Blind Clustering of Popular Music Recordings Based on Singer Voice Characteristics

Wei-Ho Tsai, Hsin-Min Wang, Dwight Rodgers, Shi-Sian Cheng, and Hung-Ming Yu

Institute of Information Science, Academia Sinica

Nankang, 115, Taipei, Taiwan, Republic of China

{wesley,whm,dwight,sscheng,donny}@iis.sinica.edu.tw

Abstract

This paper presents an effective technique for automatically clustering undocumented music recordings based on their associated singer. This serves as an indispensable step towards indexing and content-based information retrieval of music by singer. The proposed clustering system operates in an unsupervised manner, in which no prior information is available regarding the characteristics of singer voices, nor the population of singers. Methods are presented to separate vocal from non-vocal regions, to isolate the singers' vocal characteristics from the background music, to compare the similarity between singers' voices, and to determine the total number of unique singers from a collection of songs. Experimental evaluations conducted on a 200-track pop music database confirm the validity of the proposed system.

1 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 the corresponding increase in our appetite for accessing them efficiently and conveniently. As a result, content-based retrieval of music has become an attractive topic of research, and efforts have been made to develop automatic classifiers or identifiers of music by melody (Durey and Clements, 2002; Akeroyd *et al*., 2002), genre (Tzanetakis and Cook, 2002), singer (Kim and Whitman, 2002; Liu and Huang, 2002), and other means (Byrd and Crawford, 2002). As an independent capability or as part of a music information retrieval system, techniques to automatically organize a collection of music recordings based on the associated singer are needed in order to lessen or replace human documentation efforts. This study addresses the general task of singer-based clustering of unknown music recordings, when neither singer information nor populations are available.

The most obvious application of singer-based clustering is in tools for expediently organizing and labeling unlabeled – or insufficiently well labeled – music collections. For instance, many rock music bands have a lead singer who sings the majority of all the band's songs, but a minority of songs will be sung by the guitarist, drummer, or other band-members. In such cases singer-based clustering may be used to identify those songs not sung by the lead singer. Furthermore, lead singers in both rock and pop music are known to quit, do solo albums, start new bands, or join other bands. Since the vast majority of documented music data is labeled by artist (band name), singer-based clustering may be useful for those wishing to find the full works of artists like Phil Collins, Sting, Ozzy Osbourne, or even Michael Jackson[1].

Singer-based clustering with support for multiple singers in a song may be able to identify guest appearances. For instance Queen's hit song "Under Pressure" included vocals by David Bowie, and Shaggy's hit song "Mr. Lover" featured Janet Jackson. Even when music databases are labeled, these appearances are often omitted, especially in live concert recordings and bootlegs (unauthorized amateur recordings of a professional live concert). In addition, many of the methods developed for singer-based clustering can be trivially applied to problems such as music recommendation systems – such a system could suggest music by singers with similar voices.

2 Problem formulation

Given a set $\mathcal{M}$ of unlabeled music recordings, each performed by one singer from a set $\mathcal{P}$, where $|\mathcal{M}| \geq |\mathcal{P}|$, and $|\mathcal{P}|$ is unknown, the system must partition $\mathcal{M}$ into K clusters such that $K = |\mathcal{P}|$ and each cluster consists exclusively of recordings from only one singer p from $\mathcal{P}$.[2] Performance of

[1] Each of these artists was in a band prior to becoming famous for solo work: Genesis, The Police, Black Sabbath, and The Jackson Five, respectively.

[2] This formulation is not applicable to recordings containing background vocals, or multiple singers, unless these recordings are pre-segmented into singer-homogenous regions. In this research we limit ourselves to single-singer recordings.

the clustering is evaluated on the basis of average cluster purity (Solomonoff *et al*., 1998), defined as

$$\bar{\rho} = \frac{1}{|\mathcal{M}|}\sum_{k=1}^{K} n_k \rho_k , \quad (1)$$

and

$$\rho_k = \sum_{p \in \Phi} \frac{n_{kp}^2}{n_k^2}, \quad (2)$$

where ρ_k is the purity of the cluster k, n_k is the total number of recordings in the cluster k, and n_{kp} is the number of recordings in the cluster k that were performed by singer p.

3 Method overview

The purpose of singer-based clustering is to cluster the recordings on the basis of the singer's voice rather than the background music, musical genre, or other characteristics of the recording. Therefore, it is necessary to extract, model, and compare the characteristic features of the singers' voices without interference from non-singer features. Toward this end, a three stage process as shown in Fig. 1 is proposed: (1) segmentation of each recording into vocal and non-vocal segments, where a vocal segment consists of concurrent singing and accompaniment, whereas non-vocal segments consist of accompaniment only; (2) distillation of the singer's stochastic vocal characteristics from the vocal segments by specifically suppressing the characteristics of the background (non-vocal segments); and (3) clustering of the recordings based on singer characteristic similarity. Details of each of the stages will be described in the subsequent sections.

