Automatic Singer Identification of Popular Music Recordings via Estimation and Modeling of Solo Vocal Signal

Wei-Ho Tsai, *Hsin-Min Wang*, and *Dwight Rodgers*

Institute of Information Science, Academia Sinica, Taiwan, Republic of China

wesley@iis.sinica.edu.tw

Abstract

This study presents an effective technique for automatically identifying the singer of a music recording. Since the vast majority of popular music contains background accompaniment during most or all vocal passages, directly acquiring isolated solo voice data for extracting the singer's vocal characteristics is usually infeasible. To eliminate the interference of background music for singer identification, we leverage statistical estimation of a piece's musical background to build a reliable model for the solo voice. Validity of the proposed singer identification system is confirmed via the experimental evaluations conducted on a 23-singer pop music database.

1. Introduction

For more than three decades, automatic extraction of information from human voices has been an important research task for facilitating human-machine communication. Copious amounts of research have gone into the problem of both speech recognition and speaker recognition (e.g. [1,2]) – decoding the underlying linguistics messages in an utterance, and determining the identity of the person speaking an utterance. In keeping with this research target, this study addresses the problem of singer recognition, which aims at determining the identity of the singer performing a particular song. Singer recognition can be viewed as the ultimate exercise in robust, content independent speaker recognition. Here, not only is the speaker's vocal characteristics modulated by an unknown utterance, prosody, and melody, but the entire vocal signal is superimposed on a loud, non-stationary background of musical accompaniment.

Automatic singer identification (singer ID) is an emerging field [3, 4] spurred by the rapid proliferation of popular music on the Internet. In contrast to classification of complete songs based on genre or other means [5], singer ID can be used to find cameo's or guest appearances in live concert recordings, to identify the singers in a movie's musical interludes, to distinguish between an original song and a cover-band, or otherwise to obtain singer identity information where it may be undocumented or difficult to find. Furthermore, singer ID may also enable 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.

2. A Preliminary Experiment on Singer ID

2.1. System Configuration

The basic strategy applied here is to adapt statistical methods developed in the speaker identification realm [2,6] to singer ID. Our baseline system consists of a front-end signal processor that converts music recordings from their digital waveform representations into streams of spectrum-based feature vectors, followed by a backend statistical processor that performs modeling and matching. It operates in two phases, training and testing.

During training, each waveform from a set of training songs from various singers is manually segmented into vocal regions and instrumental regions, where vocal regions typically include both singing and accompaniment. The resulting vocal segments pertaining to each of the singers are then used to form a Gaussian mixture model (GMM). The main attraction of the GMM arises from its ability to provide smooth approximations to arbitrarily-shaped densities of long-term spectrum that are considered to be related to the characteristics of the singer's voice rather than the specific lyrics or tune. Under the GMM classifier framework, a set of P singers is represented by voice models $\lambda_{s,1}$, $\lambda_{s,2}$, ..., $\lambda_{s,P}$ using feature vectors extracted from training data. On the other hand, all of the instrumental regions in the recordings together with a separate instrumental-only database are grouped and modeled by a GMM λ_m. The reason for use of the extra music database is the desire to increase the generality of the instrumental model so that it is capable of handling a greater variety of instrumental music. Parameters of the GMMs are initialized via k-means clustering and iteratively adjusted via expectation -maximization (EM) [7].

In the testing phase, the system takes as input the T-length feature vectors $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T\}$ extracted from an unknown test recording, and produces as outputs the frame log-likelihoods for all the singer models and the instrumental model. Prior to determining the singer of a particular recording, the regions of music in which singing actually occurs must be identified. Our method is to compare, for K successive frames, the accumulated log-likelihoods of the voice models with that of the instrumental model. The decision rule for hypothesizing whether a K–length analysis segment is vocal or instrumental is expressed as

$$\max_{1\le i\le P}\left\{\sum_{k=1}^{K}\log p(\boldsymbol{x}_{t+k}\mid\lambda_{s,i})\right\} - \sum_{k=1}^{K}\log p(\boldsymbol{x}_{t+k}\mid\lambda_m) \underset{\text{instrumental}}{\overset{\text{vocal}}{\gtrless}} \eta, \quad (1)$$

