5-2009

A Novel Framework for Efficient Automated Singer Identification in Large Music Databases

Jialie SHEN
Singapore Management University, jlshen@smu.edu.sg

John Shepherd
University of New South Wales

Bin CUI
Peking University

Kian-Lee TAN
National University of Singapore

DOI: https://doi.org/10.1145/1508850.1508856

Published in ACM Transactions on Information Systems, May 2009, Volume 27, Issue 3, Article number 18.
http://doi.org/10.1145/1508850.1508856

A Novel Framework for Efficient Automated Singer Identification in Large Music Databases

JIALIE SHEN
Singapore Management University
JOHN SHEPHERD
The University of New South Wales
BIN CUI
Peking University
and
KIAN-LEE TAN
National University of Singapore

Over the past decade, there has been explosive growth in the availability of multimedia data, particularly image, video, and music. Because of this, content-based music retrieval has attracted attention from the multimedia database and information retrieval communities. Content-based music retrieval requires us to be able to automatically identify particular characteristics of music data. One such characteristic, useful in a range of applications, is the identification of the singer in a musical piece. Unfortunately, existing approaches to this problem suffer from either low accuracy or poor scalability. In this article, we propose a novel scheme, called *Hybrid Singer Identifier* (HSI), for efficient automated singer recognition. HSI uses multiple low-level features extracted from both vocal and nonvocal music segments to enhance the identification process; it achieves this via a hybrid architecture that builds profiles of individual singer characteristics based on statistical mixture models. An extensive experimental study on a large music database demonstrates the superiority of our method over state-of-the-art approaches in terms of effectiveness, efficiency, scalability, and robustness.

Portions of this work appeared as Shen, J., Cui, B., Shepherd, J., and Tan, K.-L. "Towards efficient automated singer identification in large music databases," *Proceedings of ACM SIGIR'06.* ACM Press, New York, NY, 59–66.
Author's addresses: J. Shen, School of Information Systems, Singapore Management University, 80 Stamford Road, Singapore 178902; email: jshen@smu.edu.sg; J. Shepherd, School of Computer Science and Engineering, The University of New South Wales, Sidney, NSW 2052, Australia; email: jas@cse.unsw.edu.au; B. Cui, Department of Computer Science, Peking University, Science Building 1, Room 1628, Beijing, China 100871; email: bin.cui@pku.edu.cn; K.-L. Tan, Department of Computer Science, National University of Singapore, Kent Ridge, Singapore 117543; email: tankl@comp.nus.edu.sg.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation efficiency and effectiveness*; H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Modeling*; J.5 [**Arts and Humanities**]: Performing arts—*(e.g, dance, music)*

General Terms: Algorithms, Experimentation, Performance

Additional Key Words and Phrases: Music retrieval, singer identification, Gaussian mixture models, classification, EM algorithm, statistical modeling, evaluation

ACM Reference Format:

Shen, J., Shepherd, J., Cui, B., and Tan, K.-L. 2009. A novel framework for efficient automated singer identification in large music databases. ACM Trans. Inform. Syst. 27, 3, Article 18 (May 2009), 31 pages. DOI = 10.1145/1508850.1508856 http://doi.acm.org/10.1145/1508850.1508856

1. INTRODUCTION

With the continued advances in data storage and communication technology, there has been an explosive growth in the volume of music data stored in digital form. Consequently, there has been increasing interest in the multimedia database and information systems communities [Lam and Tan 2001; Easley et al. 2003; Pachet 2003; Pardo 2006; Pinto and Haus 2007] to study techniques for content-based music retrieval. Many techniques have recently been developed to support automatic classification or recognition of music based on instrument, genre, and other characteristics [Li et al. 2003; Liu and Huang 2002; Zhang 2003; Li and Ogihara 2004]. In particular, techniques for automatic artist identification are gaining in importance due to potential applications such as music indexing and retrieval, copyright management, and music recommendation systems. The development of singer identification techniques enables the effective management and exploration of large music collections based on "singer similarity." With this technology, songs performed by a particular singer can be automatically clustered for easy management or searching.

Currently, the most popular and naive approach to support singer identification is to manually embed artist information into the music database with the assistance of music professionals. In this case, traditional text retrieval techniques can be directly applied for music querying. The obvious shortcoming of the approach is that it requires a significant amount of time and domain expertise to label each music item. Note that, while many formats of music data have embedded artist information, we cannot rely on the existence of such data (for example, if a snippet of music is recorded from the radio and we wish to find the performer). It is clear that songs performed by the same singer share certain audio characteristics; the singer's voice most likely contains similar audio patterns over all the songs they perform. Also, artists tend to perform within a single genre, and thus the audio characteristics of their work may contain common features (e.g., instrumentation). This suggests the feasibility of singer identification based on audio content.

By *automatic singer identification*, we refer to the task of determining, for a given song, which singer among S candidate artists performed it. The essential

part of this process is an S-class categorization. The effectiveness of solutions to this problem relies heavily on their ability to capture salient information for separating one signal from others. While traditional speech recognition techniques [Rabiner and Juang 1993; Becchetti et al. 1999] could be easily applied to this task, they are likely not to perform well. This is because the vocal track is intertwined with the nonstationary background signal from the backing instruments. It is rare that we might acquire a pure solo voice track without instrumentation (unless we had access to the original multitrack data for the song). Furthermore, there is a semantic gap between low-level acoustic features and high-level notions such as music genre. Superior query accuracy cannot be expected without using well-discriminating low-level features to distinguish the songs of one artist from those of another.

Several approaches have been recently proposed to apply statistical models or machine learning techniques for automatic singer classification/identification [Liu and Huang 2002; Zhang 2003; Tsai and Wang 2006]. In general, these methods consist of two main steps: singer characteristic modeling based on voice data, and class label identification via machine learning algorithms. In singer characteristic modeling, acoustic signal information is extracted to represent the music. Then specific mechanisms (i.e., statistical models or machine learning algorithms) are constructed to assign songs to one of the predefined singer categories based on their extracted acoustic features. Unfortunately, existing attempts at implementing this approach have been unable to achieve acceptable classification accuracy. The main reasons are (1) they focus on a single, nonrepresentative feature (voice data), (2) they do not explore effective ways of combining multiple features, and (3) they do not take account of information available in the accompanying music track.

We now expand on the above points in more detail: (1) music consists of a complex collection of features (beat, timbre, etc.), with vocal features comprising just one of many possible features that might be used for classification. Employing a single type of feature to represent an artist, as many previous systems have done, is unlikely to lead to good performance by an identification system. Moreover, the acoustic data for an artist's singing carries more information than plain speech data. For example, a singer's unique formant structure may be reflected in the songs that (s)he sings. Speech features alone may not always represent such characteristics effectively. (2) No existing work has addressed the underlying multifeature integration model that can be used to explore the conjunctive effects among the different acoustic characteristics for effective singer identification. And (3) most previous work has focused on vocal features and does not consider the influence of accompanying music. However, nonvocal components generally carry a large amount of information about music style and genre. The styles of song performed by a singer could be relatively steady during a certain period. This information is potentially useful in assisting to identify songs performed by that artist. Of course, artists do perform pieces outside their "normal style," which would be difficult to recognize if not included in the training data. However, ignoring musical information altogether does not help improve the accuracy of the identification process.

Motivated by these concerns, we introduce a novel system, called *Hybrid Singer Identifier* (HSI), for effective automated singer recognition in large music databases. HSI considers polyphonic music as input and uses a two-layer structure consisting of a preprocessing module and a singer modeling module. The main contributions of our approach can be summarized as follows:

—Instead of considering only the singer's voice, we propose a novel singer identification framework using multiple features extracted from both vocal and nonvocal components to improve the system's effectiveness on singer characteristic modeling.

—A probabilistic singer characteristic modeling method is designed based on mixture models and a score fusion scheme based on logistic regression to bridge the "semantic gap." In addition, distinguished from previous methods that only rely on single type of low-level acoustic feature, our approach can effectively integrate multiple kinds of sound information to enhance the identification process with a hybrid architecture and associated learning algorithm.

—Most previous studies in the field have focused primarily on improving the effectiveness of the identification process. While this is clearly a desirable goal, not enough previous work has also considered the problems of efficiency and scalability, which are very important issues in a practical system with large datasets. Our approach achieves both effectiveness and efficiency/scalabilty by its use of a layered component structure.

—The proposed system has been fully implemented and tested. An extensive range of tests has been designed to investigate different factors that affect the performance of the HSI approach and its competitors. The results, based on a large dataset, demonstrate the superiority of our method over other state-of-the-art approaches. The tests address a range of factors, including effectiveness, scalability, efficiency, and robustness against audio distortion and other kinds of noise.

The remainder of this article is organized as follows: Section 2 gives a brief overview of related work in the area of singer/artist identification, including the assumptions and limitations of each. In Section 3, we review our proposed architecture, giving the detailed structure of its component modules and its learning algorithms. Section 4 reports our experimental configuration and results. Finally, we give our conclusions and directions for future work in Section 5.

2. RELATED WORK

Automated singer identification is an important research problem with numerous applications in multimedia information systems. In recent years, there have been many efforts to develop frameworks for singer/artist identification and associated topics. While different kinds of data, such as text-based captions, can be applied for the task, in the following survey, we focus on feature-based approaches. Among the earliest of such systems, Minnowmatch mainly focuses

on artist rather than singer identification using mel-frequency cepstral coefficients (MFCCs), which is a feature adapted from the classical speech recognition and speaker identification mechanisms [Whitman et al. 2001]. The best identification accuracy achieved on a small dataset, containing a 10-artist set, was approximately 70%. However, with a larger set of 21 artists, the best case accuracy dropped to 50%.

In Berenzweig and Ellis [2001], vocal music was used as an input to a speech recognition system, achieving a success rate of up to 80% in isolating vocal regions. In Berenzweig et al. [2002], the authors used a neural network trained on radio recordings to similarly segment songs into vocal and nonvocal regions. By focusing on voice regions alone, they improved artist identification by 15%. The system presented here also attempts to perform segmentation of vocal regions prior to singer identification. After segmentation, the classifier uses features drawn from voice coding based on Linear Predictive Coding. Kim and Whitman [2002] developed a scheme to automatically construct the identity of a singer using acoustic features extracted from the vocal parts of popular music. The classification experiment was carried out with two different classifiers—the Gaussian Mixture Model (GMM) and Support Vector Machines (SVMs). The best accuracy achieved was 45.3% based on a small test set using the SVMs. In followup work, Kim et al. [2006] studied the "album effect" in singer identification. In the real world, consistency of audio production techniques is high in the same album, but could be very different between different albums. This can have a great impact on singer identification systems. The research indicates that accuracy can be improved greatly when systems are trained and tested based on music items from the same album. The main problem for this study is that the size of the test collection and other detailed information were not available in the article.