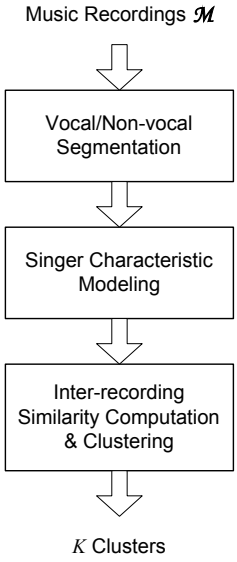

Figure 1: Block diagram of the proposed singer-based clustering.

4 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, in analogy with the study by Berenzweig and Ellis (2001). However, in contrast to their work, which uses a speech recognizer for detecting singing voices, we propose to construct a statistical classifier with parametric models trained using accompanied singing voices rather than normal speech. As shown in Fig. 2, the classifier consists of a front-end signal processor that converts digital waveforms into spectrum-based feature vectors, followed by a backend statistical processor that performs modeling, matching and decision making. It 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. The use of GMMs is motivated by the desire for modeling various broad acoustic classes by a combination of Gaussian components. These broad acoustic classes reflect some general vocal tract and instrumental configurations. It has been shown that GMMs have a strong ability to provide smooth approximations to arbitrarily-shaped densities of spectrum over a long time span (Reynolds and Rose, 1995). 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 and iteratively adjusted via expectation-maximization (EM) (Dempster *et al*., 1977).

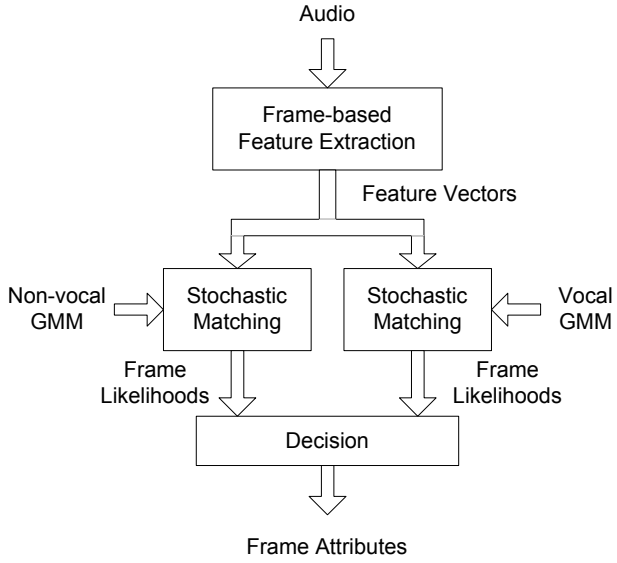

Figure 2: Vocal/non-vocal segmentation.

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

A. Frame-based decision.

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

$$\log p(\boldsymbol{x}_t \mid \lambda_V) - \log p(\boldsymbol{x}_t \mid \lambda_N) \underset{\text{non - vocal}}{\overset{\text{vocal}}{\gtrless}} \eta, \tag{3}$$

where η is a threshold. Since singing tends to be continuous, these results may be smoothed in the time domain. For smoothing, the frames are divided 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.

An improvement of the smoothing above can be made by assigning a single classification per segment directly, using:

$$\frac{1}{W}\left(\sum_{i=0}^{W-1}\log p(\boldsymbol{x}_{tW+i} \mid \lambda_V) - \sum_{i=0}^{W-1}\log p(\boldsymbol{x}_{tW+i} \mid \lambda_N)\right) \underset{\text{non - vocal}}{\overset{\text{vocal}}{\gtrless}} \eta. \tag{4}$$

where t 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.

