Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 51979, 15 pages
doi:10.1155/2007/51979

Research Article

Instrument Identification in Polyphonic Music: Feature Weighting to Minimize Influence of Sound Overlaps

Tetsuro Kitahara,[1] Masataka Goto,[2] Kazunori Komatani,[1] Tetsuya Ogata,[1] and Hiroshi G. Okuno[1]

[1] *Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Sakyo-Ku, Kyoto 606-8501, Japan*
[2] *National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8568, Japan*

Received 7 December 2005; Revised 27 July 2006; Accepted 13 August 2006

Recommended by Ichiro Fujinaga

We provide a new solution to the problem of feature variations caused by the overlapping of sounds in instrument identification in polyphonic music. When multiple instruments simultaneously play, partials (harmonic components) of their sounds overlap and interfere, which makes the acoustic features different from those of monophonic sounds. To cope with this, we weight features based on how much they are affected by overlapping. First, we quantitatively evaluate the influence of overlapping on each feature as the ratio of the within-class variance to the between-class variance in the distribution of training data obtained from polyphonic sounds. Then, we generate feature axes using a weighted mixture that minimizes the influence via linear discriminant analysis. In addition, we improve instrument identification using musical context. Experimental results showed that the recognition rates using both feature weighting and musical context were 84.1% for duo, 77.6% for trio, and 72.3% for quartet; those without using either were 53.4, 49.6, and 46.5%, respectively.

1. INTRODUCTION

While the recent worldwide popularization of online music distribution services and portable digital music players has enabled us to access a tremendous number of musical excerpts, we do not yet have easy and efficient ways to find those that we want. To solve this problem, efficient music information retrieval (MIR) technologies are indispensable. In particular, automatic description of musical content in a universal framework is expected to become one of the most important technologies for sophisticated MIR. In fact, frameworks such as MusicXML [1], WEDELMUSIC Format [2], and MPEG-7 [3] have been proposed for describing music or multimedia content.

One reasonable approach for this music description is to transcribe audio signals to traditional music scores because the music score is the most common symbolic music representation. Many researchers, therefore, have tried automatic music transcription [4–9], and their techniques can be applied to music description in a score-based format such as MusicXML. However, only a few of them have dealt with identifying musical instruments. Which instruments are used is important information for two reasons. One is that it is necessary for generating a complete score. Notes for different instruments, in general, should be described on different staves in a score, and each stave should have a description of instruments. The other reason is that the instruments characterize musical pieces, especially in classical music. The names of some musical forms are based on instrument names, such as "piano sonata" and "string quartet." When a user, therefore, wants to search for certain types of musical pieces, such as piano sonatas or string quartets, a retrieval system can use information on musical instruments. This information can also be used for jumping to the point when a certain instrument begins to play.

This paper, for these reasons, addresses the problem of which facilitates the above-mentioned score-based music annotation, in audio signals of polyphonic music, in particular, classical Western tonal music. Instrument identification is a sort of pattern recognition that corresponds to speaker identification in the field of speech information processing. Instrument identification, however, is a more difficult problem than noiseless single-speaker identification because, in most musical pieces, multiple instruments simultaneously

play. In fact, studies dealing with polyphonic music [7, 10–13] have used duo or trio music chosen from 3–5 instrument candidates, whereas those dealing with monophonic sounds [14–23] have used 10–30 instruments and achieved the performance of about 70–80%. Kashino and Murase [10] reported a performance of 88% for trio music played on piano, violin, and flute given the correct fundamental frequencies (F0s). Kinoshita et al. [11] reported recognition rates of around 70% (70–80% if the correct F0s were given). Eggink and Brown [13] reported a recognition rate of about 50% for duo music chosen from five instruments given the correct F0s. Although a new method that can deal with more complex musical signals has been proposed [24], it cannot be applied to score-based annotation such as MusicXML because the key idea behind this method is to identify instrumentation instead of instruments at each frame, not for each note. The main difficulty in identifying instruments in polyphonic music is the fact that acoustical features of each instrument cannot be extracted without blurring because of the overlapping of partials (harmonic components). If a clean sound for each instrument could be obtained using sound separation technology, the identification of polyphonic music would become equivalent to identifying the monophonic sound of each instrument. In practice, however, a mixture of sounds is difficult to separate without distortion.

In this paper, we approach the above-mentioned *overlapping problem* by weighting each feature based on how much the feature is affected by the overlapping. If we can give higher weights to features suffering less from this problem and lower weights to features suffering more, it will facilitate robust instrument identification in polyphonic music. To do this, we quantitatively evaluate the influence of the overlapping on each feature as the ratio of the *within-class variance* to the *between-class variance* in the distribution of training data obtained from polyphonic sounds because greatly suffering from the overlapping means having large variation when polyphonic sounds are analyzed. This evaluation makes the feature weighting described above equivalent to dimensionality reduction using *linear discriminant analysis* (LDA) on training data obtained from polyphonic sounds. Because LDA generates feature axes using a weighted mixture where the weights minimize the ratio of the within-class variance to the between-class variance, using LDA on training data obtained from polyphonic sounds generates a subspace where the influence of the overlapping problem is minimized. We call this method DAMS (discriminant analysis with mixed sounds). In previous studies, techniques such as time-domain waveform template matching [10], feature adaptation with manual feature classification [11], and the missing feature theory [12] have been tried to cope with the overlapping problem, but no attempts have been made to give features appropriate weights based on their robustness to the overlapping.

In addition, we propose a method for improving instrument identification using musical context. This method is aimed at avoiding musically unnatural errors by considering the temporal continuity of melodies; for example, if the identified instrument names of a note sequence are all "flute" except for one "clarinet," this exception can be considered an error and corrected.

The rest of this paper is organized as follow. In Section 2, we discuss how to achieve robust instrument identification in polyphonic music and propose our feature weighting method, DAMS. In Section 3, we propose a method for using musical context. Section 4 explains the details of our instrument identification method, and Section 5 reports the results of our experiments including those under various conditions that were not reported in [25]. Finally, Section 6 concludes the paper.

2. INSTRUMENT IDENTIFICATION ROBUST TO OVERLAPPING OF SOUNDS

In this section, we discuss how to design an instrument identification method that is robust to the overlapping of sounds. First, we mention the general formulation of instrument identification. Then, we explain that extracting harmonic structures effectively suppresses the influence of other simultaneously played notes. Next, we point out that harmonic structure extraction is insufficient and we propose a method of feature weighting to improve the robustness.