In Liu and Huang [2002], a novel scheme was designed and developed to automatically classify music objects according to their singers. First, the coefficients extracted from the output of polyphase filters are used to compute the music features for segmentation. Based on these features, a music object can be decomposed into a sequence of notes (or phonemes). Then for each phoneme in the training set, its music feature is extracted and used to train a k-nearest neighbor classifier which can identify the singer of an unknown input music object. An approximately 65% identification accuracy was achieved based on a set of 10 male and 10 female singers.

Zhang [2003] developed a system for automatic singer identification which recognizes the singer of a song by analyzing the music signal. The proposed scheme follows the framework of classical speaker identification systems, but special efforts are made to distinguish the singing voice from the background instrumental sounds in a song. A statistical model is trained for each singer's voice with typical song(s) of the singer. Then, for a song to be identified, the starting point of the singing voice is detected and a portion of the song is excerpted from that point. Audio features are extracted and matched with singers' voice models in the database. The song is assigned to the model having the best match. Accuracy rates of around 80% were achieved in a tiny database with 45 songs. Meanwhile, in Bartsch and Wakefield [2004], a singer

identification method was developed based on the spectral envelope estimation using a composite transfer function (CTF), which is calculated from the instantaneous amplitude and frequency of the signal's harmonic partials. Unfortunately, this method only examines a very limited case in which audio samples only contain the singer's voice—solo performances of Italian arias, without any accompaniment. Thus, the technique could be less effective for songs that also involve instruments. On the other hand, Li and Ogihara [2004] proposed an artist style identification method using both lyrics and acoustic features via a semisupervised learning approach. The best identification accuracy achieved was 78.8%. The corresponding test data set contains 43 artists selected from 56 albums provided by All Music Guide.[1]

Tsai et al. [Tsai et al. 2003; Tsai and Wang 2006] proposed a solo voice modeling framework to capture singers' vocal characteristics. The technique first separates vocal from nonvocal regions and then models the singers' vocal characteristics based on stochastic properties of the background music. The system is spectrum based and its main weakness is that uses only a single type of acoustic feature for vocal portions (20-dimensional MFCC features) to profile different singers. Furthermore, the approach's scalability is poor. This is because when new singer information is added to the system, the whole framework needs to be retrained and the corresponding process could be very expensive in terms of reconstruction time. Moreover, based on their experiment on a small dataset which contained 230 popular music songs, the accuracy of identification achieved was only 71%.

The 2004 ISMIR conference [ISMIR 2004] saw the first major open evaluation of techniques in the MIR domain. Analogous to the TREC series in text retrieval, the aim was to compare approaches for a range of different tasks in music information retrieval (MIR). Artist identification was one of the task tracks, with the specific task being to recognize the performers given three songs per artist after the system has been trained using seven songs per artist. The training and development sets contained a total of 105 artists selected from the USPOP2002 collection [Berenzweig et al. 2004]. For each singer, the training and development set included seven songs and three songs individually. The features provided were MFCCs. The evaluation set included about 200 artists, which were exclusive from the USPOP2002 collection. Based on information provided by the contest organizers, using all 200 artists to test the algorithm was impossible due to technical limitations.[2] Thus, only 30 and 40 artists were used in the evaluation. The task contest attracted two participants, neither of which had an accuracy greater than 34%.

Starting in 2005, the annual Music Information Retrieval Evaluation eXchange (MIREX) activity, organized by IMIRSEL,[3] has become an important technical forum for evaluating current research and development in the area of music retrieval [Downie et al. 2005b; Downie 2006]. Each year, MIREX organizes a range of music retrieval tasks, and various research groups from around

[1] www.allmusic.com.

[2] http://ismir2004.ismir.net/genre_contest/index.htm.

[3] International Music Information Retrieval Systems Evaluation Laboratory.

Table I. Summary of Identification Methods' Accuracy from Artist Identification Track at MIREX 2005

Methods	Magnatune	USPOP
ME	76.60%	68.30%
BCE (1)	77.26%	59.88%
BCE (2)	74.45%	58.96%
E. Pampalk	66.36%	56.20%
West and Lamere	55.45%	41.04%
G. Tzanetakis	53.43%	28.64%
B. Logan	37.07%	14.83%

the world submit their systems for benchmarking. The aim is to establish a common MIR evaluation forum. It provides excellent examples for state-of-the-art music retrieval, data management, and modeling techniques, and artist/singer identification is one of main focuses for the evaluation study carried out in this event. In 2005, eight teams participated in the contest, but only seven teams' submissions completed the task in the required time. Most systems followed the same basic approach, consisting of two main steps: acoustic feature extraction and class label identification via a machine learning algorithm. Two different music databases have been used in the MIREX evaluations: *Magnatune*, based on a collection of 1800 songs[4] and USPOP, a subset of the USPOP2002 collection [Berenzweig et al. 2004]. Magnatune provided 1158 training files and 642 testing files, while USPOP provided 1158 training files and 653 testing files.

The results of the seven systems evaluated at MIREX 2005 [MIREX 2005] are presented in Table I. Two systems were clearly superior to the others: the system of Bergstra et al. (BCE) and the system of Mandel et al. (ME). The BCE(1) variation of BCE performed best (77% accuracy) on the Magnatune data, while ME performed best (68% accuracy) on the USPOP data. The BCE system considered a large number of frame-based timbre features (RCEPS, MFCCs, linear predictive coefficients, low-frequency Fourier magnitudes, Rolloff, linear prediction error, and zero-crossing rate). The mean and variance of the features were calculated for each frame. The AdaBoost.MH method was used for boosting decision stumps (BCE(1)) and two-level trees (BCE(2)). The ME system used an acoustic feature based on 20-dimensional MFCC features extracted from complete songs. The classifier used by ME was SVMs with a KL divergence based kernel.

The artist identification task was next held at MIREX 2007 [MIREX 2007]. This time, only five teams participated. The best identification rate (48%) was achieved by the IMIRSEL M2K system developed for general-music retrieval. M2K used SVM as the classifier to identify the labels of incoming music objects [Downie et al. 2005a]. The second best system, developed by Mandel and Ellis, achieved an accuracy of 47%.

Table II summarizes the properties of previous identification methods. WHI, LIU, KIM, BA, BER, ZHANG, TSAI, BCE, and ME denote the identification methods published in Whitman et al. [2001], Liu and Huang [2002], Kim and

[4]From Magnatune.com.

Table II. Summary of State-of-the-Art Identification Methods' Properties

Identification Methods	Multifeature Integration	Size of Testbed	Vocal Based	Nonvocal Based	System Scalability
WHI	No	Small	No	No	Poor
LIU	No	Small	No	No	Poor
KIM	No	Small	Yes	No	Poor
BA	No	Small	No	No	Poor
BER	No	Small	Yes	No	Poor
ZHANG	No	Small	Yes	No	Poor
TSAI	No	Small	Yes	Yes	Poor
BCE	No	Small	No	No	Poor
ME	No	Small	No	No	Poor

Whitman [2002], Bartsch and Wakefield [2004], Berenzweig et al. [2002], Zhang [2003], Tsai et al. [2003],Tsai and Wang [2006], MIREX [2005], and MIREX [2005], respectively. As discussed above, these methods either use a single type of acoustic feature to represent music objects, or base singer identification on a speech-recognition approach. All of them have been tested only on small datasets, which makes it difficult to estimate their applicability to large real-life datasets. In addition, none of the above approaches adapts well to the addition of new classes of data (we call this property *scalability*). In all cases, when a new singer was added, the system has to be retrained, leading to high system reconstruction costs.

3. THE HYBRID SINGER IDENTIFIER (HSI) SYSTEM

In this section, we present the HSI method to facilitate automated singer recognition in large music databases. The architecture of the system, as illustrated in Figure 1, comprises two major component layers: a preprocessing module and a statistical singer modeling module. The major functionality of the preprocessing module is to separate an incoming song into vocal and nonvocal segments, and to extract audio features from those segments. The second layer contains a collection of statistical models, one for each singer. A statistical model for one singer consists of a series of Gaussian Mixture Models (GMMs), each constructed using one kind of acoustic feature. To identify a song, different feature vectors are first extracted from the vocal and nonvocal segments. The feature vectors are then fed into the statistical models, generating a set of likelihood values. The likelihood values generated by each model are combined, using a novel fusion scheme based on logistic regression, to form an overall relevance score. Finally, the query song is assigned to the singer with the highest overall relevance score. In the following subsections, we will give details of the modules and algorithms used in the system. The notation used in this article is defined in Table III.

3.1 Music Preprocessing—Vocal/Nonvocal Segmentation

In the first stage of the HSI identification process, vocal and nonvocal segments are identified and labeled via the preprocessing module. This process can be treated as a problem of vocal boundary detection and the detail steps are

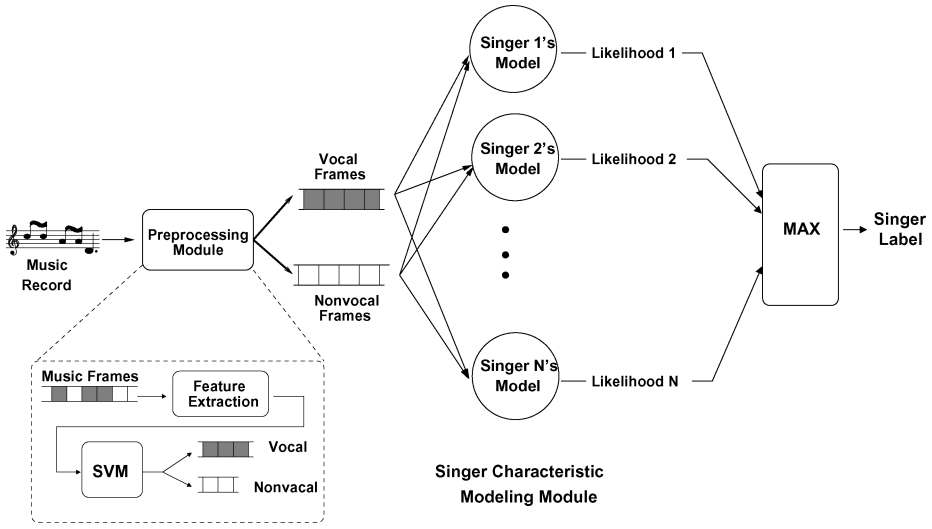


Fig. 1. Architecture of the HSI singer identification system.

