Automatic Singer Recognition of Popular Music Recordings via Estimation and Modeling of Solo Vocal Signals

Wei-Ho Tsai and *Hsin-Min Wang*

Institute of Information Science, Academia Sinica
Nankang, 115, Taipei, Taiwan, Republic of China
E-mail: {wesley, whm}@iis.sinica.edu.tw
Phone:+886-2-27883799 ext. 2403, 1714
Fax:+886-2-27824814

Abstract

In this paper, we investigate the problem of automatic singer identification, detection and tracking in popular music recordings with one or multiple singers. This problem reflects an important issue in multimedia applications that require the transcription and indexing of music data to meet the increasing demand for content-based information retrieval. The major challenges for this study arise from the fact that a singer's voice tends to be arbitrarily altered from time to time and is inextricably intertwined with the signal of the background accompaniment. To determine who is singing, or whether or when a particular singer is present in a music recording, methods are presented for separating vocal from non-vocal regions, for isolating singers' vocal characteristics from background music, and for distinguishing singers from one another. Experimental evaluations conducted on a pop music database consisting of solo and duet tracks confirm the validity of the proposed methods.

Index Terms:

singer identification, singer detection, singer tracking, music information retrieval

EDICS:

2. AUDIO AND ELECTROACOUSTICS (2-MUSI Signal Processing for Music)

I. INTRODUCTION

Supported by the rapid progress in computer and network technology, popular music is rapidly becoming one of the most prevalent data types carried by the Internet. With the increased circulation of music data comes a corresponding increase in our appetite for accessing it more efficiently and conveniently. As a result, content-based retrieval of music has become an attractive topic for research, and efforts have been made to develop automatic classifiers or recognizers of music by melody [1,2], instruments [3,4], genre [5,6] and other means [7,8].

As an independent capability, or as a part of a music information retrieval system, techniques for automatically recognizing the singers [9-13] or artists [14,15] in music recordings are needed in order to lessen, or replace, human efforts in documenting unlabeled or insufficiently labeled data. For instance, many rock/pop music bands have a lead singer who performs the majority of the band's songs, but a minority of songs may be sung by the guitarist, drummer, or other band-members. Since most current documented music data is only labeled by title, artist (band name) or lead singer, acquiring songs or parts of a song performed by a particular singer may require automatically locating the portions of a song that the singer performs. In this context, singer recognition is used as a general term that includes all the various tasks involved in distinguishing music data, based on a singer's voice characteristics. Although there are many approaches to the recognition issue, this study investigates three indispensable techniques for singer recognition: singer identification (SID), target singer detection (TSD) and target singer tracking (TST).

SID refers to the task of determining who among a group of candidate singers sang a given part of a song. This involves an N-class decision, where N is the number of candidate singers. To perform SID, the voices of all the candidate singers are assumed known from a labeled music database beforehand. The purpose of the second technique, TSD, is to decide whether or not a specified target singer performs in a music recording. It can be viewed as a two-class or binary classification, in which one class corresponds to the music data containing the target singer's voice, and the other to the music performed entirely by a singer, or singers, other than the target one. In our approach, only prior information about the target singer's voice is assumed available from his/her solo albums or previous recordings, while neither the vocal characteristics nor prior singing data are assumed available for the singers not specified as the target. Finally, TST is concerned with the problem of determining where in a music recording, if at all, the target singer is singing. TST can be viewed as a TSD performed as a function of time. A system built for this task must output a list of regions where singing by the target person has been located.

In addition to finding a given singer's sung portions, or cameo appearances, in live concert recordings, there are numerous other potential applications that singer recognition could be used for. For example, singer recognition could be used to distinguish between an original recording and cover versions sung by different singers, or to obtain singer identity information, particularly about amateur singers, whom it may be difficult to find. Singer recognition may also enable record companies to rapidly scan suspect websites for piracy – especially bootleg concert recordings, in which the company will typically not have a copy of the original audio data for comparison. In addition, singer recognition may help karaoke operators organize their customers' music preferences, and thereby provide a personalized service. Finally, many of the methods developed for singer recognition could be trivially applied to areas such as music recommendation systems, whereby songs performed by singers with similar voices could be suggested.

In terms of the goal, singer recognition is analogous to speaker or talker recognition [16,17], which aims to determine who is speaking. Success in solving both problems depends on the detection and exploitation of characteristic features that distinguish one person's voice from another's. However, the problem of singer recognition is particularly complicated by the fact that, in pop music, the singer's voice signal is inextricably intertwined with a loud, non-stationary background music signal. This makes it infeasible to acquire pristine solo voice data (without background accompaniment) for directly extracting the desired vocal characteristics, which is generally possible in speaker recognition. Therefore, system design approaches that consider the separation of vocal and instrumental characteristics are believed to be the key to providing better solutions to the singer-recognition problem.

Previous works on singer recognition [9-12] may have either ignored the influence of background music on singer voice characterization, or simply looked at the singer-recognition problem from a speaker-recognition standpoint. No attempt has been made to remove the interference of background music from the vocal characteristics. In [9], formant frequencies and magnitudes analyzed via warped linear prediction are used as key features to distinguish singers' voices from one another. SID is done by a Gaussian mixture model (GMM) classifier, or a support vector machine classifier, using as input the warped linear prediction coefficients computed from accompanied signals. In [10], an envelope-based detection method is proposed for recognizing the underlying phonemes in an MP3 music recording. It is assumed that each singer has his/her own phoneme set, and modified discrete cosine transform coefficients computed from these phonemes are used as features to construct a k-nearest neighbor classifier, thereby distinguishing singers. In [11], a common speaker-recognition method based on Gaussian mixture models, trained using Mel-scale frequency cepstral coefficients (MFCCs), is applied to distinguish singers. Meanwhile, in [12], Bartsch and Wakefield propose an SID method based on the so-called composite transfer function (CTF) for spectral envelope estimation. The CTF, which is derived from the instantaneous amplitude and frequency of the signal's harmonic partials, is claimed to be better at characterizing complicated vocal variations like vibrato. However, in contrast to the systems in [9-11], which operate on real performances from recorded popular music, this method only examines an ideal case in which audio samples contain only the classically-trained singer's voice, without accompaniment.

To deal with the singer-recognition problem more effectively, we propose a solo voice modeling technique for capturing singers' vocal characteristics. Our basic idea is that, in most pop songs, substantial similarities exist between the instrumental-only regions and the accompaniment of the singing regions. Therefore, the stochastic characteristics of the background music may be approximated by those of the instrumental-only regions. From the available information about the background music, the underlying solo voices can be statistically estimated and modeled from the accompanied voices by exploiting an *a priori* model for the background music. To expedite this process, we also presents an effective method for segmenting a music recording into vocal and non-vocal portions, in which a vocal portion consists of concurrent singing and accompaniment, and a non-vocal portion consists of accompaniment only.

Note that the music may be instrumental only, a solo, a duet, a trio, or even a chorus. Consequently, accomplishing one of the three tasks may require that multiple tasks be performed consecutively, or in parallel. For instance, to determine who is singing in a duet, the system may need to decide whether or not either of the candidate singers is present in the recording and decide the singer's identity at the same time. However, in this study, each of the three tasks is investigated independently during the initial development stage. Specifically, when dealing with the SID problem, we assume that the music recordings contain only one particular singer from a candidate set. When dealing with the TSD problem, we only allow one singer at a time to be specified as the target, even if there are multiple singers. Hence, the boolean (e.g. and/or) query often considered in an information retrieval application is not addressed here. With regard to TST, it is assumed that each of the test music recordings contains the voices of the target and non-target singers. TST performance is evaluated on the duet music data. Without loss of generality, the methods presented in this paper should be applicable to a wide variety of music data and appropriate combinations thereof.

The rest of this paper is organized as follows. Section II presents a statistical classifier for distinguishing vocal segments and accompaniments. Section III introduces a method for distilling the singers' vocal characteristics from the vocal regions of music recordings. In Section IV, we describe how to perform SID, TSD and TST based on the proposed singer modeling method. Section V presents our experimental results. Finally, in Section VI, we present our conclusions and the direction of our future works.

II. VOCAL/NON-VOCAL SEGMENTATION

As a first step in determining the vocal characteristics of a singer, music segments that contain vocals are located and marked as such. This task can be formulated as a problem of distinguishing between vocal segments and accompaniments, analogous to the study by Berenzweig and Ellis [18]. However, in contrast to their work, which uses a speech recognizer to detect singing voices, we propose to construct a statistical classifier with parametric models trained using accompanied singing voices rather than normal speech. This approach is based on the observation that there is a significant difference in spectral distribution between vocal and instrumental sound. Fig. 1 shows the spectrograms of two music examples. Due to the rapid vibration of the vocal folds, the singing voice is nearly always harmonic [19], and exhibits relatively large amounts of energy at integer multiples of the fundamental frequency in the low or middle frequency regions of the spectrogram. Compared to the singing voices, the instrumental-only sounds have less salient harmonics and spread their energy more widely.

As shown in Fig. 2, the vocal/non-vocal classifier consists of a front-end signal processor that converts digital waveforms into spectrum-based feature vectors, and a back-end statistical processor that performs modeling, matching and decision making. The feature vectors used here are Mel-scale frequency cepstral coefficients (MFCCs), which are typically computed using a fixed-length sliding window of 10ms to 40ms, also called a frame. This approach has been used predominantly in speech signal processing, particularly in speech recognition. Its applicability for handling music signal has been studied in [11] and [20].

