A Modeling of Singing Voice Robust to Accompaniment Sounds and Its Application to Singer Identification and Vocal-Timbre-Similarity-Based Music Information Retrieval

Hiromasa Fujihara, Masataka Goto, Tetsuro Kitahara, *Member, IEEE*, and Hiroshi G. Okuno, *Senior Member, IEEE*

***Abstract*—This paper describes a method of modeling the characteristics of a singing voice from polyphonic musical audio signals including sounds of various musical instruments. Because singing voices play an important role in musical pieces with vocals, such representation is useful for music information retrieval systems. The main problem in modeling the characteristics of a singing voice is the negative influences caused by accompaniment sounds. To solve this problem, we developed two methods, *accompaniment sound reduction* and *reliable frame selection*. The former makes it possible to calculate feature vectors that represent a spectral envelope of a singing voice after reducing accompaniment sounds. It first extracts the harmonic components of the predominant melody from sound mixtures and then resynthesizes the melody by using a sinusoidal model driven by these components. The latter method then estimates the reliability of frame of the obtained melody (i.e., the influence of accompaniment sound) by using two Gaussian mixture models (GMMs) for vocal and nonvocal frames to select the reliable vocal portions of musical pieces. Finally, each song is represented by its GMM consisting of the reliable frames. This new representation of the singing voice is demonstrated to improve the performance of an automatic singer identification system and to achieve an MIR system based on vocal timbre similarity.**

***Index Terms*—Music information retrieval (MIR), singer identification, singing voice, vocal, vocal timbre similarity.**

I. Introduction

THE singing voice is known to be the oldest musical instrument that most people have by nature and plays an important role in many musical genres, especially in popular music. When a song is heard, for example, most people use the vocals by the lead singer as a primary cue for recognizing the song. Therefore, most music stores classify music according to the singers' names (often referred to as artists' names) in addition to musical genres.

As the singing voice is important, the representation of its characteristics is useful for music information retrieval (MIR). For example, if the name of a singer can be identified without any information of the metadata of songs, users can find songs sung by a certain singer using a description of singers' names (artists' names). Most previous MIR systems based on metadata, however, have assumed that the metadata including artists' names and song titles were available: if they were not available for some songs, these songs could not be retrieved by submitting a query of their artists' names. Furthermore, detailed descriptions of the acoustical characteristics of singing voices can also play an important role in MIR systems because they are useful for systems based on vocal timbre similarity by computing acoustical similarities between singers. Hence, a user can discover new songs rendered by the singing voices they prefer.

To identify singers' name and compute similarities between singers without requiring the metadata for each song to be prepared, in this paper, we focused on the problem of representing the characteristics of the singing voice. This problem was difficult to solve because most singing voices are accompanied by other musical instruments and the feature vectors extracted from musical audio signals are influenced by the sounds of accompanying instruments. It is therefore necessary to focus on the vocals in polyphonic sound mixtures while considering the negative influences from accompaniment sounds.

We propose two methods of solving this problem: *accompaniment sound reduction* and *reliable frame selection*. Using the former, we can reduce the influence of instrumental accompaniment. We first extracted the harmonic structure of the melody from audio signals, and then, resynthesized it using a sinusoidal model. This method reduces the influence of accompaniment sounds. The latter method is used to select reliable frames that represent the characteristics of the singing voice. We also applied these techniques and implemented an automatic singer identification system and an MIR system based on vocal timbre similarity.

II. Related Studies

The novelty of this paper compared to the previous singer identification methods lies in our two methods that solve the problem of the accompaniment sounds. Tsai *et al.* [1], [2] have

Manuscript received January 01, 2009; revised November 27, 2009. Current version published February 10, 2010. This work was supported in part by CrestMuse, in part by CREST, JST. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Bertrand David.

H. Fujihara is with the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8568, Japan, and Kyoto University, Kyoto 606-8501, Japan (e-mail: h.fujihara@aist.go.jp).

M. Goto is with the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8568, Japan (e-mail: m.goto@aist.go.jp).

T. Kitahara is with Kwansei Gakuin University, Hyogo 662-8501, Japan (e-mail: t.kitahara@ksc.kwansei.ac.jp).

H. G. Okuno is with Kyoto University, Kyoto 606-8501, Japan (e-mail: okuno@i.kyoto-u.ac.jp).

Digital Object Identifier 10.1109/TASL.2010.2041386

pointed out the problem of negative influences caused by the accompaniment sounds and have tried to solve it by using a statistically based speaker-identification method for speech signals in noisy environments [3]. On the assumption that singing voices and accompaniment sounds are statistically independent, they first estimated an accompaniment-only model from interlude sections and a vocal-plus-accompaniment model from whole songs, and then estimated a vocal-only model by subtracting the accompaniment-only model from the vocal-plus-accompaniment model. However, this assumption is not always satisfied and the way of estimating the accompaniment-only model has a problem, i.e., accompaniments during vocal sections and performances (accompaniments) during interlude sections can have different acoustical characteristics. Although Mesaros *et al.* [4] have tried to solve this problem by using a vocal separation method similar to our accompaniment sound reduction method, their method did not deal with the existence of interlude sections where singing voice does not exist and they conducted experiments using the data containing only vocal sections. In other previous studies [5]–[9], the accompaniment sound problem has not been explicitly dealt with.

From the view point of content-based MIR studies, this paper is important because our system enables a user to retrieve a song based on the specific content of the music. We considered that there can be various ways of expressing the content of the music and it is practical for users to retrieve songs using similarities based on various aspects of the music. Although some studies [10], [11] attempted to develop MIR systems based on baseline similarity and instrument existence, most previous content-based MIR systems used low-level acoustic features such as the MFCCs, the sectral centroid, and rolloff and can retrieve songs based on only vague similarities [12]–[21]. Pampalk [18] pointed out such limitation of the low-level acoustic features and it is demanded to discover new features and similarity measures that can represent more detailed content of the music.

III. Representation of Singing Voice Robust to Accompaniment Sounds

The main difficulty in modeling the characteristics of a singing voice in polyphonic music lies in the negative influences of accompaniment sounds. Since singing voice is usually accompanied by musical instruments, the acoustical features that are directly extracted from the singing voice will depend on the accompaniment sounds. When such features as cepstral coefficients or linear prediction coefficients (LPC) are extracted, which are commonly used in music-modeling and speech-modeling studies, those obtained from musical audio signals will not solely represent the singing voice but a mixture of the singing voice and the accompaniment sounds. Therefore, it is essential to cope with this accompaniment sound problem.

One possible solution to this problem is to use data influenced by accompaniment sounds for both training and identification. In fact, most of the previous studies [5]–[8] adopted this approach. However, this often fails because accompaniment sounds usually have different acoustical features from song to song. For example, the acoustics between two musical pieces that are accompanied by a piano solo and a full band will not be sufficiently similar, even if they are sung by the same singer.

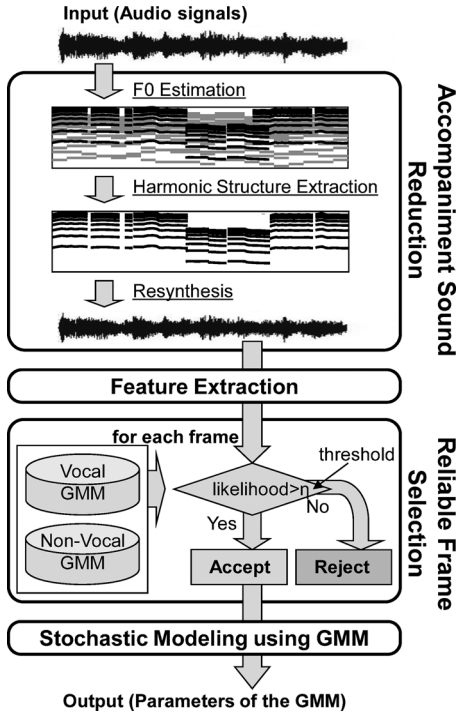

Fig. 1. Overview of our method.