C. Homogeneous-segment-based decision.

Further improvement can be made by merging adjacent segments if they do not cross a vocal/non-vocal boundary. In this study, vector clustering is employed on the set of all frame feature vectors and each frame is assigned the cluster index associated with that frame's feature vector. Each segment is then assigned the majority index of its constituent frames, and adjacent segments are merged as a homogeneous segment if they have the same index. Classification is then made per homogeneous-segment using:

$$\frac{1}{W_k}\left(\sum_{i=0}^{W_k-1}\log p(\boldsymbol{x}_{s_k+i} \mid \lambda_V) - \sum_{i=0}^{W_k-1}\log p(\boldsymbol{x}_{s_k+i} \mid \lambda_N)\right) \underset{\text{non - vocal}}{\overset{\text{vocal}}{\gtrless}} \eta. \tag{5}$$

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

5 Singer characteristic modeling

To cluster recordings by singer, the characteristics of the singer's voice must be distilled from the mixture of voice and accompaniment. Let $\mathbf{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_T\}$ represent the feature vectors from a vocal region. $\mathbf{V}$ can be modeled as a mixture of a solo voice $\mathbf{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_T\}$ and a background accompaniment $\mathbf{B} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_T\}$, where $\mathbf{S}$ and $\mathbf{B}$ are unobservable but $\mathbf{B}$'s stochastic characteristics can be approximated from the non-vocal segments. This section presents a method of obtaining a stochastic model λ_s for the solo signal $\mathbf{S}$.

Based on the techniques developed in robust speech and speaker recognition (Rose *et al*., 1994; Nadas *et al*., 1989), it is assumed that the solo signal and background music are, respectively, drawn randomly and independently according to GMMs $\lambda_s = \{w_{s,i}, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i} \mid 1 \le i \le I\}$, and $\lambda_b = \{w_{b,j}, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j} \mid 1 \le j \le J\}$, where $w_{s,i}$ and $w_{b,j}$ are mixture weights, $\boldsymbol{\mu}_{s,i}$ and $\boldsymbol{\mu}_{b,j}$ mean vectors, and $\boldsymbol{\Sigma}_{s,i}$ and $\boldsymbol{\Sigma}_{b,j}$ covariance matrices. If the accompanied signal is formed from a generative function $\boldsymbol{v}_t = f(\boldsymbol{s}_t, \boldsymbol{b}_t)$, $1 \le t \le T$, the probability of $\mathbf{V}$, given λ_s and λ_b can be represented by

$$p(\mathbf{V} \mid \lambda_s, \lambda_b) = \prod_{t=1}^{T}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} w_{s,i} w_{b,j}\, p(\boldsymbol{v}_t \mid i, j, \lambda_s, \lambda_b)\right\}, \tag{6}$$

where

$$p(\boldsymbol{v}_t \mid i, j, \lambda_s, \lambda_b) = \iint_{\boldsymbol{v}_t = f(\boldsymbol{s}_t, \boldsymbol{b}_t)} \mathcal{N}(\boldsymbol{s}_t; \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i})\mathcal{N}(\boldsymbol{b}_t; \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}) d\boldsymbol{s}_t d\boldsymbol{b}_t. \tag{7}$$

It is desired to estimate the solo voice model λ_s, given the accompanied voice $\mathbf{V}$ and the background music model λ_b. This can be done in a maximum likelihood manner as follows:

$$\lambda_s^* = \arg\max_{\lambda_s} p(\mathbf{V} \mid \lambda_s, \lambda_b). \tag{8}$$

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

$$Q(\lambda_s, \hat{\lambda}_s) = \sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{J} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) \log p(i, j, \boldsymbol{v}_t \mid \hat{\lambda}_s, \lambda_b), \tag{9}$$

where

$$p(i, j, \boldsymbol{v}_t \mid \hat{\lambda}_s, \lambda_b) = w_{s,i} w_{b,j}\, p(\boldsymbol{v}_t \mid i, j, \hat{\lambda}_s, \lambda_b), \tag{10}$$

and

$$p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) = \frac{w_{s,i} w_{b,j}\, p(\boldsymbol{v}_t \mid i, j, \lambda_s, \lambda_b)}{\sum_{m=1}^{I}\sum_{n=1}^{J} w_{s,m} w_{b,n}\, p(\boldsymbol{v}_t \mid m, n, \lambda_s, \lambda_b)}. \tag{11}$$

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

$$\hat{w}_{s,i} = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b), \tag{12}$$

$$\hat{\boldsymbol{\mu}}_{s,i} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{N} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) \cdot E\{\boldsymbol{s}_t \mid \boldsymbol{v}_t, i, j, \lambda_s, \lambda_b\}}{\sum_{t=1}^{T}\sum_{j=1}^{N} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b)}, \tag{13}$$

$$\hat{\boldsymbol{\Sigma}}_{s,i} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{J} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) \cdot E\{\boldsymbol{s}_t \boldsymbol{s}_t' \mid \boldsymbol{v}_t, i, j, \lambda_s, \lambda_b\}}{\sum_{t=1}^{T}\sum_{j=1}^{J} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b)} - \boldsymbol{\mu}_{s,i}\boldsymbol{\mu}_{s,i}', \tag{14}$$

where prime denotes vector transpose, and $E\{\cdot\}$ denotes expectation.