The back-end statistical processor operates in two phases: training, and testing. During training, a music database with manual vocal/non-vocal transcriptions is used to form two separate Gaussian mixture models (GMMs): a vocal GMM, and a non-vocal GMM. Each model consists of several mixture weights, mean vectors and covariance matrices. The use of GMMs is motivated by the wish to model the spectral distribution of various broad acoustic classes by a combination of Gaussian components. These broad acoustic classes reflect some general vocal and instrumental configurations. It has been shown that GMMs have a strong ability to provide smooth approximations to arbitrarily-shaped densities of a spectrum over a long time span [17]. We denote the vocal GMM as λ_V, and the non-vocal GMM λ_N. Parameters of the GMMs are initialized via k-means clustering [21] and iteratively adjusted via expectation-maximization (EM) [22].

In the testing phase, the classifier takes as input the T_x-frame feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T_x}\}$ extracted from an unknown recording, and produces as output the frame log-likelihoods log $p(\mathbf{x}_t|\lambda_V)$ and log $p(\mathbf{x}_t|\lambda_N), 1 \leq t \leq T_x$, for the vocal and non-vocal GMM, respectively. The attribute of each frame is then hypothesized according to a decision rule made on the frame log-likelihoods. Depending on the choice of analysis interval, there are many variations and combinations in decision-making. In this study, we compare several possibilities, including a frame-based decision, a fixed-length-segment-based decision, and a homogeneous-segment-based decision.

A. Frame-based Decision

The recognizer may trivially hypothesize whether the frame $\mathbf{x}_t$ is vocal or not by using

$$\log\, p(\mathbf{x}_t|\lambda_V) \quad \underset{\text{non-vocal}}{\overset{\text{vocal}}{\substack{>\\ \leq}}} \quad \log\, p(\mathbf{x}_t|\lambda_N). \tag{1}$$

Since singing tends to be continuous for several frames (i.e. continuous for at least 1 sec before the next rest or pause), these results may be smoothed in the time domain. For smoothing, a sliding window is applied to divide the frame feature vectors into a sequence of consecutive, non-overlapping, fixed-length segments. The majority hypothesis for each segment is then assigned to each frame of that segment.

B. Fixed-length-segment-based Decision

The above smoothing may be improved by directly assigning a single classification per segment by:

$$\sum_{i=0}^{W-1} \log\, p(\mathbf{x}_{kW+i}|\lambda_V) \quad \underset{\text{non-vocal}}{\overset{\text{vocal}}{\substack{>\\ \leq}}} \quad \sum_{i=0}^{W-1} \log\, p(\mathbf{x}_{kW+i}|\lambda_N), \tag{2}$$

where k is the segment index and W is the segment length. In general, accumulating the frame log-likelihoods over a longer period is more statistically reliable for decision-making. However, as with smoothing, long segments could run the risk of crossing multiple vocal/non-vocal change boundaries. In view of the possibility of a short singing duration,the length of a segment is preferably less than 1 sec.

C. Homogeneous-segment-based Decision

An improvement of the segment-based decision above may be made by merging adjacent segments into longer homogeneous ones, if those adjacent segments do not cross a vocal/non-vocal boundary. To do this, vector clustering is first applied to all frame feature vectors and each frame is assigned the cluster index associated with that frame's feature vector. As a result, a music recording is tokenized as a cluster index stream, which is then divided into a sequence of consecutive, non-overlapping, fixed-length segments. Each segment is then assigned the majority cluster index of its constituent frames, and adjacent segments are merged as a homogeneous segment if they have the same cluster index. Finally, classification is made per homogeneous-segment by:

$$\sum_{i=0}^{W_k-1} \log\, p(\mathbf{x}_{s_k+i}|\lambda_V) \quad \underset{\text{non-vocal}}{\overset{\text{vocal}}{\substack{>\\ \leq}}} \quad \sum_{i=0}^{W_k-1} \log\, p(\mathbf{x}_{s_k+i}|\lambda_N), \tag{3}$$

where W_k and s_k represent, respectively, the length and starting frame of the k-th homogeneous-segment.

However, this approach is not without drawbacks, compared to the frame-based and fixed-length-based decisions. The major problem is that the inevitable errors arising from the determination of homogeneous segments can propagate to the final vocal/non-vocal hypothesis test, and cause a larger range of mis-classification. Moreover, as vector clustering needs to be performed online, the computational cost can be much higher than that of the other approaches.

III. SINGER CHARACTERISTIC MODELING

Viewed as a problem of pattern recognition [23], a promising way to design a reliable singer-recognition scheme is the construction of stochastic models for summarizing some of the most relevant aspects of a singer's voice characteristics. Since the vast majority of popular music contains background accompaniment during most or all vocal passages, directly acquiring isolated solo voice data for modeling the singer's vocal characteristics is usually infeasible. To eliminate the interference of background music for singer voice characterization, we leverage statistical estimation of a piece's musical background to build a reliable model for the solo voice.

To begin, assume that an accompanied voice in feature representation $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T\}$ is the mix of a solo voice $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_T\}$ and background music $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_T\}$, where $\mathbf{V}$ can be obtained directly from the vocal segments of the music recording, but both $\mathbf{S}$ and $\mathbf{B}$ are unobservable. Our aim is to distill $\mathbf{S}$ from $\mathbf{V}$, such that the underlying singer's vocal characteristics can be parametrically represented. The proposed solution is basically adapted from the techniques developed in robust speech recognition under noisy environments [24,25]. We assume that the solo voice and background music are drawn randomly according to GMM $\lambda_s = \{w_{s,i}, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i} | 1 \leq i \leq I\}$, and GMM $\lambda_b = \{w_{b,j}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j} | 1 \leq j \leq J\}$, respectively, where $w_{s,i}$ and $w_{b,j}$ are mixture weights, $\boldsymbol{\mu}_{s,i}$ and $\boldsymbol{\mu}_{b,j}$ are mean vectors, and $\boldsymbol{\Sigma}_{s,i}$ and $\boldsymbol{\Sigma}_{b,j}$ are covariance matrices. While robust speech recognition mostly models background noise as a uni-Gaussian density, the use of a mixture of Gaussians for modeling background music is considered necessary. This is because the background music is often rather complex, compared to the normally-stable background noise, such as the white noise, usually considered in speech recognition. A Gaussian mixture allows the background music to be represented by some general instrumental configurations in the same way that singing is modeled.

If the accompanied voice is formed from a generative function $\mathbf{v}_t = f(\mathbf{s}_t, \mathbf{b}_t)$, $1 \leq t \leq T$, then the probability of observing $\mathbf{V}$, given λ_s and λ_b can be represented by:

$$p(\mathbf{V}|\lambda_s, \lambda_b) = \prod_{t=1}^{T} \left\{ \sum_{i=1}^{I} \sum_{j=1}^{J} w_{s,i} w_{b,j} \; p(\mathbf{v}_t | \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}) \right\}, \tag{4}$$

where $p(\mathbf{v}_t | \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j})$ is the probability of one possible combination of the underlying solo voice and the background music that can form an instant accompanied voice $\mathbf{v}_t$. To compute $p(\mathbf{v}_t | \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j})$ efficiently, we further assume that the solo voice and background music are statistically independent [1], and hence

$$p(\mathbf{v}_t | \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}) = \iint_{\mathbf{v}_t = f(\mathbf{s}_t, \mathbf{b}_t)} \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}) \mathcal{N}(\mathbf{b}; \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}) d\mathbf{s} d\mathbf{b}, \tag{5}$$

where $\mathcal{N}(\cdot)$ denotes a Gaussian density function.

Although the background music $\mathbf{B}$ is unobservable, in most popular music substantial similarities exist between the non-vocal regions and the accompaniment of the vocal regions. Therefore, $\mathbf{B}$'s stochastic characteristics may be approximated by those of the non-vocal regions. Based on this approximation, the background music model λ_b can be created directly, using the feature vectors computed from the non-vocal regions. Then, with the available background music model λ_b and the observable accompanied voice $\mathbf{V}$, it is sufficient to derive the solo voice model λ_s via a maximum likelihood estimation as follows:

$$\lambda_s^* = \arg \max_{\lambda_s} p(\mathbf{V}|\lambda_s, \lambda_b). \tag{6}$$

[1] Strictly speaking, the solo voice and background music are usually arranged to fit together harmonically. Thus, a joint Gaussian density may be more suitable than the two marginal Gaussian densities in Eq. (5) to represent the combination of the solo voice and background music. However, when considering implementation feasibility, we ignore the inter-dependence between the solo voice and background music at this stage.