We propose a method that can reduce the negative influence of accompaniment sounds directly from a given musical audio signal to solve this problem. This feature vector represents vocal characteristics better than features vector like MFCCs that only represents a mixture of accompaniment sounds and the singing voice.

This method consists of the following four parts: accompaniment sound reduction, feature extraction, reliable frame selection, and stochastic modeling. To reduce the negative influence of accompaniment sounds, the accompaniment sound reduction part first segregates and resynthesizes the singing voice from polyphonic audio signals on the basis of its harmonic structure. The feature extraction part then calculates the feature vectors from the segregated singing voice. The reliable frame selection part chooses reliable vocal regions (frames) from the feature vectors and removes unreliable regions that do not contain vocals or are greatly influenced by accompaniment sounds. The stochastic modeling part represents the selected features as parameters of the Gaussian mixture model (GMM). Fig. 1 shows an overview of this method.

A. Accompaniment Sound Reduction

For the accompaniment sound reduction part, we used a melody resynthesis technique that consisted of the following three steps:

1) estimating the fundamental frequency (F0) of the vocal melody using Goto's PreFEst [22];
2) extracting the harmonic structure corresponding to the melody;
3) resynthesizing the audio signal corresponding to the melody using sinusoidal synthesis.

1) F0 Estimation: We used Goto's PreFEst [22] to estimate the F0 of the melody line. PreFEst can estimate the most predominant F0 in frequency-range-limited sound mixtures. Since the melody line tends to have the most predominant harmonic structure in middle- and high-frequency regions, we can estimate the F0 of the melody line by applying PreFEst with adequate frequency-range limitations.

The following is a summary of PreFEst. After this, x is the log-scale frequency denoted in units of cents (a musical-interval measurement), and t is discrete time. Although a cent originally represented a tone interval (relative pitch), we use it as a unit of absolute pitch using $440\times 2^{(3/12)-5}$ Hz as a criterion, according to Goto [22]. The conversion from hertz to cent is expressed as

$$f_{\text{cent}} = 1200\log_2 \frac{f_{\text{Hz}}}{440\times 2^{\frac{3}{12}-5}} \tag{1}$$

where f_{cent} represents frequency in cents and f_{Hz} represents it in hertz.

Given the power spectrum, $\Psi_p^{(t)}(x)$, where x denotes frequency in cents and (t) denotes frame number, we first apply a bandpass filter (BPF) that was designed so that it would cover most of the dominant harmonics of typical melody lines. The filtered frequency components can be represented as $BPF(x)\Psi_p^{(t)}(x)$, where $BPF(x)$ is the BPF's frequency response to the melody line. In this paper, we designed the BPF according to Goto's specifications [22]. To make it possible to apply statistical methods, we represent each of the bandpass-filtered frequency components as a probability density function (pdf), called an observed pdf, $p_\Psi^{(t)}(x)$

$$p_\Psi^{(t)}(x) = \frac{BPF(x)\Psi_p^{(t)}(x)}{\int_{-\infty}^{\infty} BPF(x)\Psi_p^{(t)}(x)dx}. \tag{2}$$

Then, we deem each observed pdf to have been generated from a weighted-mixture model of the tone models of all the possible F0s, which is represented as

$$p(x|\theta^{(t)}) = \int_{\text{Fl}}^{\text{Fh}} w^{(t)}(F)p(x|F)dF \tag{3}$$

$$\theta^{(t)} = \left\{w^{(t)}(F)|\text{Fl}\le F\le \text{Fh}\right\} \tag{4}$$

where $p(x|F)$ is the pdf of the tone model for each F0, and Fh and Fl are defined as the lower and upper limits of the possible (allowable) F0 range, and $w^{(t)}(F)$ is the weight of a tone model that satisfies

$$\int_{\text{Fh}}^{\text{Fl}} w^{(t)}(F)dF = 1. \tag{5}$$

A tone model represents a typical harmonic structure and indicates where the harmonics of the F0 tend to occur. Then, we estimate $w^{(t)}(F)$ using an EM algorithm and regard it as the F0's pdf. Finally, we track the dominant peak trajectory of F0s from $w^{(t)}(F)$ using a multiple agent architecture.

2) Harmonic Structure Extraction: By using the estimated F0, we then extract the amplitude of the fundamental frequency component and harmonic components. For each component, we allow r cent error and extract the local maximum amplitude in the allowed area. The frequency $F_l^{(t)}$ and amplitude $A_l^{(t)}$ of the lth overtone $(l=1,\ldots,L)$ at time (t) can be represented as

$$F_l^{(t)} = \arg\max_F \left|S^{(t)}(F)\right| \times \left(l\overline{F}^{(t)}\left(1-2^{\frac{r}{1200}}\right)\le F\le l\overline{F}^{(t)}\left(1+2^{\frac{r}{1200}}\right)\right) \tag{6}$$

$$A_l^{(t)} = \left|S^{(t)}(F_l)\right| \tag{7}$$

where $S^{(t)}(F)$ denotes the complex spectrum, and $\overline{F}^{(t)}$ denotes F0 estimated by the PreFEst. In our experiments, we set r to 20.

3) Resynthesis: Finally, we use a sinusoidal model to resynthesize the audio signal of the melody by using the extracted harmonic structure, $F_l^{(t)}$ and $A_l^{(t)}$. Changes in phase are approximated using a quadratic function so that the frequency can change linearly. Changes in amplitude are also approximated using a linear function. Hereafter, k represents continuous time in units of seconds and K represents the duration between two consecutive frames in units of seconds. The resynthesized audio signals, $s(k)$, are expressed as

$$t_k = \lfloor k/K\rfloor \tag{8}$$

$$k' = k - Kt_k \tag{9}$$

$$s(k) = \sum_{l=1}^{L} s_l(k) \tag{10}$$

$$s_l(k) = \left\{\left(A_l^{(t_k+1)} - A_l^{(t_k)}\right)\frac{k'}{K} + A_l^{(t_k)}\right\} \times \sin\left(\theta_l(k')\right) \tag{11}$$

$$\theta_l(k) = \frac{\pi\left(F_l^{(t_k+1)} - F_l^{(t_k)}\right)}{K}k'^2 + 2\pi F_l^{(t_k)}k' + \theta_l(k') \tag{12}$$

$$\theta_l(0) = 0 \tag{13}$$

where $\lfloor x\rfloor$ is the largest integer not greater than x. Note that t_k represents a (discrete) frame number where the signal at time k belongs and k' represents relative time from the beginning of the frame.

