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 words to rhythms. It is an information rich medium in itself, capable of conveying various concrete and abstract aspects such as compositions, themes, artist styles, moods, sentiments, or even ethnic culture, which cannot be exhaustively described and attached in every music recording. As music materials encoded in digital formats rapidly grow in size and number, finding the desired information from an enormous amount of available music data 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 $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ denote N music documents in a database to be searched, $\mathbf{Y}$ 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 $\mathcal{S}(\mathbf{X}_i) = \mathcal{S}(\mathbf{Y})$. 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 $\mathcal{L}(\mathbf{X}_i,\mathbf{Y})$ denote the similarity measurement taken over the music document $\mathbf{X}_i$ and the query $\mathbf{Y}$, in which a large value of $\mathcal{L}(\mathbf{X}_i,\mathbf{Y})$ indicates high similarity, and $\mathcal{R}\{\mathcal{L}(\mathbf{X}_i,\mathbf{Y})\}$ denote the rank of $\mathcal{L}(\mathbf{X}_i,\mathbf{Y})$ among $\mathcal{L}(\mathbf{X}_1,\mathbf{Y})$, $\mathcal{L}(\mathbf{X}_2,\mathbf{Y})$, …, $\mathcal{L}(\mathbf{X}_N,\mathbf{Y})$ in

[1] Part of the idea is similar to the one in [7]

descending order. A music document $\mathbf{X}_i$ is hypothesized as relevant to the query $\mathbf{Y}$ if

$$\mathcal{R}\{\mathcal{L}(\mathbf{X}_i,\mathbf{Y})\} < \theta, \tag{1}$$

where θ 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 intervention. 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), λ_V and λ_N, respectively, for characterizing the vocal and non-vocal classes. Parameters of the GMMs are initialized via *k*-means clustering and iteratively adjusted via expectation-maximization (EM) [12]. During testing, the classifier takes as input the T_x-length feature vectors $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_{Tx}\}$ extracted from an unknown recording, and produces as outputs the frame log-likelihoods $\log p(\boldsymbol{x}_t|\lambda_V)$ and $\log p(\boldsymbol{x}_t|\lambda_N)$, $1 \le t \le T_x$. Since singing tends to be continuous, classification is preferably made in a segment-by-segment manner rather than a frame-by-frame manner. To reduce the risk of crossing multiple vocal/non-vocal boundaries, a segment is selected and examined in the following way. First, vector clustering is employed on all the frame feature vectors, and each frame is assigned a cluster index associated with that frame's feature vector. Then, each segment is assigned the majority index of its constituent frames, and adjacent segments are merged as a homogeneous segment if they have the same index. Finally, classification is made per homogeneous segment using:

$$\frac{1}{W_k}\left(\sum_{i=0}^{W_k-1}\log p(\boldsymbol{x}_{s_k+i}\,|\,\lambda_V) - \sum_{i=0}^{W_k-1}\log p(\boldsymbol{x}_{s_k+i}\,|\,\lambda_N)\right) \begin{array}{c}\text{vocal}\\ \gtrless \\ \text{non - vocal}\end{array} \eta, \tag{2}$$

where W_k and s_k represent, respectively, the length and starting frame of the *k*-th homogeneous segment, and η 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 in cepstrum-based feature representation $\mathbf{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_T\}$ is a mixture of a solo voice $\mathbf{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_T\}$ and a background music $\mathbf{B} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_T\}$. 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 λ_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 $\lambda_s = \{w_{s,i}, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i} \mid 1 \le i \le I\}$ and $\lambda_b = \{w_{b,j}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j} \mid 1 \le j \le J\}$, where $w_{s,i}$ and $w_{b,j}$ are mixture weights, $\boldsymbol{\mu}_{s,i}$ and $\boldsymbol{\mu}_{b,j}$ mean vectors, and $\boldsymbol{\Sigma}_{s,i}$ and $\boldsymbol{\Sigma}_{b,j}$ covariance matrices. If the signal $\mathbf{V}$ is formed from a generative function $\boldsymbol{v}_t = f(\boldsymbol{s}_t, \boldsymbol{b}_t)$, $1 \le t \le T$, the probability of $\mathbf{V}$, given λ_s and λ_b, can be represented by