Using the EM algorithm, an initial model λ_s is created, and a new model $\hat{\lambda}_s$ is then estimated by maximizing the auxiliary function

$$Q(\lambda_s, \hat{\lambda}_s) = \sum_{t=1}^{T} \sum_{i=1}^{I} \sum_{j=1}^{J} p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b) \cdot \log p(i, j, \mathbf{v}_t | \hat{\lambda}_s, \lambda_b), \tag{7}$$

where

$$p(i, j, \mathbf{v}_t | \hat{\lambda}_s, \lambda_b) = \hat{w}_{s,i} w_{b,j} \; p(\mathbf{v}_t | \hat{\boldsymbol{\mu}}_{s,i}, \hat{\boldsymbol{\Sigma}}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}), \tag{8}$$

and

$$p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b) = \frac{w_{s,i} w_{b,j} \; p(\mathbf{v}_t | \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j})}{\sum_{m=1}^{I} \sum_{n=1}^{J} w_{s,m} w_{b,n} p(\mathbf{v}_t | \boldsymbol{\mu}_{s,m}, \boldsymbol{\Sigma}_{s,m}, \boldsymbol{\mu}_{b,n}, \boldsymbol{\Sigma}_{b,n})}. \tag{9}$$

Letting $\nabla Q(\lambda_s, \hat{\lambda}_s) = 0$ with respect to each of the parameters in $\hat{\lambda}_s$ to be estimated, we have

$$\hat{w}_{s,i} = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{J} p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b), \tag{10}$$

$$\hat{\boldsymbol{\mu}}_{s,i} = \frac{\sum_{t=1}^{T} \sum_{j=1}^{J} p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b) \cdot E\{\mathbf{s}_t | \mathbf{v}_t, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}\}}{\sum_{t=1}^{T} \sum_{j=1}^{J} p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b)}, \tag{11}$$

$$\hat{\boldsymbol{\Sigma}}_{s,i} = \frac{\sum_{t=1}^{T} \sum_{j=1}^{J} p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b) \cdot E\{\mathbf{s}_t \mathbf{s}'_t | \mathbf{v}_t, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}\}}{\sum_{t=1}^{T} \sum_{j=1}^{J} p(i, j | \mathbf{v}_t, \lambda_s, \lambda_b)} - \hat{\boldsymbol{\mu}}_{s,i} \hat{\boldsymbol{\mu}}'_{s,i}, \tag{12}$$

where prime ($'$) denotes the vector transpose, and the conditional expectations $E\{\mathbf{s}_t | \mathbf{v}_t, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}\}$ and $E\{\mathbf{s}_t \mathbf{s}'_t | \mathbf{v}_t, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}\}$ are, respectively, obtained from

$$E\{\mathbf{s}_t | \mathbf{v}_t, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}\} = \frac{\iint_{\mathbf{v}_t = f(\mathbf{s}_t, \mathbf{b}_t)} \mathbf{s} \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}) \mathcal{N}(\mathbf{b}; \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}) d\mathbf{s} d\mathbf{b}}{p(\mathbf{v}_t | \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j})}, \tag{13}$$

and

$$E\{\mathbf{s}_t \mathbf{s}'_t | \mathbf{v}_t, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}\} = \frac{\iint_{\mathbf{v}_t = f(\mathbf{s}_t, \mathbf{b}_t)} \mathbf{s}\mathbf{s}' \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}) \mathcal{N}(\mathbf{b}; \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}) d\mathbf{s} d\mathbf{b}}{p(\mathbf{v}_t | \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j})}. \tag{14}$$

The new model $\hat{\lambda}_s$ then becomes λ_s for the next iteration and the re-estimation process is repeated until the likelihood converges to a local maximum.

To facilitate the implementation, the probability $p(\mathbf{v}_t | \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j})$ and the conditional expectations $E\{\mathbf{s}_t | \mathbf{v}_t, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}\}$ and $E\{\mathbf{s}_t \mathbf{s}'_t | \mathbf{v}_t, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}\}$ must be explicitly expressed as closed forms. Suppose that $\mathbf{V}$, $\mathbf{S}$ and $\mathbf{B}$ are log-spectrum features (cepstral features), and the singing and background music are added in the time domain or linear-spectrum domain. The accompanied voice can then be approximately represented by $\mathbf{v}_t = \log(\exp(\mathbf{s}_t) + \exp(\mathbf{b}_t)) \approx \max(\mathbf{s}_t, \mathbf{b}_t)$, $1 \leq t \leq T$, according to Nadas' MIXMAX model [24]. For greater efficiency, the covariance matrices of GMM used in this study are assumed to be diagonal, i.e. $\mathbf{\Sigma}_{s,i} = \{\sigma^2_{s,i,d}\}_{d=1}^{D}$ and $\mathbf{\Sigma}_{b,j} = \{\sigma^2_{b,j,d}\}_{d=1}^{D}$, where D is the dimension of the feature vector. Each vector component involved can thus be operated independently. We compute $p(\mathbf{v}_t|\boldsymbol{\mu}_{s,i}, \mathbf{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \mathbf{\Sigma}_{b,j})$ using:

$$p(\mathbf{v}_t|\boldsymbol{\mu}_{s,i}, \mathbf{\Sigma}_{s,i}, \boldsymbol{\mu}_{b,j}, \mathbf{\Sigma}_{b,j}) = \prod_{d=1}^{D} p(v_{t,d}|\mu_{s,i,d}, \sigma^2_{s,i,d}, \mu_{b,j,d}, \sigma^2_{b,j,d}), \tag{15}$$

where $v_{t,d}$, $\mu_{s,i,d}$, and $\mu_{b,j,d}$ are, respectively, the *d-th* component of $\mathbf{v}_t$, $\boldsymbol{\mu}_{s,i}$, and $\boldsymbol{\mu}_{b,j}$. For ease of discussion, we drop the component index d and focus on the scalar operations. For an arbitrary component,v_t, of $\mathbf{v}_t$, the probability $p(v_t|\mu_{s,i}, \sigma^2_{s,i}, \mu_{b,j}, \sigma^2_{b,j})$ can be computed as:

$$\begin{aligned} p(v_t|\mu_{s,i}, \sigma^2_{s,i}, \mu_{b,j}, \sigma^2_{b,j}) &= \iint_{v_t \approx \max(s_t, b_t)} \mathcal{N}(s; \mu_{s,i}, \sigma^2_{s,i}) \mathcal{N}(b; \mu_{b,j}, \sigma^2_{b,j}) ds db \\ &= \mathcal{N}(v_t; \mu_{s,i}, \sigma^2_{s,i}) \int_{-\infty}^{v_t} \mathcal{N}(b; \mu_{b,j}, \sigma^2_{b,j}) db \\ &\quad + \mathcal{N}(v_t; \mu_{b,j}, \sigma^2_{b,j}) \int_{-\infty}^{v_t} \mathcal{N}(s; \mu_{s,i}, \sigma^2_{s,i}) ds \\ &= \mathcal{N}(v_t; \mu_{s,i}, \sigma^2_{s,i}) \Phi(\frac{v_t - \mu_{b,j}}{\sigma_{b,j}}) + \mathcal{N}(v_t; \mu_{b,j}, \sigma^2_{b,j}) \Phi(\frac{v_t - \mu_{s,i}}{\sigma_{s,i}}), \end{aligned} \tag{16}$$

where

$$\Phi(\tau) = \int_{-\infty}^{\tau} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} d\xi. \tag{17}$$

The value of $\Phi(\tau)$ can be obtained using a table of the error function.

The conditional expectation in Eq. (13), with respect to an arbitrary vector component s_t of $\mathbf{s}_t$, can be derived as follows:

$$\begin{aligned} E\{s_t|v_t, \mu_{s,i}, \sigma^2_{s,i}, \mu_{b,j}, \sigma^2_{b,j}\} &= p(s_t = v_t|\mu_{s,i}, \sigma^2_{s,i}, \mu_{b,j}, \sigma^2_{b,j}) \cdot v_t \\ &\quad + \Big[1 - p(s_t = v_t|\mu_{s,i}, \sigma^2_{s,i}, \mu_{b,j}, \sigma^2_{b,j})\Big] \\ &\quad \cdot E\{s_t|s_t < v_t, \mu_{s,i}, \sigma^2_{s,i}, \mu_{b,j}, \sigma^2_{b,j}\}, \end{aligned} \tag{18}$$

where

$$p(s_t = v_t|\mu_{s,i}, \sigma_{s,i}^2, \mu_{b,j}, \sigma_{b,j}^2) \quad = \quad \frac{\mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2)\Phi(\frac{v_t-\mu_{b,j}}{\sigma_{b,j}})}{\mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2)\Phi(\frac{v_t-\mu_{b,j}}{\sigma_{b,j}}) + \mathcal{N}(v_t; \mu_{b,j}, \sigma_{b,j}^2)\Phi(\frac{v_t-\mu_{s,i}}{\sigma_{s,i}})}, \tag{19}$$

and

$$\begin{aligned} E\{s_t|s_t < v_t, \mu_{s,i}, \sigma_{s,i}^2, \mu_{b,j}, \sigma_{b,j}^2\} &= \frac{\int_{-\infty}^{v_t} s\mathcal{N}(s; \mu_{s,i}, \sigma_{s,i}^2)ds}{\Phi(\frac{v_t - \mu_{s,i}}{\sigma_{s,i}})} \\ &= \mu_{s,i} - \sigma_{s,i}^2 \cdot \frac{\mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2)}{\Phi(\frac{v_t - \mu_{s,i}}{\sigma_{s,i}})}. \end{aligned} \tag{20}$$