4) Evaluation: To evaluate accompaniment sound reduction, we calculated a difference in the average spectral distortion (SD) between original signals and segregated signals. Given the spectrum of a vocal-only signal $S_v^{(i)}$, an original polyphonic signal $S_a^{(i)}$, and a segregated signal $S_r^{(i)}$, we define the difference of the average SD by using the following equation:

$$\frac{1}{I}\sum_{i=1}^{I}\left\{D\left(S_v^{(i)}, S_a^{(i)}\right) - D\left(S_v^{(i)}, S_r^{(i)}\right)\right\} \tag{14}$$

where $D(S_1, S_2)$ denotes the SD (in dB) of 2 spectra S_1 and S_2, I denotes the total number of frames that include a singing voice, and i denotes frame number. The difference in the average SD of 40 songs used in the experiments in Section IV-B was −4.77 dB on average. Note that the vocal-only signals are obtained from the multitrack data of these songs. This value represents the harmonic component of the accompaniment sound

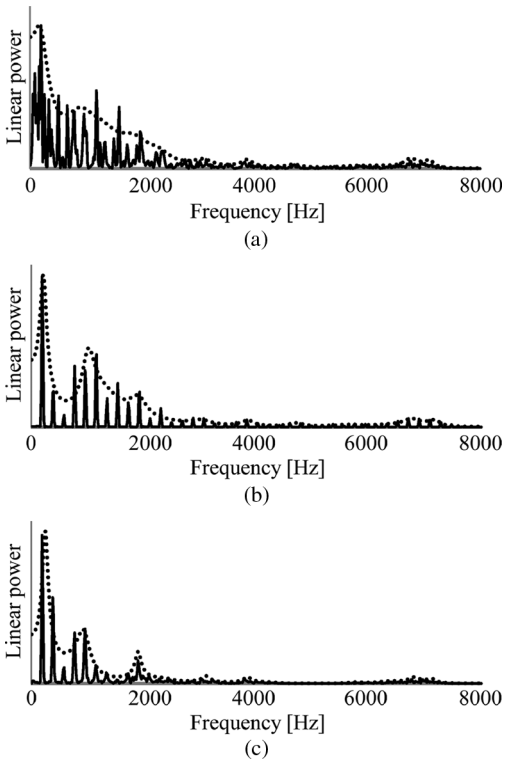

Fig. 2. Example harmonic structure extraction. (a) An original spectrum and its envelope. (b) An extracted spectrum and its envelope. (c) A spectrum of vocal-only signal and its envelope.

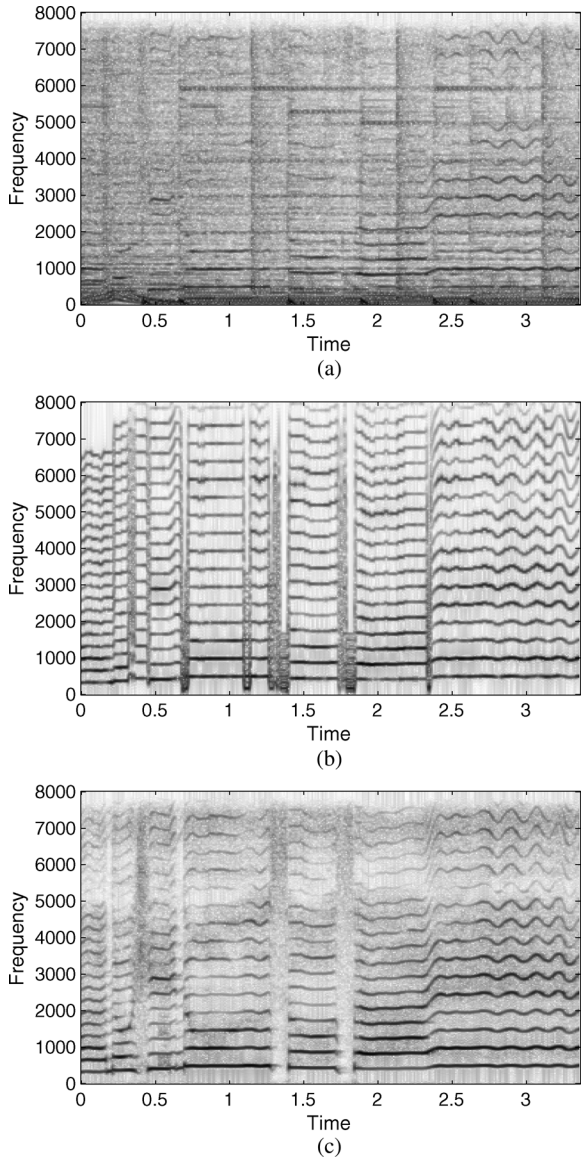

Fig. 3. Example of accompaniment sound reduction. (a) A spectrogram of polyphonic signals. (b) A spectrogram of segregated signals. (c) A spectrogram of vocal-only signals.

that is reduced by our method, and indicates that this method functions effectively.

Fig. 2 shows an example of the harmonic structure extraction. Fig. 2(a)–(c) shows an original spectrum and its envelope, an extracted spectrum and its envelope, and a spectrum of vocal-only data and its envelope, respectively. The envelopes were calculated by using the linear prediction coding (LPC). As seen in the figures, a spectral envelope of extracted spectrum precisely represents formants of singing voice, compared with that of original spectrum.

To clarify the effectiveness of accompaniment sound reduction, we show a spectrogram of polyphonic musical audio signals, that of the audio signals segregated by the accompaniment sound reduction method, and that of original (ground-truth) vocal-only signals in Fig. 3. It can be seen that harmonic components of accompaniment sound are decreased by executing the accompaniment sound reduction method. Note that some errors of F0 estimation that can be seen in the figure will be removed by after-mentioned reliable frame selection method.

B. Feature Extraction

We calculate feature vectors consisting of two features, from the resynthesized audio signals.

1) LPC-Derived Mel Cepstral Coefficients (LPMCCs): It is known that the individual characteristics of speech signals are expressed in their spectral envelopes. LPMCCs are mel-cepstral coefficients of a LPC spectrum [23], [24], which is the method to estimate the transfer function of vocal tract. Cepstral analysis on the LPC spectrum plays a role of orthogonalization and is known to be effective in pattern recognition.

2) $\Delta F0s$: We use ΔF0s which represent the dynamics of F0's trajectory, because a singing voice tends to have temporal variations in its F0 as a consequence of vibrato and such temporal information is expected to express the singer's characteristics.

C. Reliable Frame Selection

Because the F0 of the melody is simply estimated as the most predominant F0 in each frame [22], the resynthesized audio signals may contain both vocal sound in singing sections and other instrument sounds in interlude sections. The feature vectors obtained from them therefore include unreliable regions (frames) where other accompaniment sounds are predominant. The reliable frame selection part removes such unreliable regions and

TABLE I
TRAINING DATA FOR RELIABLE FRAME SELECTION

Name	Gender	Piece Number
Shingo Katsuta	M	027
Yoshinori Hatae	M	037
Masaki Kuehara	M	032, 078
Hiroshi Sekiya	M	048, 049, 051
Katsuyuki Ozawa	M	015, 041
Masashi Hashimoto	M	056, 057
Satoshi Kumasaka	M	047
Oriken	M	006
Konbu	F	013
Eri Ichikawa	F	020
Tomoko Nitta	F	026
Kaburagi Akiko	F	055
Yuzu Iijima	F	060
Reiko Sato	F	063
Tamako Matsuzaka	F	070
Donna Burke	F	081, 089, 091, 093, 097

makes it possible to use only the reliable regions for modeling the singing voice.

1) Procedure: To achieve this, we introduce two kinds of GMMs, a vocal GMM λ_{V} and a nonvocal GMM λ_{N}. The vocal GMM λ_{V} is trained on feature vectors extracted from the singing sections, and the nonvocal GMM λ_{N} is trained on those extracted from the interlude sections. Given a feature vector $\mathbf{x}$, the likelihoods for the two GMMs, $p(\mathbf{x}|\lambda_{\mathrm{V}})$ and $p(\mathbf{x}|\lambda_{\mathrm{N}})$, correspond to how the feature vector $\mathbf{x}$ is like a vocal or a (nonvocal) instrument, respectively. We therefore determine whether the feature vector $\mathbf{x}$ is reliable or not by using the following equation:

$$\log p(\mathbf{x}|\lambda_{\mathrm{V}}) - \log p(\mathbf{x}|\lambda_{\mathrm{N}}) \underset{\text{not}-\text{reliable}}{\overset{\text{reliable}}{\gtrless}} \eta \qquad (15)$$

where η is a threshold.