where η is the threshold. In this case, misjudging an instrumental segment is more detrimental than falsely rejecting a vocal segment, and hence η is tuned such that the miss detection rate of instrumental segments is as low as possible. After instrumental regions have been removed, log-likelihoods associated with each singer GMM are accumulated for all the remaining vocal segments. According to the maximum likelihood decision rule, the identifier should decide in favor of a singer S^* satisfying

$$S^* = \arg\max_{1 \le i \le P} \left\{ \sum_{\boldsymbol{x}_t \in \text{vocal}} \log p(\boldsymbol{x}_t \mid \lambda_{s,i}) \right\}. \quad (2)$$

For implementation efficiency, GMMs with diagonal covariance matrices are used throughout this study.

2.2. Database

Musical data used in this study consisted of 230 tracks from Mandarin pop music CDs performed by 13 female and 10 male singers, in which 10 distinct songs per singer were selected. These songs were down-sampled from CD sampling rate of 44.1 kHz to 22.05 kHz, to exclude the high frequency components beyond the range of normal singing voices. The database was then divided into two subsets, one for training, and another for testing. Each subset comprises five songs per singer. In light of the characteristic repetition within pop music, we only used the first minute of each song. This avoided redundancy in the training material and increased the speed of the identification procedure. An instrumental-only database consisting of around twenty minutes of music was added to the training data for the instrumental GMM.

2.3. Experimental Results

Our first experiments were conducted to examine the validity of vocal/instrumental segmentation. Choosing appropriate voice features and determining the number of mixture components in the GMMs are two indispensable steps toward better performance. In this experiment, we evaluated several different feature measurements, including Mel-scale Frequency Cepstral Coefficients (MFCCs) and Perceptual Linear Prediction (PLP), both with and without their first-order derivatives. Feature vectors of each type, consisting of 20 coefficients, were extracted from the musical data for every 32-ms Hamming-windowed frame with 10-ms frame shifts. Cepstral mean normalization (CMN) was also experimentally applied here in an attempt to minimize channel-induced perturbations.

To begin with, the performance of various GMM mixture counts was evaluated, with the feature type fixed as MFCCs. The segment length *K* mentioned above was empirically set to be 40 frames, and the accuracy was computed by comparing the hypothesized classification of each segment with the manual labels. † Fig.1 shows the results of vocal/instrumental discrimination. We found that the best result is achieved by using 64-mixture singer GMMs along with a 96-mixture instrumental GMM.

Next, we evaluated the performance of various feature types, with the mixture counts held constant at 64 and 96. Table 1 summarizes the experimental results. We found that MFCC performed better than PLP, and neither feature improved from the addition of delta coefficients. In addition, we found that CMN did not improve accuracy over raw MFCC. These results indicate that only MFCC is adequately robust for vocal/instrumental discrimination.

The confusion probability matrix from the discrimination results of MFCCs is shown in Table 2. 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 ID.

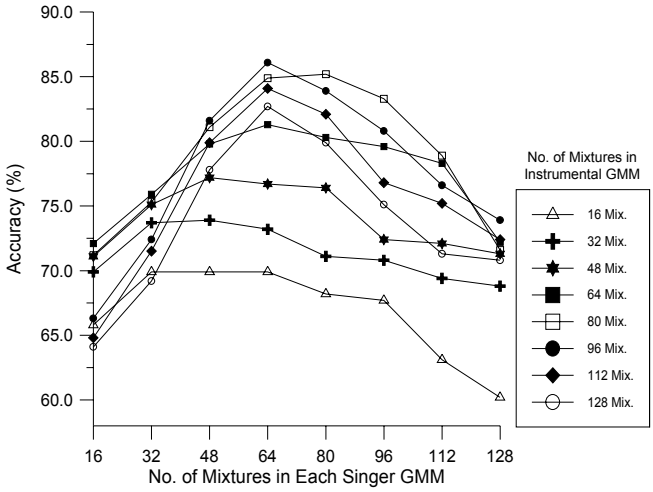

Figure 1: Accuracy of vocal/instrumental discrimination with respect to the number of mixtures in GMM.