2.1. General formulation of instrument identification

In our instrument identification methodology, the instrument for each note is identified. Suppose that a given audio signal contains K notes, $n_1, n_2, \ldots, n_k, \ldots, n_K$. The identification process has two basic subprocesses: feature extraction and a posteriori probability calculation. In the former process, a feature vector consisting of some acoustic features is extracted from the given audio signal for each note. Let $\mathbf{x}_k$ be the feature vector extracted for note n_k. In the latter process, for each of the target instruments, $\omega_1, \ldots, \omega_m$, the probability $p(\omega_i \mid \mathbf{x}_k)$ that the feature vector $\mathbf{x}_k$ is extracted from a sound of the instrument ω_i is calculated. Based on the Bayes theorem, $p(\omega_i \mid \mathbf{x}_k)$ can be expanded as follows:

$$p(\omega_i \mid \mathbf{x}_k) = \frac{p(\mathbf{x}_k \mid \omega_i)\, p(\omega_i)}{\sum_{j=1}^{m} p(\mathbf{x}_k \mid \omega_j)\, p(\omega_j)}, \tag{1}$$

where $p(\mathbf{x}_k \mid \omega_i)$ is a probability density function (PDF) and $p(\omega_i)$ is the a priori probability with respect to the instrument ω_i. The PDF $p(\mathbf{x}_k \mid \omega_i)$ is trained using data prepared in advance. Finally, the name of the instrument maximizing $p(\omega_i \mid \mathbf{x}_k)$ is determined for each note n_k. The symbols used in this paper are listed in Table 1.

2.2. Use of harmonic structure model

In speech recognition and speaker recognition studies, features of spectral envelopes such as Mel-frequency cepstrum coefficients are commonly used. Although they can reasonably represent the general shapes of observed spectra, when a signal of multiple instruments simultaneously playing is analyzed, focusing on the component corresponding to each instrument from the observed spectral envelope is difficult. Because most musical sounds except percussive ones have

Table 1: List of symbols.

$n_1, \ldots, n_K$	Notes contained in a given signal
$\mathbf{x}_k$	Feature vector for note n_k
$\omega_1, \ldots, \omega_m$	Target instruments
$p(\omega_i \mid \mathbf{x}_k)$	A posteriori probability
$p(\omega_i)$	A priori probability
$p(\mathbf{x}_k \mid \omega_i)$	Probability density function
$s_{\mathrm{h}}(n_k), s_{\mathrm{l}}(n_k)$	Maximum number of simultaneously played notes in higher or lower pitch ranges when note n_k is being played
$\mathcal{N}$	Set of notes extracted for context
c	Number of notes in $\mathcal{N}$
f	Fundamental frequency (F0) of a given note
$f_{\mathbf{x}}$	F0 of feature vector $\mathbf{x}$
$\boldsymbol{\mu}_i(f)$	F0-dependent mean function for instrument ω_i
Σ_i	F0-normalized covariance for instrument ω_i
χ_i	Set of training data of instrument ω_i
$p(\mathbf{x} \mid \omega_i; f)$	Probability density function for F0-dependent multivariate normal distribution
$D^2(\mathbf{x}; \boldsymbol{\mu}_i(f), \Sigma_i)$	Squared Mahalanobis distance

harmonic structures, previous studies on instrument identification [7, 9, 11] have commonly extracted the harmonic structure of each note and then extracted acoustic features from the structures.

We also extract the harmonic structure of each note and then extract acoustic features from the structure. The harmonic structure model $\mathcal{H}(n_k)$ of the note n_k can be represented as the following equation:

$$\mathcal{H}(n_k) = \{(F_i(t), A_i(t)) \mid i = 1, 2, \ldots, h,\ 0 \le t \le T\}, \quad (2)$$

where $F_i(t)$ and $A_i(t)$ are the frequency and amplitude of the ith partial at time t. Frequency is represented by relative frequency where the temporal median of the fundamental frequency, $F_1(t)$, is 1. Above, h is the number of harmonics, and T is the note duration. This modeling of musical instrument sounds based on harmonic structures can restrict the influence of the overlapping of sounds of multiple instruments to the overlapping of partials. Although actual musical instrument sounds contain nonharmonic components, which can be factors characterizing sounds, we focus only on harmonic ones because nonharmonic ones are difficult to reliably extract from a mixture of sounds.

2.3. Feature weighting based on robustness to overlapping of sounds

As described in the previous section, the influence of the overlapping of sounds of multiple instruments is restricted to the overlapping of the partials by extracting the harmonic structures. If two notes have no partials with common frequencies, the influence of one on the other when the two notes are simultaneously played may be ignorably small. In practice, however, partials often overlap. When two notes with the pitches of C4 (about 262 Hz) and G4 (about 394 Hz) are simultaneously played, for example, the 3 ith partials of the C4 note and the 2 ith partials of the G4 note overlap for every natural number i. Because note combinations that can generate harmonious sounds cause overlaps in many partials in general, coping with the overlapping of partials is a serious problem.

One effective approach for coping with this overlapping problem is feature weighting based on the robustness to the overlapping problem. If we can give higher weights to features suffering less from this problem and lower weights to features suffering more, it will facilitate robust instrument identification in polyphonic music. Concepts similar to this feature weighting, in fact, have been proposed, such as the missing feature theory [12] and feature adaptation [11].

(i) Eggink and Brown [12] applied the missing feature theory to the problem of identifying instruments in polyphonic music. This is a technique for canceling unreliable features at the identification step using a vector called a mask, which represents whether each feature is reliable or not. Because masking a feature is equivalent to giving a weight of zero to it, this technique can be considered an implementation of the feature weighting concept. Although this technique is known to be effective if the features to be masked are given, automatic mask estimation is very difficult in general and has not yet been established.

(ii) Kinoshita et al. [11] proposed a feature adaptation method. They manually classified their features for identification into three types (additive, preferential, and fragile) according to how the features varied when partials overlapped. Their method recalculates or cancels the features extracted from overlapping components according to the three types. Similarly to Eggink's work, canceling features can be considered an implementation of the feature weighting concept. Because this method requires manually classifying features in advance, however, using a variety of features is difficult. They introduced a feature weighting technique, but this technique was performed on monophonic sounds, and hence did not cope with the overlapping problem.

(iii) Otherwise, there has been Kashino's work based on a time-domain waveform template-matching technique with adaptive template filtering [10]. The aim was the robust matching of an observed waveform and a mixture of waveform templates by adaptively filtering the templates. This study, therefore, did not deal with feature weighting based on the influence of the overlapping problem.

The issue in the feature weighting described above is how to quantitatively design the influence of the overlapping problem. Because training data were obtained only from monophonic sounds in previous studies, this influence could not be evaluated by analyzing the training data. Our DAMS method quantitatively models the influence of the overlapping problem on each feature as the ratio of the within-class variance to the between-class variance in the distribution

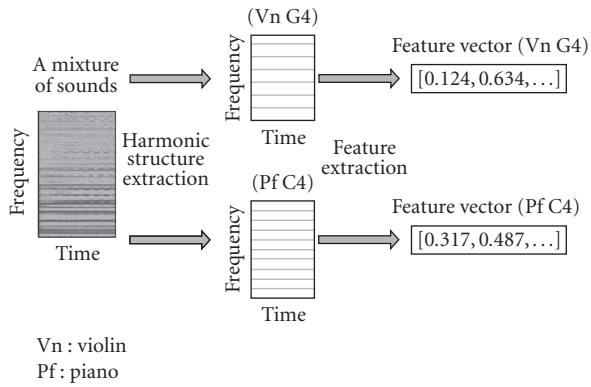

Figure 1: Overview of process of constructing mixed-sound template.