For the solo voices that are less mixed with the background music, i.e. $p(s_t = v_t|\mu_{s,i}, \sigma_{s,i}^2, \mu_{b,j}, \sigma_{b,j}^2) \approx 1$, the conditional expectation $E\{s_t|v_t, \mu_{s,i}, \sigma_{s,i}^2, \mu_{b,j}, \sigma_{b,j}^2\}$ is determined by the accompanied voice v_t. Conversely, if the solo voices are submerged by the loud background music, i.e. $p(s_t = v_t|\mu_{s,i}, \sigma_{s,i}^2, \mu_{b,j}, \sigma_{b,j}^2) \approx 0$, the conditional expectation $E\{s_t|v_t, \mu_{s,i}, \sigma_{s,i}^2, \mu_{b,j}, \sigma_{b,j}^2\}$ will approximate to $\mu_{s,i}$, which means the accompanied voice v_t will not contribute to the re-estimation of the new model mean . Similarly, the conditional expectation in Eq. (14) is computed using

$$\begin{aligned} E\{s_t^2|v_t, \mu_{s,i}, \sigma_{s,i}^2, \mu_{b,j}, \sigma_{b,j}^2\} \quad = \quad & p(s_t = v_t|\mu_{s,i}, \sigma_{s,i}^2, \mu_{b,j}, \sigma_{b,j}^2) \cdot v_t^2 \\ & + \Big[1 - p(s_t = v_t|\mu_{s,i}, \sigma_{s,i}^2, \mu_{b,j}, \sigma_{b,j}^2)\Big] \\ & \cdot E\{s_t^2|s_t < v_t, \mu_{s,i}, \sigma_{s,i}^2, \mu_{b,j}, \sigma_{b,j}^2\}, \end{aligned} \tag{21}$$

where

$$\begin{aligned} E\{s_t^2|s_t < v_t, \mu_{s,i}, \sigma_{s,i}^2, \mu_{b,j}, \sigma_{b,j}^2\} &= \frac{\int_{-\infty}^{v_t} s^2\mathcal{N}(s; \mu_{s,i}, \sigma_{s,i}^2)ds}{\Phi(\frac{v_t - \mu_{s,i}}{\sigma_{s,i}})} \\ &= \mu_{s,i}^2 + \sigma_{s,i}^2 - (\mu_{s,i} + v_t) \cdot \sigma_{s,i}^2 \cdot \frac{\mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2)}{\Phi(\frac{v_t - \mu_{s,i}}{\sigma_{s,i}})}. \end{aligned} \tag{22}$$

Note that if the number of mixtures in the background music GMM is zero, then our solo voice modeling method degenerates to directly modeling the observed vocal signal, without taking the background music into account. This serves as a baseline to examine the effectiveness of our solo voice modeling method.

IV. SINGER RECOGNITION

A block diagram of the proposed singer-recognition system for singer identification, target singer detection and target singer tracking is shown in Fig. 3. The operation of this system can be divided into two phases, namely: training and testing. During training, music recordings from a training set are segmented into vocal and non-vocal regions, according to the method described in Section II. The resulting non-vocal regions are then used to form a GMM which approximately simulates the acoustic characteristics of the background accompaniment. The background music GMM, together with the segmented vocal regions pertaining to a particular singer or singer class, is then used to estimate a solo voice model, according to the method described in Section III. In the testing phase, a background music GMM is created on-line, using the segmented non-vocal regions of the test music recording. The system then hypothesizes who is singing, or whether or when a specified singer is singing, by evaluating the conditional probability of the test vocal signal, given the background music GMM and a solo voice model of interest, i.e. using Eq. (4). The required models and hypotheses differ from each other, depending on the tasks being performed. Each of the tasks are described below.

A. Singer identification (SID)

Given a test music recording $\mathbf{X}$, the objective of SID is to determine which among a group of P singers $\{R_1, R_2, \ldots, R_P\}$ performed $\mathbf{X}$. Under the solo voice modeling framework, P candidate singers are in turn represented by P singer-specific models $\{\lambda_{s,1}, \lambda_{s,2}, \ldots, \lambda_{s,P}\}$. SID can be viewed as a problem of choosing one of the P models that best matches $\mathbf{X}$. According to the maximum likelihood decision rule, the identifier should decide in favor of a singer R^* for the recording $\mathbf{X}$ satisfying

$$R^* = \arg \max_{1 \leq i \leq P} \log p(\mathbf{X}_V | \lambda_{s,i}, \tilde{\lambda}_b), \tag{23}$$

where $\mathbf{X}_V$ denotes a collection of the vocal regions in $\mathbf{X}$, and $\tilde{\lambda}_b$ is the background music GMM created by using the non-vocal regions of $\mathbf{X}$.

B. Target singer detection (TSD)

The purpose of TSD is to determine whether or not a specified target singer is present in a given test music recording $\mathbf{X}$. First, consider the case that recording $\mathbf{X}$ contains only one particular singer's voice. TSD can be viewed as a problem of judging if the recording $\mathbf{X}$ is performed by the target singer. Using the solo voice modeling method, two models are created in the training phase, namely, the target singer model, which is trained using the recordings performed solely by the target singer, and the universal singer model, which represents a generic vocal characteristic of a non-target singer and is trained using the music recordings performed by a plurality of singers other than the target one. We denote the target singer model and the universal singer model as λ_s^T and λ_s^U, respectively. Accordingly, the problem of judging if $\mathbf{X}$ is performed by the target singer can be further converted into a hypothesis test of whether the attribute of $\mathbf{X}$ is target or non-target. The required decision rule for this hypothesis test can be expressed as:

$$\frac{1}{T_V}\left[\log\, p(\mathbf{X}_V|\lambda_s^T,\tilde{\lambda_b}) - \log\, p(\mathbf{X}_V|\lambda_s^U,\tilde{\lambda_b})\right] \quad \overset{\text{target}}{\underset{\text{non-target}}{\substack{> \\ \leq}}} \quad \theta_{TSD}, \tag{24}$$

where θ_{TSD} is the threshold, and T_V is the total length of the vocal regions, $\mathbf{X}_V$, in $\mathbf{X}$. Eq. (24) essentially measures how well the target singer model matches the test recording, compared to the universal singer model.

Due to binary decision, two types of errors exist in TSD. One is missed detection (MD) error, which occurs when a test music recording performed by the target singer is hypothesized as non-target. The other is false alarm (FA) error, which occurs when a test music recording not performed by the target singer is hypothesized as the target. MD and FA are subject to trade-off, and the number of errors can be adjusted by setting the threshold θ_{TSD}. In some applications, a lower occurrence of MD may be more important than that of FA, or vice versa. However, when the application is unknown, an appropriate threshold may be set in such a way that the number of MD and FA errors is equal. In addition, the log-likelihood provided by the universal singer model normalizes the acoustic variations in the test recording, and makes it easier to set a stable decision threshold.

Extending the above hypothesis test framework to deal with the case when the test recording $\mathbf{X}$ contains multiple singers' voices, TSD can be intuitively performed by first segmenting the entire recording into singer-homogeneous regions, and then determining the attribute of each of those regions. However, our study does not investigate this approach, because it involves a further problem concerning how to automatically locate singer-homogeneous regions. Moreover, it may be necessary to judge whether or not a test recording contains multiple singers's voices before the singer-homogeneous regions are located. To sidestep this problem, we want to examine if Eq. (24) remains applicable for multi-singer music recordings, without any extra process. It is assumed that if a test music recording contains the target singer's voice, the target singer model will better match the test recording than the universal singer model, no matter whether any other singers are present in the test recording or not. Therefore, instead of examining the singer-homogeneous regions, we use the same TSD method for solo and multi-singer music, in which all the vocal portions within a music recording are examined as a whole.

C. Target singer tracking (TST)

TST aims to determine where the target singer is present in a test music recording $\mathbf{X}$. Here, we assume that the test recording contains the target singer's voice, and hence the problem is to separate the target singer's voice regions from the non-target singers' voice regions, if applicable. The system first locates the vocal regions of $\mathbf{X}$ and then divides each of the continuous vocal regions into a sequence of consecutive, fixed-length, non-overlapping segments. Each of the segments is assumed homogeneous in terms of the singer and can thus be hypothesized as either target or non-target according to a comparison of the log-likelihoods for the target singer model and for the universal singer model, i.e.

$$\frac{1}{W}\left[\sum_{i=0}^{W-1}\log\, p(\mathbf{x}_{tW+i}|\lambda_s^T,\tilde{\lambda_b}) - \sum_{i=0}^{W-1}\log\, p(\mathbf{x}_{tW+i}|\lambda_s^U,\tilde{\lambda_b})\right] \begin{array}{c}\text{target}\\ > \\ \leq \\ \text{non-target}\end{array} \theta_{TST}, \tag{25}$$

where θ_{TST} is the global threshold, and W is the segment length. The principle of Eq. (25) basically resembles that of Eq. (24), except that the examination unit in Eq. (25) is a short vocal segment, rather than the whole vocal portions used in Eq. (24).

V. EXPERIMENTAL RESULTS

A. Music Data

Extensive computer simulations have been carried out to evaluate the performance of our proposed methods. The music data used in this study consisted of 242 solo tracks, 22 duet tracks and 174 instrumental-only tracks from Mandarin pop music CDs. All the tracks were manually labeled with singer identity and the vocal/non-vocal boundaries. The length of the tracks ranged from 135 to 391 seconds. In consideration of the normal range of singing voices as well as data storage, all the tracks were down-sampled from the CD sampling rate of 44.1 kHz to 22.05 kHz. This excludes the high frequency components containing sparse vocal information. Feature vectors, each consisting of 20 Mel-scale frequency cepstral coefficients (MFCCs), were extracted from this data, using a 32-ms Hamming-windowed frame with 10-ms frame shifts.

The 242 solo music tracks were grouped into two sets by singer, denoted as DB-S-1 and DB-S-2. The DB-S-1 comprised 200 tracks performed by 10 male and 10 female singers, with 10 distinct songs per singer. The DB-S-2 comprised the remaining 42 tracks, involving 13 female and 8 male singers, none of whom appeared in DB-S-1, with each of the singers performing two distinct songs. The DB-S-1 set was then divided into two subsets, one for training singer-specific models, and another for evaluation purposes. The former, denoted as DB-S-1-T, contained five tracks per singer, while the latter, denoted as DB-S-1-E, contained the remaining five tracks per singer. The DB-S-2 set was used to create a universal singer model for the TSD and TST experiments, while the 22 duet music tracks, denoted as DB-D, were used for TSD and TST evaluation. Each of the singers in the DB-S-1 set was present on at least one track of the DB-D set. In addition, the 174 instrumental-only tracks, denoted as DB-I, were used for training the non-vocal model. A summary of our music data is given in Table I.

