# SINGER IDENTIFICATION BASED ON ACCOMPANIMENT SOUND REDUCTION AND RELIABLE FRAME SELECTION

**Hiromasa Fujihara,[†] Tetsuro Kitahara,[†] Masataka Goto,[‡]**
**Kazunori Komatani,[†] Tetsuya Ogata,[†] and Hiroshi G. Okuno[†]**

[†]Dept. of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
{fujihara, kitahara, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp

[‡]National Institute of Advanced Industrial
Science and Technology (AIST)
Tsukuba, Ibaraki 305-8568, Japan
m.goto@aist.go.jp

## ABSTRACT

This paper describes a method for automatic singer identification from polyphonic musical audio signals including sounds of various instruments. Because singing voices play an important role in musical pieces with a vocal part, the identification of singer names is useful for music information retrieval systems. The main problem in automatically identifying singers 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 method makes it possible to identify the singer of a singing voice after reducing accompaniment sounds. It first extracts harmonic components of the predominant melody from sound mixtures and then resynthesizes the melody by using a sinusoidal model driven by those components. The latter method then judges whether each frame of the obtained melody is reliable (i.e. little influenced by accompaniment sound) or not by using two Gaussian mixture models for vocal and non-vocal frames. It enables the singer identification using only reliable vocal portions of musical pieces. Experimental results with forty popular-music songs by ten singers showed that our method was able to reduce the influences of accompaniment sounds and achieved an accuracy of 95%, while the accuracy for a conventional method was 53%.

**Keywords:** Singer identification, artist identification, melody extraction, singing detection, similarity-based MIR

## 1 INTRODUCTION

Singing voice is known as the oldest musical instrument that everyone has 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 vocal part by the lead singer's voice as a primary cue for recognizing (identifying) the song (name). Therefore, most music stores classify music according to singers' names (often referred to as artist names) in addition to musical genres.

As known from the above importance of the singing voice, the description of singer names of songs is useful for music information retrieval (MIR). When a user wants to find songs sung by a certain singer, an MIR system can use the description of singer names (artist names). Furthermore, detailed descriptions of acoustical characteristics of singing voices can also play an important role in MIR because they are useful for computing acoustical similarities between singers. Most previous MIR systems, however, assumed that the metadata including artist names and song titles were available: if they were not available for some songs, those songs cannot be retrieved by submitting a query of their artist names.

To achieve such singing-voice-based MIR and compute artist similarities without requiring the metadata for every song to be prepared, in this paper, we focus on the problem of identifying singers for songs, *automatic singer identification* problem. This problem is difficult because most singing voices are accompanied by other musical instruments. It is therefore necessary to focus on the vocal part in polyphonic sound mixtures while considering the negative influences from accompaniment sounds. In other words, feature vectors extracted from musical audio signals are influenced by the sounds of accompanying instruments. Although speaker identification problem for (non-music) speech signals has been dealt with by many studies in the field of speech information processing, their results cannot be directly applied to the singer identification problem for singing voices with accompaniments because most existing speaker-identification techniques assume speech signals presented without other simultaneous sounds. On the other hand, Tsai *et al.* [1, 2] have pointed out this problem 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 the 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: accompaniments during vocal sections and performances (accompaniments) during interlude sections can have different acoustical characteristics. In other previous studies [4, 5, 6, 7], the accompaniment sound problem has not explicitly been dealt with.

To solve this problem, we propose two methods: *accompaniment sound reduction* and *reliable frame selection*. Using the former method, we reduce the influence of accompaniment. We first extracted the harmonic structure of the melody from audio signals, and then, resynthesize it using a sinusoidal model. This method reduces the influence of accompaniment sounds. The latter method selects frames that are reliable enough for classification.

The rest of this paper is organized as follows. In the next section, we describe our method for a singer identification task. In Section 3, we describe the implementation of our system. In Section 4, we describe our experiments and present the results. In Section 5, we draw conclusions and point out future directions.

## 2 SINGER IDENTIFICATION ROBUST TO ACCOMPANIMENT SOUNDS

This paper describes an automatic singer identification system, which is the system for determining a singer's name of given musical audio signals. The target data are real-world musical audio signals such as popular music CD recordings that contain singing voices of a single singer and accompaniment sounds.