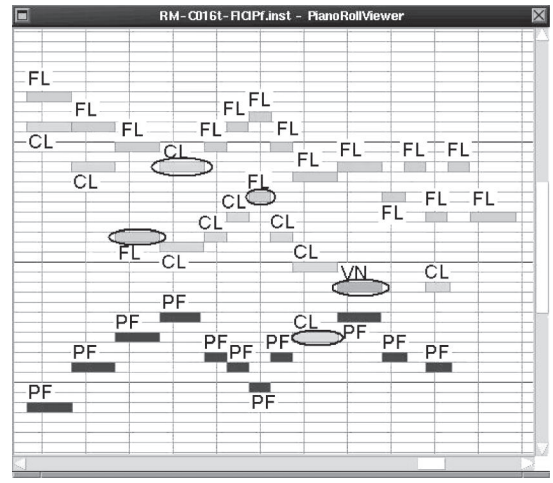

Figure 2: Example of musically unnatural errors. This example is excerpted from results of identifying each note individually in a piece of trio music. Marked notes are musically unnatural errors, which can be avoided by using musical context. PF, VN, CL, and FL represent piano, violin, clarinet, and flute.

of training data obtained from polyphonic sounds. As described in the introduction, this modeling makes weighting features to minimize the influence of the overlapping problem equivalent to applying LDA to training data obtained from polyphonic sounds.

Training data are obtained from polyphonic sounds through the process shown in Figure 1. The sound of each note in the training data is labeled in advance with the instrument name, the F0, the onset time, and the duration. By using these labels, we extract the harmonic structure corresponding to each note from the spectrogram. We then extract acoustic features from the harmonic structure. We thus obtain a set of many feature vectors, called a *mixed-sound template*, from polyphonic sound mixtures.

The main issue in constructing a mixed-sound template is to design an appropriate subset of polyphonic sound mixtures. This is a serious issue because there are an infinite number of possible combinations of musical sounds due to the large pitch range of each instrument.[1] The musical feature that is the key to resolving this issue is a tendency of intervals of simultaneous notes. In Western tonal music, some intervals such as minor 2nds are more rarely used than other intervals such as major 3rds and perfect 5ths because minor 2nds generate dissonant sounds in general. By generating polyphonic sounds for template construction from the scores of actual (existing) musical pieces, we can obtain a data set that reflects the tendency mentioned above.[2] We believe that this approach improves instrument identification even if the pieces used for template construction are different from the piece to be identified for the following two reasons.

(i) There are different distributions of intervals found in simultaneously sounding notes in tonal music. For example, three simultaneous notes with the pitches of C4, C#4, and D4 are rarely used except for special effects.

(ii) Because we extract the harmonic structure from each note, as previously mentioned, the influence of multiple instruments simultaneously playing is restricted to the overlapping of partials. The overlapping of partials can be explained by two main factors: which partials are affected by other sounds, related to *note combinations*, and how much each partial is affected, mainly related to *instrument combinations*. Note combinations can be reduced because our method considers only relative-pitch relationships, and the lack of instrument combinations is not critical to recognition as we find in an experiment described below. If the intervals of note combinations in a training data set reflect those in actual music, therefore, the training data set will be effective despite a lack of other combinations.

3. USE OF MUSICAL CONTEXT

In this section, we propose a method for improving instrument identification by considering musical context. The aim of this method is to avoid unusual events in tonal music, for example, only one clarinet note appearing in a sequence of notes (a melody) played on a flute, as shown in Figure 2. As mentioned in Section 2.1, the a posteriori probability $p(\omega_i \mid \mathbf{x}_k)$ is given by $p(\omega_i \mid \mathbf{x}_k) = p(\mathbf{x}_k \mid \omega_i)p(\omega_i)/\sum_j p(\mathbf{x}_k \mid \omega_j)p(\omega_j)$. The key idea behind using musical context is to apply the a posteriori probabilities of n_k's temporally neighboring notes to the a priori probability $p(\omega_i)$ of the note n_k (Figure 3). This is based on the idea that if almost all notes around the note n_k are identified as the instrument ω_i, n_k is also probably played on ω_i. To achieve this, we have to resolve the following issue.

[1] Because our data set of musical instrument sounds consists of 2651 notes of five instruments, $C(2651, 3) \approx 3.1$ billion different combinations are possible even if the number of simultaneous voices is restricted to three. About 98 years would be needed to train all the combinations, assuming that one second is needed for each combination.

[2] Although this discussion is based on tonal music, this may be applicable to atonal music by preparing the scores of pieces of atonal music.

Issue: distinguishing notes played on the same instrument as n_k from neighboring notes

Because various instruments are played at the same time, an identification system has to distinguish notes that are played on the same instrument as the note n_k from notes played on other instruments. This is not easy because it is mutually dependent on musical instrument identification.

We resolve this issue as follows.

Solution: take advantage of the parallel movement of simultaneous parts.

In Western tonal music, voices rarely cross. This may be explained due to the human's ability to recognize multiple voices easier if they do not cross each other in pitch [26]. When they listen, for example, to two simultaneous note sequences that cross, one of which is descending and the other of which is ascending, they cognize them as if the sequences approach each other but never cross. Huron also explains that the pitch-crossing rule (parts should not cross with respect to pitch) is a traditional voice-leading rule and can be derived from perceptual principles [27]. We therefore judge whether two notes, n_k and n_j, are in the same part (i.e., played on the same instrument) as follows: let $s_{\mathrm{h}}(n_k)$ and $s_{\mathrm{l}}(n_k)$ be the maximum number of simultaneously played notes in the higher and lower pitch ranges when the note n_k is being played. Then, the two notes n_k and n_j are considered to be in the same part if and only if $s_{\mathrm{h}}(n_k) = s_{\mathrm{h}}(n_j)$ and $s_{\mathrm{l}}(n_k) = s_{\mathrm{l}}(n_j)$ (Figure 4). Kashino and Murase [10] have introduced musical role consistency to generate music streams. They have designed two kinds of musical roles: the highest and lowest notes (usually corresponding to the principal melody and bass lines). Our method can be considered an extension of their musical role consistency.

3.1. 1st pass: precalculation of a posteriori probabilities

For each note n_k, the a posteriori probability $p(\omega_i \mid \mathbf{x}_k)$ is calculated by considering the a priori probability $p(\omega_i)$ to be a constant because the a priori probability, which depends on the a posteriori probabilities of temporally neighboring notes, cannot be determined in this step.

3.2. 2nd pass: recalculation of a posteriori probabilities

This pass consists of three steps.