It is difficult to determine a universal constant threshold for a variety of songs because if the threshold is too high for some songs, there are too few reliable frames to appropriately calculate the similarities. We therefore determine the threshold that is dependent on songs so that the $\alpha\%$ of all the frames in each song are selected as reliable frames. Note that most of the nonvocal frames are rejected in this selection step.

2) Evaluation: We evaluated the reliable frame selection method by conducting experiments to confirm the following two facts: 1) the method can reject nonvocal frames and 2) the method can select frames which are less influenced by the accompaniment sound. We trained GMM for vocal and nonvocal using songs listed in Table I and used the 40 songs listed in Table II for evaluation. All of these data are the same as those used in the experiments in Section IV-B. First, to confirm 1), we evaluated a precision rate and a recall rate of the method and Fig. 4 shows the results. When α is 0.15, the precision rate is approximately 79% and, thus, we can confirm that many nonvocal sections are rejected. Then, to confirm 2), Fig. 5 shows a dependency on α of spectral distortion of frames that are selected by reliable frame selection. The average SD is positively correlated with α. Therefore, we could confirm that the method can select frames that are less influenced by the accompaniment sound, by setting α to a smaller value.

TABLE II
SONGS USED FOR EVALUATION. NUMBERS IN TABLE ARE PIECE NUMBERS IN RWC-MDB-P-2001

	Name	Gender	D_1	D_2	D_3	D_4
a	Kazuo Nishi	M	012	029	036	043
b	Hisayoshi Kazato	M	004	011	019	024
c	Kousuke Morimoto	M	038	039	042	044
d	Shinya Iguchi	M	082	084	088	090
e	Jeff Manning	M	085	087	095	098
f	Hiromi Yoshii	F	002	017	069	075
g	Tomomi Ogata	F	007	028	052	080
h	Rin	F	014	021	050	053
i	Makiko Hattori	F	065	067	068	077
j	Betty	F	086	092	094	096

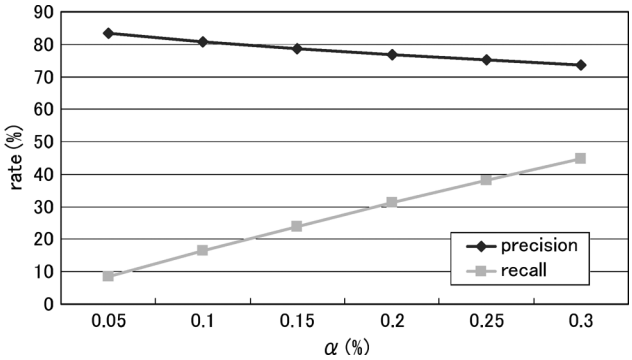

Fig. 4. Precision rate and recall rate of reliable frame selection.

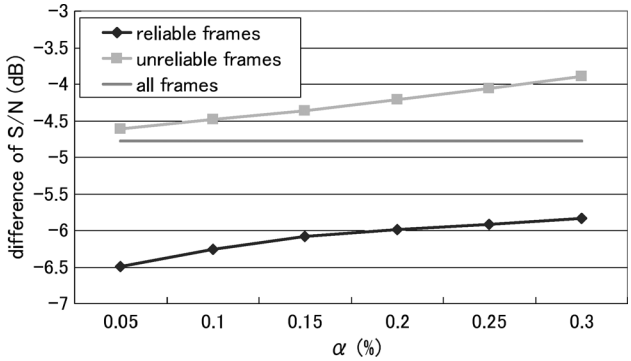

Fig. 5. Dependency of spectral distortion of selected frames on α.

D. Stochastic Modeling

Finally, we model a probability distribution of the feature vectors for a song using GMM and estimate the parameters of the GMM with the EM algorithm. In our experiments, we set the number of Gaussians to 64.

IV. SINGER IDENTIFICATION

This section describes one of the applications of our vocal modeling techniques, i.e., the system for identifying the singer by determining a singer's name from given musical audio signals. The target data are real-world musical audio signals such as popular music CD recordings that contain the singing voices of single singers and accompaniment sounds.

A. Determination of Singer

First, we prepare the audio signals of target singers as training data and calculate the GMMs for all singers by using the method described in Section III. Given input audio signals, we also calculate the GMM for the song. Then, the name of the singer is determined through the following equation:

$$s = \arg\max_i \frac{1}{T} \sum_{t=1}^{T} \log p(\mathbf{x}_t|\lambda_i). \tag{16}$$

B. Experiments Using RWC Music Database

We conducted experiments to evaluate our singer identification system.

1) Condition and Results: We conducted experiments on singer identification using the "RWC Music Database: Popular Music (RWC-MDB-P-2001)" [25] under the following four conditions to find out how effective our methods of accompaniment sound reduction and reliable frame selection were as follows:

1) without either reduction or selection (baseline);
2) with reduction, without selection;
3) without reduction, but with selection;
4) with both reduction and selection (ours).

We used 40 songs by ten different singers (five were males and five were females), listed in Table II, taken from the RWC-MDB-P-2001. Using these data, we conducted the four-fold cross validation, that is, we first divided all the data into four groups, D_i ($i = 1, 2, 3, 4$) in Table II, and then repeated the following step four times; each time, we left out one of the four groups for training and used the one we had omitted for testing. We used 25 songs of 16 different singers listed in Table I, also taken from the RWC-MDB-P-2001, which differ from the singers used for evaluation, as the training data for the reliable frame selection. We set α to 15%, using the experiment described in Section IV-B2 as a reference. To evaluate the performance of the LPMCCs, we use both the LPMCCs and the MFCCs as the feature vectors and compare the result. Acuracy was defined by a ratio of the number of correctly identified song to the number of songs used for evaluation.

Fig. 6 shows the results of the experiments. As seen in the table, accompaniment sound reduction and reliable frame selection improved the accuracy of singer identification. When these two methods were used together, in particular, the accuracy was significantly improved from 55% to 95%.

Fig. 7 shows the confusion matrices of the experiments when the LPMCCs are used. As can be seen, confusion between males and females decreased by using the reduction method. This means that, under conditions 2) and 4), the reduction method decreased the influence of accompaniment sound, and the system could correctly identify the genders. However, without the reduction method [conditions 1) and 3)], the influences of accompaniment sound prevented the system from correctly identifying even the genders of the singers.

When we compare the MFCCs and the LPMCCs, we can find that the accuracies of the LPMCCs exceed those of the MFCCs in all the conditions. This is particularly remarkable when we use both the accompaniment sound reduction method and the reliable frame selection method. We can confirm that the LPMCCs represent the characteristics of the singing voice well.

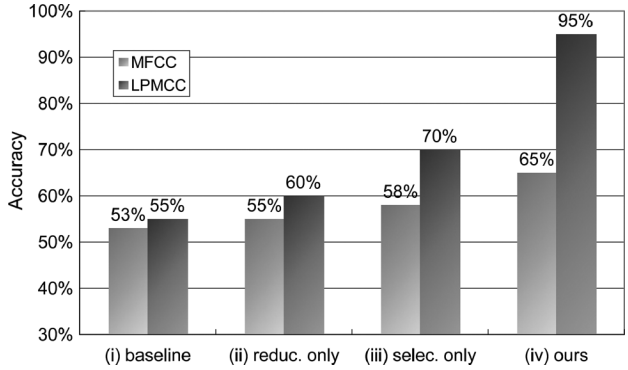

Fig. 6. Results of the experiments using RWC Music Database, where "reduc." and "selec." correspond to accompaniment sound reduction and reliable frame selection, respectively.