The main difficulty in achieving automatic singer identification lies in the negative influences of accompaniment sounds. Since a singing voice usually exists with accompaniment sounds at the same time, acoustical features that are extracted from such a singing voice will be dependent on the accompaniment sounds. When features that are commonly used in speaker identification studies, such as cepstral coefficients or linear prediction coefficients (LPC), are extracted, in fact, those to be obtained from musical audio signals will represent not solely the singing voice but a mixture of the singing voice and the accompaniment sounds. To achieve accurate singer identification, therefore, it is indispensable to cope with this accompaniment sound problem.

One possible solution to this problem may be to use data influenced by accompaniment sounds for both training and identification. In fact, most of the previous studies [4, 5, 6, 7] adopted this approach. However, it often fails because accompaniment sounds usually have different acoustical features from song to song. For example, the acoustical similarity for two musical pieces, the accompaniments of which are on a piano solo and a full band, respectively, will not become high enough, even if they are sung by the same singer.

To solve the problem, we developed two methods: accompaniment sound reduction and reliable frame selection. Figure 1 shows an overview of our methods.

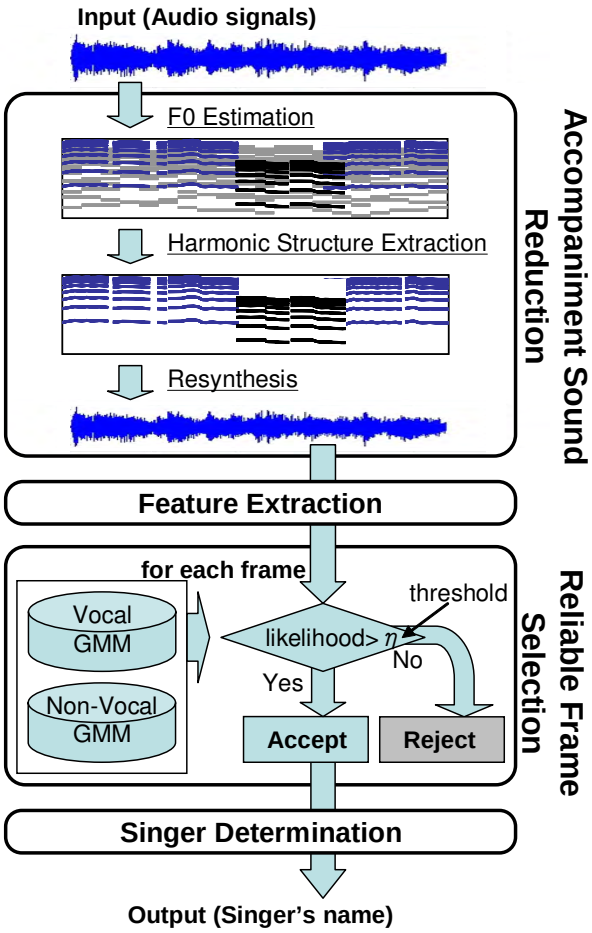



Figure 1: Method overview

### 2.1 Accompaniment Sound Reduction

One of the best solutions to the accompaniment sound influence is to reduce the accompaniment sounds from a given audio signal. In order to achieve this, we use a melody resynthesis technique based on the harmonic structure that consists of the following three parts:

1. Estimating the fundamental frequency (F0) of the melody using Goto's PreFEst [8].
2. Extracting the harmonic structure corresponding to the melody.
3. Resynthesizing the audio signal (waveform) corresponding to the melody using a sinusoidal synthesis.

Thus, we obtain the waveform corresponding only to the melody. Note that the melody's waveform obtained with this method contains instrument (*i.e.*, non-vocal) sounds in interlude sections as well as voices in singing sections, because the melody is just defined as the most predominant note in each frame [8]. It may therefore be considered necessary to detect singing sections. In practice, however,

we can omit this detection, which is a difficult problem, by using reliable frame selection described below.