(1) Finding notes played on the same instrument

Notes that satisfy $\{n_j \mid s_{\mathrm{h}}(n_k) = s_{\mathrm{h}}(n_j) \cap s_{\mathrm{l}}(n_k) = s_{\mathrm{l}}(n_j)\}$ are extracted from notes temporally neighboring n_k. This extraction is performed from the nearest note to farther notes and stops when c notes have been extracted (c is a positive integral constant). Let $\mathcal{N}$ be the set of the extracted notes.

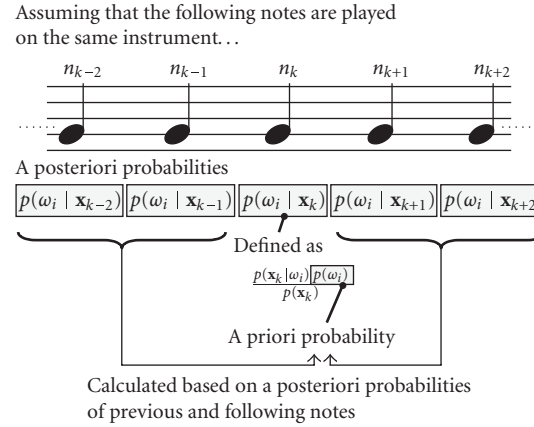

Figure 3: Key idea for using musical context. To calculate a posteriori probability of note n_k, a posteriori probabilities of temporally neighboring notes of n_k are used.

(2) Calculating a priori probability

The a priori probability of the note n_k is calculated based on the a posteriori probabilities of the notes extracted in the previous step. Let $p_1(\omega_i)$ and $p_2(\omega_i)$ be the a priori probabilities calculated from musical context and other cues, respectively. Then, we define the a priori probability $p(\omega_i)$ to be calculated here as follows:

$$p(\omega_i) = \lambda p_1(\omega_i) + (1 - \lambda) p_2(\omega_i), \tag{3}$$

where λ is a confidence measure of musical context. Although this measure can be calculated through statistical analysis as the probability that the note n_k will be played on instrument ω_i when all the extracted neighboring notes of n_k are played on ω_i, we use $\lambda = 1 - (1/2)^c$ for simplicity, where c is the number of notes in $\mathcal{N}$. This is based on the heuristics that as more notes are used to represent a context, the context information is more reliable. We define $p_1(\omega_i)$ as follows:

$$p_1(\omega_i) = \frac{1}{\alpha} \prod_{n_j \in \mathcal{N}} p(\omega_i \mid \mathbf{x}_j), \tag{4}$$

where $\mathbf{x}_j$ is the feature vector for the note n_j and α is the normalizing factor given by $\alpha = \sum_{\omega_i} \prod_{n_j} p(\omega_i \mid \mathbf{x}_j)$. We use $p_2(\omega_i) = 1/m$ for simplicity.

(3) Updating a posteriori probability

The a posteriori probability is recalculated using the a priori probability calculated in the previous step.

4. DETAILS OF OUR INSTRUMENT IDENTIFICATION METHOD

The details of our instrument identification method are given below. An overview is shown in Figure 5. First, the spectrogram of a given audio signal is generated. Next, the

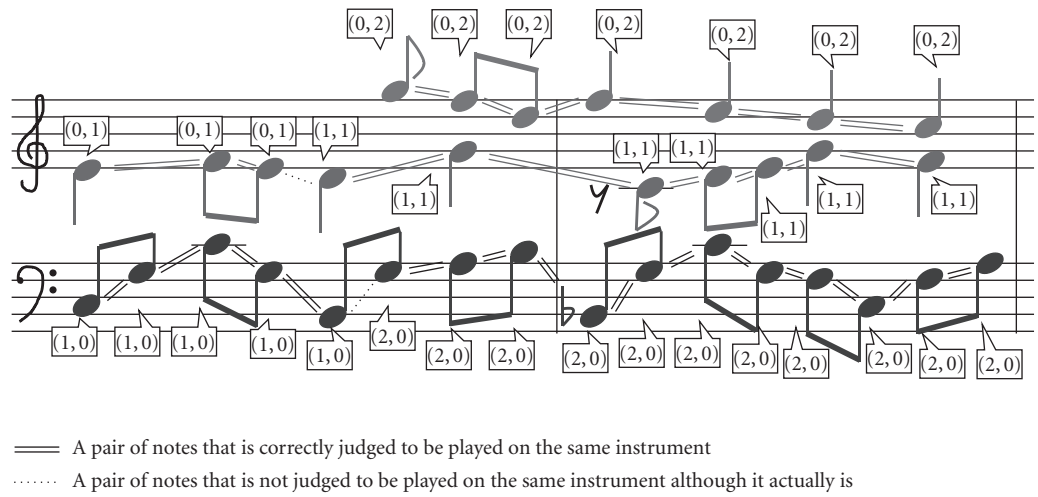

Figure 4: Example of judgment of whether notes are played on the same instrument. Each tuple (a,b) represents $s_h(n_k)$ = a and $s_l(n_k)$ = b.

harmonic structure of each note is extracted based on data on the F0, the onset time, and the duration of each note, which are estimated in advance using an existing method (e.g., [7, 9, 28]). Then, feature extraction, dimensionality reduction, a posteriori probability calculation, and instrument determination are performed in that order.

4.1. Short-time Fourier transform

The spectrogram of the given audio signal is calculated using the short-time Fourier transform (STFT) shifted by 10 milliseconds (441 points at 44.1 kHz sampling) with an 8192-point Hamming window.

4.2. Harmonic structure extraction

The harmonic structure of each note is extracted according to note data estimated in advance. Spectral peaks corresponding to the first 10 harmonics are extracted from the onset time to the offset time. The offset time is calculated by adding the duration to the onset time. Then, the frequency of the spectral peaks is normalized so that the temporal mean of F0 is 1.

Next, the harmonic structure is trimmed because training and identification require notes with fixed durations. Because a mixed-sound template with a long duration is more stable and robust than a template with a short one, trimming a note to keep it as long as possible is best. We therefore prepare three templates with different durations (300, 450, and 600 milliseconds), and the longest usable, as determined by the actual duration of each note, is automatically selected and used for training and identification.[3] For example, the 450-millisecond template is selected for a 500-millisecond note. In this paper, the 300-milliseconds, 450-millisecond, and 600-millisecond templates are called *Template Types I, II,* and *III*. Notes shorter than 300 milliseconds are not identified.

4.3. Feature extraction

Features that are useful for identification are extracted from the harmonic structure of each note. From a feature set that we previously proposed [19], we selected 43 features (for Template Type III), summarized in Table 2, that we expected to be robust with respect to sound mixtures. We use 37 features for Template Type II and 31 for I because of the limitations of the note durations.

4.4. Dimensionality reduction

Using the DAMS method, the subspace minimizing the influence of the overlapping problem is obtained. Because a feature space should not be correlated to robustly perform the LDA calculation, before using the DAMS method, we obtain a noncorrelative space by using principal component analysis (PCA). The dimensions of the feature space obtained with PCA are determined so that the cumulative proportion value is 99% (20 dimensions in most cases). By using the DAMS method in this subspace, we obtain an $(m-1)$-dimensional space (m: the number of instruments in the training data).