B. Vocal/non-vocal Segmentation Results

In our first experiment we tested the validity of vocal/non-vocal segmentation. The vocal model was trained using the vocal segments of all the tracks in DB-S-1-T, and the non-vocal model was trained using DB-I together with the non-vocal segments of all the tracks in DB-S-1-T. The test data used here was DB-S-1-E [2] and DB-D. Performance was evaluated on the basis of frame classification accuracy computed by comparing the hypothesized attribute of each frame with the manual label, i.e.

$$\text{Frame classification accuracy (in \%)} = \frac{\#\text{correctly-classified frames}}{\#\text{total frames}} \times 100\%.$$

However, in view of the limited precision with which the human ear detects vocal/non-vocal changes, all frames that occurred within 0.5 seconds of a perceived switch-point were ignored in the computation.

Table II summarizes the results of vocal/non-vocal segmentation, using a 64-mixture vocal GMM and an 80-mixture non-vocal GMM (empirically the most accurate configuration). The

[2] In the experiments reported in this paper, the vocal segments in the training and testing data contained singing voices from the same singer set. This is because in our singer-recognition tasks, vocal samples from the target singer must be acquired beforehand. These vocal samples can be used not only for extracting a target singer's voice characteristics, but also for training the vocal GMM. Further, such a strategy has been shown to yield slightly better performance in vocal/non-vocal segmentation, compared to those of our prior experiments [26], in which the music data used for training the vocal GMM did not contain the voice of the singer in the testing data.

table shows that the homogeneous-segment-based method is superior to the other methods when an adequate number of clusters are used. The best accuracy achieved here was 82.3%. This served as a basis for front-end processing for subsequent experiments. Table III shows the confusion probability matrix obtained by the homogeneous-segment-based decision. The rows of the matrix correspond to the ground-truth of the segments, while the columns indicate the hypotheses. We can see that the majority of errors are caused by the misidentification of vocal segments. Further analysis of our results showed that more than 80% of the falsely-classified vocal segments had unusually loud background music or unusually quiet vocals. However, due to the high background to vocal ratio, we believe that such false judgments may actually benefit singer recognition.

C. Singer Identification Results

In our SID experiments, the data used for the training of singer-specific models and testing was, respectively, DB-S-1-T and DB-S-1-E. To evaluate SID performance with different lengths of test recordings, each of the tracks in DB-S-1-E was divided into several overlapping segments of L feature vectors. A 10-sec segment corresponded to 1000 feature vectors, and the overlap of two consecutive segments was 500 feature vectors. Each segment was treated as a separate music recording. The SID experiment was conducted in a segment-by-segment manner, and the SID accuracy was computed as the percentage of correctly-identified segments over the total number of test segments. In the training phase, the number of mixture components used in each of the solo voice models and the background music models was empirically determined to be 48 and 16, respectively. In the testing phase, the online-created background music model was empirically set to have 4 mixture components, if the number of the non-vocal frames exceeded 200; otherwise, no background music model was used.

Fig. 4 shows the SID results with respect to $L = 1000$ (10 sec), 3000 (30 sec), 6000 (60 sec), and the entire track, in which the L-length segments that were fully labeled as non-vocal were not used for testing. As expected, the SID accuracy improves as the segment length increases.

Furthermore, the performance of the proposed solo voice modeling method is significantly better than that of the direct GMM method without background music modeling. The superiority of solo voice modeling over the direct GMM method is particularly clear when testing long recordings, where more information about the background music can be exploited. From Fig. 4, an interesting observation can be made when we compare the SID performance using the manual vocal/non-vocal segmentation with the automatic segmentation. Intuitively, the SID performance achieved with the manual vocal/non-vocal segmentation should serve as an upper bound for that obtained using automatic segmentation. However, the results contradict that intuition. The major reason for this phenomenon is that automatic segmentation is actually advantageous for pruning some feature vectors that are manually labeled as vocal, but heavily mixed with the loud background music. Despite some loss of information, pruning such vocal frames can prevent non-singer features interfering with SID.

Table IV shows the confusion matrix obtained by the solo voice modeling method using automatic vocal/non-vocal segmentation as the front-end process, in which all the different-length trials above are taken into account. The singers indexed by 1 to 10 are female, whereas the singers indexed by 11 to 20 are male. We can see from Table IV that a male singer is rarely mis-identified as a female one, and vice versa. In addition, some singers (e.g. 5, 15, and 18) frequently tend to be mis-identified as particular singers . However, no specific pair of singers could be confused with each other. We speculate that this might be attributable to the higher variations of observed vocal characteristics in the training material of those singers associated with a lower SID accuracy. Such speculation is supported by the observation that the models of the singers associated with a lower SID accuracy usually had lower likelihoods, $[\log p(\mathbf{V}|\lambda_s, \lambda_b)]/T$, (see Eq. (4)), during the training process, compared to those of other singers.

Another experiment related to SID was conducted to investigate the problem concerning the correlation between the background music and the singer. Since most pop artists have their own musical style, it is possible that the affect of the background music on the vocal signal sometimes improves the SID performance, rather than fully degrading it. For example, some singers habitually use a guitar as the main instrument, making it reasonably easy to distinguish these singers from those who always sing to a piano accompaniment. To examine this, we performed an artist-identification (AID) experiment using the non-vocal regions of music recordings, instead of the vocal regions used in SID. In the training phase, all the manually-segmented non-vocal regions of the tracks in DB-S-1-T, belonging to a particular singer, were grouped and used to form an artist-specific music GMM, thereby creating 20 models. During a test, non-vocal regions extracted manually from an unknown track in DB-S-1-E were evaluated on each of the artist-specific music GMMs. The artist whose model output the highest likelihood was taken as the artist of the test recording.

Table V shows the AID results using different numbers of GMM mixture components. The best accuracy achieved was 23.0%, which was much better than the chance probability (5%). This result implies that a correlation between the background music and the singer does exist. To ascertain how background music affects SID, we compared the SID accuracy (the case of L = entire track) and AID accuracy with respect to each of the singers (artists). As shown in Fig. 5, singers associated with a lower SID accuracy tend to have a lower AID accuracy, and vice versa. We deduced that the background music is helpful for the SID when it is consistent in both training and testing conditions, but is detrimental when the background music in the test recording is "unseen" by the training materials. From Fig. 5, it can also be seen that SID based on solo voice modeling is relatively immune to background music, compared to the direct GMM-based SID.

D. Target Singer Detection Results

In our TSD experiments, the data used for the training of the target singer models and the universal singer model was, respectively, DB-S-1-T and DB-S-2. The test set included DB-S-1-E and DB-D. Each of the tracks in the test set was uniformly segmented into three non-overlapping music clips, and TSD was performed separately on each of the clips. The experiment was run in a leave-one-out manner, which uses each of the 20 singers in DB-S-1 as a target one at a time and rotates through all the singers. This produced 300 test samples treated as target singer trials and 5,700 test samples treated as non-target singer trials in DB-S-1-E [3]. There were also 72 test samples treated as target singer trails and 1,153 test samples treated as non-target singer trials in DB-D [4].

Assessment of the TSD performance was based on the miss detection rate (MDR) and false alarm rate (FAR), calculated by:

$$\text{MDR (in \%)} = \frac{\text{\#clips labeled as target but undetected}}{\text{\#clips labeled as target}} \times 100\%,$$

and

$$\text{FAR (in \%)} = \frac{\text{\# clips falsely-detected as target}}{\text{\#clips detected as target}} \times 100\%.$$

To show the trade-off between MDR and FAR, the results were reported on the detection error trade-off (DET) plot [27], which represents miss detection and false alarm according to their corresponding Gaussian deviates. In addition, the assessment of TSD performance can be alternatively represented as a single number via *d-prime measure* [e.g. 28]. This measure takes into account the difference between the probability of correct detection and the probability of false alarm. Given a set of testing music recordings, the log-likelihood difference on the left-hand side of Eq. (24) is computed for each of the recordings. *d-prime* can then be estimated

[3] Each of the singers in DB-S-1-E performed 5 distinct songs, and each song was uniformly segmented into 3 music clips. Therefore, whenever one singer was specified as the target, there were 15 (5 × 3) test samples performed by that singer and 285 (19 × 5 × 3) test samples not performed by that singer. Since the singer population in DB-S-1-E was 20, there were 300 (20 × 15) test samples treated as target trials and 5,700 (20 × 285) test samples treated as non-target trials.

[4] Each of the singers in DB-S-1 was present on one to three of the 22 tracks in DB-D. Some of the tracks in DB-D contained two singers in DB-S-1.

using

$$d' = \frac{|m_t - m_n|}{\sqrt{(\nu_t + \nu_n)/2}}, \tag{26}$$

where m_t and ν_t are, respectively, the mean and variance of the log-likelihood differences computed for the target singer trials, and m_n and ν_n are the mean and variance of the log-likelihood differences computed for the non-target singer trials. The larger the value of d', the better will be the performance.

Fig. 6 shows the TSD results obtained with automatic vocal/non-vocal segmentation. Here, the number of mixtures used in the target singer model, universal singer model, and background music model was empirically determined to be 48, 48, and 8, respectively. We can see that TSD in solo music is much easier than in duet music. Compared to the performance yielded by the direct GMM method without background music modeling, the effectiveness of the proposed solo voice modeling method was clearly demonstrated. The best equal error rate (MDR = FAR) for testing the solo music tracks and the duet music tracks were 12.4% and 19.6%, respectively.

