum_{k=1}^{K} \log p(x_{i,k} | \lambda_i) - \sum_{v=1}^{W_i} \log p(x_{i,v} | \lambda_b) \geq \eta.
\]

where \( W_i \) and \( x_i \) represent, respectively, the length and starting frame of the \( k \)-th homogeneous segment, and \( \eta \) is the decision threshold.

After locating the singing portions, the next step is to probe the solo voice signal underlying the singing portions and thereby extract the singer voice characteristics. The basic strategy applied here is to exploit the non-vocal music segments as a prior knowledge of the background accompaniments to construct a reliable model for the solo voice signal. Suppose that an accompanied voice signal representation \( \mathbf{V} = \{v_1, v_2, ..., v_T\} \) is a mixture of a solo voice \( \mathbf{S} = \{s_1, s_2, ..., s_T\} \) and a background music \( \mathbf{B} = \{b_1, b_2, ..., b_T\} \). Both \( \mathbf{S} \) and \( \mathbf{B} \) are unobservable, but \( \mathbf{B} \)'s stochastic characteristics may be estimated from the non-vocal segments, based on the assumption that substantial similarities exist between the accompaniments of singing regions and instrumental-only regions. With available information from \( \mathbf{V} \) and \( \mathbf{B} \), it is sufficient to build a stochastic model \( \lambda_s \) for the solo voice \( \mathbf{S} \). To this end, we further assume that \( \mathbf{S} \) and \( \mathbf{B} \) are, respectively, drawn randomly and independently according to GMMs
\[
\lambda_s = \{w_{si}, \mu_{si}, \Sigma_{si} \mid 1 \leq i \leq l \}, \lambda_b = \{w_{bi}, \mu_{bi}, \Sigma_{bi} \mid 1 \leq j \leq J \},
\]
where \( w_{si, j} \) and \( w_{bi, j} \) are mixture weights, \( \mu_{si} \) and \( \mu_{bi} \) mean vectors, and \( \Sigma_{si} \) and \( \Sigma_{bi} \) covariance matrices. If the signal \( \mathbf{V} \) is a generator function \( v_t = f(s_t, b_t) \), the probability of \( \mathbf{V} \), given \( \lambda_s, \lambda_b \), is defined as
\[
p(\mathbf{V} | \lambda_s, \lambda_b) = \prod_{i=1}^{l} \sum_{j=1}^{J} w_{si, j} n(x_i | \mu_{si, j}, \Sigma_{si, j}),
\]
where
\[
p(v_i | \mu_{si, j}, \Sigma_{si, j}) = \int \mathcal{N}(x_i | \mu_{si, j}, \Sigma_{si, j}) \mathcal{N}(b_i | \mu_{bi, j}, \Sigma_{bi, j}) \, \text{d}b_i.
\]

To build \( \lambda_s \), a maximum-likelihood estimation is made as
\[
\lambda_s^* = \arg \max \ p(\mathbf{V} | \lambda_s, \lambda_b).
\]

Using the EM algorithm, a new model \( \lambda_s^* \) is iteratively estimated by maximizing the auxiliary function
\[
Q(\lambda_s, \lambda_b) = \sum_{i=1}^{l} \sum_{j=1}^{J} p(i, j | \lambda_s, \lambda_b) \log \ p(i, j | v_i, \lambda_s, \lambda_b),
\]
where
\[
p(i, j | v_i, \lambda_s, \lambda_b) = \hat{w}_{i,j} n(v_i | \hat{\mu}_{i,j}, \hat{\Sigma}_{i,j}),
\]
and
\[
p(i, j | v_i, \lambda_s, \lambda_b) = \sum_{v_w} w_{i,v_w} n(v_i | \mu_{iw, j}, \Sigma_{iw, j}).
\]

Letting \( \lambda_s^* = 0 \) with respect to each parameter to be re-estimated, we have
\[
\hat{w}_{i,j} = \frac{1}{T} \sum_{t=1}^{T} p(i, j | v_t, \lambda_s^*, \lambda_b^*).
\]
The details of Eqs. (9)-(11) can be found in [11,13].}

As shown in Fig. 2, the characteristic voice model, and then use this model to evaluate each of the music documents. As shown in Fig. 1, a collection of N music documents $X_1, X_2, ..., X_N$ is represented by solo voice models $\lambda_{i1}, \lambda_{i2}, ..., \lambda_{iN}$ according to the method described in Sec. 3. The characteristic similarity $L(X_i, Y)$, $1 \leq i \leq N$, is then evaluated by computing the log-probability (likelihood) that $Y$ tests against $\lambda_{i,j}$, i.e.,

$$L(X_i, Y) = \log p(Y_i | \lambda_{i,j}, \lambda_{b,j}),$$ \hspace{1cm} (12)

where $Y_i$ is the vocal portion of $Y$, and $\lambda_{b,j}$ is the background music GMM trained using the non-vocal portion of $Y$.

Once the singer voice characteristics are modeled, the similarity between document and query can be measured in many ways. Here, we exemplarily examine four possibilities.

**Method I:**

As shown in Fig. 1, a collection of $N$ music documents $X_1, X_2, ..., X_N$ is represented by solo voice models $\lambda_{i1}, \lambda_{i2}, ..., \lambda_{iN}$ according to the method described in Sec. 3. The characteristic similarity $L(X_i, Y)$, $1 \leq i \leq N$, is then evaluated by computing the log-probability (likelihood) that $Y$ tests against $\lambda_{i,j}$, i.e.,

$$L(X_i, Y) = \log p(Y_i | \lambda_{i,j}, \lambda_{b,j}),$$ \hspace{1cm} (12)

where $Y_i$ is the vocal portion of $Y$, and $\lambda_{b,j}$ is the background music GMM trained using the non-vocal portion of $Y$.