### 2.2 Reliable Frame Selection

Another solution to the accompaniment sound influence is to select frames that are less influenced by the accompaniment sounds and to use only them for identification. We call this approach reliable frame selection. In order to achieve this, we introduce two kinds of Gaussian mixture models (GMMs), a vocal GMM $\lambda_{\mathrm{V}}$ and a non-vocal GMM $\lambda_{\mathrm{N}}$. The vocal GMM $\lambda_{\mathrm{V}}$ is trained on feature vectors extracted from singing sections, and the non-vocal GMM $\lambda_{\mathrm{N}}$ is trained on those extracted from interlude sections. Given a feature vector $\mathbf{x}$, the likelihoods for the two GMMs, $p(\mathbf{x}|\lambda_{\mathrm{V}})$ and $p(\mathbf{x}|\lambda_{\mathrm{N}})$, represent how the feature vector $\mathbf{x}$ is like a vocal or a (non-vocal) instrument, respectively. If the feature vector $\mathbf{x}$ is less influenced by accompaniment sounds (*i.e.*, more reliable), $p(\mathbf{x}|\lambda_{\mathrm{V}})$ will be higher and $p(\mathbf{x}|\lambda_{\mathrm{N}})$ will be lower. We therefore determine whether the feature vector $\mathbf{x}$ is reliable or not based on the following equation:

$$\log p(\mathbf{x}|\lambda_{\mathrm{V}}) - \log p(\mathbf{x}|\lambda_{\mathrm{N}}) \underset{\text{not-reliable}}{\overset{\text{reliable}}{\gtrless}} \eta, \qquad (1)$$

where $\eta$ is a threshold. In our experiments, we use 64-mixture GMMs. It is difficult to decide a universal threshold for a variety of songs because we cannot select enough feature vectors for classification from a song which have few reliable frames. We therefore determine the threshold dependent on songs so that $\alpha\%$ of the whole frames in the song are selected as reliable frames. Note that most of the non-vocal frames are rejected in this selection step. This means that we can avoid detecting singing sections by using this reliable frame selection.

## 3 IMPLEMENTATION

In this section, we describe the implementation of our system. As described above, our system consists of the following four phases: accompaniment sound reduction, feature extraction, reliable frame selection and classification.

### 3.1 Pre-Processing

Given an audio signal, it is monauralized and downsampled to 16 kHz. Then, the spectrogram is calculated using the short-time Fourier transform shifted by 10.0 ms (160 points) with a 2048-point (128.0 ms) Hamming window.

### 3.2 Accompaniment Sound Reduction

Using the method described in Section 2.1, we reduce accompaniment sounds as follows:

#### *3.2.1 F0 Estimation*

We use Goto's PreFEst [8] for estimating the F0s of the melody. PreFEst estimates 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 F0s of the melody by applying PreFEst with reliable frequency-range limitation.

We will describe a summary of PreFEst below. Hereafter, $x$ is the log-scale frequency denoted in units of cents (a musical-interval measurement), and $(t)$ means time. Given the power spectrum $\Psi_p^{(t)}(x)$, we first apply a band-pass filter (BPF) that is designed so that it covers 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 for the melody line. To enable the application of statistical methods, we represent each of the bandpass-filtered frequency components as a probability density function (PDF), called an observed PDF, $p_\Psi^{(t)}(x)$:

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

Then, we consider 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 follows:

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

$$\theta^{(t)} = \{w^{(t)}(F)|\mathrm{Fl} \le F \le \mathrm{Fh}\}, \qquad (4)$$

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

$$\int_{\mathrm{Fh}_i}^{\mathrm{Fl}_i} w^{(t)}(F)dF = 1. \qquad (5)$$

Tone model represents a typical harmonic structrue and indicates where the harmonics of the F0 tend to occur. Then, we estimate $w^{(t)}(F)$ using EM algorithm and regard it as the F0's PDF. Finally, we obtain the most dominant F0s $\overline{F^{(t)}}$ by the following equation:

$$\overline{F^{(t)}} = \underset{F}{\mathrm{argmax}}\, w^{(t)}(F) \qquad (6)$$

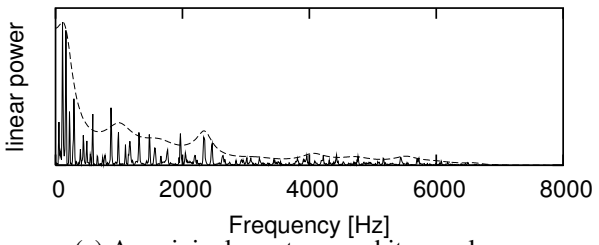
#### *3.2.2 Harmonic Structure Extraction*

Based on the estimated F0, we extract the power and the phase of fundamental frequency component and harmonic components. For each component, we allow $|r|$ cent error and extract the peak in the allowed area. The power $A_l$, the phase $\theta_l$ and frequency $F_l$ of $l$ th overtone $(l = 1,\ldots,20)$ can be represented as