4.5. A posteriori probability calculation

For each note n_k, the a posteriori probability $p(\omega_i \mid \mathbf{x}_k)$ is calculated. As described in Section 2.1, this probability can be calculated using the following equation:

$$p(\omega_i \mid \mathbf{x}_k) = \frac{p(\mathbf{x}_k \mid \omega_i)\, p(\omega_i)}{\sum_j p(\mathbf{x}_k \mid \omega_j)\, p(\omega_j)}. \tag{5}$$

[3] The template is selected based on the fixed durations instead of the tempo because temporal variations of spectra, which influence the dependency of features on the duration, occur on the absolute time scale rather than in the tempo.

Table 2: Overview of 43 features.

Spectral features	
1	Spectral centroid
2	Relative power of fundamental component
3–10	Relative cumulative power from fundamental to ith components ($i = 2, 3, \ldots, 9$)
11	Relative power in odd and even components
12–20	Number of components whose durations are p% longer than the longest duration ($p = 10, 20, \ldots, 90$)
Temporal features	
21	Gradient of straight line approximating power envelope
22–30*	Average differential of power envelope during t-second interval from onset time ($t = 0.15, 0.20, 0.25, \ldots, 0.55$ (s))
31–39*	Ratio of power at t second after onset time
Modulation features	
40, 41	Amplitude and frequency of AM
42, 43	Amplitude and frequency of FM

*In Template Types I and II, some of these features have been excluded due to the limitations of the note durations.

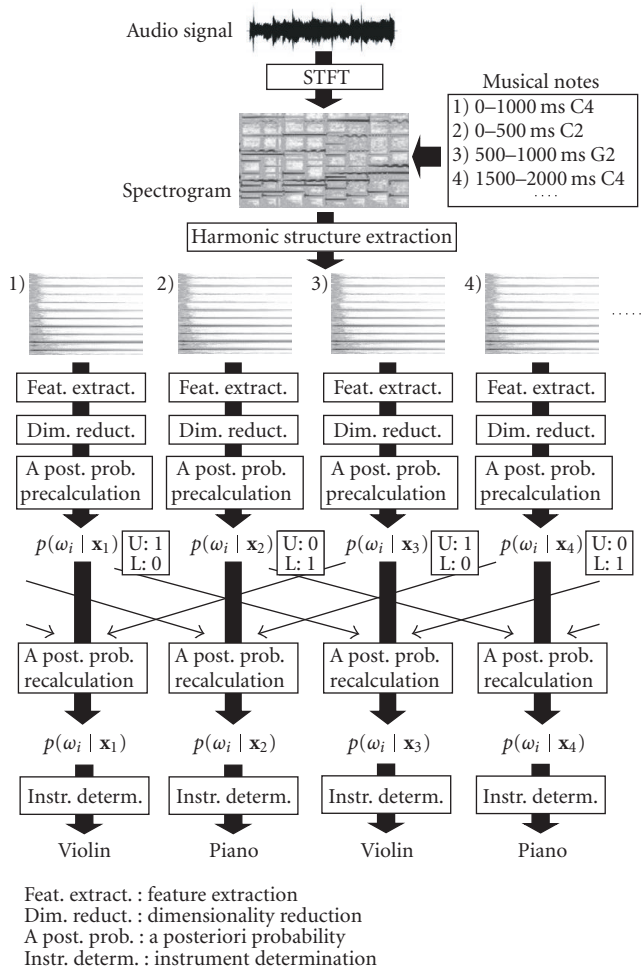

Figure 5: Flow of our instrument identification method.

The PDF $p(\mathbf{x}_k \mid \omega_i)$ is calculated from training data prepared in advance by using an F0-dependent multivariate normal distribution, as it is defined in our previous paper [19]. The F0-dependent multivariate normal distribution is designed to cope with the pitch dependency of features. It is specified by the following two parameters.

(i) F0-dependent mean function $\mu_i(f)$

For each element of the feature vector, the pitch dependency of the distribution is approximated as a function (cubic polynomial) of F0 using the least-square method.

(ii) F0-normalized covariance Σ_i

The F0-normalized covariance is calculated using the following equation:

$$\Sigma_i = \frac{1}{|\chi_i|} \sum_{\mathbf{x} \in \chi_i} (\mathbf{x} - \boldsymbol{\mu}_i(f_{\mathbf{x}}))(\mathbf{x} - \boldsymbol{\mu}_i(f_{\mathbf{x}}))', \tag{6}$$

where χ_i is the set of the training data of instrument ω_i, $|\chi_i|$ is the size of χ_i, $f_{\mathbf{x}}$ denotes the F0 of feature vector $\mathbf{x}$, and $'$ represents the transposition operator.

Once these parameters are estimated, the PDF is given as

$$p(\mathbf{x}_k \mid \omega_i; f) = \frac{1}{(2\pi)^{d/2} \left|\Sigma_i\right|^{1/2}} \exp\left\{ -\frac{1}{2} D^2(\mathbf{x}_k; \boldsymbol{\mu}_i(f), \Sigma_i) \right\}, \tag{7}$$

where d is the number of dimensions of the feature space and D^2 is the squared Mahalanobis distance defined by

$$D^2(\mathbf{x}_k; \boldsymbol{\mu}_i(f), \Sigma_i) = (\mathbf{x}_k - \boldsymbol{\mu}_i(f))' \Sigma_i^{-1} (\mathbf{x}_k - \boldsymbol{\mu}_i(f)). \tag{8}$$

The a priori probability $p(\omega_i)$ is calculated on the basis of the musical context, that is, the a posteriori probabilities of neighboring notes, as described in Section 3.

4.6. Instrument determination

Finally, the instrument maximizing the a posteriori probability $p(\omega_i \mid \mathbf{x}_k)$ is determined as the identification result for the note n_k.

Table 3: Audio data on solo instruments.

Instr. no.	Name	Pitch range	Variation	Dynamics	Articulation	no. of data
01	Piano (PF)	A0–C8	1, 2, 3	Forte, mezzo, and piano	Normal only	792
09	Classical guitar (CG)	E2–E5	〃			702
15	Violin (VN)	G3–E7	〃			576
31	Clarinet (CL)	D3–F6	〃			360
33	Flute (FL)	C4–C7	1, 2			221

Table 4: Instrument candidates for each part. The abbreviations of instruments are defined in Table 3.

Part 1	PF, VN, FL
Part 2	PF, CG, VN, CL
Part 3	PF, CG
Part 4	PF, CG

5. EXPERIMENTS

5.1. Data for experiments

We used audio signals generated by mixing audio data taken from a solo musical instrument sound database according to standard MIDI files (SMFs) so that we would have correct data on F0s, onset times, and durations of all notes because the focus of our experiments was solely on evaluating the performance of our instrument identification method by itself.