Table 1: Results of vocal/instrumental discrimination using various feature types

Feature type	Accuracy (%)
MFCC	86.1
MFCC + Delta MFCC	85.7
PLP	81.2
PLP + Delta PLP	80.3
MFCC; CMN	82.6

Table 2: Confusion matrix of vocal/instrumental discrimination using MFCCs.

Actual	Hypothesized	
	Vocal	Instrumental
Vocal	0.84	0.16
Instrumental	0.09	0.91

† Accuracy (%) = # correctly-identified frames / # total frames × 100%

Lastly, performance of the singer ID was evaluated based on the MFCC features. For comparison, we also performed the singer-ID experiments using manual extraction of vocal regions on the test data. Table 3 lists the experimental results. The best singer-ID accuracies of 83.5% and 75.3% were yielded by manual segmentation and automatic segmentation on test data, respectively.

Table 3: Baseline singer-ID accuracy (%)

Segmentation	No. of mixtures			
	16	32	48	64
Manual	80.1	83.5	83.5	83.5
Automatic	71.9	75.3	75.3	75.3

3. Stochastic Modeling of Solo Signal from Accompanied Voices

3.1. Motivation

The baseline system introduced above suffers from a major problem in that the singer-specific models are not created solely using the singer's voice, but instead contain the singer's voice, accompanied by background music. This impure training data leads to an imperfect model, and hence deteriorates the overall performance. However, for most applications, it is impossible or impractical to acquire solo voice data for every singer involved. Fortunately, in most pop songs, substantial similarities exist between the instrumental regions and the accompaniment of the vocal regions, and therefore the stochastic characteristics of the background music can be approximated by those of the instrumental-only regions. This motivates us to estimate and model the solo signal from the voices with accompaniment signal by exploiting an *a priori* model for the background music.

3.2. Methodology

Assume that an accompanied voice $\mathbf{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_T\}$ is the mix of a solo voice $\mathbf{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_T\}$ and a background music $\mathbf{B} = \{\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_T\}$, where $\mathbf{V}$ can be obtained directly from the vocal segments of a song, and $\mathbf{B}$ is unobservable but its stochastic characteristics can be estimated via the instrumental segments. The objective is to distill the underlying solo signal within a musical recording, and thereby extract the associated singer information. Our proposed solution is built upon other previous efforts in robust speech and speaker recognition [8,9]. As a first step toward this end, the solo signal and background music are, respectively, assumed to be drawn randomly and independently according to GMMs $\lambda_s = \{w_i, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i} \mid 1 \le i \le M\}$, and $\lambda_b = \{ q_j, \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j} \mid 1 \le j \le N\}$, where w_i and q_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 $\mathbf{V} = f(\mathbf{S},\mathbf{B})$, the probability of $\mathbf{V}$, given λ_s and λ_b can be represented by

$$p(\mathbf{V} \mid \lambda_s, \lambda_b) = \prod_{t=1}^{T} \left\{ \sum_{i=1}^{M} \sum_{j=1}^{N} w_i q_j \, p(\boldsymbol{v}_t \mid i, j, \lambda_s, \lambda_b) \right\}, \tag{3}$$

where

$$p(\boldsymbol{v}_t \mid i, j, \lambda_s, \lambda_b) \tag{4}$$
$$= \iint_{\mathbf{V}=f(\mathbf{S},\mathbf{B})} \mathcal{N}(\boldsymbol{s}_t; \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}) \mathcal{N}(\boldsymbol{b}_t; \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}) d\boldsymbol{s}_t d\boldsymbol{b}_t.$$

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:

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

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

$$Q(\lambda_s, \overline{\lambda}_s) = \sum_{t=1}^{T} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) \log p(i, j, \boldsymbol{v}_t \mid \overline{\lambda}_s, \lambda_b), \tag{6}$$

where

$$p(i, j, \boldsymbol{v}_t \mid \overline{\lambda}_s, \lambda_b) = w_i q_j \, p(\boldsymbol{v}_t \mid i, j, \overline{\lambda}_s, \lambda_b), \tag{7}$$

and

$$p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) = \frac{w_i q_j \, p(\boldsymbol{v}_t \mid i, j, \lambda_s, \lambda_b)}{\sum_{m=1}^{M} \sum_{n=1}^{N} w_m q_n \, p(\boldsymbol{v}_t \mid m, n, \lambda_s, \lambda_b)}. \tag{8}$$