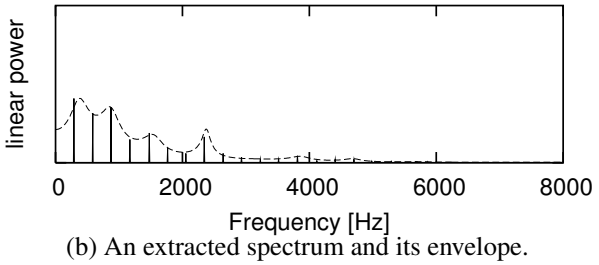
$$F_l = \underset{F}{\mathrm{argmax}}\,|S(F)| \quad (l\overline{F}\cdot(1-2^{\frac{r}{1200}}) \le F \le l\overline{F}\cdot(1+2^{\frac{r}{1200}})), \qquad (7)$$

$$A_l = |S(F_l)|, \qquad (8)$$

$$\theta_l = \arg S(F_l), \qquad (9)$$



(a) An original spectrum and its envelope.



(b) An extracted spectrum and its envelope.

Figure 2: Example of F0 estimation and harmonic structure extraction. Envelopes of the spectrums are calculated using linear prediction (LP) analyses.

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

Figure 2 shows an example of F0 estimation and harmonic structure extraction. Figure 2 (a) shows an original spectrum and its envelope and Figure 2 (b) shows an extracted spectrum and its envelope. As seen in the figures, a spectral envelope of extracted spectrum precisely represents formants of singing voice, compared with that of original spectrum.

#### 3.2.3 Resynthesis

We resynthesize the audio signals of the melody from the extracted harmonic structure by using a sinusoidal model [9]. Resynthesized audio signals are expressed as

$$s(t) = \sum_{l=1}^{L} A_l \cos(\omega_l t + \theta_l), \tag{10}$$

where $A_l$, $\theta_l$, $F_l$ represent the power, the phase and the frequency of the $l$ th overtone and $t$ is time.

### 3.3 Feature Extraction

We calculate feature vectors from the resynthesized audio signals. It is known that the individual characteristics of speech signals are expressed in their spectral envelopes. In the field of speech recognition studies, in fact, various methods have been proposed [10] for calculating feature vectors concerning spectral envelopes. Here, we compare some of them, which are commonly used in speech recognition studies.

#### 3.3.1 Mel-frequency Cepstral Coefficients (MFCC)

MFCCs [11, 12] are cepstral coefficients calculated on a mel-frequency scale. Cepstral analysis is the method to separate envelope of spectrum from fine structure. In order to compute cepstral coefficients [10], we take the log-magnitude descrete cosine transform (DCT) from the power spectrum. The envelops are represented in lower order of the cepstral coefficients, while the fine structures are in higher order. Mel-frequency is a logarithmic frequency scale fitted to the characteristics of the human auditory sense. For the MFCC computations, mel-filterbank analysis is applied first. Then, we obtain the MFCC from the log-magnitude DCT. In this paper, we use 15 dimensional MFCC, caluculated via 20 dimensional mel-filterbank analysis.

#### 3.3.2 Linear Prediction Coefficients (LPC)

Linear prediction (LP) analysis [13, 14] is a method for estimating the transfer function of vocal tract, assuming that input audio signal contains only human voice. In the LP model, given a signal $s(n)$, we predict the signal as a linear combination of its previous samples. The predicted value $s_W(n)$ is given by

$$s_W(n) = \sum_{i=1}^{p} \alpha_i s_W(n-i) + g(n), \tag{11}$$

where $p$ represents the order of the predictor, $\alpha_i$s are defined as the linear prediction coefficients (LPC), and $g(n)$ represents the error in the model. The LPCs are determined by minimizing the mean squared prediction error of $g(n)$. We use 20th-order LPC in this paper.

#### 3.3.3 LP-derived Cepstral Coefficients (LPCC)

LPCCs [13] are cepstral coefficients of a LPC spectrum. Cepstral analysis on the LPC spectrum plays a role of orthogonalization and is known to be effective in pattern recognition. The LPCCs $c(n)$ is directly obtained from the LPC with the following equation:

$$c(n) = \begin{cases} \log \sigma^2 & (n=0) \\ \alpha_n + \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) c(k)\alpha_{n-k} & (1 \le n \le p) \\ \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) c(k)\alpha_{n-k} & (n > p) \end{cases}, \tag{12}$$

where $\sigma^2$ represents the power of the signal, $\alpha_n$ represent the LPCs, and $p$ represents an order of the LPC. We set the order of the LPCC to 15 in this paper.

