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 by 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 \( \mathcal{S}(\cdot) \) 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 \( \mathcal{S}(X_i) = \mathcal{S}(Y) \). 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 \( \mathcal{L}(X_i, Y) \) denote the similarity measurement taken over the music document \( X_i \) and the query \( Y \), in which a large value of \( \mathcal{L}(X_i, Y) \) indicates high similarity, and \( R\{\mathcal{L}(X_i, Y)\} \) denote the rank of \( \mathcal{L}(X_i, Y) \) 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]
A music document $X_i$ is hypothesized as relevant to the query $Y$ if
\[
R(L(X,Y)) < \theta,
\]
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 of examining the existing methods, we apply the RF implicitly by incorporating the inter-document relationships into the similarity measurement between document and query. This approach enables a retrieval process to be performed efficiently and without user interaction. In addition, the scope of this study is restricted to dealing with the music recordings (both document and query) that are performed by only one singer. For the music containing multiple singers, the proposed framework remains applicable if the recordings are pre-segmented into singer-homogeneous regions.

### 3. Singer characteristic extraction

As a prerequisite to singer-based music document retrieval, singing voices in a music recording must be detected and characterized. Following our previous work in [11], music segments that contain vocals are first identified by using a vocal/non-vocal classifier, and the identified vocal regions are then stochastically represented as a parametric model which isolates the characteristics of the underlying solo voices from the background accompaniments.

The vocal/non-vocal classifier consists of a front-end signal processor that converts digital waveforms to cepstrum-based feature vectors, followed by a backend statistical processor that performs modeling and matching. It operates in two phases, training and testing. During training, a music database with manual vocal/non-vocal transcriptions is used to create two Gaussian mixture models (GMMs), $\lambda_v$ and $\lambda_b$, respectively, for characterizing the vocal and non-vocal classes. Parameters of the GMMs are initialized independently according to GMMs $\lambda_v = \{w_{v,i}, \mu_{v,i}, \Sigma_{v,i} \mid 1 \leq i \leq I \}$ and $\lambda_b = \{w_{b,j}, \mu_{b,j}, \Sigma_{b,j} \mid 1 \leq j \leq J \}$, where $w_{v,i}$ and $w_{b,j}$ are mixture weights, $\mu_{v,i}$ and $\mu_{b,j}$ mean vectors, and $\Sigma_{v,i}$ and $\Sigma_{b,j}$ covariance matrices. If the signal $V$ is formed from a generative function $V = f(s, b)$, $1 \leq i \leq T$, the probability of $V$, given $\lambda_v$ and $\lambda_b$, can be represented by
\[
p(V | \lambda_v, \lambda_b) = \prod_{i=1}^T \left\{ \sum_{v=1}^V \sum_{b=1}^B w_{v,i} w_{b,i} p(v_i | \mu_{v,i}, \Sigma_{v,i}, \mu_{b,i}, \Sigma_{b,i}) \right\}, \tag{3}
\]
where
\[
p(v_i | \mu_{v,i}, \Sigma_{v,i}, \mu_{b,j}, \Sigma_{b,j}) = \int p(v_i | \mu_{v,i}, \Sigma_{v,i})N(\mu_{b,j}, \Sigma_{b,j})\text{d}\mu_{b,j}. \tag{4}
\]
To build $\lambda_v$, a maximum-likelihood estimation is made as
\[
\hat{\lambda}_{v} = \arg \max_{\lambda_v} p(V | \lambda_v, \lambda_b). \tag{5}
\]
Using the EM algorithm, a new model $\hat{\lambda}_{v}$ is iteratively estimated by maximizing the auxiliary function
\[
Q(\lambda_{v,i}, \hat{\lambda}_{v}) = \sum_{i=1}^T \sum_{j=1}^J \sum_{k=1}^K p(i,j) p(v_i | \hat{\lambda}_{v}) \log p(i,j,v_i | \hat{\lambda}_{v}), \tag{6}
\]
where
\[
p(i,j,v_i | \hat{\lambda}_{v}, \lambda_{v}) = \tilde{w}_{v,i} w_{v,j} p(v_i | \tilde{\mu}_{v,j}, \tilde{\Sigma}_{v,j}, \mu_{v,i}, \Sigma_{v,i}), \tag{7}
\]
and
\[
p(i,j,v_i | \lambda_{v,i}, \hat{\lambda}_{v}) = \frac{w_{v,i} w_{b,j} p(v_i | \mu_{v,i}, \Sigma_{v,b}, \mu_{b,j}, \Sigma_{b,j})}{\sum_{i=1}^T \sum_{j=1}^J w_{v,i} w_{b,j} p(v_i | \mu_{v,i}, \Sigma_{v,b}, \mu_{b,j}, \Sigma_{b,j})}. \tag{8}
\]
Letting $\nabla Q(\lambda_{v,i}, \hat{\lambda}_{v}) = 0$ with respect to each parameter to be re-estimated, we have
\[
\tilde{w}_{v,i} = \frac{1}{T} \sum_{i=1}^T \sum_{j=1}^J p(i,j | v_i, \hat{\lambda}_{v}, \lambda_{v,i}). \tag{9}
\]
\[
\hat{\mu}_{ij} = \frac{\sum_{i=1}^{N_{i}} \sum_{j=1}^{N_{j}} p(i,j|\lambda_{i,\lambda_{j}}) E[\mathbf{v}_{i,j}|\mathbf{v}_{i,j},\lambda_{i,\lambda_{j}},\Sigma_{i,j},\mu_{i,j},\Sigma_{i,j}]}{\sum_{i=1}^{N_{i}} \sum_{j=1}^{N_{j}} p(i,j|\lambda_{i,\lambda_{j}})}
\]

where prime denotes vector transpose, and \(E[\cdot]\) expectation. The details of Eqs. (9)-(11) can be found in [11,13].

### 4. Similarity computation

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 \(X_{1}, X_{2}, \ldots, X_{N}\) is represented by solo voice models \(\lambda_{s,1}, \lambda_{s,2}, \ldots, \lambda_{s,N}\) according to the method described in Sec. 3. The characteristic similarity \(\mathcal{L}(X,Y)\), \(1 \leq i \leq N\), is then evaluated by computing the log-probability (likelihood) that \(Y\) tests against \(\lambda_{s,j}\), i.e.,

\[
\mathcal{L}(X,Y) = \log p(Y|\lambda_{s,i}, \lambda_{b,j}),
\]

where \(Y\) 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:**

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 \(R_{k}\) be the modified similarity measurement between document \(X_{i}\) and query \(Y\) is computed using

\[
\hat{\mathcal{L}}(X_{i}, Y) = \mathcal{L}(X_{i}, Y) + \sum_{k=1}^{N} \alpha_{k} \mathcal{L}(X_{i}, X_{k})
\]

where \(\mathcal{L}(\cdot)\) can be computed using Method I, II, or III, and \(\alpha\) is a constant assigned to be smaller than one if the value of \(\mathcal{L}(\cdot)\) 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,
The recall-precision rate obtained in this study was 72.6%. Incorporating the inter-document relations into the similarity improvement in the retrieval performance can be obtained by Method IV was obviously the best approach. Here, the value and Method II, outperformed either Method I or Method II. 54.7% and 54.1%. Method III, a combination of Method I and the equal recall-precision (RR=PR) rates are, respectively, see that Method I and Method II performed almost equally, determined to be 32, 32, 8, 32, and 8, respectively. We can 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 λs,i, λs,i, λs,i, λs,i, and λs,i, 1 ≤ i ≤ 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 α 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%.

![Figure 3: Results of the singer-based music document retrieval obtained with the various similarity measurements.](image-url)