2) Dependence of Accuracy on α: We conducted experiments by setting α to various values to investigate how dependent the accuracies were on α, which represents the percentage of frames determined to be reliable by using the reliable frame selection method. These experiments used the same dataset as that used in the previous experiments. The experimental results in Fig. 8 indicate that the accuracy of classification was not affected by small changes in α. We can also see that the value of α that yielded the highest accuracy differed. The reason for this is as the follow: accompaniment sound reduction method reduced the influence of accompaniment sounds and emphasized the differences between reliable and unreliable frames. Thus, if we increased α excessively, the system selected many unreliable frames and the performance of the system decreased.

3) Combination of Accompaniment Sound Reduction and Reliable Frame Selection: To confirm an effectiveness of reliable frame selection in combination with the accompaniment sound reduction method, we conducted experiments under the following three conditions.

1) Only hand-labeled vocal sections are used. We execute accompaniment sound reduction using ground-truth F0s. We do not execute reliable frame selection.
2) Only hand-labeled vocal sections are used. We execute accompaniment sound reduction using F0s estimated by Prefest. We do not execute reliable frame selection.
3) An entire region of a song are used. We execute accompaniment sound reduction using F0s estimated by PreFEst and reliable frame selection.

Table III shows the results of the experiments. When we compare condition 3) with condition 2), the accuracy was improved by 12 points (from 83% to 95%). This fact indicates that the reliable frame selection method can achieve higher accuracy than the manual removal of nonvocal sections. When we compare condition 3) with condition 1), the accuracy was improved by 7 points (88% to 95%). This fact indicates that, even if there were no F0 estimation errors, it was difficult to achieve high accuracy without reliable frame selection. In contrast, this fact also indicates that some F0 estimation errors did not degrade system performance if we use a reliable frame selection method because

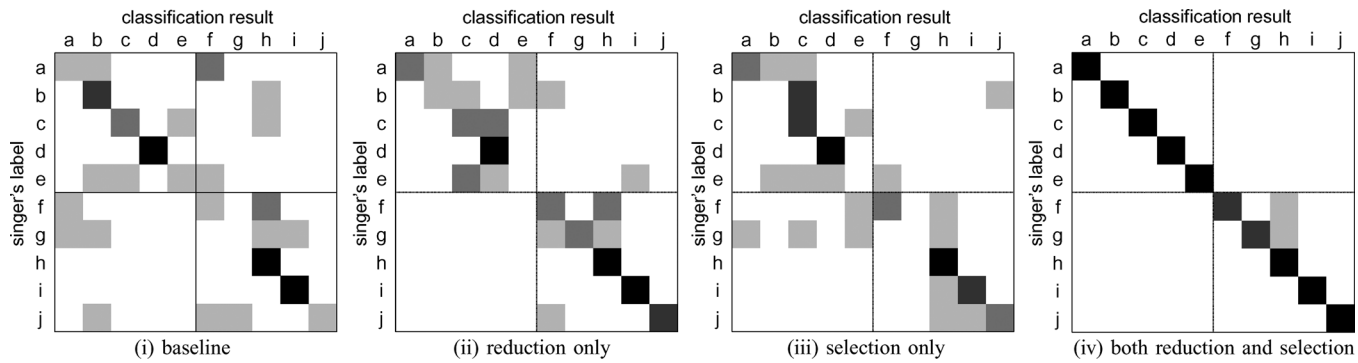

Fig. 7. Confusion matrices. Center lines in each figure are boundaries between males and females. Note that confusion between males and females decreased by using the accompaniment sound reduction method.

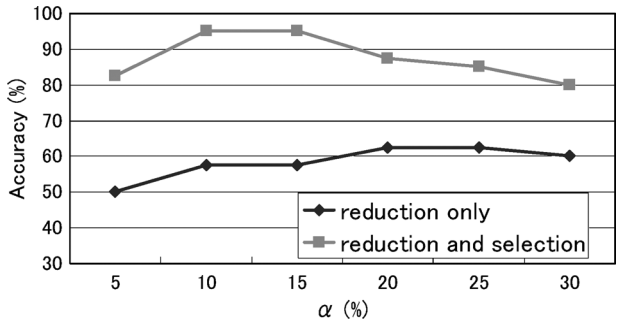

Fig. 8. Experimental results for dependence of accuracy on α. $\alpha\%$ of all frames was determined to be reliable.

TABLE III
EVALUATION OF A COMBINATION OF ACCOMPANIMENT SOUND REDUCTION AND RELIABLE FRAME SELECTION, WHERE "SELEC." MEANS RELIABLE FRAME SELECTION

Condition	(i)	(ii)	(iii)
selec.	Not used		Used
Data	Vocal section		Whole sections
F0	Ground truth	PreFEst	
Acuracy	88%	83%	95%

the method rejected the region in which PreFEst failed to estimate correct F0s.

C. Experiments Using Commercial CD Recordings

We also conducted experiments using commercial CD recordings available in Japan. The experiments were done under the four conditions described in Section IV-B. 246 songs by 20 singers (8 males and 12 females) listed in Table IV were used in these experiments. These artists were selected from the Japanese best-seller list of CD in 2004. The same as in the previous experiments, the 25 songs by 16 singers listed in Table I were used as the training data for reliable frame selection. Using these data, we conducted the fourfold cross validation.

The bar chart in Fig. 9 shows the results of these experiments. We confirmed that the accuracy improved by approximately 12% by using both methods, while accuracy improved by approximately 8% by using each of the two methods.

Fig. 10 shows the confusion matrices. As can be seen, the system more often misidentified songs by female singers than those by males. We consider this is because the pitch of female singing is generally higher than that of male singing. We found spectral envelopes estimated from high-pitched sounds by using cepstrum or LPC analysis are strongly affected by spectral valleys between adjacent harmonic components.

TABLE IV
ARTISTS SELECTED FROM COMMERCIAL CD RECORDINGS

	Artist Name	Gender	Tracks
A	Asian Kung-fu generation	M	11
B	Bump of Chicken	M	10
C	Ken Hirai	M	10
D	Noriyuki Makihara	M	12
E	Naotaro Moriyama	M	11
F	Mr.Children	M	12
G	PornoGraffitti	M	13
H	QUEEN	M	16
I	Aiko	F	13
J	Avril Lavigne	F	14
K	BoA	F	12
L	Ayumi Hamasaki	F	8
M	Ayaka Hirahara	F	10
N	Mai Kuraki	F	16
O	Mika Nakashima	F	13
P	Ai Otsuka	F	11
Q	Hiromi Shimatani	F	15
R	Kou Shibasaki	F	12
S	Hikaru Utada	F	15
T	Hitomi Yaida	F	12

TABLE V
QUERY SONGS AND RETRIEVED CORRESPONDING SONGS USED FOR SUBJECTIVE EXPERIMENT: THREE-DIGIT NUMBER INDICATES THE PIECE NUMBER OF THE RWC-MDB-P-2001. GIVEN EACH QUERY SONG, TOP-RANKED SONG BY BASELINE METHOD (MFCC) AND TOP-RANKED SONG BY OUR METHOD ARE SHOWN ON SAME LINE