#### 3.3.4 Linear Prediction Mel Cepstral Coefficients (LPMCC)

LPMCCs are mel-cepstral coefficients of LPC spectrum. In addition to the role of orthogonalization, the LPMCCs are superior to the LPC in terms of suitability to the human auditory sense, which is a benefit of the mel-frequency scale. We derive the LPMCC by computing the MFCC from the LPC spectrum because of simplicity of implementation. We set the order of the LPMCC to 15 in this paper.

Table 1: 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 |

Table 2: Songs used for evaluation. The numbers written in the table are piece numbers of 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 |

### 3.4 Reliable Frame Selection

We select frames that are reliable and influenced a little by accompaniment sounds based on the method described in Section 2.2.

### 3.5 Singer Determination

The name of the singer is determined based on 64-mixture GMMs. Let $\mathbf{X} = \{\mathbf{x}_t | t = 1, \ldots, T\}$ be a time series of feature vectors selected in the reliable frame selection phase, and $\lambda_s$ be the GMM for the singer $s$. Then, the name of the singer is determined through the following equation:

$$s = \underset{i}{\operatorname{argmax}} \frac{1}{T} \sum_{t=1}^{T} \log p(\mathbf{x}_t | \lambda_i). \tag{13}$$

## 4 EXPERIMENTS

In this section, we describe the experiments that were conducted to evaluate our system.

### 4.1 Effectiveness of the whole system

To confirm the effectiveness of our methods, accompaniment sound reduction and reliable frame selection, we conducted experiments on singer identification under the following four conditions:

Table 3: Experimental results that show effectiveness of the whole system, where "reduc." and "selec." mean accompaniment sound reduction and reliable frame selection, respectively.

| | (i) baseline | (ii) reduc. only | (iii) selec. only | (iv) ours |
|---|---|---|---|---|
| a | 1/4 | 2/4 | 2/4 | 4/4 |
| b | 3/4 | 1/4 | 3/4 | 4/4 |
| c | 2/4 | 2/4 | 3/4 | 4/4 |
| d | 4/4 | 4/4 | 4/4 | 4/4 |
| e | 1/4 | 0/4 | 0/4 | 4/4 |
| f | 1/4 | 2/4 | 2/4 | 3/4 |
| g | 0/4 | 2/4 | 0/4 | 3/4 |
| h | 4/4 | 4/4 | 4/4 | 4/4 |
| i | 4/4 | 4/4 | 3/4 | 4/4 |
| j | 1/4 | 3/4 | 2/4 | 4/4 |
| Total | 53% | 60% | 58% | 95% |

**(i)** without both the reduction and the selection (baseline),

**(ii)** without the reduction, with the selection,

**(iii)** with the reduction, without the selection, and

**(iv)** with both the reduction and the selection (ours).

We used forty songs by ten different singers (five were male and five were female), listed in Table 2, taken from "RWC Music Database: Popular" [15]. Using these data, we conducted the 4-fold cross validation, that is, we first divided the whole data into four groups, $D_i$ ($i = 1, 2, 3, 4$) in Table 2, and then repeated the following step four times: each time, we left out one of the four groups for training and used the omitted one for testing. As the training data for the reliable frame selection, we used twenty-five songs of sixteen different singers listed in Table 1, also taken from "RWC Music Data: Popular", which differ from the singers used for evaluation. We set $\alpha$ to 15%, in reference to the experiment described in Section 4.2. As a feature vector, in response to the experiment described in Section 4.3, we use the LPMCC with the reduction and the MFCC without. We adopt the MFCC for the experiment without the reduction, because, as described in Section 3.3.2, the LPMCC that is based on LPC can be applied only to human voice.

Table 3 shows results of the experiments. As seen in the table, accompaniment sound reduction and reliable frames selection improved the accuracy of singer identification. When these two methods were used together, in particular, the accuracy was significantly improved: from 53% to 95%.

Figure 3 shows confusion matrices of the experiments. As seen in the figure, confusions between male and female decreased by using the reduction method. It means that, in the cases of (ii) and (iv), the reduction method reduced the influences of accompaniment sound, and the system could correctly identify the genders. On the other hand, in the cases without the reduction method (Conditions (i) and (iii)), the influences of accompaniment sound prevented the system from correctly identifying even the genders of the singers.