Table III. Summary of Symbols and Definitions

Symbols	Definitions
S	Number of singers in the database
s	Notation of singer s
T	Number of blocks for segmenting input objects
F	Number of features extracted
M	Number of training examples for logistic fusion function
J	Number of mixture components in GMMs
C^s	Score combination function for singer s
v_{bf}	Feature vector extracted from block b for feature type f
V_f	Set of feature vectors extracted from different blocks for feature type f
L^s	Final score generated by logistic combination function for singer s
l^s_f	Likelihood value generated by category s's profile model using feature type f
W^s	Fusion weight vector of Logistic combination function for singer s

illustrated in Algorithm 1. We use a learning approach based on SVMs whose inputs are acoustic feature vectors. This approach is effective because there is a significant difference between the spectral features of segments containing vocal and instrument data and those containing only instrumental data.

Algorithm 1. Algorithm for preprocessing music.

Input : Song s, Frame Length l
Output : Vocal and non-vocal segments

1. Segment song s into frames with length l;
2. Calculate spectral features for each frame;
3. Classify music frames into two categories, vocal and non-vocal, with SVMs;
4. Return vocal and non-vocal segments;

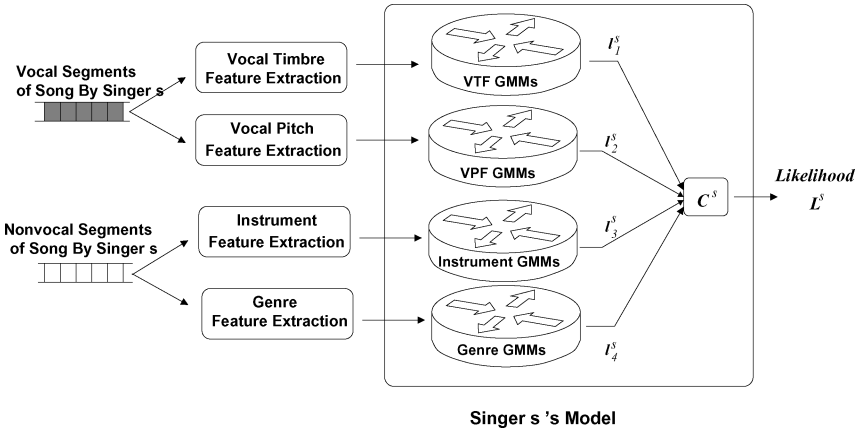


Fig. 2. Statistical singer modeling for singer s.

The preprocessing phase is similar to the approach in Lu et al. [2003], and consists of two subprocesses: feature extraction and classification with SVM. When a song comes in, it is divided into many short time frames of a predefined length (line 1). In our implementation, we set the length of frame to 0.5 s as this yielded the best performance in our experiments. The following acoustic features are calculated from each frame (line 2): MFCC features, spectral centroid, spectral flux, zero crossings, and low energy. These audio features are given as input to a SVM, which classifies each frame as containing vocals or not (line 3). We employ a SVM because it has been demonstrated to be effective on a range of categorization problems. The basic function of the SVM is that it can nonlinearly map the input features into a high-dimensional feature space; then a linear classifier is constructed to use optimal hyper planes to separate positive patterns and negative patterns with maximum margin. SVM performs well for binary classification, such as in this task. We have used the LIBSVM [Chang and Lin 2001] library in our work.

3.2 Multifeature Statistical Singer Modeling

In the second layer of our HSI system, there are multiple singer characteristic models, one for each singer. Each model is made up of two parts: feature extractors and multiple mixture models. There are specialized feature extractors that work on vocal and nonvocal frames. Each feature is fed into a specialized GMM, and the GMMs for all features are combined to build an overall model for one singer. The structure of this layer is presented in Figure 2 and we will describe each component in detail below.

3.2.1 *Feature Extraction.* To effectively represent the complex musical content of a specific artist's song, HSI extracts different features from both the vocal and nonvocal segments. Four different features are extracted: the vocal timbre feature (VTF), the vocal pitch feature (VPF), the genre-based feature (GBF),

and the instrument-based feature (IBF).[5] The VTF and VPF capture information from the vocal segments performed by the singer. The genre-based feature represents the music style. The instrument feature is used to model the characteristics of a typical instrument configuration for songs performed by the artist. This is motivated by research studies such as Xu et al. [2005] which indicate that a similar instrument configuration is found on most of the songs performed by a given singer. The details of the features used by the HSI framework are as follows:

—*Vocal timbre feature* (VTF). This feature conveys the timbre information of the vocal component. It is particularly important as the voice is a special instrument generated with flesh and bone and its timbre is unique for each singer. In this study, we extracted LPCCs (linear prediction-based cepstral coefficients) [Rabiner and Schafer 1978] from vocal segments to represent this information (LPCCs are linear prediction coefficients (LPCs) represented in the cepstrum domain). The advantage of an LPCC-based feature is that it provides a more consistent representation of an singer's vocal tract characteristics; the peaks of the vocal spectrum can be tracked by the envelop generated from the LPCCs. To compute the feature, HSI uses a FIR filter to preemphasize the audio input. The LPC analysis is carried out and then a recursion formula is applied to calculate cepstral coefficients based on the LPC parameters. Those parameters form the final VTF vector. The dimensionality of this feature vector is 16. In our current implementation, the order of the LPC analysis is 12.

—*Vocal pitch feature* (VPF). This feature describes the harmonic and structural information related to each singer's voice. The algorithm proposed by Tolonen and Karjalainen [2000] is used as the feature extractor and used to model pitch features in songs. The main advantages of this method is its computational efficiency and its effectiveness in capturing human auditory perception. With this method, the raw signal is first processed with a bandpass filter, where the lower and upper passband limits can have any values between 0 and 4500 Hz (the limits of the frequencies achievable by the human voice). Then, amplitude envelopes are extracted for different frequencies and summed to construct a pitch histogram which is used for describing prosodic features. We derive an 18-dimensional feature vector consisting of the amplitude and periods of the maximum six peaks in the histogram, a pitch interval between the six most prominent peaks, and the overall sums of the histograms. We use this feature to build the singer's prosodic model.

—*Genre-based feature* (GBF). This feature contains music information about genre. Because artists tend to perform in a limited range of genres (perhaps just one), this information helps to improve singer identification accuracy. In this study, genre information was summarized in Daubechies wavelet coefficient Histograms (DWCHs) [Li et al. 2003]. With DWCHs, local and global temporal information inside a music signal can be captured at the same time. To extract DWCHs, a sound file is treated as a kind of oscillation

[5]Note that our method can be easily extended to consider more acoustic features.

waveform in the time domain and can be considered as a two-dimensional entity of the amplitude over time, in the form of $M(t) = D(A, t)$, where A is the amplitude. It first uses wavelets to decompose the music signal into different subbands. Then, a histogram for each subband is constructed. Finally, the first three moments of each histogram and energy for each subband are calculated to form DWCHs. This is currently the state-of-the-art feature extraction technique for content-based genre classification. It produces a 40-dimensional feature vector.

—*Instrument-based feature* (IBF). This feature captures instrument configuration information for each song. Recent results on music understanding show that there is an association between the instrument configuration [Xu et al. 2005] and the singer (because singers tend to work with a particular set of backing instruments). Thus, an instrument-based feature ought to be helpful in enhancing singer identification effectiveness. In this study, MFCC features were used to represent information about instrument configuration. This is because MFCCs have been widely used to model timbre for instrument identification [Livshin and Rodet 2004]. To obtain IBF, we apply the logarithm of the amplitude spectrum based on a short-term Fourier transform on each signal frame. Then the frequencies are divided into 13 bins using Mel-frequency scaling. After taking the logarithm of the amplitude spectrum, the frequency bins are clustered and smoothed according to Mel-frequency scaling after conducting the logarithm on the amplitude spectrum. The final features are obtained by decorrelating the Mel-spectral vectors using a discrete cosine transform (DCT). This produces a 13-dimensional IBF vector.

3.2.2 *Statistical Singer Profiling with Mixture Models.* For the purpose of effective singer identification, HSI constructs a statistical model for each singer based on multiple features using GMMs. GMMs have recently received significant attention due to their superior performance for speech identification [Rabiner and Juang 1993]. In principle, a GMM combines the benefits of both the parametric and nonparametric density models. Like a parametric model, it employs a trainable model that does not require all the training data to make a classification. On the other hand, like a nonparametric model, GMMs have sufficiently high degrees of freedom to approximate any distribution with arbitrary accuracy, without expensive computation and storage demands. In addition, one of the main advantages of GMMs is that they are fast in terms of computational and training speed.

In our framework, the individual features of the music signal are extracted, and then individual profiling models for one singer are built up based on each feature.[6] The statistical singer profiling module of HSI aims to capture statistical properties of different features with finite mixture models. The probability of a singer label s can be modeled as a random variable drawn from a probability distribution for a certain feature type f. Given a parameter

[6]In this study, since four different features were extracted, the number of profiling models for each singer was four.

set Θ_f^s estimated based on feature f, it can be presented as a mixture of multivariate component densities:

$$P_f^s(V_f|\Theta_f^s) = \prod_{t=1}^{T} \left\{ \sum_{j=1}^{J} w_{fj}^s p_f^s\left(\boldsymbol{v}_{tf} | \boldsymbol{\mu}_{fj}^s, \boldsymbol{\Sigma}_{fj}^s\right) \right\}, \tag{1}$$

where $\boldsymbol{V}_f = \{\boldsymbol{v}_{1f}, \boldsymbol{v}_{2f}, ..., \boldsymbol{v}_{Tf}\}$ is a set of feature vector. Assume that Gaussian density is used as multivariate component in this study, according to GMM $\Theta_f^s = \{w_{fj}^s, \boldsymbol{\mu}_{fj}^s, \boldsymbol{\Sigma}_{fj}^s | \ where \ 1 < j < J\}$, where w_{fj}^s, $\boldsymbol{\mu}_{fj}^s$ and $\boldsymbol{\Sigma}_{fj}^s$ denote mixture weights, mean vectors, and covariance matrices, respectively. Also, $p_f^s(\boldsymbol{v}_{tf} | \boldsymbol{\mu}_{fj}^s, \boldsymbol{\Sigma}_{fj}^s)$ is the probability of a singer label s based on feature f extracted from segment t. Given data $\boldsymbol{v}_{tf}$, it can be easily calculated using Gaussian density function and associated parameters $\{\boldsymbol{\mu}_{fj}^s, \boldsymbol{\Sigma}_{fj}^s\}$.