The SMFs we used in the experiments were three pieces taken from RWC-MDB-C-2001 (Piece Nos. 13, 16, and 17) [29]. These are classical musical pieces consisting of four or five simultaneous voices. We created SMFs of duo, trio, and quartet music by choosing two, three, and four simultaneous voices from each piece. We also prepared solo-melody SMFs for template construction.

As audio sources for generating audio signals of duo, trio, and quartet music, an excerpt of RWC-MDB-I-2001 [30], listed in Table 3, was used. To avoid using the same audio data for training and testing, we used 011PFNOM, 151VN-NOM, 311CLNOM, and 331FLNOM for the test data and the others in Table 3 for the training data. We prepared audio signals of all possible instrument combinations within the restrictions in Table 4, which were defined by taking the pitch ranges of instruments into account. For example, 48 different combinations were made for quartet music.

5.2. Experiment 1: leave-one-out

The experiment was conducted using the leave-one-out cross-validation method. When evaluating a musical piece, a mixed-sound template was constructed using the remaining two pieces. Because we evaluated three pieces, we constructed three different mixed-sound templates by dropping the piece used for testing. The mixed-sound templates were constructed from audio signals of solo and duo music (S+D) and solo, duo, and trio music (S + D + T). For comparison, we also constructed a template, called a solo-sound template, only from solo musical sounds. The number of notes in each template is listed in Table 5. To evaluate the effectiveness of F0-dependent multivariate normal distributions and using musical context, we tested both cases with and without each technique. We fed the correct data on the F0s, onset times, and durations of all notes because our focus was on the performance of the instrument identification method alone.

Table 5: Number of notes in mixed-sound templates (Type I). Templates of Types II and III have about 1/2 and 1/3–1/4 times the notes of Type I (details are omitted due to a lack of space). S + D and S + D + T stand for the templates constructed from audio signals of solo and duo music, and from those of solo, duo, and trio music, respectively.

Number	Name	S + D	S + D + T	Subset*
No. 13	PF	31,334	83,491	24,784
	CG	23,446	56,184	10,718
	VN	14,760	47,087	9,804
	CL	7,332	20,031	4,888
	FL	4,581	16,732	3,043
No. 16	PF	26,738	71,203	21,104
	CG	19,760	46,924	8,893
	VN	12,342	39,461	8,230
	CL	5,916	16,043	3,944
	FL	3,970	14,287	2,632
No. 17	PF	23,836	63,932	18,880
	CG	17,618	42,552	8,053
	VN	11,706	36,984	7,806
	CL	5,928	16,208	3,952
	FL	3,613	13,059	2,407

*Template used in Experiment III.

The results are shown in Table 6. Each number in the table is the average of the recognition rates for the three pieces. Using the DAMS method, the F0-dependent multivariate normal distribution, and the musical context, we improved the recognition rates from 50.9 to 84.1% for duo, from 46.1 to 77.6% for trio, and from 43.1 to 72.3% for quartet music on average.

We confirmed the effect of each of the DAMS method (mixed-sound template), the F0-dependent multivariate normal distribution, and the musical context using

Table 6: Results of Experiment 1. ◯: used, ×: not used; bold font denotes recognition rates of higher than 75%.

Template		Solo sound				S + D				S + D + T			
F0-dependent		×	×	◯	◯	×	×	◯	◯	×	×	◯	◯
Context		×	◯	×	◯	×	◯	×	◯	×	◯	×	◯
Duo	PF	53.7%	63.0%	70.7%	**84.7%**	61.5%	63.8%	69.8%	**78.9%**	69.1%	70.8%	71.0%	**82.7%**
	CG	46.0%	44.6%	50.8%	42.8%	50.9%	67.5%	70.2%	**85.1%**	44.0%	57.7%	71.0%	**82.9%**
	VN	63.7%	**81.3%**	63.1%	**75.6%**	68.1%	**85.5%**	70.6%	**87.7%**	65.4%	**84.2%**	67.7%	**88.1%**
	CL	62.9%	70.3%	53.4%	56.1%	**81.8%**	**92.1%**	**81.9%**	**89.9%**	**84.6%**	**95.1%**	**82.9%**	**92.6%**
	FL	28.1%	33.5%	29.1%	38.7%	67.6%	**84.9%**	67.6%	**78.8%**	56.8%	70.5%	61.5%	74.3%
	Av.	50.9%	58.5%	53.4%	59.6%	66.0%	**78.8%**	72.0%	**84.1%**	64.0%	**75.7%**	70.8%	**84.1%**
Trio	PF	42.8%	49.3%	63.0%	**75.4%**	44.1%	43.8%	57.0%	61.4%	52.4%	53.6%	61.5%	68.3%
	CG	39.8%	39.1%	40.0%	31.7%	52.1%	66.8%	68.3%	**82.0%**	47.2%	62.8%	68.3%	**82.8%**
	VN	61.4%	**76.8%**	62.2%	72.5%	67.0%	**81.8%**	70.8%	**83.5%**	60.5%	**80.6%**	68.1%	**82.5%**
	CL	53.4%	55.7%	46.0%	43.9%	69.5%	**77.1%**	72.2%	**78.3%**	71.0%	**82.8%**	**76.2%**	**82.8%**
	FL	33.0%	42.6%	36.7%	46.5%	68.4%	**77.9%**	68.1%	**76.9%**	59.1%	69.3%	64.0%	71.5%
	Av.	46.1%	52.7%	49.6%	54.0%	60.2%	69.5%	67.3%	**76.4%**	58.0%	69.8%	67.6%	**77.6%**
Quartet	PF	38.9%	46.0%	54.2%	64.9%	38.7%	38.6%	50.3%	53.1%	46.1%	46.6%	53.3%	57.2%
	CG	34.3%	33.2%	35.3%	29.1%	51.2%	62.7%	64.8%	**75.3%**	51.2%	64.5%	65.0%	**79.1%**
	VN	60.2%	74.3%	62.8%	73.1%	70.0%	**81.2%**	72.7%	**82.3%**	67.4%	**79.2%**	69.7%	**79.9%**
	CL	45.8%	44.8%	39.5%	35.8%	62.6%	66.8%	65.4%	69.3%	68.6%	74.4%	70.9%	74.5%
	FL	36.0%	50.8%	40.8%	52.0%	69.8%	**76.1%**	69.9%	**76.2%**	61.7%	69.4%	64.5%	70.9%
	Av.	43.1%	49.8%	46.5%	51.0%	58.5%	65.1%	64.6%	71.2%	59.0%	66.8%	64.7%	72.3%

Table 7: Results of McNemar's test for quartet music (Corr. = correct, Inc. = incorrect).