Query song			Retrieved (top ranked) song	
Piece #	Gender	Language	MFCC	Our method
004	M	Japanese	031	082
010	F	Japanese	016	054
029	M	Japanese	017	012
035	F	Japanese	036	094
045	M	Japanese	090	042
053	F	Japanese	062	014
072	M	Japanese	071	076
077	F	Japanese	071	067
092	F	English	024	086
098	M	English	009	085

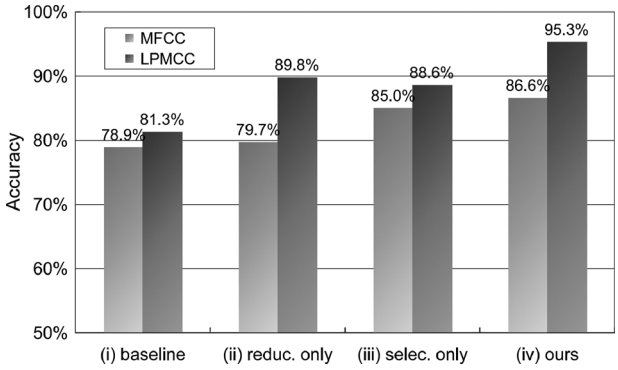

Fig. 9. Results of the experiments using commercial CD-recordings, where "reduc." and "selec." mean accompaniment sound reduction and reliable frame selection, respectively.

The accuracy of the baseline method (Condition 1) in the experiments using the commercial CDs was higher than that of the RWC Music Database by approximately 15%. This is because songs on the same album tend to use the same instruments and be homogenous in sound quality. Berenzweig *et al.* [6] called this phenomenon the "Album effect" and they pointed out that the performance of a singer identification system depends on what kind of dataset is used. On the other hand, since the RWC Music Database consists of a variety of genres and instruments even for songs by the same singer, the accuracy of the baseline method was only 55%. However, since the proposed method (Condition 4) was extremely accurate for the experiments using the RWC Music Database, we found that our method can identify the singers' names correctly even if there were a variety of songs in the database.

V. MIR Based on Vocal Timbre Similarity

We also applied our technique to a new MIR system based on vocal timbre similarity and developed a system named *VocalFinder*. In this paper, the term vocal timbre means a shape of a spectral envelope of the singing voice. By using this system, we could find a song by using its musical content in addition to traditional bibliographic information. This kind of retrieval is called content-based MIR, and our system, which focuses on singing voices as content, falls into this category.

A. Similarity Calculation

We chose symmetric Kullback–Leibler divergence [26] to be the similarity measure between two songs. Since it is difficult to calculate this similarity measure in a closed form, we approximate it as follows (this approximation is called "cross-likelihood ratio test"[26]); the similarity between songs X and Y $d_{\mathrm{CE}}(X,Y)$ is calculated by

$$d_{\mathrm{CE}}(X,Y) = \log \prod_i \frac{\mathcal{N}_{\mathrm{GMM}}(x_i;\theta_X)}{\mathcal{N}_{\mathrm{GMM}}(x_i;\theta_Y)} + \log \prod_j \frac{\mathcal{N}_{\mathrm{GMM}}(y_j;\theta_Y)}{\mathcal{N}_{\mathrm{GMM}}(y_j;\theta_X)} \quad (17)$$

where $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ correspond to the feature vectors of reliable frames, which could be MFCCs or LPMCCs, in songs X and Y, respectively, θ_X and θ_Y correspond to the GMM parameters of songs X and Y, respectively, and $\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x};\theta)$ represents the likelihood of GMM with parameter θ.

B. System Operation

Fig. 11 shows a screenshot of the system. As the training data for the vocal and nonvocal GMMs, we used the same 25 songs listed and used in the experiences in Section IV-B. We registered the other 75 songs from the RWC-MDB-P-2001 in the system database, which were not used to construct these GMMs. In the figure, the song "PROLOGUE" (RWC-MDB-P-2001 No.7) sung by the female singer "Tomomi Ogata" is given as a query. Given a query song, it took about 20 seconds to calculate similarities and output a ranked list of retrieved songs. As seen in the Fig. 11, the retrieval results list the ranking, the song titles, the artists' names, and similarities.

In most of songs retrieved given various queries, the vocal timbres of the top ten songs were generally similar to that of each query song in our experience. For example, in Fig. 11, the top 21 songs were sung by female singers, and the vocal timbres of the top 15 songs in this figure were similar to the query song. Note that four songs by "Tomomi Ogata" who was the singer in the query took first, second, ninth, and twelfth places. This is because the singing styles of the ninth and twelfth songs were different from those of the first and second songs and the query.

C. Subjective Experiment

We conducted a subjective experiment to compare our system using the proposed vocal-based feature vector with a baseline system using the traditional MFCCs of the input sound mixtures.

Six university students (two males and four females) participated in this experiment. They had not received any professional training in music. They first listened to a set of three songs—a query song (song X), the top-ranked song retrieved by our system (songs A/B), and the top-ranked song retrieved by the baseline system (songs B/A)—, and then judged which song was more similar to the query song (Fig. 12). They did not know which song was retrieved by our system and the song order of A and B was randomized. We allowed them to listen to these songs in any order for as long as they liked.

We selected ten query songs from the system database taking into consideration that these songs were sung by different genders and in different genres. For each query song, we asked the subjects the following questions.

- Question 1: When comparing the singing voice timbres of songs A and B, which song resembles song X?
- Question 2: When comparing the overall timbres of songs A and B, which song resembles song X?

Figs. 13 and 14 show the results of the experiment. On average, 80% of the responses for ten songs judged that the timbre of the singing voice obtained by our method was more similar to that of the query song (Fig. 13). A binominal test, in which significance was set at 0.05, was performed on these results and the degree of significance was 0.0000086, which indicates that there were significant differences between our method and the conventional method. On the other hand, 70% of the responses judged that the overall timbre obtained by the baseline method was more similar to that of the query song (Fig. 14).

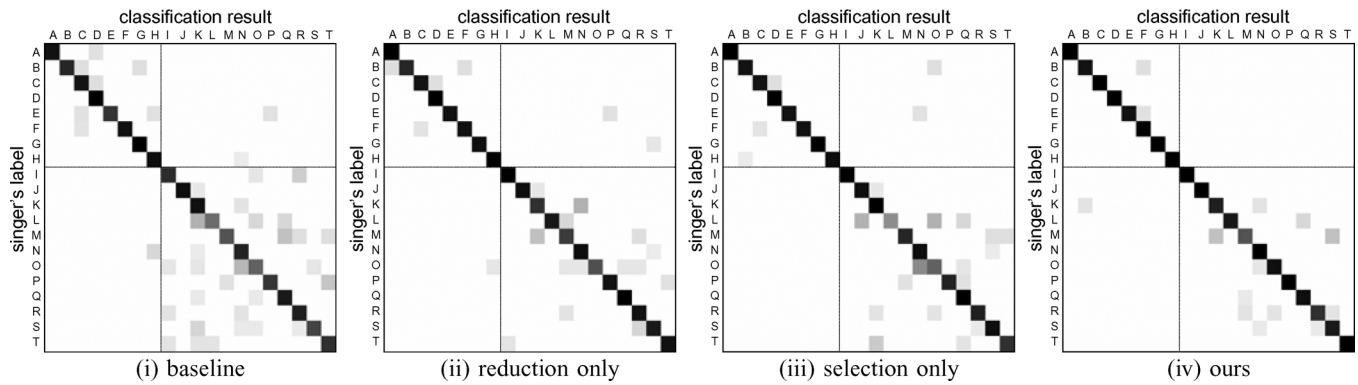

Fig. 10. Confusion matrices of experiments using commercial CD-recordings. Center lines in each figure are boundaries between males and females.