The learning examples used to train GMMs are randomly selected from the original datasets and cover all subclasses. After the training process, we can obtain a set of model parameters for each class's GMMs and the likelihood value generated, based on feature type f for input feature vector $\boldsymbol{V}_f$, can be given as

$$l_f^s = log\left(P_f^s(\boldsymbol{V}_f|\Theta_f^s)\right) = \sum_{t=1}^{T} log\left(\left\{\sum_{j=1}^{J} w_{fj}^s p_f^s\left(\boldsymbol{v}_{tf} | \boldsymbol{\mu}_{fj}^s, \boldsymbol{\Sigma}_{fj}^s\right)\right\}\right) \tag{2}$$

and we can derive an overall likelihood value based on the features for singer s, expressed as

$$L^s = C^s\left(l_f^s, w_f^s\right), \tag{3}$$

where w_f^s is the combination weight and C^s is likelihood value combination function. L^s can be used to quantify the universal similarity distance between an input song and a singer label s.

3.2.3 *Model Selection in the Statistical Singer Profiling Module.* In HSI, the well-known Expectation Maximization (EM) algorithm is used to determine a set of model parameters. EM is a widely used standard algorithm for parameter estimation in statistics which uses an iterative hill-climbing procedure to drive the process of estimation. The goal is to derive an optimal parameter set Θ_f^s via a maximum likelihood estimation:

$$\left(\Theta_f^s\right)' = \underset{\Theta_f^s}{argmax} \ P_f^s\left(\left\{\boldsymbol{V}_f|\Theta_f^s\right\}\right). \tag{4}$$

In the first step, Θ_f^s is initialized with random values. Then, the value of the parameter set is reestimated in each iteration of the EM algorithm according to the following two steps: the *Expectation* step (E step) and the *Maximization* step (M step). The new model $\widehat{\Theta_f^s}$ is obtained with the auxiliary function in the E step:

$$\boldsymbol{Q}\left\{\Theta_f^s; \widehat{\Theta_f^s}\right\} = \prod_{t=1}^{T} \left\{ \sum_{j=1}^{J} w_{fj}^s p_f^s\left(j|\boldsymbol{v}_{tf}\Theta_f^s\right) log \ p_f^s\left(j, \boldsymbol{v}_{tf}|\widehat{\Theta_f^s}\right) \right\}, \tag{5}$$

where

$$p_f^s\left(j, \boldsymbol{v}_{tf}|\widehat{\Theta_f^s}\right) = \widehat{w_{fj}^s} p_f^s\left(\boldsymbol{v}_{tf}|\widehat{\boldsymbol{\mu}_{fj}^s}, \widehat{\boldsymbol{\Sigma}_{fj}^s}\right) \tag{6}$$

and

$$p_f^s\left(j, |\boldsymbol{v}_{tf}, \widehat{\Theta_f^s}\right) = \frac{w_{fj}^s p_f^s\left(\boldsymbol{v}_{tf}|\boldsymbol{\mu}_{fj}^s, \boldsymbol{\Sigma}_{fj}^s\right)}{\sum_{j=1}^J w_{fj}^s p_f^s\left(\boldsymbol{v}_{tf}|\boldsymbol{\mu}_{fj}^s, \boldsymbol{\Sigma}_{fj}^s\right)}. \tag{7}$$

For the M step, we update the parameter using the following estimation:

$$\widehat{w_{fj}^s} = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J} p_f^s\left(j|\boldsymbol{v}_{tf}, \Theta_f^s\right), \tag{8}$$

$$\widehat{\boldsymbol{\mu}_{fj}^s} = \frac{\sum_{t=1}^T \sum_{j=1}^J \left\{p_f^s\left(j|\boldsymbol{v}_{tf}, \Theta_f^s\right) v_{tfj}\right\}}{\sum_{t=1}^T \sum_{j=1}^J p_f^s\left(j|\boldsymbol{v}_{tf}, \Theta_f^s\right)}, \tag{9}$$

$$\widehat{\boldsymbol{\Sigma}_{fj}^s} = \frac{\sum_{t=1}^T \sum_{j=1}^J \left\{p_f^s\left(j|\boldsymbol{v}_{tf}, \Theta_f^s\right)\left(v_{tfj} - \widehat{\boldsymbol{\mu}_{fj}^s}\right)\left(v_{tfj} - \widehat{\boldsymbol{\mu}_{fj}^s}\right)\right\}}{\sum_{t=1}^T \sum_{j=1}^J p_f^s\left(j|\boldsymbol{v}_{tf}, \Theta_f^s\right)}. \tag{10}$$

The updating procedure is repeated until the log-likelihood value is increased by less than a predefined threshold from one iteration to the next. Since HSI considers four different features, the overall training procedure will be repeated four times, once for each feature.

3.3 Fusion Weight Estimation via Learning

To develop the statistical singer model for each singer in HSI, a logistic function is used as a combination function C^s to derive an overall likelihood score. Logistic functions have been widely used in the statistical and machine learning community and play an important role in one of the most popular statistical algorithms—logistic regression (LR) [Blum 1990; Jordan 1995; Vapnik 1998; Hastie et al. 2001]. The main reason for using LR to estimate parameters is that few statistical assumptions are required for its use and relatively it reqiures a low computational cost in terms of training. In addition, the output of logistic functions can be mapped into a probabilistic output in [0,1]. With logistic functions, Equation (3) can be reformulated as

$$L^s = C^s\left(l_f^s, w_f^s\right) = \frac{1}{1 + exp(-y_s \sum_{f=1}^F w_f^s l_f^s)}, \tag{11}$$

where $y_s = 1$ if this input object belongs to singer s, $y_s = -1$ otherwise, and w_f^s is the weight for singer s's likelihood value generated based on feature f. F is the size of input score and equals the number of features extracted in the first layer. L^s denotes the overall relevancy score—conditional probability of singer s. Based on Equation (11), the likelihood value occurring in the learning samples is

$$\prod_{m=1}^{M} \frac{1}{1 + exp(-y_s \sum_{f=1}^F w_f^s l_f^s)}, \tag{12}$$

Algorithm 2. Logistic regression based training algorithm to determine weights of score fusion for singer s.

Input: Matrix $MA \in [-1, 1]^{M \times F}$ where $MA_{mf} = y_m l_f^m$
N_p is number of the positive training examples
N_n is number of the negative training examples
Output: Weight vector $\vec{W}^s = (w_1^s, w_2^s \ldots, w_F^s)$

1. $\vec{W}^s = (0, 0, \ldots, 0)$ and $q_0 = (0.5, 0.5, \ldots, 0.5)$;
2. **for** $t = 1, 2 \ldots$ **do**
3. $\quad q_{t+1,m} = q_{t,m}\, exp(-\sum_{f=1}^{N} \delta_{t,f} MA_{mf})$;
4. $\quad$ **for** *each positive training example* **do**
5. $\quad\quad q_{t,m} = \frac{N_n q_{t,m}}{N_n + N_p}$;
6. $\quad$ **for** *each negative training example* **do**
7. $\quad\quad q_{t,m} = \frac{N_p q_{t,m}}{N_n + N_p}$;
8. $\quad j_t = \underset{f}{argmax} \left| \sum_{m=1}^{M} q_{t,m} MA_{mf} \right|$;
9. $\quad r_t = \sum_{m=1}^{M} q_{t,m} MA_{mf}$;
10. $\quad z_t = \sum_{m=1}^{M} q_{t,m}$;
11. $\quad a_t = \frac{1}{2} ln(\frac{z_t + r_t}{z_t - r_t})$;
12. $\quad \delta_{t,f} = \begin{cases} a_t & if \quad j = f_t \\ 0 & otherwise; \end{cases}$
13. $\quad$ update with $\vec{W}_{t+1}^s = \vec{W}_t^s + \vec{\delta}_t$;
14. Return $\vec{W}^s = \{w_1^s, w_2^s, \ldots . w_F^s\}$;

where M is the number of training examples. It is easy to see that the goal of the training process is to maximize the overall likelihood value. Thus, the goal is to "learn" $\vec{W}^s$ to minimize the log loss of the model, and the associated function can be denoted as

$$\sum_{m=1}^{M} ln \left(1 + exp \left(-y_s \sum_{f=1}^{F} w_f^s \, l_f^s \right) \right). \tag{13}$$

To achieve this, we apply a modified version of the sequential-update optimization algorithm proposed by Collins et al. [2000].[7] The pseudocode for our algorithm is given in Algorithm 2. The algorithm is equivalent to the AdaBoost scheme [Freund and Schapire 1997; Lebanon and Lafferty 2001]. The basic principle is that, on each iteration t during the training, the algorithm updates the distribution $q_{t,m}$ to increase the weights of misclassified training examples (lines 14–15). In our implementation, the algorithm counts the distribution between positive and negative learning examples. We revise the algorithm to give $q_{t,m}$ weight (lines 4–9). The distribution weights can be computed with $q_{t,m} = \frac{N_n q_{t,m}}{N_n + N_p}$ for positive examples and $q_{t,m} = \frac{N_p q_{t,m}}{N_n + N_p}$ for negative examples.

[7]For a detailed derivation, please refer to Collins et al. [2000].

Algorithm 3. Algorithm for automatic singer identification.

Input : A song with unknown singer s
Output : Singer label of the song
1. Process the song to get vocal and non-vocal segments;
2. Feature extraction;
3. Derive relevance scores using statistical singer profiling module;
4. Return singer label with Equation 14;

3.4 Identification Process with HSI

The goal of the HSI is to identify the singer of a song presented only as audio data (i.e., no metadata annotations). We now describe the entire process, using the architecture introduced previously. We start with a training database where we know the singers of all songs in the database. In the training stage, we process the songs in the database, and extract features for each singer. According to the characteristics of the singers, we generate the parameters for the multifeature statistical modeling module. Then, we use Algorithm 2 to determine the fusion weights for combining the scores.

After the training phase, we can conduct the task of singer identification. The basic procedure is shown in Algorithm 3 and consists of four steps. For a given music item, at the initial stage of the process, the system partitions the song into vocal and non-vocal segments (line 1). After that, the feature extraction procedure generates four different kinds of features using the techniques described in Section 3.2.1. Next, the features are fed into the statistical modules for individual singers in the second layer of the HSI. The likelihood score based on a particular feature can be generated based on Equation (2). Then, relevance scores, one from each singer statistical model, can be calculated with Equation (11) (line 3). Those scores quantify the similarity between the incoming song and the singer label. In the final step, a singer's label for the song can be assigned based on those relevance scores.

$$s^* = \underset{1 \geq s \geq S}{argmax}\ L^s. \tag{14}$$

3.5 Major Advantages of HSI

The HSI architecture presented in the previous sections enjoys several advantages over the other competing approaches:

—*Comprehensiveness*. In our singer characteristic model, several kinds of features, extracted from both vocal and nonvocal segment, are taken into consideration. The combination of those features enables us to summarize the music content from each artist more precisely.