Letting $\nabla Q(\lambda_s, \overline{\lambda}_s) = 0$ with respect to each parameter to be reestimated, we have

$$\overline{w}_i = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{N} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b), \tag{9}$$

$$\overline{\boldsymbol{\mu}}_{s,i} = \frac{\sum_{t=1}^{T} \sum_{j=1}^{N} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) \cdot E[\boldsymbol{s}_t \mid \boldsymbol{v}_t, i, j, \lambda_s, \lambda_b]}{\sum_{t=1}^{T} \sum_{j=1}^{N} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b)}, \tag{10}$$

$$\overline{\boldsymbol{\Sigma}}_{s,i} = \frac{\sum_{t=1}^{T} \sum_{j=1}^{N} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b) \cdot E[\boldsymbol{s}_t \boldsymbol{s}_t' \mid \boldsymbol{v}_t, i, j, \lambda_s, \lambda_b]}{\sum_{t=1}^{T} \sum_{j=1}^{N} p(i, j \mid \boldsymbol{v}_t, \lambda_s, \lambda_b)} - \boldsymbol{\mu}_{s,i} \boldsymbol{\mu}_{s,i}', \tag{11}$$

where prime denotes vector transpose, and $E[\cdot]$ denotes expectation. Suppose that $\mathbf{V}$, $\mathbf{S}$ and $\mathbf{B}$ are cepstral features, and the background music is additive in the time domain or linear-spectrum domain. The accompanied signal can then be approximately expressed by $\mathbf{V} \approx \max(\mathbf{S}, \mathbf{B})$, according to Nadas' MIXMAX model [9]. Based on this assumption, we approximate

$$\sum_{i=1}^{M} \sum_{j=1}^{N} w_i q_j \, p(\boldsymbol{v}_t \mid i, j, \lambda_s, \lambda_b) \tag{12}$$
$$\approx \max\left\{ \sum_{i=1}^{M} w_i \mathcal{N}(\boldsymbol{v}_t; \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}), \sum_{j=1}^{N} q_j \mathcal{N}(\boldsymbol{v}_t; \boldsymbol{\mu}_{b,j}, \boldsymbol{\Sigma}_{b,j}) \right\}.$$

The formulas (10) and (11), required for implementation, can be, respectively, rewritten as

$$\overline{\boldsymbol{\mu}}_{s,i} = \frac{\sum_{t=1}^{T} \gamma_t(i) \cdot \boldsymbol{v}_t}{\sum_{t=1}^{T} \gamma_t(i)}, \tag{13}$$

and

$$\overline{\boldsymbol{\Sigma}}_{s,i} = \frac{\sum_{t=1}^{T} \gamma_t(i) \boldsymbol{v}_t \boldsymbol{v}_t'}{\sum_{t=1}^{T} \gamma_t(i)} - \boldsymbol{\mu}_{s,i} \boldsymbol{\mu}_{s,i}', \tag{14}$$

where

$$\gamma_t(i) = \frac{w_i \mathcal{N}(\boldsymbol{v}_t; \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i})}{\max\left\{\sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{v}_t; \boldsymbol{\mu}_{s,m}, \boldsymbol{\Sigma}_{s,m}), \sum_{n=1}^{N} q_n \mathcal{N}(\boldsymbol{v}_t; \boldsymbol{\mu}_{b,n}, \boldsymbol{\Sigma}_{b,n})\right\}}. \tag{15}$$

During a test, a background music model $\lambda_{\tilde{b}}$ is created on-line using the instrumental collection of the test recording, and the identifier then hypothesizes the most likely singer S^* using

$$S^* = \arg\max_{1 \le i \le P} p(\tilde{\mathbf{V}} \mid \lambda_{s,i}, \lambda_{\tilde{b}}). \tag{16}$$

An entire singer-ID system that implements the above algorithm is illustrated in Fig. 2.