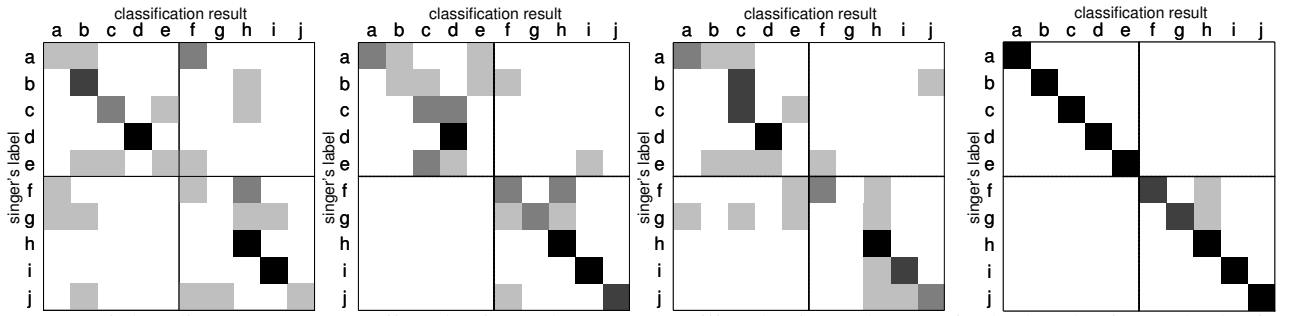



Figure 3: Confusion matrices. Center lines in each figure are boundaries between male and female. Note that confusion between male and female decreased by using the accompaniment sound reduction method.

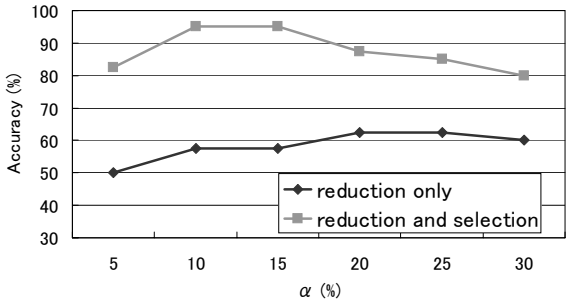



Figure 4: Experimental results that show dependency of accuracy on $\alpha$. $\alpha$% of all frames was judged as reliable.

Table 4: Investigation of accompaniment sound influence and comparison of features. The feature name written in bold font is the one that gives the highest accuracy. "corr. F0s" and "est. F0s" mean using correct F0s and estimated F0s, respectively, and "reduc." means accompaniment sound reduction.

| | (i) | (ii) | (iii) | (iv) | (v) |
|---|---|---|---|---|---|
| | Using vocal-only | | Using mixed-down data | | |
| | w/o reduc. | with reduc. | with reduc. corr. F0s | with reduc. est. F0s | w/o reduc. |
| MFCC | **98**% | 95% | 78% | 75% | **75**% |
| LPC | 83% | 88% | 50% | 58% | 48% |
| LPCC | 95% | **98**% | 75 % | 75% | 63% |
| LPMCC | **98**% | **98**% | **88**% | **83**% | 68% |

### 4.2 Dependency of accuracy on $\alpha$

We conducted experiments with setting $\alpha$ to various values to investigate the dependency of accuracies on $\alpha$. Experimental results shown in Figure 4 show that classification accuracy was not affected by small changes of $\alpha$. It is also noticeable that a value of $\alpha$ that gives the highest accuracy differed. The reason of this fact is the following: accompaniment sound reduction method reduced the influences of accompaniment sounds and emphasized the differences between reliable and unreliable frames. Thus, if we raised $\alpha$ too much, the system selected many unreliable frames and the system performance decreases. Without the reduction method, on the other hand, reliability of each frame did not make much difference because of the influences of accompaniment sounds. Therefore, it was profitable to use many frames for classification by setting $\alpha$ comparatively higher. However, in this case, we could not attain sufficient classification accuracy because of the influences of accompaniment sounds.