—*Effectiveness*. The multifeature statistical singer characteristic model, in conjunction with the probabilistic likelihood fusion scheme with logistic function, enables us to capture effects of different features more effectively. Consequently this improves the final retrieval/query accuracy and robustness of

the whole framework. The experimental results further verify that the system with the decision model scheme is more robust against training flaws and raw training example problems.

—*Scalability*. The statistical singer characteristic model for each artist is constructed independently. Thus, the whole system does not need to be rebuilt when music items performed by new singers are integrated into the database. Therefore, the system naturally has superior scalability over other schemes. Moreover, there is significant potential for parallelism in the identification process.

—*Efficiency*. Another advantage of HSI is its simplicity. This characteristic directly leads to faster identification of singer labels. In addition, the layering architecture leads to a great saving on process and maintenance cost. Furthermore, the approach is relatively easy to implement.

In next section, we will demonstrate those advantages empirically. Certainly, the advantages of this approach need to be balanced against the potential problem of mis-labelling. As our experimental results show, HSI is more robust against this problem compared to other approaches.

4. AN EXPERIMENTAL STUDY

This section presents an experimental study to evaluate the proposed HSI technique. First, we give the experimental configuration including test datasets and the performance metrics to evaluate different methods. Then, we present an experimental study to examine the performance improvement of HSI over several existing schemes, namely, TSAI, BER, LIU, BCE, and ME. The results demonstrate the superiority of HSI over the current best approaches in the areas of accuracy of singer identification, scalability to accommodate different sizes of data, robustness against various kinds of noise, and efficiency in terms of the response time.

4.1 Experimental Configuration

In this section, we present the experimental settings for the performance evaluation, including competitive systems, testing datasets, and performance metrics. All methods have been implemented and tested on a Pentium III, 450-MHz PC running the Linux operating system [Shen et al. 2006].

4.1.1 *Data Sets.* We used four datasets for the our experimental study. Dataset I contained 230 songs from 13 female and 10 male artists with 10 songs per artist. The dataset has been used in Tsai et al. [2003] and Tsai and Wang [2006]. Dataset II consisted of 8000 songs covering 90 different singers. This dataset was constructed from a CD collection of the authors. It included 45 male singers (such as Van Morrison, Michael Jackson, Elton John, Michael Bolton, etc) and 45 female singers (such as Kylie Minogue, Madonna, Jennifer Lopez, etc). The length of each music item in the two datasets was set as 30 s and there was no overlap between the two datasets. For both datasets, the sound files were converted to 22,050-Hz, 16-bit, mono audio files. For singer

identification, we used 20% of each dataset for training purposes and the remaining songs to evaluate the performance of all the schemes studied. Datasets III and IV were the Magnatune and USPOP collections used in MIREX 2005 artist identification contest.

4.1.2 *Evaluation Metrics.* As discussed above, the main goal of the system is to identify the artist who performs an incoming query song. Thus, our evaluation method focuses on how accurate the identification process is with different approaches for a particular database. We used the *accuracy* as the metric for evaluation:

$$Accuracy = \frac{N_C}{N}, \tag{15}$$

where N_C is the number of songs correctly identified and N is the number of songs used in the evaluation. We also measured the average response time to evaluate the efficiency of the different techniques:

$$AvgResponse = \frac{Total\ Query\ Response\ Time}{N}, \tag{16}$$

where *Total Query Response Time* is the total time required for the system to identify all N songs used in the evaluation. It represents the average time required for identifying a single query song. A lower *AvgResponse* is preferred as it implies faster identification process, and hence better query efficiency.

4.2 Experimental Results and Analysis

In this section, we compare HSI with five existing methods including BER, LIU, TSAI, BCE, and ME. The notations defined in Section 2 are applied to describe the methods. To ensure a fair comparison, we selected the same set of data for training all systems, and used the rest of the data for performance evaluation. In addition, the experimental results presented below were obtained based on GMMs without tuning the parameters. In Section 4.2.3, we analyze the influence of parameter tuning.

4.2.1 *On Vocal/Nonvocal Segmentation.* Accurate vocal/nonvocal segmentation is important to system performance. In the first experiment, we evaluated the accuracy of our scheme in identifying vocal and nonvocal segments. The method was trained using both vocal and nonvocal segments from the datasets described in Section 4.1.1. The size of the training data was 20% of the original test collections. We evaluated the classification performance on the basis of frame classification accuracy. Table IV shows the confusion matrix of the vocal/nonvocal segmentation. In this table, the rows indicate the ground-truth of the segments and the columns correspond to the hypotheses. The results show that a major part of missegmentation was due to errors of identifying vocal segments. Throughout our study, we also found that nearly 90% of the misidentified vocal segments contained relatively loud background music or other noise.

Table IV. Confusion Matrix of the Vocal/Nonvocal Segmentation

Actual	Hypothesized (%)			
	Dataset I		Dataset II	
	Vocal	Nonvocal	Vocal	Nonvocal
Vocal	92.4	7.6	87.1	12.9
Nonvocal	6.6	93.4	10.3	89.7

Actual	Hypothesized (%)			
	Dataset III		Dataset IV	
	Vocal	Nonvocal	Vocal	Nonvocal
Vocal	93.5	6.5	92.1	7.9
Nonvocal	9.6	90.4	8.3	91.7

Table V. Identification Accuracy Comparison (Results are given by five-fold cross-validation. HSI-L denotes linear combination of likelihood score with same weight and HSI-V denotes HSI only considering vocal component.)

Singer Identification Methods	Identification Accuracy(%)							
	Dataset I			Dataset II			Dataset III	Dataset IV
	Female	Male	Ave.	Female	Male	Ave.		
HSI	**86.3**	**88.3**	**87.3**	**77.0**	**75.2**	**76.1**	**89.9**	**84.2**
HSI-L	78.2	80.2	79.2	69.8	70.4	70.1	85.9	80.2
HSI-V	76.2	79.3	77.2	68.5	68.5	68.5	82.4	79.4
TSAI	73.0	71.4	72.2	61.0	63.4	62.2	79.5	75.2
LIU	66.5	66.1	66.3	55.2	56.4	55.8	77.3	70.9
BER	65.3	65.3	65.3	56.2	55.4	55.8	74.5	69.2
BCE	66.2	66.1	66.1	56.8	56.2	56.5	77.6	67.8
ME	65.8	65.6	65.7	55.3	55.1	55.2	76.7	68.2

4.2.2 *On Identification Accuracy.* In this section, we describe a comparative study on the accuracy of the various singer identification schemes. Table V summarizes the results for the four datasets. The bottom four rows show how the BER, LIU, BCE, and ME performed. It is worth noting that, since both TSAI and HSI are based on GMM, their performance is sensitive to parameter settings. In this experiment, we randomly initialized the GMM parameters. Parameter tuning and its impact on performance is studied in Section 4.2.3. The experiments show that HSI and TSAI were the two most accurate methods, although TSAI was consistently less accurate than HSI. This may be explained by the fact that TSAI only considers MFCC-based acoustic characteristics inside the vocal segments, while HSI takes more acoustic characteristics into account and considers both vocal and nonvocal segments.

Overall, the experimental results show that HSI significantly outperformed other approaches. For example, Table VI shows that, compared to TSAI, the HSI method improved the identification precision from 72.5% to 87.3% for dataset I, 62.1% to 76.2% for dataset II, 79.5% to 89.9% for dataset III, and 75.2% to 84.2% for dataset IV. While the improvement over TSAI was significant, the imporvement over LIU and the other methods was even more substantial. On average, around 30% improvement can be observed for the four datasets. To enhance the stability and robustness of the empirical study, we also validated the approach using K-fold cross-validation with K set to 5. This showd a similar level of improvement; the results are given in Table V.

Table VI. Identification Accuracy Comparison

Singer Identification Methods	Identification Accuracy(%)							
	Dataset I			Dataset II			Dataset III	Dataset IV
	Female	Male	Ave.	Female	Male	Ave.		
HSI	**86.4**	**88.2**	**87.3**	**77.4**	**75.0**	**76.2**	**89.9**	**84.2**
TSAI	73.3	71.3	72.5	61.1	63.1	62.1	79.5	75.2
LIU	66.4	66.0	66.2	55.2	56.4	55.8	77.3	70.4
BER	65.4	65.2	65.3	56.0	55.4	56.7	74.5	69.2
BCE	66.2	66.1	66.1	56.8	56.2	56.5	77.3	59.9
ME	65.8	65.6	65.7	55.3	55.7	55.5	76.6	68.3

Table VII. Factors Affecting HSI Accuracy

Singer Identification Methods	Identification Accuracy(%)							
	Dataset I			Dataset II			Dataset III	Dataset IV
	Female	Male	Ave.	Female	Male	Ave.		
HSI	**86.4**	**88.2**	**87.3**	**77.4**	**75.0**	**76.2**	**89.9**	**84.2**
HSI-LIR	80.2	82.2	81.2	72.7	71.5	72.1	86.9	81.0
HSI-L	78.2	80.2	79.2	69.7	70.5	70.1	85.9	80.2
HSI-V	76.2	79.3	77.2	68.7	68.9	68.8	82.4	79.4

HSI has two advantages over the other competitive schemes. First, HSI extracts both vocal and nonvocal features from songs. The use of multiple features from both vocal and nonvocal components can result in more comprehensive statistical models for songs by a particular singer and hence result in better identification effectiveness. Second, the weights for fusing likelihood scores derived via logistic regression can capture joint effects among various acoustic characteristics. This naturally raises the question of how much each of these factors contributes toward improving HSI's accuracy. In a second experiment, we examined how the accuracy of HSI was affected by these factors. Several variations on HSI were developed and evaluated using the same test data as above; the results are given in Table VII. The HSI-LIR variation uses linear regression, rather than logistic regression, to estimate fusion weights. HSI-L is even simpler, and uses linear fusion weights. HSI-V uses only the vocal segments and the features extracted from these to build singer characteristic models. As expected, the nonlinear likelihood value fusion weights generation scheme presented in Section 3.3 plays an important role in the whole identification procedure and can bring significant improvement in identification accuracy. Table VII shows that HSI based on logistic regression is significantly more accurate than HSI using linear weights or linear regression to estimate weights. Table VII shows that the use of nonvocal information in addition to vocal information also leads to a significant improvement in accuracy. This strengthens the claim that nonvocal information helps to improve identification effectiveness.