E. Target Singer Tracking Results

Finally, performance of TST was evaluated on DB-D. The target singer set and the model configurations were the same as those used in the TSD experiments. The results were also reported in MR and FR. They were computed using

$$\text{MDR (in \%)} = \frac{\text{\#frames labeled as target but undetected}}{\text{\#frames labeled as target}} \times 100\%,$$

and

$$\text{FAR (in \%)} = \frac{\text{\#frames falsely-detected as target}}{\text{\#frames detected as target}} \times 100\%.$$

After discarding the frames that occurred within 0.5 seconds of a labeled switch-point, there were 232,746 test frames treated as target trials, 120,576 test frames treated as non-target trials, and 160,710 test frames treated as non-vocal trials. Among the 232,746 target trials, we found that 112,198 trials were from the frames containing only the target singer's voice, while 120,548 trials were from the frames containing the overlapping voices of the target singer and another simultaneous singer. In our evaluation, the overlapping voice frames are ignored in the error rate computation, because it is ambiguous to assign a singer attribute to them.

Fig. 7 shows the TST results obtained using an empirically-optimal segment length, $W = 200$ frames. It is clear that the benefit of the solo voice modeling method was demonstrated once again. The best equal error rate was 29.6%. From the results shown in Figs. 6 and 7, we can see that TST is more difficult than TSD. This is mainly because in TST, the vocal portions within a test recording are divided into relatively shorter segments and the examination is performed in a segment-by-segment manner, whereas in TSD, all the vocal portions within a test recording are examined as a whole, so more information about singers' voices can be exploited.

VI. CONCLUSIONS AND FUTURE WORK

We have examined the feasibility of automatic singer recognition in a pop music recording, and shown that the characteristics of a singer's voice can be extracted from music via vocal segment detection, followed by statistical analysis of the vocal signal. In particular, we have proposed a reliable model to eliminate the interference of background music for solo singer identification by leveraging statistical estimation of a piece's musical background. The solo voice modeling technique has also been utilized for solving the problems of target singer detection and target singer tracking. Experimental evaluations conducted on multi-singer music data have demonstrated the superiority of the proposed singer characteristic modeling over direct Gaussian mixture modeling, without taking background music into account.

Although this work shows that singers in pop music recordings can be distinguished from one another, the proposed solutions to the three singer-recognition tasks may only be regarded as a preliminary investigation of various potential applications. More work is needed to validate the practicality of the proposed methods based on current performance profiles. In particular, it is necessary to scale up the current singer-recognition experiments with a larger singer population. This would enable us to further examine the proposed methods with regard to mis-identified singers, and may also help in classifying singers based on similar singing voice characteristics. To scale up the system, a wider variety of music data, including various music styles, genres, singing languages, etc. must be acquired. Moreover, as Mandarin and other Asian pop music often sounds like the vocals are mixed louder than in Western music, it is worth comparing the singer-recognition results conducted on Asian and Western pop music.

As mentioned in Section I, this work does not deal with the problem in many practical applications that require singer identification, detection and tracking to be performed consecutively, or in parallel. For such a problem, there is a need to investigate how to best combine the techniques related to these three tasks within a unified framework. On the other hand, the problem concerning the correlation between the background music and the singer may need to be investigated further by using some cover-version music data. A popular song made famous by one artist is often performed repeatedly or re-recorded by many other artists. Such cover-versions usually have the same melody and similar accompaniment as the original version. Therefore, singer-recognition experiments conducted on cover-version music data could largely exclude the factors that may affect the objectivity of assessment.

Finally, our future work will extend the current singer-recognition methods to deal with background vocals or simultaneous singers. From the viewpoint of information retrieval, a music segment that contains both the target singer's voice and non-target singers' voices should be treated as relevant. However, we have found in our preliminary experiments that such music segments tend to poorly match the individual solo voice models of their singers. Specific techniques for detecting and identifying simultaneous-singer music are therefore required.

Acknowledgements

This work was partially supported by the National Science Council, Taiwan, ROC, under Grant No. NSC92-2422-H-001-093. The authors are very grateful to the anonymous reviewers and the associate editor, Dr. Mike Davies, for their careful reading of this article and their constructive suggestions.

References

[1] A. S. Durey and M. A. Clements, “Features for melody spotting using hidden Markov models,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1765-1768, Orlando, Florida, 2002.

[2] M. A. Akeroyd, B. C. J. Moore, and G. A. Moore, “Melody recognition using three types of dichotic-pitch stimulus,” *The Journal of the Acoustical Society of America*, vol. 110, no. 3, pp. 1498-1504, 2001.

[3] P. Herrera, X. Amatriain, E. Batlle, and X. Serra, “Towards instrument segmentation for music content description: a critical review of instrument classification techniques,” *Proceedings of 1st International Symposium on Music Information Retrieval*, Plymouth, Massachusetts, 2000.

[4] A. Eronen, “Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs,” *Proceedings of 7th International Symposium on Signal Processing and Its Applications*, pp. 133-136, Paris, France, 2003.

[5] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, 2002.

[6] C. Xu, N. C. Maddage, X. Shao; F. Cao, and Q. Tian, “Musical genre classification using support vector machines,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.429-432, Hong Kong, 2003.

[7] D. Byrd and T. Crawford, “Problems of music information retrieval in the real word,” *Information Processing and Management*, vol. 38, pp. 249-272, 2002.

[8] J. L. Hsu, C. C. Liu, and A. L. P. Chen, "Discovering nontrivial repeating patterns in music data," *IEEE Transactions on Multimedia*, vol. 3, no. 3, pp. 311-325, 2001.

[9] Y. E. Kim and B. Whitman, "Singer identification in popular music recordings using voice coding features," *Proceedings of 3rd International Conference on Music Information Retrieval*, pp. 164-169, Paris, France, 2002.

[10] C. C. Liu, and C. S. Huang, "A singer identification technique for content-based classification of MP3 music objects," *Proceedings of International Conference on Information and Knowledge Management*, pp. 438-445, McLean, Virginia, 2002.

[11] T. Zhang, "Automatic Singer Identification," *Proceedings of IEEE International Conference on Multimedia and Expo*, Baltimore, Maryland, 2003.

[12] M. A. Bartsch, and G. H. Wakefield, "Singing voice identification using spectral envelope estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 100-109, 2004.

[13] W. H. Tsai, H. M. Wang, and D. Rodgers, "Automatic singer identification of popular music recordings via estimation and modeling of solo vocal signal," *Proceedings of 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003.

[14] B. Whitman, G. Flake, and S. Lawrence, "Artist detection in music with minnowmatch," *Proceedings of IEEE Workshop on Neural Networks for Signal Processing*, pp. 559-568, Falmouth, Massachusetts, 2001.

[15] A. Berenzweig, D. P. W. Ellis, and S. Lawrence, "Using voice segments to improve artist classification of music," *Proceedings of International Conference on Virtual, Synthetic, and Entertainment Audio*, Espoo, Finland, 2002.

[16] J. P. Campbell, “Speaker recognition: a tutorial,” *PROCEEDINGS OF THE IEEE*, vol. 85, no. 9, pp. 1437-1462, 1997.

[17] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995.

[18] A. L. Berenzweig and D. P. W. Ellis, “Locating singing voice segments within music signals,” *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.119-122, New York, 2001.

[19] P. R. Cook, “Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing,” Ph.D. Thesis. Stanford University, Stanford, CA, 1990.

[20] B. Logan, “Mel frequency cepstral coefficients for music modeling,” *Proceedings of 1st International Symposium on Music Information Retrieval*, Plymouth, Massachusetts, 2000.

[21] M. R. Anderberg, “Cluster analysis for applications,” *Academic Press*, 1973.

[22] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, vol. 39, pp. 1-38, 1977.

[23] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical pattern recognition: a review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, 2000.

[24] A. Nadas, D. Nahamoo, and M. A. Picheny, “Speech recognition using noise-adaptive prototypes,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 10, pp. 1495-1503, 1989.

[25] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, “Integrated models of signal and background with application to speaker identification in noise,” IEEE Transactions on Speech and Audio Processing, vol. 2, no. 2, pp. 245-257, 1994.

[26] 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, *Proceedings of 4th International Conference on Music Information Retrieval*, Baltimore, 2003.

[27] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," *Proceedings of 5th European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997.

[28] D. M. Green, and J. A. Swets, "Signal Detection Theory and Psychophysics," Los Altos, Calif.: Peninsula Publishing, 1988.