Suppose that **V**, **S** and **B** are log-spectrum features, and the background music is additive in the time domain or linear-spectrum domain. The accompanied signal can then be approximately expressed by $\boldsymbol{v}_t \approx \max(\boldsymbol{s}_t, \boldsymbol{b}_t)$, $1 \le t \le T$, according to Nadas' MIXMAX model (Nadas *et al*., 1989). For implementation efficiency, the covariance matrices of the GMMs used in this study are assumed diagonal, and thus each vector component involved can be operated on independently. Denoting by $\sigma_{s,i}^2$ and $\sigma_{b,j}^2$, respectively, an arbitrary diagonal component of $\boldsymbol{\Sigma}_{s,i}$ and $\boldsymbol{\Sigma}_{b,j}$, and focusing on scalar observations for ease of discussion, we compute Eq. (7) using

$$\begin{aligned} & p(v_t \mid i, j, \lambda_s, \lambda_b) \\ &= \iint_{v_t = f(s_t, b_t)} \mathcal{N}(s_t; \mu_{s,i}, \sigma_{s,i}^2)\mathcal{N}(b_t; \mu_{b,j}, \sigma_{b,j}^2) ds_t db_t \\ &= \mathcal{N}(s_t; \mu_{s,i}, \sigma_{s,i}^2) \int_{-\infty}^{v_t} \mathcal{N}(b_t; \mu_{b,j}, \sigma_{b,j}^2) db_t \\ &\quad + \mathcal{N}(b_t; \mu_{b,j}, \sigma_{b,j}^2) \int_{-\infty}^{v_t} \mathcal{N}(s_t; \mu_{s,i}, \sigma_{s,i}^2) ds_t \\ &= \mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2)\Phi\left(\frac{v_t - \mu_{b,j}}{\sigma_{b,j}}\right) \\ &\quad + \mathcal{N}(v_t; \mu_{b,j}, \sigma_{b,j}^2)\Phi\left(\frac{v_t - \mu_{s,i}}{\sigma_{s,i}}\right), \end{aligned} \tag{15}$$

where

$$\Phi(\tau) = \int_{-\infty}^{\tau} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw. \tag{16}$$

The value of $\Phi(\tau)$ can be obtained using a table of the error function. On the other hand, the conditional expectation in Eq. (13) can be shown in the following form:

$$\begin{aligned} E\{s_t \mid v_t, i, j, \lambda_s, \lambda_b\} &= p(s_t = v_t \mid i, j, \lambda_s, \lambda_b) \cdot v_t \\ &\quad + \left(1 - p(s_t = v_t \mid i, j, \lambda_s, \lambda_b)\right) \cdot E\{s_t \mid s_t < v_t, i, j, \lambda_s, \lambda_b\}, \end{aligned} \tag{17}$$

where

$$\begin{aligned} & p(s_t = v_t \mid i, j, \lambda_s, \lambda_b) \\ &= \frac{\mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2)\Phi\left(\frac{v_t - \mu_{b,j}}{\sigma_{b,j}}\right)}{\mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2)\Phi\left(\frac{v_t - \mu_{b,j}}{\sigma_{b,j}}\right) + \mathcal{N}(v_t; \mu_{b,j}, \sigma_{b,j}^2)\Phi\left(\frac{v_t - \mu_{s,i}}{\sigma_{s,i}}\right)}, \end{aligned} \tag{18}$$

and

$$E\{s_t \mid s_t < v_t, i, j, \lambda_s, \lambda_b\} = \mu_{s,i} - \sigma_{s,i} \frac{\mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2)}{\Phi\left(\frac{v_t - \mu_{s,i}}{\sigma_{s,i}}\right)}. \tag{19}$$

Similarly, the conditional expectation in Eq. (14) is computed using

$$\begin{aligned} E\{s_t^2 \mid v_t, i, j, \lambda_s, \lambda_b\} &= p(s_t = v_t \mid i, j, \lambda_s, \lambda_b) \cdot v_t^2 \\ &\quad + \left(1 - p(s_t = v_t \mid i, j, \lambda_s, \lambda_b)\right) \cdot E\{s_t^2 \mid s_t < v_t, i, j, \lambda_s, \lambda_b\}, \end{aligned} \tag{20}$$

where

$$\begin{aligned} & E\{s_t^2 \mid s_t < v_t, i, j, \lambda_s, \lambda_b\} \\ &= \mu_{s,i}^2 + \sigma_{s,i}^2 - (\mu_{s,i} + v_t)\sigma_{s,i} \frac{\mathcal{N}(v_t; \mu_{s,i}, \sigma_{s,i}^2)}{\Phi\left(\frac{v_t - \mu_{s,i}}{\sigma_{s,i}}\right)}. \end{aligned} \tag{21}$$

Note that if the number of mixtures in the background music GMM is zero then this 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 modeling method.