**Method II:**

An alternative way is to represent a music query as a solo voice model, and then use this model to evaluate each of the music documents. As shown in Fig. 2, the characteristic similarity $L(X_i, Y)$, $1 \leq i \leq N$, is computed using

$$L(X_i, Y) = \frac{1}{T_i} \left[ \log p(X_i, Y | \lambda_{i,j}, \lambda_{b,j}) \right] - \log p(Y_i | \lambda_{i,j}, \lambda_{b,j}),$$ \hspace{1cm} (13)

where $X_i$ is the vocal portion of document $X_i$, $T_i$ is the length of $X_i$, $\lambda_{i,j}$ is the background music GMM trained using the non-vocal portion of document $X_i$, $\lambda_{b,j}$ is the solo voice model trained using the music query $Y$, and $\lambda_{b,j}$ is the solo voice model trained using all the music documents. Eq. (13) is basically a likelihood ratio test following the Neyman-Pearson Lemma [13]. The model $\lambda_{b,j}$ simulates the alternative hypothesis, which evaluates the extent of $S(X_i) \neq S(Y)$, as opposed to the null hypothesis provided by $\lambda_{i,j}$, which evaluates the extent of $S(X_i) = S(Y)$. Compared to Method I, this method is computationally less extensive in the offline indexing phase, but could be costly expensive in the online retrieving phase, since a solo voice modeling process must be carried out whenever a query incomes.

**Method III:**

Without taking the computational complexity into consideration, the retrieval performance may be improved by combining Method I and Method II. A straightforward combination is to sum up their similarity measurements:

$$L(X, Y) = \frac{1}{T_y} \left[ \log p(Y_i | \lambda_{i,j}, \lambda_{b,j}) \right] - \log p(Y_i | \lambda_{i,j}, \lambda_{b,j}),$$ \hspace{1cm} (14)

where $T_y$ is the length of $Y_i$.

**Method IV:**

Instead of simply measuring the similarity between each music document and a query, a further improvement may be made by incorporating the relation between documents into the similarity measurement. The basic idea is that the documents relevant to a particular query are supposed to resemble each other, and therefore the inter-document similarity can be exploited as supplementary information for document-query similarity measurement. Let $R_k \{ L(X, Y) \} = R_k$, $1 \leq k \leq N$, the modified similarity measurement $\hat{L}(X_i, Y)$ between document $X_i$ and query $Y$ is computed using

$$\hat{L}(X, Y) = L(X, Y) + \sum_{k=1}^{N} \alpha_k \hat{L}(X_i, X_k),$$ \hspace{1cm} (15)

where $L(\cdot)$ can be computed using Method I, II, or III, and $\alpha$ is a constant assigned to be smaller than one if the value of $L(\cdot)$ is positive and larger than one otherwise.

**5. Experimental results**

The music data used in this study consisted of 416 tracks from Mandarin pop music CDs. All the tracks were manually labeled with the singer identity and the vocal/non-vocal boundaries for serving as the ground truth. The database was divided into two subsets, denoted as DB1 and DB2,
respectively. The DB1 comprised 200 tracks performed by 10 female and 10 male singers, with 10 distinct songs per singer. DB2 contained the remaining 216 tracks, involving 13 female and 8 male singers, none of whom appeared in DB1. All music data were down-sampled from the CD sampling rate of 44.1 kHz to 22.05 kHz, to exclude the high frequency components beyond the range of normal singing voices. Feature vectors, each consisting of 20 Mel-scale frequency cepstral coefficients, were extracted from these data using a 32-ms Hamming-windowed frame with 10-ms frame shifts.

In our experiments, DB1 was used for examining the validity of the proposed methods, while DB2 was used for training the vocal model $\lambda_{\text{v}}$, non-vocal model $\lambda_{\text{nv}}$, and alternative hypothesis model $\lambda_{\text{a}}$. Performance of the vocal/non-vocal segmentation has been evaluated and reported on [11], where the result with the best segmentation accuracy of 79.8% was adopted here as a front-end processing for our subsequent experiments.

Experiments for the music document retrieval were conducted in a leave-one-out manner, which used the first minute of each track in DB1 as a query once at a time to retrieve the remaining 199 tracks in DB1, and then rotated through all the tracks. Fig. 3 shows the precision rates (PR) and the recall rates (RR) with respect to the number of documents presented to the user. The number of mixtures in $\lambda_{\text{v}}$, $\lambda_{\text{nv}}$, $\lambda_{\text{a}}$, and $\lambda_{\text{a}}$, $1 \leq t \leq 199$, were empirically determined to be 32, 32, 8, 32, and 8, respectively. We can see that Method I and Method II performed almost equally, and the equal recall-precision (RR–PR) rates are, respectively, 54.7% and 54.1%. Method III, a combination of Method I and Method II, outperformed either Method I or Method II. Method IV was obviously the best approach. Here, the value of $\alpha$ in Eq. (15) was empirically set to be 1.2, and Method III was used to compute $L(X,Y)$. It is clear that a significant improvement in the retrieval performance can be obtained by incorporating the inter-document relations into the similarity measurement between document and query. The best equal recall-precision rate obtained in this study was 72.6%.

6. Conclusions

This study has examined the feasibility of retrieving a specified singer’s music documents through the use of query-by-example. We have shown that the relevant documents can be determined by extracting the singer voice characteristics in a music recording, followed by comparing the similarity between document and query. Future work will further improve the retrieval performance by applying the relevant feedback or some other sophisticated methods, and will extend the current framework to handle music data containing multiple singers, recorded under adverse environments, or encoded in various formats with loss.

7. References