(a) Template comparison (with both F0-dependent and context)

		Solo sound	
		Corr.	Inc.
S + D	Corr.	233	133
	Inc.	25	109

$\chi_0^2 = (133 - 25)^2/(133 + 25) = 73.82$

		Solo sound	
		Corr.	Inc.
S + D + T	Corr.	224	148
	Inc.	34	94

$\chi_0^2 = (148 - 34)^2/(148 + 34) = 71.40$

		S + D	
		Corr.	Inc.
S + D + T	Corr.	347	25
	Inc.	19	109

$\chi_0^2 = (25 - 19)^2/(25 + 19) = 1.5$

(b) With versus without F0-dependent (with S + D + T template and context)

		w/o F0-dpt.	
		Corr.	Inc.
w/ F0-dpt.	Corr.	314	58
	Inc.	25	103

$\chi_0^2 = (58 - 25)^2/(58 + 25) = 13.12$

(c) With versus without context (with S + D + T template and F0-dependent model)

		w/o Context	
		Corr.	Inc.
w/ Context	Corr.	308	64
	Inc.	27	101

$\chi_0^2 = (64 - 27)^2/(64 + 27) = 15.04$

McNemar's test. McNemar's test is usable for testing whether the proportions of A-labeled ("correct" in this case) data to B-labeled ("incorrect") data under two different conditions are significantly different. Because the numbers of notes are different among instruments, we sampled 100 notes at random for each instrument to avoid the bias. The results of McNemar's test for the quartet music are listed in Table 7 (those for the trio and duo music are omitted but are basically the same as those for the quartet), where the χ_0^2 are test statistics. Because the criterion region at $\alpha = 0.001$ (which is the level of significance) is $(10.83, +\infty)$, the differences except for S + D versus S + D + T are significant at $\alpha = 0.001$.

Other observations are summarized as follows.

(i) The results of the S+D and S+D+T templates were not significantly different even if the test data were from quartet music. This means that constructing a template from polyphonic sounds is effective even if the sounds used for the template construction do not have the same complexity as the piece to be identified.

(ii) For PF and CG, the F0-dependent multivariate normal distribution was particularly effective. This is because these instruments have large pitch dependencies due to their wide pitch ranges.

(iii) Using musical context improved recognition rates, on average, by approximately 10%. This is because, in the musical pieces used in our experiments, pitches in the melodies of simultaneous voices rarely crossed.

(iv) When the solo-sound template was used, the use of musical context lowered recognition rates, especially for CL. Because our method of using musical context calculates the a priori probability of each note on the basis of the a posteriori probabilities of temporally neighboring notes, it requires an accuracy sufficient for precalculating the a posteriori probabilities of the temporally neighboring notes. The lowered recognition rates are because of the insufficient accuracy of this precalculation. In fact, this phenomenon did not occur when the mixed-sound templates, which improved the accuracies of the precalculations, were used. Therefore, musical context should be used together with some technique of improving the pre-calculation accuracies, such as a mixed-sound template.

(v) The recognition rate for PF was not high enough in some cases. This is because the timbre of PF is similar to that of CG. In fact, even humans had difficulty distinguishing them in listening tests of sounds resynthesized from harmonic structures extracted from PF and CG tones.

5.3. Experiment 2: template construction from only one piece

Next, to compare template construction from only one piece with that from two pieces (i.e., leave-one-out), we conducted an experiment on template construction from only one piece. The results are shown in Table 8. Even when using a template made from only one piece, we obtained comparatively high recognition rates for CG, VN, and CL. For FL, the results of constructing a template from only one piece were not high (e.g., 30–40%), but those from two pieces were close to the results of the case where the same piece was used for both template construction and testing. This means that a variety of influences of sounds overlapping was trained from only two pieces.

5.4. Experiment 3: insufficient instrument combinations

We investigated the relationship between the coverage of instrument combinations in a template and the recognition rate. When a template that does not cover instrument combinations is used, the recognition rate might decrease. If this decrease is large, the number of target instruments of the template will be difficult to increase because $O(m^n)$ data are needed for a full-combination template, where m and n are the number of target instruments and simultaneous voices. The purpose of this experiment is to check whether such a decrease occurs in the use of a reduced-combination template. As the reduced-combination template, we used one that contains the combinations listed in Table 9 only. These combinations were chosen so that the order of the combinations was $O(m)$. Similarly to Experiment 1, we used the leave-one-out cross-validation method. As we can see from Table 10, we did not find significant differences between using the full instrument combinations and the reduced combinations. This was confirmed, as shown in Table 11, through McNemar's test, similarly to Experiment 1. Therefore, we expect that the number of target instruments can be increased without the problem of combinational explosion.

Table 8: Template construction from only one piece (Experiment 2). Quartet only due to lack of space (unit: %).

	S + D				S + D + T			
	13	16	17	⋆	13	16	17	⋆
PF	(57.8)	32.3	38.4	36.6	(67.2)	33.2	45.1	39.7
CG	(73.3)	**78.1**	**76.2**	**76.7**	(**76.8**)	**84.3**	**80.3**	**82.1**
13 VN	(**89.5**)	59.4	**87.5**	**86.2**	(**87.2**)	58.0	**85.2**	**83.1**
CL	(68.5)	70.8	62.2	73.8	(72.3)	72.3	68.6	**75.9**
FL	(**85.5**)	40.2	74.9	**82.7**	(**86.0**)	38.9	68.8	**80.8**
PF	74.1	(64.8)	61.1	71.2	**79.6**	(67.1)	73.0	**78.3**
CG	**79.2**	(**77.9**)	**78.9**	74.3	70.4	(**82.6**)	74.0	**75.2**
16 VN	**89.2**	(**85.5**)	**87.0**	**87.0**	**86.0**	(**83.5**)	**84.7**	**85.0**
CL	68.1	(**78.9**)	68.9	**76.1**	72.4	(**82.8**)	**76.3**	**82.1**
FL	**82.0**	(**75.9**)	72.5	**77.3**	**77.9**	(72.3)	35.7	69.2
PF	53.0	39.4	(51.2)	51.6	52.2	40.6	(55.7)	53.7
CG	73.7	69.0	(**75.8**)	**75.0**	**76.0**	74.3	(**78.4**)	**80.0**
17 VN	**79.5**	61.2	(**78.3**)	73.6	**77.4**	58.0	(**78.7**)	71.7
CL	51.3	60.5	(57.1)	57.9	61.1	62.6	(66.9)	65.4
FL	65.0	35.0	(73.1)	68.7	58.6	34.7	(70.9)	62.6

* Leave-one-out. Numbers in left column denote piece numbers for test, those in top row denote piece numbers for template construction.

Table 9: Instrument combinations in Experiment 3.

Solo	PF, CG, VN, CL, FL
Duo	PF–PF, CG–CG, VN–PF, CL–PF, FL–PF
Trio	Not used
Quartet	Not used

Table 10: Comparison of templates whose instrument combinations were reduced (subset) and not reduced (full set).