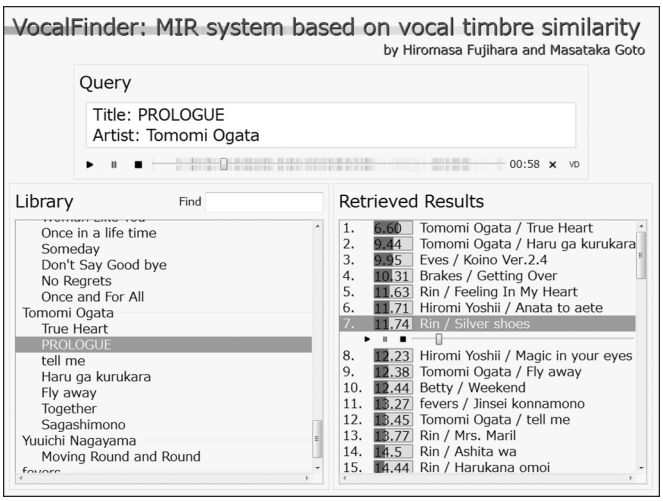

Fig. 11. Screenshot of the system.

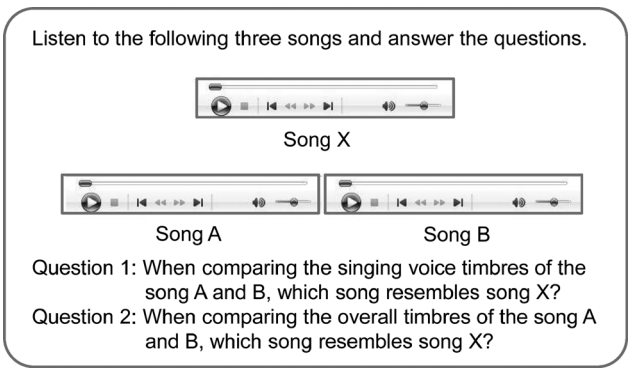

Fig. 12. Interface used for subjective experiment.

We also preformed a binominal test and the degree of significance was 0.033. Therefore, we confirmed that our method can reduce the influence of accompaniment sounds and find songs by using vocal timbres. We also found that our method finds not only songs with similar vocal timbres (or by same singer) but also songs with similar singing styles. For example, when song RWC-MDB-P-2001 No.53 was used as a query, both our method and the baseline method retrieved the top-ranked songs by the singer to be the same as that in the query, but 5 out of 6

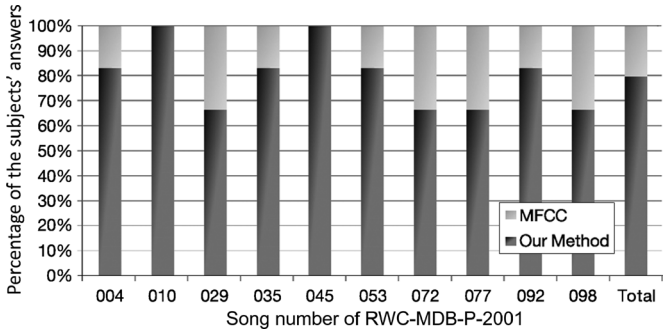

Fig. 13. Evaluation results: Question 1: singing voice timbre.

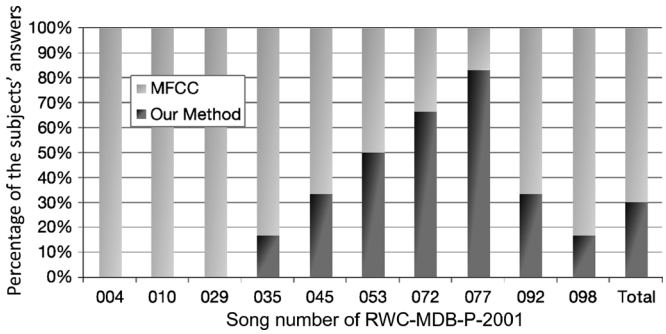

Fig. 14. Evaluation results: Question 2: overall timbre.

subjects judged that the song obtained by our method was more similar to the query song in terms of the vocal timbre similarity.

VI. Discussion

This section discusses the novelty and effectiveness of the method proposed in this paper.

A. Novelty and Effectiveness of Accompaniment Sound Reduction

We clarified the problem caused by accompaniments when modeling the singing voice, which has not effectively been dealt with except for a few attempts. We provided two effective solutions, i.e., accompaniment sound reduction and reliable frame selection. The accompaniment sound reduction method

is characterized by the way it dealt with accompaniment sound: it segregated the singing voice directly from the spectrum of the singing voice without modeling the accompaniment sound. Although the conventional method dealt with this problem by modeling the accompaniment sound, it was generally difficult to model the accompaniment sound.

In this paper, we conducted two disparate experiments to confirm the effectiveness of this method. First, we evaluated the difference in the average spectral distortion and found that the method reduced the spectral distortion by 4.77 dB. Second, the results of the experiments on singer identification showed that the method improved identification accuracy from 70% to 95% for the data taken from the RWC Music Database and from 88.6% to 95.3% for the data taken from commercial CD recordings.

B. Novelty and Effectiveness of Accompaniment Sound Reduction

The reliable frame selection method made it possible to consistently select reliable frames that represented the characteristics of the singing voice. It needs to be noted that this method even rejected unreliable vocal frames as well as nonvocal frames to improve the robustness. Although similar methods were used in previous studies, they focused on distinguishing vocal and nonvocal frames; they did not consider the reliability of each frame.

The effectiveness of the reliable frame selection method is confirmed by the following two comparisons. First, we compared the spectral distortion between frames selected and those rejected by the method. The spectral distortion of the former was smaller than the letter, and therefore, we could say that the method can select frames that are less influenced by the accompaniment sounds. Then, we compared the results of experiments on signer identification. The accuracy of the experiment in which the hand-labeled vocal sections are used without the selection method was 83%, while that with the selection method was 95%. This result indicates that it is important not only to detect the vocal regions but also to select reliable frames.

C. Effectiveness of a Combination of the Two Methods

It needs to be noted that the reliable frame selection method is robust to the error of the accompaniment sound reduction method because the reliable frame selection method can reject frames in which the singing voice was not properly segregated by the accompaniment sound reduction method. This is confirmed by the experiments using the ground-truth F0 for the accompaniment sound reduction method. Though methods similar to the accompaniment sound reduction have been used to improve the noise robustness in the field of speech recognition [27], this is the first paper that proposed a method that can be used in combination with the accompaniment sound reduction method and increase robustness to F0 estimation errors.

VII. Conclusion

We described two methods that work in combination to model the characteristics of the singing voice. To deal with the singing voice including sound mixtures of various musical instruments, our method solved the problem of the accompaniment sound influences. We developed an automatic singer identification system and an MIR system based on vocal timbre similarity by applying the new representation of the singing voice, and tested and confirmed the effectiveness of these systems by conducting objective and subjective experiments.

In the future, we plan to extend our method to represent singing styles of singers in addition to the vocal timbre by modeling F0's trajectories of the singing voices. We also plan to integrate this system with content-based MIR methods based on other musical elements to give users a wider variety of retrieval methods.

References

[1] W.-H. Tsai and H.-M. Wang, "Automatic detection and tracking of target singer in multi-singer music recordings," in *Proc. 2004 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2004)*, 2004, pp. 221–224.