6 Similarity computation & clustering

Finally, to compare and cluster the singers, each recording is evaluated against each singer's solo model in a method extended from (Tsai *et al*., 2001). From section 5, a solo model $\lambda_{s,i}$ and a background music model $\lambda_{b,i}$ is generated for each of the M recordings to be clustered, $1 \le i \le M$. The log-likelihood, $L_{i,j} = \log p(\mathbf{V}_i|\lambda_{s,j},\lambda_{b,i})$, $1 \le i, j \le M$, that the vocal portion of the recording $\mathbf{V}_i$ tests against the model $\lambda_{s,j}$, is then computed using Eq. (6) and (15). A large log-likelihood $L_{i,j}$ should indicate that the singer of song i is similar to the singer of song j. Singer-based clustering can be formulated as a conventional vector clustering algorithm by assigning the characteristic vector $\mathbf{L}_i = [L_{i,1}, L_{i,2}, \ldots, L_{i,M}]'$, $1 \le i \le M$, to each recording i, and computing the similarity between two recordings using the Euclidean distance: $\| \mathbf{L}_i - \mathbf{L}_j \|$.

The clustering quality may be improved by emphasizing the larger likelihoods and suppressing the smaller ones. To achieve this, the $L_{i,j}$ for each recording i are ranked in descending order of likelihood. Let the rank of $L_{i,j}$ be denoted by $R_{i,j}$. Then, the characteristic vectors $\mathbf{F}_i = [F_{i,1}, F_{i,2}, \ldots, F_{i,M}]'$, $1 \le i \le M$, are formed using

$$F_{i,j} = \begin{cases} 1.0 & , \ j = i \\ \exp\{\alpha(L_{i,j} - L_{i,\varphi})\} & , \ j \ne i \text{ and } R_{i,j} \le \theta \\ 0.0 & , \ j \ne i \text{ and } R_{i,j} > \theta \end{cases}, \tag{22}$$

and

$$\varphi = \arg\max_{k \ne i} L_{i,k}, \tag{23}$$

where α is a positive constant for scaling, and θ is an integer constant for pruning the lower log-likelihoods. Example characteristic vectors computed for a collection of 25 music recordings with 5 singers are shown as columns in Fig. 3. Dark regions in the figure represent large values, while light regions represent small ones. This figure shows that the vectors $\mathbf{F}_i$ more clearly distinguish between different singers than vectors $\mathbf{L}_i$.

To solve the vector clustering problem, we use the k-means algorithm, which starts with a single cluster and recursively splits clusters in attempts to minimize the within-cluster variances. A choice must be made as to how many clusters should be created. If the number of clusters is low, a single cluster is likely to include recordings from multiple singers. On the other hand, if the number of clusters is too high, a single singer's recordings will be split across multiple clusters.

Clearly the optimal number of clusters K is equal to the number of singers, which is unknown.

In this study, the Bayesian Information Criterion (BIC) (Schwarz, 1978) is employed to decide the best value of K. The BIC assigns a value to a stochastic model based on how well the model fits a data set, and how simple the model is, specifically

$$\mathrm{BIC}(\Lambda) = \log p(\boldsymbol{\mathcal{D}} \mid \Lambda) - \frac{1}{2}\gamma\, d \log |\boldsymbol{\mathcal{D}}|, \tag{24}$$

where d is the number of free parameters in model Λ, $|\boldsymbol{\mathcal{D}}|$ is the size of the data set $\boldsymbol{\mathcal{D}}$, and γ is a penalty factor. A K-clustering can be modeled as a collection of Gaussian distributions (one per cluster). The BIC may then be computed by:

$$\mathrm{BIC}(K) = -\sum_{k=1}^{K}\left(\frac{n_k}{2}\log|\boldsymbol{\Sigma}_k|\right) - \frac{1}{2}\gamma\, K\left(M + \frac{1}{2}M(M+1)\right)\log M, \tag{25}$$

where n_k is the number of elements of the cluster k, and $\boldsymbol{\Sigma}_k$ the covariance matrix of the characteristic vectors in the cluster k,. The BIC value should increase as the splitting improves conformity of the model, but should decline significantly after an excess of clusters are created. A reasonable number of clusters can be determined by