TABLE I

DATABASE DESCRIPTION

Music data			Purpose
DB-S (242 solo tracks)	DB-S-1 (10 male & 10 female singers; 10 tracks/singer)	DB-S-1-T (5 tracks/singer)	Training of the vocal GMM & singer-specific models
		DB-S-1-E (5 tracks/singer)	Evaluation
	DB-S-2 (13 female & 8 male singers; 2 tracks/singer)		Training of the universal singer model
DB-D (22 duet tracks)			Evaluation
DB-I (174 instrumental-only tracks)			Training of the non-vocal GMM

TABLE II

RESULTS OF THE VOCAL/NON-VOCAL SEGMENTATION

(a) FRAME-BASED DECISION

Smoothing window (# frames)	1 (no smooth)	20	40	60	80
Accuracy (%)	70.3	73.4	77.3	78.2	77.6

(a) FIXED-LENGTH-SEGMENT-BASED DECISION

Segment length (# frames)	20	40	60	80
Accuracy (%)	74.7	78.9	80.8	80.1

(c) HOMOGENEOUS-SEGMENT-BASED DECISION (SMOOTHING WINDOW = 60 FRAMES)

# clusters for tokenization	2	4	8	16	32
Accuracy (%)	40.8	65.1	76.1	82.3	80.4

TABLE III

CONFUSION PROBABILITY MATRIX OF THE VOCAL/NON-VOCAL DISCRIMINATION

Actual	Hypothesized	
	Vocal	Non-vocal
Vocal	0.78	0.22
Non-vocal	0.09	0.91

TABLE IV

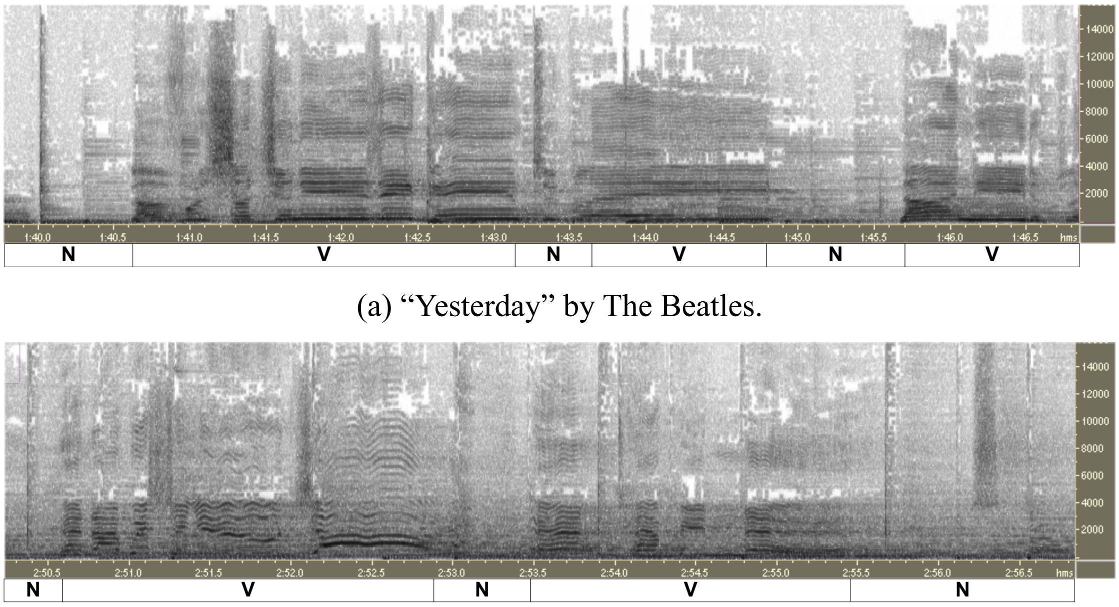
CONFUSION MATRIX OF THE SINGER IDENTIFICATION

Actual Singer Index	Hypothesized Singer Index																				Accuracy (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	102	0	0	0	0	0	0	2	2	0	0	1	0	0	0	0	0	0	0	0	95.3
2	0	80	1	4	0	0	4	0	0	3	0	0	0	0	0	0	0	0	0	0	87.0
3	0	1	127	1	0	0	2	1	1	16	0	0	0	0	0	0	0	0	0	1	84.7
4	0	0	0	81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100.0
5	0	0	12	0	81	0	8	0	2	13	2	0	0	0	0	0	0	0	0	0	68.6
6	2	1	2	5	0	66	0	5	6	2	0	0	0	0	0	0	0	0	0	0	74.2
7	2	0	2	0	2	0	104	0	0	8	0	0	0	0	0	0	0	0	0	0	88.1
8	7	0	6	2	2	1	7	98	1	4	0	0	0	0	0	0	0	0	2	0	75.4
9	1	0	5	1	6	1	0	0	94	2	0	1	0	1	0	0	0	0	0	2	82.5
10	0	0	2	1	1	0	3	4	0	100	0	0	0	0	0	0	0	0	0	0	90.1
11	1	0	0	0	0	0	1	0	0	1	105	0	0	1	1	0	0	0	0	0	95.5
12	1	0	0	0	1	0	0	2	0	0	6	109	0	8	2	0	0	0	0	0	84.5
13	1	0	0	0	0	0	0	0	2	0	0	3	78	3	0	0	0	0	0	0	90.7
14	3	0	4	0	0	0	0	0	1	0	4	3	6	59	0	0	0	0	0	0	73.8
15	0	0	0	0	0	0	0	0	0	0	20	0	0	1	98	1	0	0	4	0	79.0
16	0	0	0	0	0	1	0	3	0	0	0	0	0	5	0	102	0	0	0	0	91.9
17	0	0	2	0	0	2	0	1	4	0	2	1	0	1	0	0	71	2	5	0	78.0
18	2	0	0	0	0	0	1	0	0	0	2	16	3	7	0	0	0	71	5	0	66.4
19	2	1	2	0	1	0	4	1	1	2	8	0	1	1	2	0	0	0	85	0	76.6
20	1	0	0	0	0	0	0	0	0	0	3	4	1	0	6	3	0	2	2	62	73.8

TABLE V

RESULTS OF THE ARTIST IDENTIFICATION PERFORMED ON THE NON-VOCAL REGIONS OF THE MUSIC RECORDINGS

	No. of Mixture Components			
	8	16	32	48
Accuracy (in %)	18.0	16.0	23.0	20.0



(a) “Yesterday” by The Beatles.

(b) “I Will Always Love You” by Whitney Houston.

Fig. 1. Spectrogram of two music examples. The regions marked by “V” contain singing with accompaniments, whereas the regions marked by “N” contain instrumental sounds only.

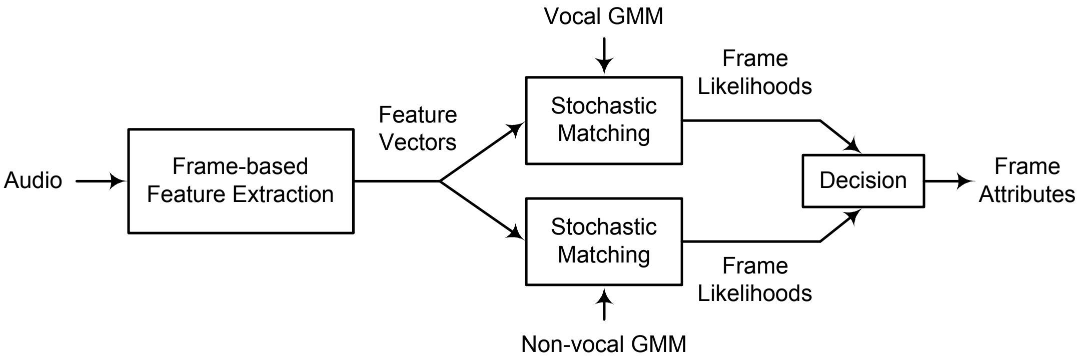


Fig. 2. Vocal/non-vocal segmentation.

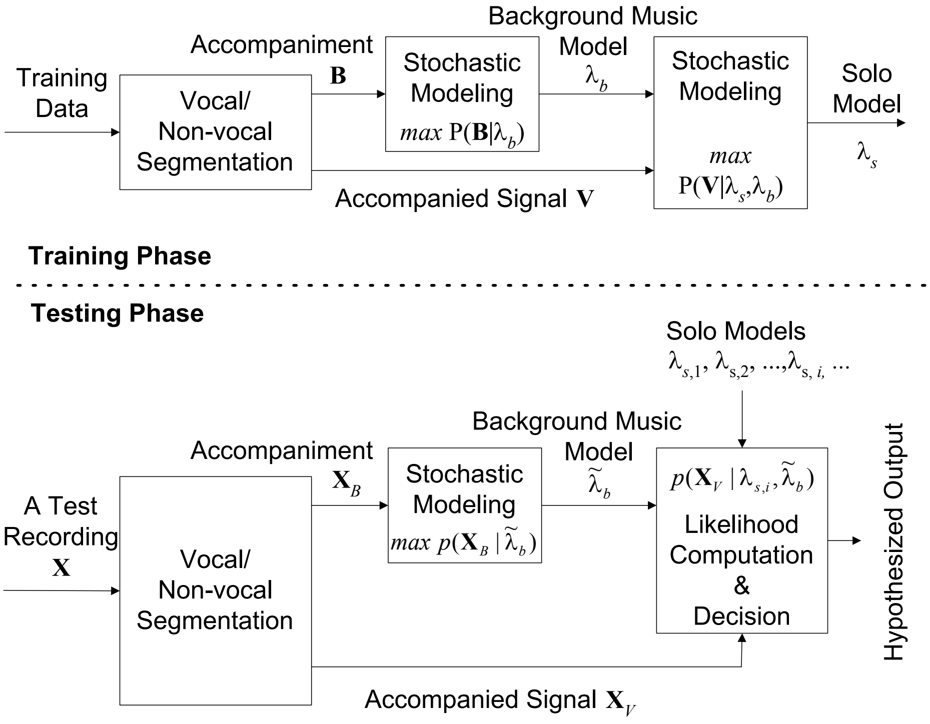


Fig. 3. Block diagram of the proposed singer-recognition system.