4.2.3 *On GMM Parameter Tuning.* Gaussian Mixture Models (GMMs) are among the most statistically popular methods for data modeling. Each GMM specifies the number of mixture components J that affects how well the model can yield a concise, accurate data representation for a given input. Ideally, the number of mixture components corresponds to the number of groups present

Table VIII. Effect of GMM Parameter Tuning

Singer Identification Methods	Identification Accuracy (%)							
	Dataset I			Dataset II			Dataset III	Dataset IV
	Female	Male	Ave.	Female	Male	Ave.		
HSI-T	91.2	93.2	92.2	81.7	83.5	82.6	92.4	89.5
HSI	86.4	88.2	87.3	77.4	75.0	76.2	89.9	84.2
TSAI-T	89.1	90.3	89.7	72.1	73.1	72.6	83.4	80.4
TSAI	73.3	71.3	72.5	61.1	63.1	62.1	79.5	75.2

in the input. However, a larger J leads to more expensive computation, and so there is a tradeoff between performance and accurate modeling. For this study, we applied the minimum description length (MDL) principle as a criterion for the selection of J [Rissanen 1978], adopting an idea that has been used for still-image processing [Carson et al. 2002; Greenspan et al. 2001]. The goal of our procedure is to pick a J that maximizes the following equation:

$$logL\left(\Theta^{s}_{f\,ML}|\boldsymbol{V}_f\right) - \tfrac{l_k}{2} logN, \tag{17}$$

where $\Theta^{s}_{f\,ML}$ is the parameter set for a J-mixtures GMM, L is the likelihood function, and l_k is the number of free parameters for a model containing j mixture components. For a Gaussian mixture having full covariance matrices, we can have

$$l_k = (j-1) + jd + j\tfrac{d(d+1)}{2}. \tag{18}$$

Using the above principle, we may find two models with different j values but with the same data modeling quality. In this case, the simpler model will be selected. Based on our experimental results, the value of J can range from 2 to 7.

After introducing the basic principle on the parameter tuning in our HSI, now we proceed to study the effect of this procedure empirically. Our basic methodology is to compare the accuracy improvement of the GMM based approaches due to the parameter tuning. Table VIII shows how HSI and TSAI perform if parameter tuning is considered (HSI and TSAI are the untuned versions of the methods; HSI-T and TSAI-T are versions with tuning applied). The performance gain observed in this experiment verifies the effectiveness of parameter tuning (between 5% and 26% for two methods). However, it is clear that even after tuning TSAI's identification accuracy was still significantly lower than HSI's on large datesets. For TSAI, the identification accuracy after the optimization was very close to results achieved by Tsai and Wang [2006]. Also, TSAI's performance improvement was much higher than that achieved by HSI. Similarly to the reason given in Section 4.2.2, this was due to the effect of the logistic regression-based decision module. On the other hand, it also implies that TSAI is more sensitive to the optimization procedure.

4.2.4 *On Query Efficiency.* For large music databases, response time is an important aspect of system performance. Although, as shown in the last section, the statistical singer modeling module and extra decision module in the HSI improve the accuracy, they might introduce query cost overhead. In this experiment, we showed how the extra mechanisms in HSI affect its performance relative to other approaches.

Table IX. Identification Efficiency Comparison

Singer Identification Methods	Query Time(s)							
	Dataset I			Dataset II			Dataset III	Dataset IV
	Female	Male	Ave.	Female	Male	Ave.		
HSI	**0.269**	**0.241**	**0.255**	**1.199**	**1.112**	**1.156**	**0.294**	**0.304**
TSAI	0.277	0.213	0.235	1.731	1.631	1.681	0.318	0.320
LIU	0.216	0.226	0.221	1.780	1.764	1.772	0.305	0.301
BER	0.227	0.235	0.231	1.756	1.804	1.775	0.311	0.321
BCE	0.217	0.227	0.222	1.734	1.744	1.739	0.313	0.309
ME	0.227	0.235	0.231	1.722	1.794	1.758	0.321	0.311

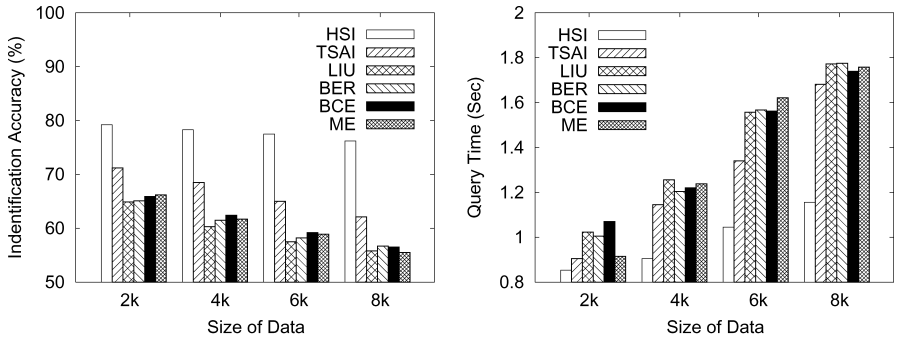
Table IX shows the total response time for the different singer identification schemes over the four datasets for the same set of tests used in Section 4.2.2. From the experimental results summarized in the table, we can see that, for the small test collections (datasets I, III and IV), HSI has similar efficiency compared to the other methods. However, in the evaluation on the large data set (dataset II), HSI was significantly faster than the other five methods. The identification tests on dataset II (containing 8000 songs) required between 1.6 and 1.8 s for BER, LIU, TSAI, BCE, and ME. In contrast, HSI required between 1.1 and 1.2 s for the same set of tests (around a 30% improvement). Another interesting observation is that, unlike the other five approaches, the proportional increase in response time for HSI between the different sizes of data was much smaller than the other methods.

4.2.5 *On Scalability Comparison.* Scalability is particularly important for large music information systems, because such systems can potentially contain thousands of songs. As the number of music items increases, the performance of a system may degrade due to noise and to the presence of more similar songs in the database. Another important aspect of scalability is the cost of incrementally adding new music items to an existing music database. In this section, we examine the behavior of our scheme as the data set changes. We evaluate other approaches against HSI using (1) datasets containing different numbers of singers) and (2) datasets containing a different numbers of songs.

In the first experiment, we compared the cost of "upgrading" the system as new artists are gradually added. We started with one singer, and then added one singer at a time up to 25 singers. The subset of singers and the order of singer insertion was chosen randomly, but the same order was used for all systems. We measured the cost of upgrading the classifier after adding each new singer, and we present the upgrade costs at various points in the process (e.g., going from four to five singers, from 9 to 10, and so on). Table X gives the results for this experiment. The results show that, compared to other methods, HSI consumed much less construction time. One thing worth noting is that when the number of singers was fewer than five, all other methods used less time to complete the construction than HSI did. This was because, as well as building GMMs for the new singer, HSI's construction cost also includes training time for relative LR analysis. This overhead makes HSI less efficient in terms of construction cost when the number of classes is small. From Table X, we also observe that there was no significant increase in reconstruction time when the system included

Table X. Construction Time Comparison on Dataset II—Different Singer Numbers

Singer Number	Identification Methods (s)					
	HSI	TSAI	LIU	BER	BCE	ME
1	**1395**	1255	1289	1390	1360	1259
5	**1409**	1395	1408	1500	1590	1765
10	**1392**	1800	2051	2280	2170	2234
15	**1402**	2301	2450	2756	2344	2654
20	**1410**	2998	3154	3572	3245	3498
25	**1410**	3387	3754	4129	3899	4172



(a) Query accuracy (b) Query time

Fig. 3. Scalability comparison on dataset II—different sizes of data.

more object singers. The main reason is that, with the HSI approach, only one associated modeling structure needs to be built when a new singer is integrated into the database. This advantage can lead to great saving A on reconstruction time. In contrast, all other methods need to be rebuilt entirely after a new singer (i.e., a new classification class) is added into the database.

In the second experiment, we examined the effect on query accuracy and query cost against the size of the database (total number of songs). As the number of stored items increases, we might expect the performance of the system to degrade; the query cost will most likely slow, and the accuracy will most likely decrease because there are more similar songs in the system. In order to carry out this measurement, we created four different-sized databases by randomly selecting 2000, 4000, 6000, and 8000 songs from dataset II. Twenty percent of the data was used for training and the rest for testing (i.e., there was no overlap between training and testing sets). Figure 3 shows the experimental results for all the systems. We observe that the accuracy and query times for both LIU and BER degraded significantly as the data size increased. This is because the larger data sizes affect the performance of the machine learning-based classifiers in those systems. While TSAI achieved better scalability, the improvement was rather limited. Compared to the other approaches, HSI's performance was relatively robust against the volume of data. There was no dramatic decrease in accuracy or increase in query processing time with larger datasets, for example, around 79% accuracy for a large size of 2000 songs, which was only 4% higher than the same system with 8000 songs. The main reason is that HSI

is constructed using multiple feature from the songs and this enables HSI to capture the song characteristics more precisely. Furthermore, the query time cost increases relatively slowly against the increasing size of the database.

4.2.6 *On Robustness Comparison.* Real-world applications often require singer identification under less than ideal conditions. For example, music data used for the query may have been recorded live and may contain noise. Or perhaps the system was trained against a nonrepresentative sample of works by a particular artist. In this section, the robustness of our proposed HSI system is demonstrated by comparing it against the robustness of other approaches.

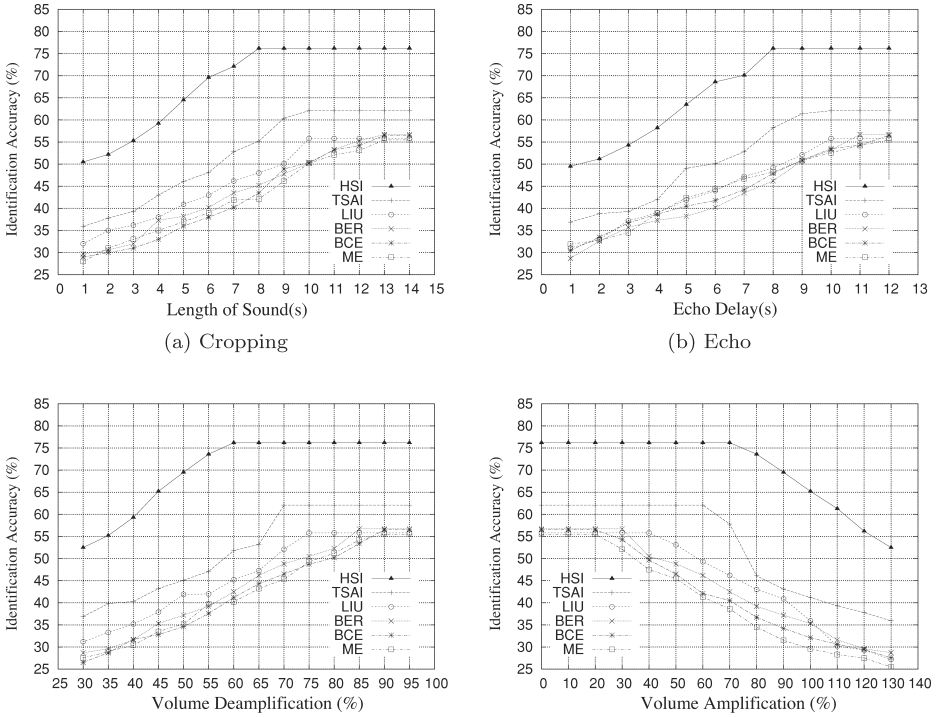
4.2.6.1 *Robustness Against Audio Alternatives.* The human auditory system has a very refined ability to identify particular sounds or music, even in the presence of moderate amounts of noise and/or distortion. In real-world applications, music retrieval systems might also have to deal with less than perfect samples of music data, either as stored items or as queries. In this section, to study the robustness of the different singer identification techniques, we examine how the accuracy of retrieval was affected by distortion in the query music data.