$$p(\mathbf{V}\,|\,\lambda_s,\lambda_b) = \prod_{t=1}^{T}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} w_{s,i}w_{b,j}\,p(\boldsymbol{v}_t\,|\,\boldsymbol{\mu}_{s,i},\boldsymbol{\Sigma}_{s,i},\boldsymbol{\mu}_{b,j},\boldsymbol{\Sigma}_{b,j})\right\}, \tag{3}$$

where

$$p(\boldsymbol{v}_t\,|\,\boldsymbol{\mu}_{s,i},\boldsymbol{\Sigma}_{s,i},\boldsymbol{\mu}_{b,j},\boldsymbol{\Sigma}_{b,j}) = \iint_{\boldsymbol{v}_t=f(\boldsymbol{s}_t,\boldsymbol{b}_t)} \mathcal{N}(\boldsymbol{s};\boldsymbol{\mu}_{s,i},\boldsymbol{\Sigma}_{s,i})\mathcal{N}(\boldsymbol{b};\boldsymbol{\mu}_{b,j},\boldsymbol{\Sigma}_{b,j})\,d\boldsymbol{s}d\boldsymbol{b}. \tag{4}$$

To build λ_s, a maximum-likelihood estimation is made as

$$\lambda_s^* = \arg\max_{\lambda_s} p(\mathbf{V}\,|\,\lambda_s,\lambda_b). \tag{5}$$

Using the EM algorithm, a new model $\hat{\lambda}_s$ is iteratively estimated by maximizing the auxiliary function

$$Q(\lambda_s,\hat{\lambda}_s) = \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{J} p(i,j\,|\,\boldsymbol{v}_t,\lambda_s,\lambda_b)\log p(i,j,\boldsymbol{v}_t\,|\,\hat{\lambda}_s,\lambda_b), \tag{6}$$

where

$$p(i,j,\boldsymbol{v}_t\,|\,\hat{\lambda}_s,\lambda_b) = \hat{w}_{s,i}w_{b,j}\,p(\boldsymbol{v}_t\,|\,\hat{\boldsymbol{\mu}}_{s,i},\hat{\boldsymbol{\Sigma}}_{s,i},\boldsymbol{\mu}_{b,j},\boldsymbol{\Sigma}_{b,j}), \tag{7}$$

and

$$p(i,j\,|\,\boldsymbol{v}_t,\lambda_s,\lambda_b) = \frac{w_{s,i}w_{b,j}\,p(\boldsymbol{v}_t\,|\,\boldsymbol{\mu}_{s,i},\boldsymbol{\Sigma}_{s,i},\boldsymbol{\mu}_{b,j},\boldsymbol{\Sigma}_{b,j})}{\sum_{m=1}^{I}\sum_{n=1}^{J} w_{s,m}w_{b,n}\,p(\boldsymbol{v}_t\,|\,\boldsymbol{\mu}_{s,m},\boldsymbol{\Sigma}_{s,m},\boldsymbol{\mu}_{b,n},\boldsymbol{\Sigma}_{b,n})}. \tag{8}$$

Letting $\nabla Q(\lambda_s,\hat{\lambda}_s) = 0$ with respect to each parameter to be re-estimated, we have

$$\hat{w}_{s,i} = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J} p(i,j\,|\,\boldsymbol{v}_t,\lambda_s,\lambda_b), \tag{9}$$

$$\hat{\boldsymbol{\mu}}_{s,i} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{N} p(i,j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) E\{\boldsymbol{s}_t \mid \boldsymbol{v}_t, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}\}}{\sum_{t=1}^{T}\sum_{j=1}^{N} p(i,j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b)}, \quad (10)$$

$$\hat{\boldsymbol{\Sigma}}_{s,i} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{J} p(i,j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) E\{\boldsymbol{s}_t \boldsymbol{s}_t' \mid \boldsymbol{v}_t, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}\}}{\sum_{t=1}^{T}\sum_{j=1}^{J} p(i,j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b)} - \hat{\boldsymbol{\mu}}_{s,i}\hat{\boldsymbol{\mu}}_{s,i}', \quad (11)$$