3.3. Experimental Results

Computer simulations were conducted to test the feasibility of the proposed singer ID system. The number of mixture components used in solo GMMs and the background music GMM were empirically set to be 48 and 4, respectively. Table 4 summarizes the singer-ID results with respect to manual and automatic segmentation of vocal/instrumental regions, where the automatic segmentation was performed by the method reported in Sec. 2.3. Compared to the results shown in Table 3, the effectiveness of the method for stochastic modeling of the solo signal is clearly demonstrated.

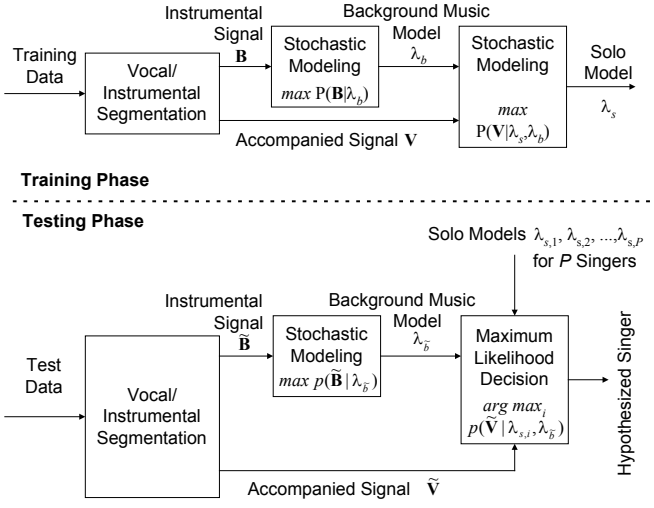

Figure 2: The proposed singer-ID system.

Table 4: Singer ID based on stochastic modeling of solo signal.

Segmentation	Accuracy (%)
Manual	87.8
Automatic	80.4

4. Conclusions and Future Work

This study demonstrated that the statistical methods applied to speaker recognition can be adapted successfully to the task of singer ID with background music. We showed that Gaussian mixture classifiers can be used to discriminate between vocal and instrumental regions of music data, and to distinguish singer's voices from one another. Furthermore, we proposed how to improve performance of singer ID by explicit statistical modeling of the solo signal from accompanied voices.

Although fairly good performance has been reported in this paper, more work is needed to validate the proposed singer-ID system. One particular problem neglected by this and most prior work in singer ID is concerned with the correlation between the background music and the singer. Since most pop artists have their own musical style, it is possible that the corruption of the vocal model by the background music is, in fact, improving performance, rather than degrading it. We are currently devising and performing experiments to determine whether this is, in fact, the case.

5. References

[1] Gauvain, J. L., and Lamel, L., "Large-vocabulary continuous speech recognition: advances and applications", *PROCEEDINGS OF THE IEEE*, Vol. 88, No. 8, 2000.

[2] Campbell, J. P., "Speaker recognition: a tutorial", *PROCEEDINGS OF THE IEEE*, Vol. 85, No. 9, 1997.

[3] Kim, Y. E., and Whitman B., "Singer identification in popular music recordings using voice coding features", *Proc. Int. Conf. Music Information Retrieval* (*ISMIR*), 2002.

[4] Liu, C. C., and Huang, C. S., "A singer identification technique for content-based classification of MP3 music objects", *Proc. Int. Conf. Information and Knowledge Management* (*CIKM*), 2002.

[5] Tzanetakis, G., and Cook, P., "Musical genre classification of audio signals", *IEEE Trans. Speech and Audio Proc.*, 10(5): 293 -302, 2002.

[6] Reynolds, D. A., and Rose, R. C., "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Trans. Speech and Audio Proc.*, 3(1): 72-83, 1995.

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

[8] Rose, R. C., Hofstetter, E. M., and Reynolds, D. A., "Intrgrated models of signal and background with application to speaker identification in noise", *IEEE Trans. Speech and Audio Proc.*, 2(2): 245-257, 1994.

[9] Nadas, A., Nahamoo, D., and Picheny, M. A., "Speech recognition using noise-adaptive prototypes", *IEEE Trans. Acoust.,Speech,and SignalProc.*, 37(10): 1495-1503, 1989.