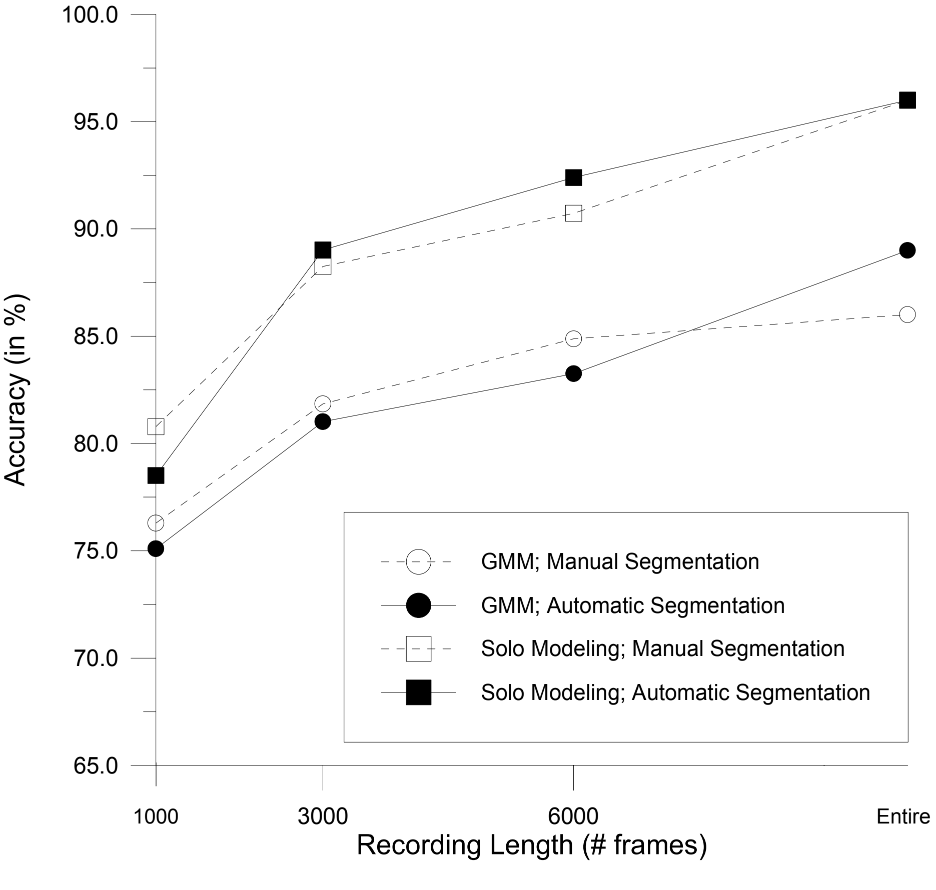


Fig. 4. Singer identification results.

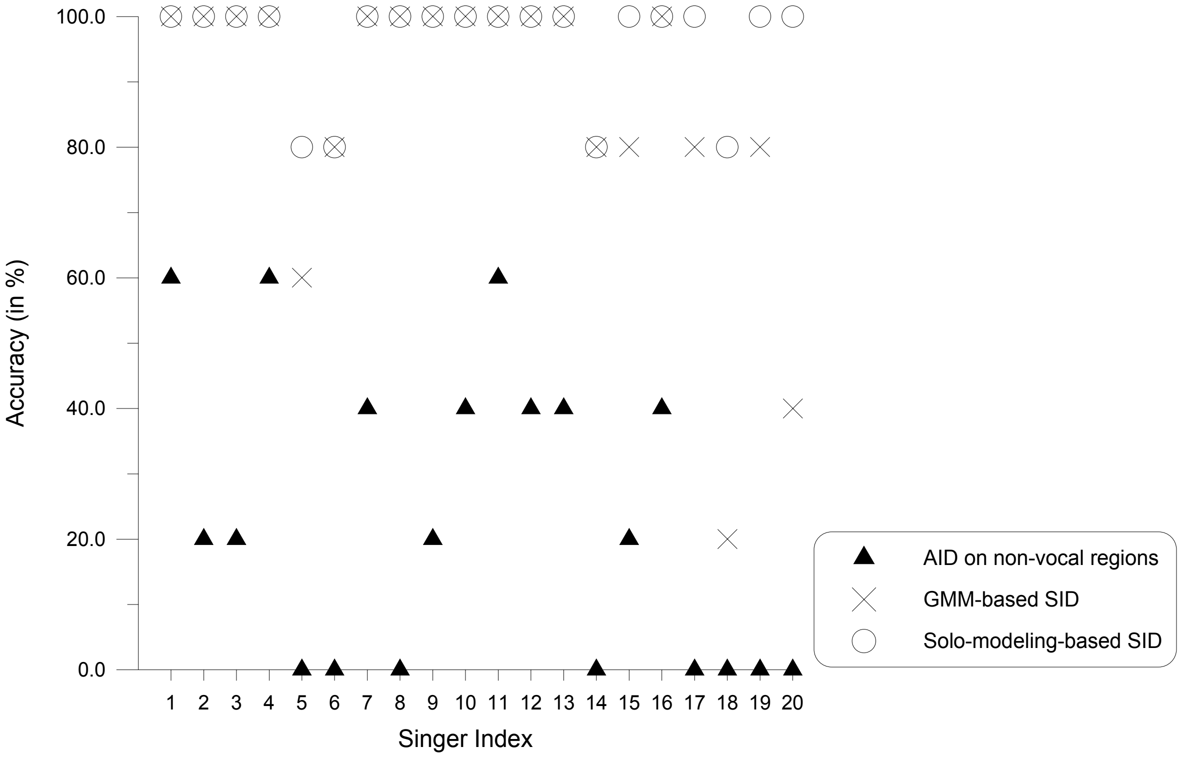


Fig. 5. SID and AID accuracies with respect to each of the 20 singers.

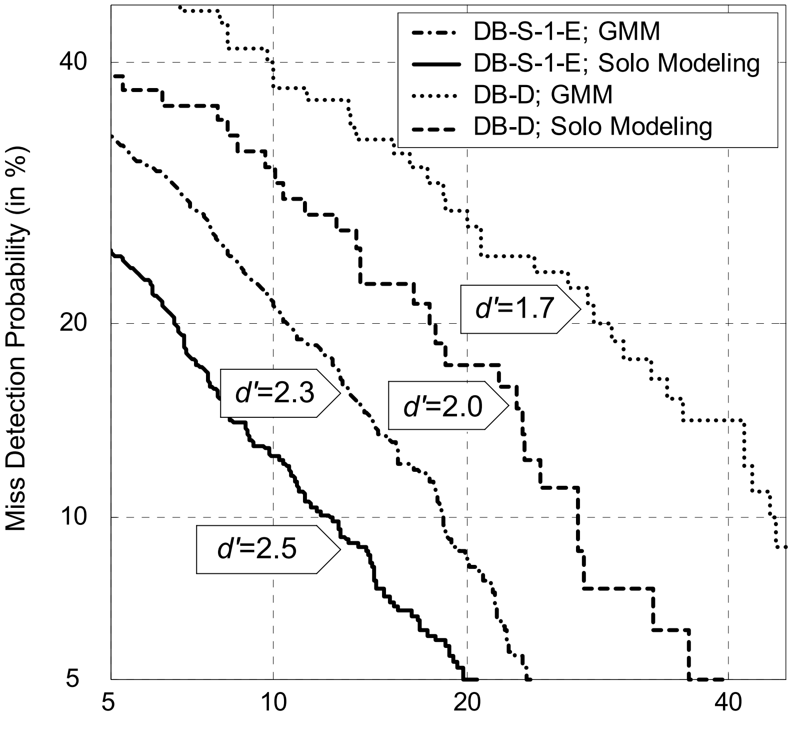


Fig. 6. Performance of the target singer detection.

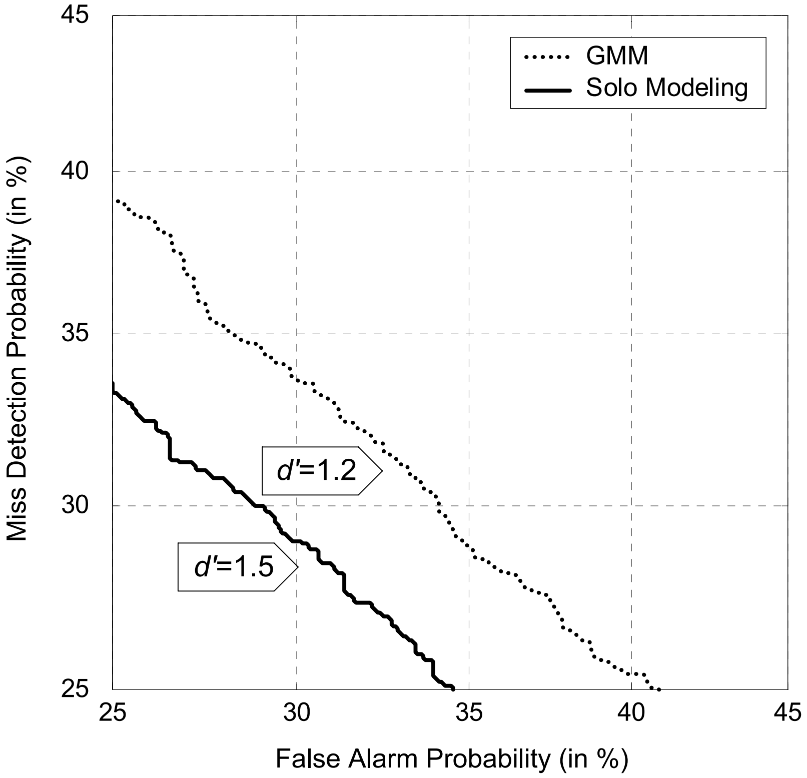


Fig. 7. Performance of the target singer tracking.