During the evaluation, we ran the same set of tests on dataset II as in the earlier accuracy experiment. However, we applied various kinds of distortion to each query song and compared the accuracy of the distorted queries with the accuracy of the nondistorted queries. Figures 4 and 5 summarize the accuracy of the six different systems in the presence of distortion. The performance on dataset I shows a similar trend but we omit the results for that experiment to save space.

The experimental results clearly demonstrate that HSI was the most robust technique. It performed significantly better than the competitors on all distortion cases. For example, HSI was robust to echo with an 8-s delay, 70% volume amplification, 60% volume deamplification, 8-s cropping, and 45-dB SNR white background noise on average.[8] In contrast, TSAI can only tolerate echo with a 10-s delay, 55-dB SNR white background noise, 10-s cropping, 60% volume amplification, and 70% volume deamplification. Thus, we can conclude that HSI is fairly robust to different levels of noise and acoustic distortion.

4.2.6.2 *Robustness Against Segmentation Length.* The first step of a typical music information retrieval system is to divide the incoming music data into segments for further processing via a sliding window. The segment length has the potential to affect the final effectiveness of the systems substantially. Shorter frames can contain more precise information (e.g., a sample of solo voice). On the other hand, having short frames (and thus more frames) results in higher processing time and potentially more expensive storage costs. Longer frames have a higher chance to contain both vocal and nonvocal components,

[8]The equation $SNR_{dB} = 10log_{10}\frac{S}{N}$ was used to calculate the signal-to-noise ratio, where S is the signal power, and N is the noise power in decibels.



(a) Cropping (b) Echo

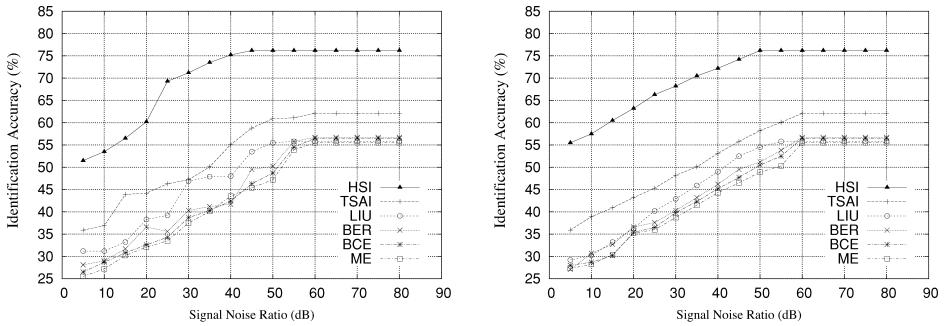
(c) Volume deamplification (d) Volume amplification

Fig. 4. Robustness comparison of singer identification methods on dataset II—different kinds of audio alternativs.

which make them less useful for capturing precise information for artist recognition.

In this study, we investigated the effects of segment lengths on the accuracy of different systems. We tested six different segment lengths from 0.1 to 1.1 s with the two datasets. Figure 6 presents identification accuracy of the systems as a function of segment length. Since ME calculates acoustic features based on an entire piece of music, the method was not included in this study. The figure shows that changes in segment length have a substantial impact on the accuracy of LIU, TSAI, BER, and BCE. The effectiveness of all those methods decreased significantly when the length of music frame increased. However, we did not observe such trends in our approach, suggesting that HSI's performance is more resilient to changes in segment length.

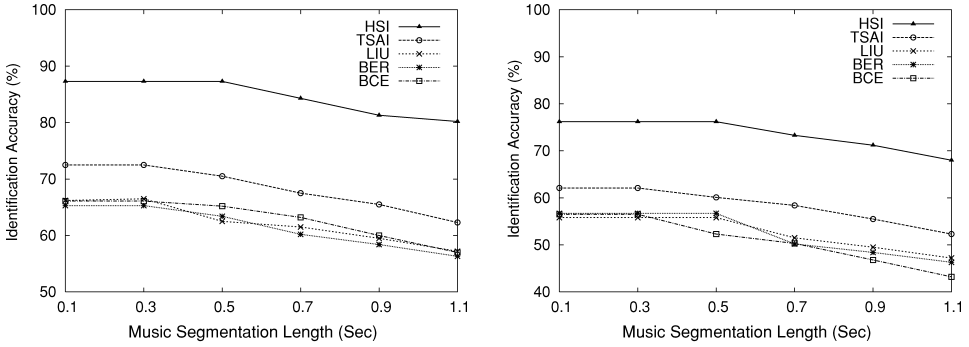
4.2.6.3 *Robustness Against Various Training Conditions.* Learning-based approaches are typically required to work under resource constraints. For example, since training can be a costly exercise, training resources can be reduced by using fewer training examples, or lower-quality training examples. On the other hand, making such sacrifices may lead to a less effective classifier. In this section, we compare HSI and other systems on their robustness under various



(a) Laplace noise addition (b) Gaussian noise addition

(c) White noise addition (d) Median noise addition

Fig. 5. Robustness comparison of singer identification methods on dataset II—different kinds of noise.



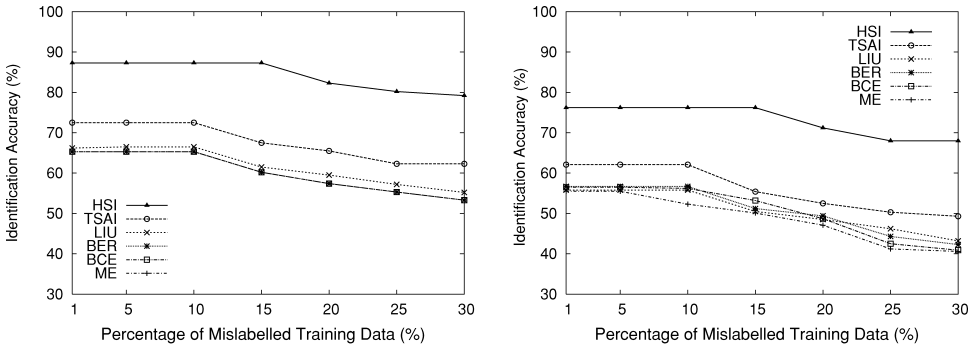
(a) Dataset I (b) Dataset II

Fig. 6. Robustness of different schemes against various segment lengths.

training conditions. These cases include the following:

—*Mislabeled training examples*. As the most crucial resource, training datasets labeled by human could contain mislabeled training examples. The system should be robust to this kind of error.

—*Minimal positive training examples*. Training example selection is expensive since it relies on a manual process. If the system can perform well with a



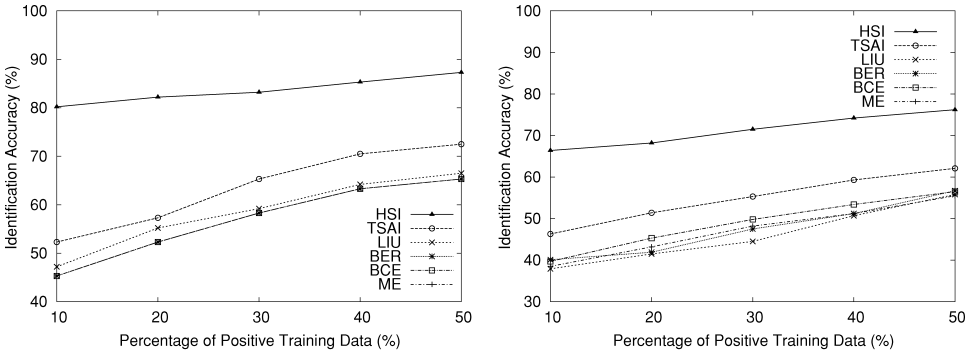
(a) Dataset I (b) Dataset II

Fig. 7. Robustness of different schemes against various sizes of mislabeled training examples.

small number of positive training examples, this will significantly reduce the training cost.

In the first experiment, the effects of mislabeled training examples were studied. To carry out experiments under different sizes of incorrect training data, 1%, 5%, 10%, 15%, 20%, 25%, and 30% training data from both datasets were randomly selected and their original labels were reversed. Figure 7 summarizes the precision rate versus the different noise settings on the various methods. The results show that HSI was superior to the other approaches when the proportion of mislabeled data was increased. As can be seen, the performance of the five other approaches degraded dramatically after the size of the mislabeled data was greater than 10% of the whole training set. In contrast, HSI maintained reasonable accuracy even with 15% incorrect training data of dataset II. From the above, we can easily see that, by taking advantage of the decision model and comprehensive singer modeling scheme, our scheme was able to achieve better robustness against mislabeled training data.

Positive training example sets are typically quite small in real-life applications. In the second experiment, we investigated the effects of training set size on the accuracy of HSI and other methods. To make the experimental results more stable, data from different categories was chosen uniformly from our data collections. Different portions of the positive examples were randomly selected, and then how HSI and other approaches performed with changes to the amount of positive training data (from 10% to 50%) was studied. From the results shown in Figure 8, we observe that the performance of all methods degraded when the size of the positive training examples was decreased to a certain threshold. However, compared to the other approaches, HSI emerged as the more robust technique when relatively small amounts of training examples were available. The superior robustness of HSI was due to the fact that the GMM based on multiple features can capture more information about objects from a particular singer, and the decision module with a logistic-based score fusion function can rectify possible misclassifications. The results corroborate the conclusions from the previous experiments that (1) HSI is a novel identification technique with good robustness against different constraint learning conditions, and (2) the



(a) Dataset I (b) Dataset II

Fig. 8. Robustness of different schemes against various sizes of positive training examples.