$$K^* = \arg\max_{1 \le K \le M} \mathrm{BIC}(K). \tag{26}$$

7 Experimental results

The music data used in this study consisted of 416 tracks from Mandarin pop music CDs. All the tracks were manually labeled with the singer identity and the vocal/non-vocal boundaries. The database was divided into two subsets, denoted as DB-1 and DB-2, respectively. The DB-1 comprised 200 tracks performed by 10 female and 10 male singers, with 10 distinct songs per singer. DB-2 contained the remaining 216 tracks, involving 13 female and 8 male singers, none of whom appeared in DB-1. All music data were down-sampled from the CD sampling rate of 44.1 kHz to 22.05 kHz, to exclude the high frequency components beyond the range of normal singing voices.

Extensive computer simulations were conducted to evaluate the performance of the proposed singer-clustering system. The vocal and non-vocal GMMs were trained using DB-2, and overall performance of the system was then evaluated using DB-1. The feature vectors used in the system were Mel-scale frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980; Logan, 2000), computed using a 32-ms Hamming-windowed frame with 10-ms frame shifts.

Our first experiments tested the validity of the vocal/non-vocal segmentation methods. Accuracy was computed by comparing the hypothesized attribute of each frame with the manual label[3]. 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 1 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 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 was 79.8%.

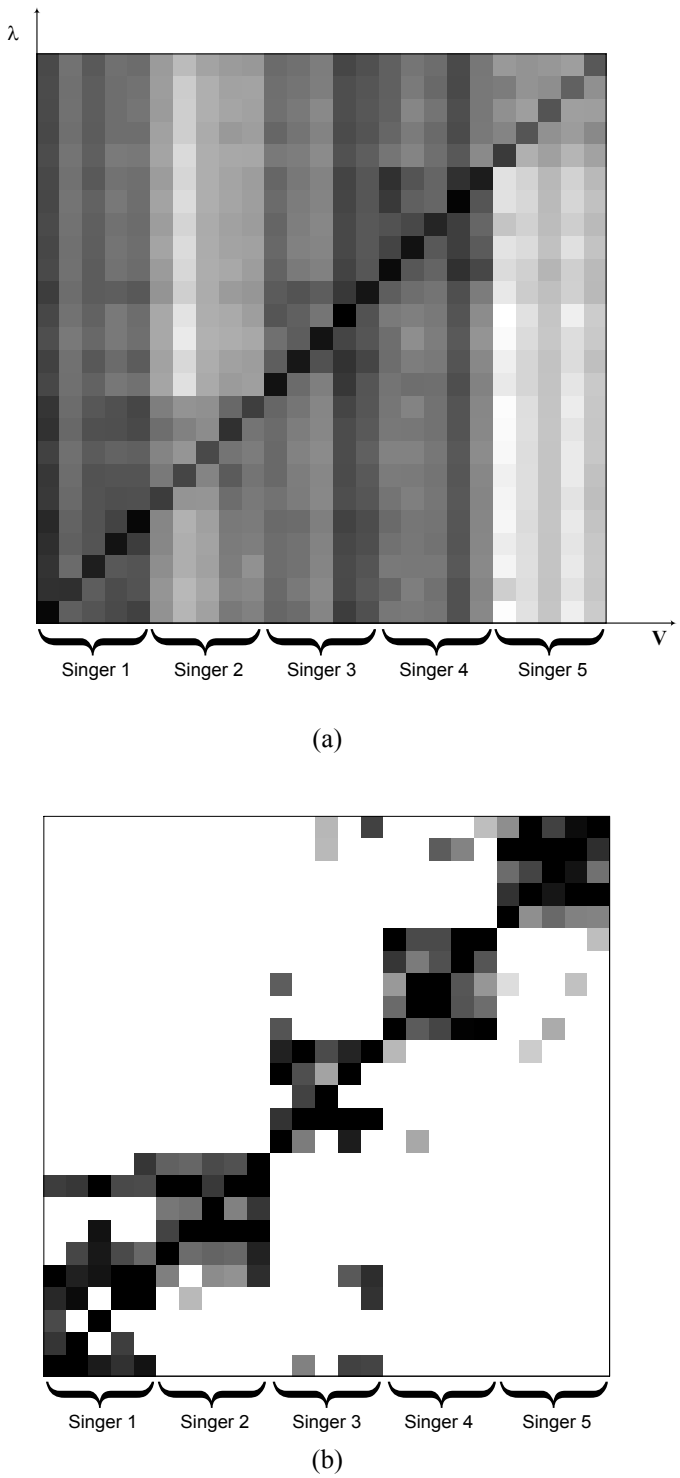

Figure 3: A gray scale representation of the log-likelihoods and characteristic vectors. (a) log-likelihoods; (b) characteristic vectors ($\theta = 6$).