[2] W.-H. Tsai and H.-M. Wang, "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 330–341, Jan. 2007.

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

[4] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *Proc. 8th Int. Conf. Music Inf. Retrieval (ISMIR 2007)*, 2007, pp. 375–378.

[5] B. Whitman, G. Flake, and S. Lawrence, "Artist detection in music with minnowmatch," in *Proc. 2001 IEEE Workshop Neural Netw. Signal Process.*, 2001, pp. 559–568.

[6] A. L. Berenzweig, D. P. W. Ellis, and S. Lawrence, "Using voice segments to improve artist classification of music," in *Proc. AES 22nd Int. Conf. Virtual, Synth., Entertainment Audio*, 2002.

[7] Y. E. Kim and B. Whitman, "Singer identificatin in popular music recordings using voice coding features," in *Proc. 3rd Int. Conf. Music Inf. Retrieval (ISMIR2002)*, 2002, pp. 164–169.

[8] T. Zhang, "Automatic singer identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME 2003)*, 2003, vol. I, pp. 33–36.

[9] W.-H. Tsai, S.-J. Liao, and C. Lai, "Automatic identification of simultaneous singers in duet recordings," in *Proc. 9th Int. Conf. Music Inf. Retrieval (ISMIR 2008)*, 2008, pp. 115–120.

[10] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrogram: Probabilistic representation of instrument existence for polyphonic music," *IPSJ J.*, vol. 48, no. 1, pp. 214–226, 2007.

[11] Y. Tsuchihashi, T. Kitahara, and H. Katayose, "Using bass-line features for content-based mir," in *Proc. 9th Int. Conf. Music Inf. Retrieval (ISMIR 2008)*, 2008, pp. 620–625.

[12] J.-J. Aucouturier and F. Pachet, "Music similarity measures: What's the use?," in *Proc. 3rd Int. Conf. Music Inf. Retrieval (ISMIR2002)*, 2002, pp. 157–163.

[13] B. Logan, "Content-based playlist generation: Exploratory experiments," in *Proc. 3rd Int. Conf. Music Inf. Retrieval (ISMIR 2002)*, 2002, pp. 295–296.

[14] E. Allamanche, J. Herre, O. Hellmuth, T. Kastner, and C. Ertel, "A multiple feature model for musical similarity retrieval," in *Proc. 4th Int. Conf. Music Inf. Retrieval (ISMIR 2003)*, 2003, pp. 217–218.

[15] A. Berenzweig, B. Logan, D. P. W. Ellis, and B. Whitman, "A large-scale evaluation of acoustic and subjective music similarity measures," in *Proc. 4th Int. Conf. Music Inf. Retrieval (ISMIR 2003)*, 2003, pp. 63–70.

[16] M. F. McKinney and J. Breebaart, "Features for audio and music classification," in *Proc. 4th Int. Conf. Music Inf. Retrieval (ISMIR 2003)*, 2003, pp. 151–158.

[17] G. Tzanetakis, J. Gao, and P. Steenkiste, "A scalable peer-to-peer system for music content and information retrieval," in *Proc. 4th Int. Conf. Music Inf. Retrieval (ISMIR 2003)*, 2003, pp. 209–214.

[18] E. Pampalk, "Computational models of music similarity and their application in music information retrieval," Ph.D. dissertation, Universitat Wien, Vienna, Austria, 2006.

[19] A. Flexer, F. Gouyou, S. Dixon, and G. Widmer, "Probabilistic combination of features for music classification," in *Proc. 7th Int. Conf. Music Inf. Retrieval (ISMIR 2006)*, 2006, pp. 628–633.
[20] T. Pohle, P. Knees, M. Schedl, and G. Widmer, "Independent component analysis for music similarity computation," in *Proc. 7th Int. Conf. Music Inf. Retrieval (ISMIR 2006)*, 2006, pp. 228–233.
[21] D. P. W. Ellis, "Clasifying music audio with timbral and chroma features," in *Proc. 8th Int. Conf. Music Inf. Retrieval (ISMIR 2007)*, 2007, pp. 339–340.
[22] M. Goto, "A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Commun.*, vol. 43, no. 4, pp. 311–329, 2004.
[23] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.
[24] K. Shikano, Evaluation of LPC spectral matching measures for phonetic unit recognition Comput. Sci. Dept. Carnegie Mellon Univ., Tech. Rep. CMU-CS-96-108, 1986.
[25] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. 3rd Int. Conf. Music Inf. Retrieval (ISMIR 2002)*, Oct. 2002, pp. 287–288.
[26] T. Virtanen and M. Helen, "Probabilistic model based similarity measures for audio query-by-example," in *Proc. 2007 IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA 2007)*, 2007, pp. 82–85.
[27] T. Nakatani and H. G. Okuno, "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Commun.*, vol. 27, pp. 209–222, 1999.

Hiromasa Fujihara received the B.S. and M.S. degrees from Kyoto University, Kyoto, Japan, in 2005 and 2007, respectively. He is currently pursuing the Ph.D. degree in the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University.

He is currently a Research Scientist of the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. His research interests include singing information processing and music information retrieval.

Mr. Fujihara was awarded the Yamashita Memorial Research Award from the Information Processing Society of Japan (IPSJ).

Masataka Goto received the Doctor of Engineering degree from Waseda University, Tokyo, Japan, in 1998.

He is currently a Leader of the Media Interaction Group, Information Technology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. He serves concurrently as a Visiting Professor in the Department of Statistical Modeling, The Institute of Statistical Mathematics, and an Associate Professor (Cooperative Graduate School Program) in the Department of Intelligent Interaction Technologies, Graduate School of Systems and Information Engineering, University of Tsukuba.

Dr. Goto received 24 awards over the past 17 years, including the Commendation for Science and Technology by the Minister of MEXT "Young Scientists' Prize," DoCoMo Mobile Science Awards "Excellence Award in Fundamental Science," IPSJ Nagao Special Researcher Award, and the IPSJ Best Paper Award.

Tetsuro Kitahara (M'07) received the B.S. degree from Tokyo University of Science, Tokyo, Japan, in 2002 and the M.S. and Ph.D. degrees from Kyoto University, Kyoto, Japan, in 2004 and 2007, respectively.

He is currently a Postdoctoral Researcher at Kwansei Gakuin University, Hyogo, Japan, for the CrestMuse Project funded by CREST, JST, Japan. His research interests include music informatics and computational auditory scene analysis.

Dr. Kitahara received several awards including the Second Kyoto University President Award.

Hiroshi G. Okuno (SM'06) received B.A. and Ph.D. from the University of Tokyo in 1972 and 1996, respectively.

He worked for NTT, JST, and Tokyo University of Science. He is currently a Professor of Graduate School of Informatics, Kyoto University, Kyoto, Japan. He was a Visiting Scholar at Stanford University, Stanford, CA, from 1986 to 1988. He has done research in programming languages, parallel processing, and reasoning mechanism in AI. He is currently engaged in computational auditory scene analysis, music scene analysis, and robot audition. He coedited *Computational Auditory Scene Analysis* (Lawrence Erlbaum Associates, 1998), *Advanced Lisp Technology* (Taylor and Francis, 2002), and *New Trends in Applied Artificial Intelligence (IEA/AIE)* (Springer, 2007).

Dr. Okuno received various awards including the 1990 Best Paper Award of the JSAI, the Best Paper Award of IEA/AIE-2001 and 2005, and the IEEE/RSJ IROS-2001 and 2006 Best Paper Nomination Finalist. He is a member of the AAAI, ACM, ASJ, ISCA, and five Japanese societies.