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 $\mathbf{X}_1$, $\mathbf{X}_2$,…, $\mathbf{X}_N$ is represented by solo voice models $\lambda_{s,1}$, $\lambda_{s,2}$,…, $\lambda_{s,N}$, according to the method described in Sec. 3. The characteristic similarity $\mathcal{L}(\mathbf{X}_i,\mathbf{Y})$, $1 \le i \le N$, is then evaluated by computing the log-probability (likelihood) that $\mathbf{Y}$ tests against $\lambda_{s,i}$; i.e.,

$$\mathcal{L}(\mathbf{X}_i,\mathbf{Y}) = \log\ p(\mathbf{Y}_v|\lambda_{s,i},\ \lambda_{b,y}), \quad (12)$$

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

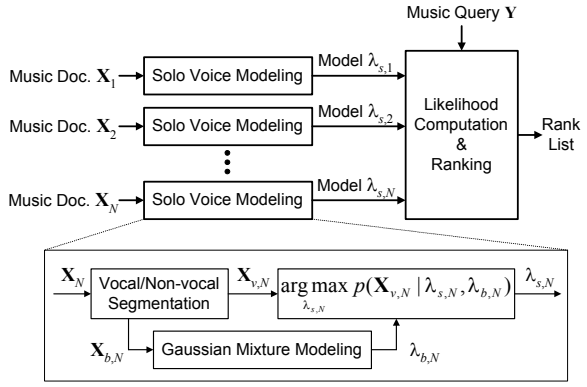

Figure 1: Method I for similarity measurement.

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 $\mathcal{L}(\mathbf{X}_i,\mathbf{Y})$, $1 \le i \le N$, is computed using

$$\mathcal{L}(\mathbf{X}_i,\mathbf{Y}) = \frac{1}{T_i}\left[\log p(\mathbf{X}_{v,i} \mid \lambda_{s,y}, \lambda_{b,i}) - \log p(\mathbf{X}_{v,i} \mid \lambda_{s,x}, \lambda_{b,i})\right], \quad (13)$$

where $\mathbf{X}_{v,i}$ is the vocal portion of document $\mathbf{X}_i$, T_i is the length of $\mathbf{X}_{v,i}$, $\lambda_{b,i}$ is the background music GMM trained using the non-vocal portion of document $\mathbf{X}_i$, $\lambda_{s,y}$ is the solo voice model trained using the music query $\mathbf{Y}$, and $\lambda_{s,x}$ 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 $\lambda_{s,x}$ simulates the *alternative hypothesis*, which evaluates the extent of $\mathcal{S}(\mathbf{X}_i) \neq \mathcal{S}(\mathbf{Y})$, as opposite to the *null hypothesis* provided by $\lambda_{s,y}$, which evaluates the extent of $\mathcal{S}(\mathbf{X}_i) = \mathcal{S}(\mathbf{Y})$. 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.

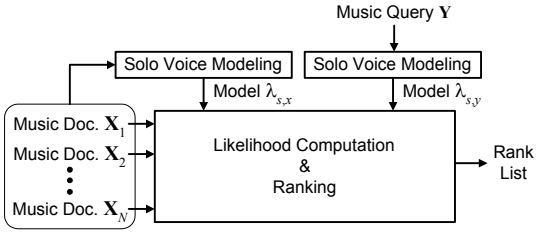

Figure 2: Method II for similarity measurement.

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:

$$\mathcal{L}(\mathbf{X}_i,\mathbf{Y}) = \frac{1}{T_y}\log p(\mathbf{Y}_v \mid \lambda_{s,i}, \lambda_{b,y}) + \frac{1}{T_i}\left[\log p(\mathbf{X}_{v,i} \mid \lambda_{s,y}, \lambda_{b,i}) - \log p(\mathbf{X}_{v,i} \mid \lambda_{s,x}, \lambda_{b,i})\right], \quad (14)$$

where T_y is the length of $\mathbf{Y}_v$.