### 4.3 Investigation of Accompaniment Sound Influence and Comparison of Features

#### *4.3.1 Conditions*

There are two purposes in the experiments here. The first one is to investigate the influence by accompaniment sounds. We investigated it by comparing our results to ones for vocal-only data. In addition, we compared estimating F0s of melodies and using given correct ones. In our experiments, we virtually generated the correct F0s by estimating ones using vocal-only tracks. Although the estimates using vocal-only tracks were not completely correct, its accuracy was sufficiently high comparing with estimating ones using mixed-down versions. The second purpose is to compare a variety of features. We used four kinds of features described in Section 3.3 and compared these results. In the experiments here, we manually cut out 60 seconds of singing regions for each song, because we omitted reliable frame selection method in order to investigate the effectiveness of the only accompaniment sound reduction method.

#### *4.3.2 Results and Discussion*

Table 4 shows the results of the experiments. When we focused on the differences between the cases (iv) and (v), the accuracies for the LPC, the LPCC, and the LPMCC were improved by the reduction method, whereas that for the MFCC was not improved. This is because the LPC etc. assume that given signal contains only a single speech. For this assumption, the accuracies for the LPC etc. in the case (v) were low because the inputs were a mixture of singing voices and accompaniment sounds. Because the case (iv) dealt with signals obtained by extracting only singing voices, the accuracies were higher than the case (v). Because the MFCC is not based on such an assump-

tion, on the other hand, the accuracy for the MFCC in the case (v) was high, but that in the case (iv) was same. Because the LPMCC models the sounding mechanism of humans' voices, it could be expected to achieve a high accuracy if the assumption is satisfied. In fact, whereas the accuracy for the LPMCC was lower than that for MFCC in the case (v), that for the LPMCC was higher than that for the MFCC in the case (iv). Whereas most previous studies used the MFCC, we achieved to adopt more robust features by reducing accompaniment sounds.

When we compared the LPC, the LPCC, and the LPMCC, the accuracy for the LPCC was 17% higher than that for the LPC, and that for the LPMCC was 8% higher than that for the LPCC. The reason why the accuracy for the LPCC was higher than that for the LPC is that the LPCCs are orthogonal features unlike the LPC. The reason why the accuracy for the LPMCC was higher than that for the LPCC is mel-frequency cepstrum allow better suppression of insignificant spectral variation in the higher frequency bands.

The accuracies for the cases (iii) and (iv) were comparatively close. This was because partial misestimation of F0s was not critically connected to errors for singer identification since the names of singers were determined based on the mean of a time series of likelihoods.

The case (iv) in experiment described in Section 4.1 was superior to the case (iv) in this experiment, even though we manually fed singing regions into the system for this experiment. This means that the selection method actually functioned not only as distinguishing vocal and non-vocal frames but also as determining whether each frame was reliable or not. The case (v) in this experiment, however, was inferior to the case (iii) in experiments described in 4.1. This result means that, to accurately select reliable frames, it is indispensable to use both the reduction method and the selection method together.

Table 5 lists an excerpt of experimental results for each singer. As seen in the table, the reduction method improved accuracies particularly for the singer (g). This was because the songs of the singer (g) have different kinds of genres such as a piano ballad and R&B. These songs are accompanied on different instruments and hence have different acoustical characteristics of accompaniment sounds. Whereas the system without the reduction method did not correctly identify the singer's name for these songs, that with our method did. This result shows that the reduction method could reduce the influence of acoustical differences in accompaniment sounds. In contrast, identification errors for the singer (e) increased by the reduction method. This is because the melodies' F0s were incorrectly estimated in some songs. We can also confirm this by the fact that identification errors did not increase when we provided the correct F0s.

Table 5: Accuracy for each singer, where "reduc." means accompaniment sound reduction.

| | (iii) with reduc., corr. F0s LPMCC | (iv) with reduc., est. F0s LPMCC | (v) w/o reduc. MFCC |
|---|---|---|---|
| a | 3/4 | 3/4 | 3/4 |
| b | 4/4 | 4/4 | 4/4 |
| c | 4/4 | 4/4 | 3/4 |
| d | 4/4 | 4/4 | 4/4 |
| e | 3/4 | 2/4 | 3/4 |
| f | 3/4 | 2/4 | 2/4 |
| g | 2/4 | 2/4 | 0/4 |
| h | 4/4 | 4/4 | 4/4 |
| i | 4/4 | 4/4 | 4/4 |
| j | 4/4 | 4/4 | 3/4 |
| Total | 88% | 83% | 75% |

## 5 CONCLUSION

We have described two methods that work in combination to automatically identify singers for music information retrieval. To identify the singer names of musical pieces including sounds of various instruments, our method solved the problem of the accompaniment sound influences. In our experiments with forty songs by ten singers, we found that our methods achieved identification accuracy of 95% and confirmed the robustness and effectiveness of those methods.