Table XI. "Album Effect" for Different Singer Identification Methods

Singer Identification Methods	Identification Accuracy(%)					
	Dataset II(A)			Dataset II		
	Female	Male	Ave.	Female	Male	Ave.
HSI	69.4	69.2	69.3	77.4	75.0	76.2
TSAI	53.3	51.3	52.5	61.1	63.1	62.1
LIU	46.4	46.0	46.2	55.2	56.4	55.8
BER	45.4	45.2	45.3	56.0	55.4	56.7
BCE	46.2	46.4	46.3	56.8	56.2	56.5
ME	45.8	45.6	45.7	55.3	55.7	55.5

decision module in conjunction with the hybrid architecture provides superior query effectiveness.

4.2.7 *On the Album Effect.* As mentioned earlier, the "album effect" can have a great impact on the performance of the singer identification process. The main cause of this problem is that the classifier actually learns the mastering and production process rather than the artist's voice. Consequently, it is the album's "style" rather than the singer that is being identified by the classifier. The performance of different systems degrades substantially when the songs used in training and test are from different albums. To study the album effect on different identification methods, we reorganized dataset II to obtain a new version—dataset II(A). In dataset II(A), for songs performed by the same singer, we made sure that the songs in the training dataset came from different albums. Similarly, test cases for a given singer were drawn from different albums. We compared the results obtained using dataset II and dataset II(A).

The experimental results are summarized in Table XI and confirm the findings of other researchers. As expected, the identification accuracies based on dataset II(A) were lower than those of dataset II for the all identification methods. However, we found that HSI appeared to be the least affected by the "album effect." For HSI, the difference between identification accuracies obtained using two datasets was 6.9%, which is less than the accuracy decrease for all other methods.

5. CONCLUSIONS AND FUTURE WORK

In recent years, the emergence and maturity of network and data storage technologies have made a significant amount of music data available in digital form. Content-based music retrieval has gained considerable momentum as a means of managing and accessing large music datasets. Although singer identification has received a large amount of research attention, traditional techniques have three basic impediments when applied in real-life applications: (i) poor scalability and expensive reconstruction cost; (ii) lack of comprehensive evaluation results based on large scale datasets; and (iii) low query accuracy.

Motivated by these concerns, we developed a novel framework, called HSI, to facilitate effective singer identification in large music databases. The system has been fully implemented and tested with different datasets. As shown in our experimental evaluation, the HSI system not only has significantly better effectiveness, scalability, and efficiency over the state-of-the-art systems, but also achieves significantly better robustness against various kinds of acoustic distortion. In addition, HSI enjoys less sensitivity to segment length and mislabeled training examples than do previous approaches.

These improvements are accomplished by the following:

—A novel singer identification framework based on the multiple feature integration and a likelihood fusion scheme using a logistic regression function.

—A novel singer characteristic modeling method based on stochastic models trained with the learning samples.

—A layered system architecture for the seamless combination of the two components—singer characteristic models and a score fusion component with superior scalability and efficiency.

In summary, the HSI framework is an effective, scalable, and robust solution for the singer identification problem. Despite the current success of HSI, there are still further directions for investigation. We plan to evaluate the framework on a larger dataset and develop advanced acoustic feature extraction methods to further improve accuracy and robustness. We will examine novel indexing structures to reduce the overall query processing cost and develop an analytic model for predicting query costs. Another promising research direction is to extend the current approach to indexing other kinds of multimedia data.

ACKNOWLEDGMENTS

We would like to thank Professor Wei-Ho Tsai at Taipei University of Technology, for kindly sharing his dataset and codes with us. Also special thanks are due to Professor C. J. Keith van Rijsbergen in the Department of Computing Science at the University of Glasgow for his valuable advice.

REFERENCES

BARTSCH, M. AND WAKEFIELD, G. 2004. Singing voice identification using spectral envelop estimation. *IEEE Trans. Speech Aud. Process. 12*, 100–109.

BECCHETTI, C., RICOTTI, L., AND RICOTTI, L. 1999. *Speech Recognition*. John Wiley, New York, NY.

BERENZWEIG, A., ELLIS, D. P. W., AND LAWRENCE, S. 2002. Using voice segments to improve artist classification of music. In *Proceedings of the AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*. 119–122.

BERENZWEIG, A., LOGAN, B., ELLIS, D., AND WHITMAN, B. 2004. A large-scale evaluation of acoustic and subjective music-similarity measures. *Comput. Mus. J. 28*, 63–76.

BERENZWEIG, A. L. AND ELLIS, D. P. W. 2001. Locating singing voice segments within music signals. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 119–122.

BLUM, A. 1990. Learning Boolean functions in an infinite attribute space. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing* (STOC'90). 64–72.

CARSON, C., BELONGIE, S., GREENSPAN, H., AND MALIK, J. 2002. Blobworld:image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. Patt. Anal. Mach. Intell. 24*, 8, 1026–1038.

CHANG, C.-C. AND LIN, C.-J. 2001. LIBSVM: A library for support vector machines. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

COLLINS, M., SCHAPIRE, R. E., AND SINGER, Y. 2000. Logistic regression, Adaboost and Bregman distances. In *Proceedings of the 13rd Annual Conference on Computational Learning Theory* (COLT'00). 158–169.

DOWNIE, J., EHMANN, A., AND HU, X. 2005a. Music-to-knowledge (M2K): A prototyping and evaluation environment for music digital library research. In *Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries* (JCDL). 376.

DOWNIE, J. S. 2006. The Music Information Retrieval Evaluation Exchange (MIREX). *D-Lib Mag. 12*, 12 (Dec.)

DOWNIE, J. S., WEST, K., EHMANN, A., AND VINCENT, E. 2005b. The 2005 Music Information Retrieval Evaluation Exchange (MIREX 2005) preliminary overview. In *Proceedings of the 6th International Conference on Music Information Retrieval* (ISMIR). 320–323.

EASLEY, R. F., MICHEL, J. G., AND DEVARAJ, S. 2003. The MP3 open standard and the music industry's response to internet piracy. *Commun. ACM 46*, 11, 90–96.

FREUND, Y. AND SCHAPIRE, R. 1997. A decision-theoretic generalzation of on-line learning and an application to boosting. *J. Comput. Syst. Sci. 55*, 1, 119–139.

GREENSPAN, H., GOLDBERGER, J., AND RIDEL, L. 2001. A continuous probabilistic framework for image matching. *Comput. Vis. Image Underst. 84*, 3, 384–406.

HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Verlag, Berlin, Germany.

ISMIR. 2004. *The Fifth International Conference on Music Information Retrieval*. `http://ismir2004.ismir.net/index.html`.

JORDAN, M. I. 1995. Why the logistic function? a tutorial discussion on probabilities and neural networks. Tech. rep. 9503. MIT, Cambridge, MA.

KIM, Y. E. AND WHITMAN, B. 2002. Singer identification in popular music recordings using voice coding features. In *Proceedings of the 3rd International Conference Music on Information Retrieval* (ISMIR). 164–169.

KIM, Y. E., WILLIAMSON, D., AND PILLI, S. 2006. Towards quantifying the album effect in artist identification. In *Proceedings of the 7th International Conference Music Information Retrieval* (ISMIR'06). 393–394.

LAM, C. K. M. AND TAN, B. C. Y. 2001. The Internet is changing the music industry. *Commun. ACM 44*, 8, 62–68.

LEBANON, G. AND LAFFERTY, J. 2001. Boosting and maximum likelihood for exponential model and Bregman distances. In *Advances in Neural Information Processing Systems 14 (Proceedings of NIPS)*. 110–121.

LI, T. AND OGIHARA, M. 2004. Music artist style identification by semisupervised learning from both lyrics and content. In *Proceedings of the 12th Annual ACM International Conference on Multimedia* . 364–367.

LI, T., OGIHARA, M., AND LI, Q. 2003. A comparative study on content-based music genre classification. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. 282–289.

LIU, C. C. AND HUANG, C. S. 2002. A singer identification technique for content-based classification of MP3 music objects. In *Proceedings of the 10th International Conference on Information and Knowledge Management* (CIKM). 506–511.

LIVSHIN, A. AND RODET, X. 2004. Musical instrument identification in continuous recordings. In *Proceedings of the 7th International Conference on Digital Audio Effects* (DAFx). 222–227.

LU, L., ZHANG, H., AND LI, S. Z. 2003. Content-based audio classification and segmentation by using support vector machines. *Multimed. Syst. 8*, 6, 482–492.

MIREX. 2005. Artist identification contest track. `http://www.music-ir.org/evaluation/mirex-results/audio-artist/index.html`.

MIREX. 2007. Artist identification contest track. `http://www.music-ir.org/mirex2007/index.php/AudioArtistIdentificationResults`.

PACHET, F. 2003. Content management for electronic music distribution. *Commun. ACM 46*, 4, 71–75.

PARDO, B. 2006. Special issue: Music information retrieval. *Commun. ACM 49*, 8.

PINTO, A. AND HAUS, G. 2007. A novel XML music information retrieval method using graph invariants. *ACM Trans. Inf. Syst. 25*, 4, 19.

RABINER, L. AND JUANG, B. 1993. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ.

RABINER, L. AND SCHAFER, R. 1978. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ.

RISSANEN, J. 1978. Modeling by shortest data description. *Automatica 14*, 465–471.

SHEN, J., SHEPHERD, J., CUI, B., AND TAN, K.-L. 2006. HSI: A novel framework for efficient automated singer identification in large music database. In *Proceedings of the 22nd International Conference on Data Engineering* (ICDE). 169.

TOLONEN, T. AND KARJALAINEN, M. 2000. A computationally efficient multipitch analysis model. *IEEE Trans. Speech Aud. Process. 8*, 4, 708–716.

TSAI, W. H. AND WANG, H. M. 2006. Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *IEEE Trans. Speech Aud. Process. 14*, 1, 330–341.

TSAI, W. H., WANG, H. M., RODGERS, D., CHENG, S. S., AND YU, H. M. 2003. Blind clustering of popular music recordings based on singer voice characteristics. In *Proceedings of the 4th international Conference on Music Information Retrieval* (ISMIR). 167–173.

VAPNIK, V. 1998. *Statistical Learning Theory*. John Wiley & Sons. New York, NY.

WHITMAN, B., FLAKE, G., AND LAWRENCE, S. 2001. Artist detection in music with Minnowmatch. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*. 559–568.

XU, C. S., MADDAGE, N., AND SHAO, X. 2005. Automatic music classification and summarization. *IEEE Trans. Speech Aud. Process. 13*, 3, 441–450.

ZHANG, T. 2003. Automatic singer identification. In *Proceedings of the 2003 International Conference on Multimedia and Expo* (ICME). 33–36.