[3] Accuracy (%) = # correctly-identified frames / # total frames × 100%

Smoothing window (# frames)	1 (no smooth)	20	40	60	80
Accuracy (%)	68.5	72.1	76.5	76.8	76.0

(a) Frame-based decision

Segment length (# frames)	20	40	60	80
Accuracy (%)	73.9	77.2	77.6	76.9

(b) Fixed-length-segment-based decision

# clusters for tokenization	2	4	8	16	32
Accuracy (%)	42.3	65.6	75.9	79.8	78.5

(c) Homogeneous-segment-based decision (Smoothing window = 60 frames)

Table 1: Results of vocal/non-vocal segmentation.

Table 2 shows the confusion probability matrix from the vocal/non-vocal discrimination results of the homogeneous-segment-based decision. The rows of the confusion matrix correspond to the ground-truth of the segments while the columns indicate the hypotheses. We can see that the majority of errors are misidentifications of vocal segments. Qualitatively, we found that many falsely identified 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 the singer clustering.

Actual	Hypothesized	
	Vocal	Non-vocal
Vocal	0.75	0.25
Non-vocal	0.13	0.87

Table 2: Confusion probability matrix of the vocal/non-vocal discrimination.

Next, the entire singer-clustering system, based on both manual segmentation and the best results of automatic segmentation, was examined on DB-1. Fig. 4 shows the average purity as a function of the number of clusters. We can see that as expected, the average purity gains sharply as the number of clusters increases in the beginning and then tends to saturate after too many clusters are created. Comparing the results with and without explicit usage of the background model in extracting the solo information, the effectiveness of the solo signal modeling over direct vocal modeling is clearly demonstrated. When the number of clusters is equal to the singer population ($K = P = 20$), the highest purities of 0.87 and 0.77 were yielded by using manual segmentation and automatic segmentation, respectively. This confirms that the system is capable of grouping the music data according to singer.

Lastly, the problem of automatically determining the number of singers was investigated. A series of clustering experiments were conducted using 50 music recordings (5 singers × 10 tracks), 100 music recordings (10 singers × 10 tracks), 150 music recordings (15 singers × 10 tracks), and 200 music recordings (20 singers × 10 tracks), respectively. Fig. 5 shows the resulting BIC values as a function of cluster number. The peak of each curve in the figure is located very close to the actual number of singers, validating the proposed BIC-based decision criterion.

8 Conclusions

This study has examined the feasibility of unsupervised clustering of music data based on their singer. It has been shown that the characteristics of a singer's voice can be extracted from music via vocal segment detection followed by solo vocal signal modeling. Singer-based clustering has been formulated and solved using a vector-clustering framework with reliable estimation of the correct number of clusters.

Although viable results have been reported in this paper, more work is needed to validate the proposed methods for a wider variety of music data, such as larger singer populations and richer songs with different genres. Furthermore, future work for singer-based clustering will extend the current system to handle duets, chorus, background vocals, or other music data with multiple simultaneous or non-simultaneous singers.

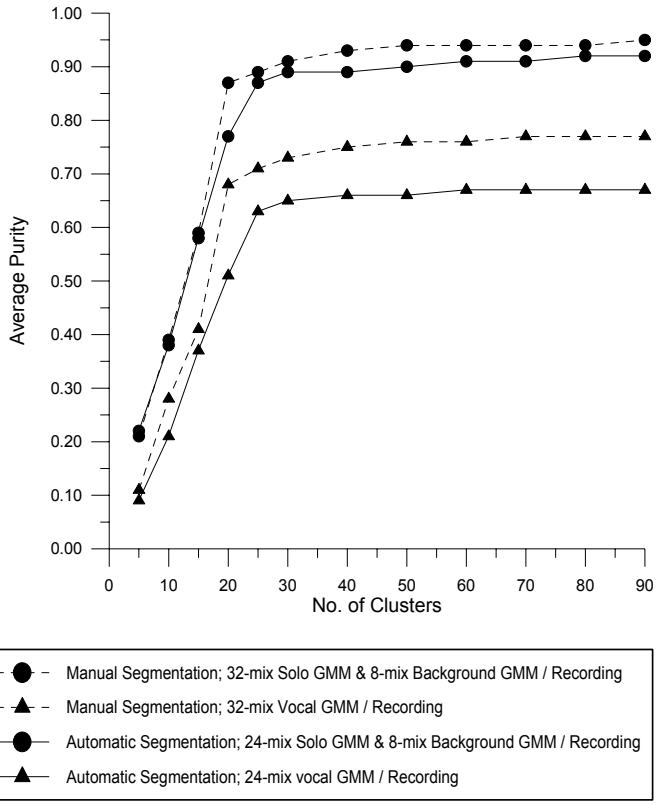

Figure 4: Results of singer-based clustering.