The main contributions of this paper can be summarized as follows:

- We clarified the problem of the accompaniment sound influence for singer identification, which has not been dealt with except for only a few attempts, and provided two effective solutions, accompaniment sound reduction and reliable frame selection.
- The use of the accompaniment sound reduction method made it possible to reduce the negative influence of accompaniment sound by extracting and resynthesizing the harmonic structure of the predominant melody. Though similar methods have been used to improve the noise robustness in the field of speech recognition [16], this is the first paper that shows its effectiveness for singer identification.
- The reliable frame selection method made it possible to select frames reliable enough for classification. Although similar methods were used in previous studies, they focused on distinguishing vocal and non-vocal frames; they did not consider the reliability of each frame. Note that our selection method rejects even unreliable vocal frames as well as non-vocal frames.
- We showed an investigation of features for singer identification. While many features have been proposed in the field of speech recognition [10], it has not been clear which feature was appropriate for singer identification. We compared various features in terms of the singer identification, and found that the LPMCC was the most robust among them. This result will contribute to the progression of singer identification research.

In the future, we plan to extend our method to calculate acoustical similarities between musical pieces in

terms of singers and apply it to music information retrieval based on singing voice similarities.

## ACKNOWLEDGEMENTS

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Scientific Research (A), No.15200015, and Informatics Research Center for Development of Knowledge Society Infrastructure (COE program of MEXT, Japan). We thank everyone who has contributed to building and distributing the RWC Music Database [15]. We also thank Kazuyoshi Yoshii and Takuya Yoshioka (Kyoto University) for their valuable discussions.

## REFERENCES

[1] Wei-Ho Tsai, Hsin-Min Wang, and Dwight Rodgers. Automatic singer identification of popular music recordings via estimation and modeling of solo vocal signal. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech2003)*, 2003.

[2] Wei-Ho Tsai and Hsin-Min Wang. Automatic detection and tracking of target singer in multi-singer music recordings. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, pages 221–224, 2004.

[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 Transactions on Speech and Audio Processing*, 2(2):245–257, 1994.

[4] Brian Whitman, Gary Flake, and Steve Lawrence. Artist detection in music with Minnowmatch. In *Proceedings of the 2001 IEEE Workshop on Neural Networks for Signal Processing*, pages 559–568, 2001.

[5] Adam L. Berenzweig, Daniel P. W. Ellis, and Steve Lawrence. Using voice segments to improve artist classification of music. In *AES 22nd International Conference on Virtual, Synthetic, and Entertainment Audio*, 2002.

[6] Youngmoo Edmund Kim and Brian Whitman. Singer identificatin in popular music recordings using voice coding features. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR2002)*, pages 164–169, 2002.

[7] Tong Zhang. Automatic singer identification. In *Proceedings of IEEE International Conference on Multimedia & Expo (ICME 2003)*, 2003.

[8] Masataka Goto. A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4):311–329, 2004.

[9] James Anderson Moorer. Signal procesing aspects of computer music: A survey. *Proceedings of the IEEE*, 65(8):1108–1137, 1977.

[10] Joseph Picone. Signal modeling techniques in speech recognition. *IEEE Proceedings*, 81(9):1215–1247, 1993.

[11] Steven B. Davis and Paul Mermelstein. Comparison of parametric representation for monosyllabic word recognition. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 28(4):357–366, 1980.

[12] Beth Logan. Mel frequency cepstral coefficients for music modelling. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2000)*, pages 23–25, 2000.

[13] Bishnu S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *the Journal of the Acoustical Society of America*, 55(6):1304–1312, 1974.

[14] K. Shikano. Evaluation of LPC spectral matching measures for phonetic unit recognition. Technical Report CMU-CS-96-108, CMU, Computer Science Department, 1986.

[15] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Popular, classical, and jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pages 287–288, October 2002.

[16] Tomohiro Nakatani and Hiroshi G. Okuno. Harmonic sound stream segregation using localization and its application to speech stream segregation. *Speech Communcations*, 27:209–222, 1999.