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 $\mathcal{R}\{\mathcal{L}(\mathbf{X}_k,\mathbf{Y})\} = R_k$, $1 \le k \le N$, the modified similarity measurement $\hat{\mathcal{L}}(\mathbf{X}_i,\mathbf{Y})$ between document $\mathbf{X}_i$ and query $\mathbf{Y}$ is computed using

$$\hat{\mathcal{L}}(\mathbf{X}_i,\mathbf{Y}) = \mathcal{L}(\mathbf{X}_i,\mathbf{Y}) + \sum_{k=1}^{N} \alpha^{R_k}\mathcal{L}(\mathbf{X}_i,\mathbf{X}_k), \quad (15)$$

where $\mathcal{L}(\cdot)$ can be computed using Method I, II, or III, and α 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,

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 λ_V, non-vocal model λ_N, and alternative hypothesis model $\lambda_{s,x}$. 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 $\lambda_{s,y}$, $\lambda_{s,x}$, $\lambda_{b,y}$, $\lambda_{s,i}$, and $\lambda_{b,i}$, $1 \leq i \leq 199$, 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 α in Eq. (15) was empirically set to be 1.2, and Method III was used to compute $\mathcal{L}(\mathbf{X}_i,\mathbf{Y})$. 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%.

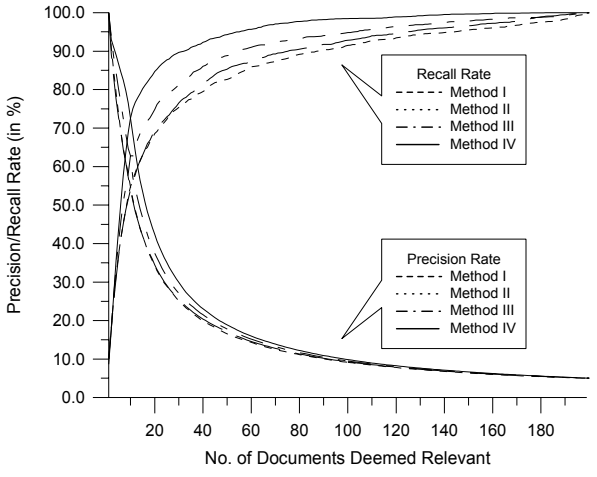

Figure 3: Results of the singer-based music document retrieval obtained with the various similarity measurements.

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

[1] A. S. Durey and M. A. Clements, "Features for melody spotting using hidden Markov models," *Proc. of ICASSP*, Florida, pp. 1765–1768, 2002.

[2] P. Herrera, X. Amatriain, E. Batlle, and X. Serra, "Towards instrument segmentation for music content description: a critical review of instrument classification techniques," *Proc. of ISMIR*, Massachusetts, 2000.

[3] G. Tzanetakis, and P. Cook, "Musical genre classification of audio signals", *IEEE Trans. Speech & Audio Proc.*, vol. 10, no. 5, pp. 293–302, 2002.

[4] Y. E. Kim, and B. Whitman, "Singer identification in popular music recordings using voice coding features", *Proc. of ISMIR*, Paris, 2002.

[5] C. C. Liu and C. S. Huang, "A singer identification technique for content-based classification of MP3 music objects," *Proc. of CIKM*, Virginia, 2002.

[6] T. Zhang, "Automatic Singer Identification," *Proc. of ICME*, Baltimore, 2003.

[7] J. Haitsma and T. Kalker, "A Highly Robust Audio fingerprinting System," *Proc. of ISMIR*, Paris, 2002.

[8] G. Salton, Automatic text processing. Reading, MA: Addison-Wesley Publishing Company. 1989.

[9] F. C. Ekmekcioglu, A. M. Robertson, and P. Willett, "Effectiveness of query expansion in ranked-output document retrieval systems," *Journal of Information Science*, vol. 18, pp. 139–147. 1992.

[10] K. Porkaew, K. Chakrabarti, and S. Mehrotra, "Query refinement for content-based multimedia retrieval in MARS," *Proc. of ACM Multimedia*, Orlando, 1999.

[11] W. H. Tsai, H. M. Wang, D. Rodgers, S. S. Cheng, and H. M. Yu, "Blind clustering of popular music recordings based on singer voice characteristics," *Proc. of ISMIR*, Baltimore, 2003.

[12] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. Royal Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[13] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise", *IEEE Trans. Speech & Audio Proc.*, vol. 2, no. 2, pp. 245–257, 1994.

[14] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans. R. Soc.*, vol. 231, pp. 289–337, 1933.