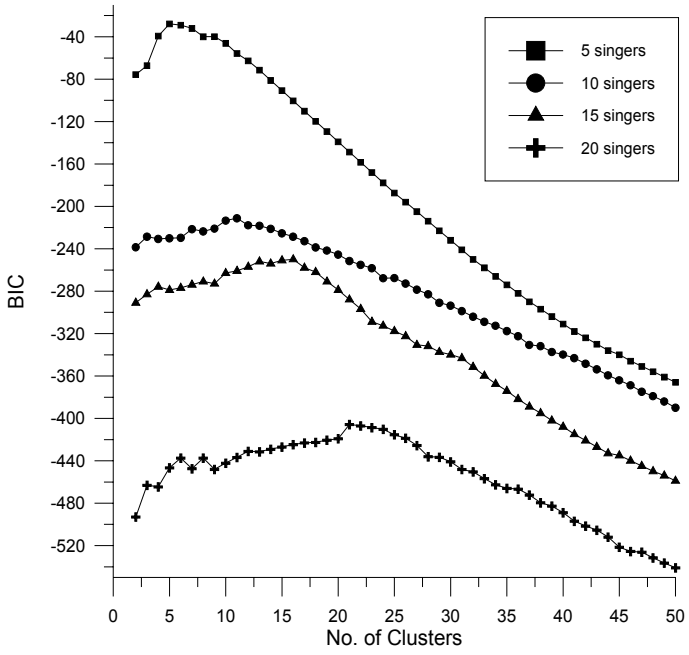

Figure 5: BIC measurements after each split.

Acknowledgements

This research was partially supported by the National Science Council, Taiwan, ROC, under Grant No. NSC92-2422-H-001-093.

References

Akeroyd, M. A., Moore, B. C. J., and Moore, G. A. (2001). Melody recognition using three types of dichotic-pitch stimulus. *The Journal of the Acoustical Society of America*, 110(3), 1498–1504.

Byrd, D. & Crawford, T. (2002). Problems of music information retrieval in the real word. *Information Processing and Management*, 38(2), 249–272.

Davis, S. & Mermelstein, P (1980) Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4), 357–366.

Dempster, A. Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39, 1–38.

Durey, A. S. & Clements, M. A. (2002). Features for melody spotting using hidden Markov models. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2 (pp. 1765–1768). Orlando, Florida.

Kim, Y. E. & Whitman, B. (2002). Singer identification in popular music recordings using voice coding features. In *Proceedings of Interational Conference on Music Information Retrieval* (pp. 164–169), Paris, France.

Liu, C. C. & Huang, C. S. (2002). A singer identification technique for content-based classification of MP3 music objects. In *Proceedings of International Conference on Information and Knowledge Management* (pp. 438–445), McLean, Virginia.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Proceedings of Interational Symposium on Music Information Retrieval*, Plymouth, Massachusetts.

Nadas, A., Nahamoo, D., and Picheny, M. A. (1989). Speech recognition using noise-adaptive prototypes. *IEEE Tranactions on Acoustics, Speech, and SignalProcessing*, 37(10), 1495−1503.

Reynolds, D. A. & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1), 72–83.

Rose, R. C., Hofstetter, E. M., and Reynolds, D. A. (1994). Integrated models of signal and background with application to speaker identification in noise. *IEEE Transactions on Speech and Audio Processing*, 2(2), 245−257.

Schwarz, G. (1978). Estimation the dimension of a model. *The Annals of Statistics*, 6, 461–464.

Solomonoff, A., Mielke, A., Schmidt, M., and Gish, H. (1998). Clustering speakers by their voices. In *Proceeding of IEEE Conference on Acoustics, Speech, and Signal Processing*, 2, (pp. 757−760), Seattle, Washington.

Tsai, W. H., Chu, Y. C., Huang, C. S., and Chang, W. W. (2001). Background learning of speaker voices for text-independent speaker identification. In *Proceedings of European Conference on Speech Communication and Technology*, Aalborg, Denmark.

Tzanetakis, G. & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), 293–302.