		Subset	Full set
Duo	PF	**85.4%**	**78.9%**
	CG	70.8%	**85.1%**
	VN	**88.2%**	**87.7%**
	CL	**90.4%**	**89.9%**
	FL	**79.7%**	**78.8%**
	Average	**82.9%**	**84.1%**
Trio	PF	73.9%	61.4%
	CG	62.0%	**82.0%**
	VN	85.7%	**83.5%**
	CL	**79.7%**	**78.3%**
	FL	**76.5%**	**76.9%**
	Average	**75.6%**	**76.4%**
Quartet	PF	68.9%	53.1%
	CG	52.4%	**75.3%**
	VN	**85.0%**	**82.3%**
	CL	71.1%	69.3%
	FL	74.5%	**76.2%**
	Average	70.4%	71.2%

Table 11: Results of McNemar's test for full-set and subset templates $\chi_0^2 = (25 - 19)^2/(25 + 19) = 1.5$.

		Subset	
		Corr.	Inc.
Full set	Corr.	341	25
	Inc.	19	115

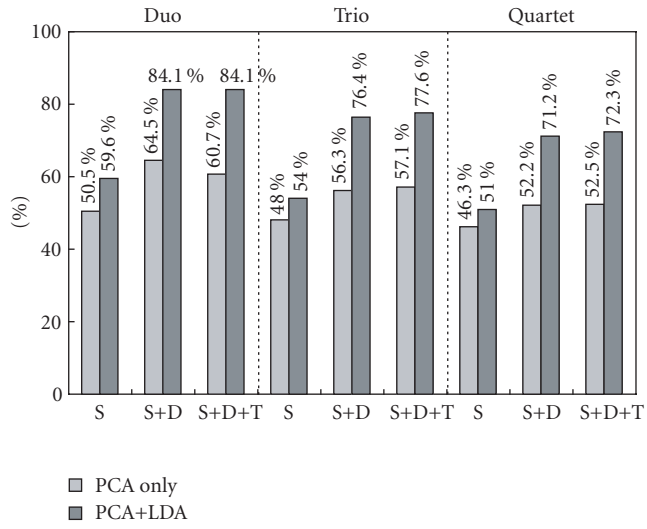

Figure 6: Comparison between using both PCA and LDA with using only PCA (Experiment 4). "Duo," "trio," and "quartet" represent pieces for test (identification). "S," "S+D," and "S+D+T" represent types of templates.

Table 12: ANOVA SS = sum of squares, DF = degrees of freedom, DR = dimensionality reduction.

Src. of var.	SS	DF	F value	P value
DR	0.336	1	102.08	1.806×10^{-13}
Template	0.302	2	45.75	7.57×10^{-12}
Interaction	0.057	2	8.75	5.73×10^{-4}
Residual	0.158	48	—	—
Total	0.855	53	—	—

5.5. Experiment 4: effectiveness of LDA

Finally, we compared the dimensionality reduction using both PCA and LDA with that using only PCA to evaluate the effectiveness of LDA. The experimental method was leave-one-out cross-validation. The results are shown in Figure 6. The difference between the recognition rates of the solo-sound template and the S + D or S + D + T template was 20–24% using PCA + LDA and 6–14% using PCA only. These results mean that LDA (or DAMS) successfully obtained a subspace where the influence of the overlapping of sounds of multiple instruments was minimal by minimizing the ratio of the within-class variance to the between-class variance. Under all conditions, using LDA was superior to not using LDA.

We confirmed that combining LDA and the mixed-sound template is effective using two-way factorial analysis of variance (ANOVA) where the two factors are dimensionality reduction methods (PCA only and PCA + LDA) and templates (S, S + D, and S + D + T). Because we tested each condition using duo, trio, and quartet versions of Piece Nos. 13, 16, and 17, there are nine results for each cell of the two-factor matrix. The table of ANOVA is given in Table 12. From the table, we can see that the interaction effect as well as the effects of dimensionality reduction methods and templates are significant at $\alpha = 0.001$. This result means that mixed-sound templates are particularly effective when combined with LDA.

5.6. Application to XML annotation

In this section, we show an example of XML annotation of musical audio signals using our instrument identification method. We used a simplified version of MusicXML instead of the original MusicXML format because our method does not include rhythm recognition and hence cannot determine note values or measures. The document-type definition (DTD) of our simplified MusicXML is shown in Figure 7. The main differences between it and the original one are that elements related to notation, which cannot be estimated from audio signals, are reduced and that time is represented in seconds. The result of XML annotation of a piece of polyphonic music is shown in Figure 8. By using our instrument identification method, we classified notes according to part and described the instrument for each part.

```
<!ENTITY % score-header
      "(work?, movement-number?, movement-title?,
        identification?, defaults?, credit*,
        part-list)">

<!ELEMENT part-list (score-part+)>
<!ELEMENT score-part
      (identification?, part-name,
      part-abbreviation?, score-instrument)>
<!ATTLIST score-part
      id ID #REQUIRED
>
<!ELEMENT score-instrument
      (instrument-name, instrument-abbreviation?)>
<!ELEMENT instrument-name (#PCDATA)>
<!ELEMENT instrument-abbreviation (#PCDATA)>

<!ELEMENT score-partwise-simple>
      (%score-header;, part+)>
<!ATTLIST score-partwise-simple
      version CDATA "1.0"
>
<!ELEMENT part (note+)>
<!ATTLIST part
      id IDREF #REQUIRED
>

<!ELEMENT note (pitch, onset, offset)>
<!ELEMENT pitch (step, alter?, octave)>
<!ELEMENT step (#PCDATA)>
<!ELEMENT alter (#PCDATA)>
<!ELEMENT octave (#PCDATA)>
<!ELEMENT onset (#PCDATA)>
<!ATTLIST onset
      unit CDATA "sec"
>
<!ELEMENT offset (#PCDATA)>
<!ATTLIST offset
      unit CDATA "sec"
>
```

Figure 7: DTD of our simplified MusicXML.

5.7. Discussion

We achieved average recognition rates of 84.1% for duo, 77.6% for trio, and 72.3% for quartet music chosen from five different instruments. We think that this performance is state of the art, but we cannot directly compare these rates with experimental results published by other researchers because different researchers used different test data in general. We also find the following two limitations in our evaluation:

(1) the correct F0s are given;

(2) nonrealistic music (i.e., music synthesized by mixing isolated monophonic sound samples) is used.

First, in most existing studies, including ours, the methods were tested under the condition that the correct F0s are manually fed [10, 13]. This is because the multiple

```
<?xml version="1.0" encoding="UTF-8"
    standalone="no" ? >
<!DOCTYPE score-partwise-simple SYSTEM
     "partwisesimple.dtd">
<score-partwise-simple>
   <part-list>
    <score-part id="P1">
     <part-name>Part 1</part-name>
     <score-instrument>Piano</score-instrument>
    </score-part>
    <score-part id="P2">
     <part-name>Part 3</part-name>
     <score-instrument>Violin</score-instrument>
    </score-part>
            . . . . . .
   </part-list>
   <part id="P1">
    <note>
     <pitch>
      <step>G</step>
      <alter>+1</alter>
      <octave>3</octave>
     </pitch>
     <onset>1.0</onset>
     <offset>2.0</offset>
    </note>
    <note>
     <pitch>
      <step>G</step>
       <octave>3</octave>
     </pitch>
     <onset>2.0</onset>
     <offset>2.5</offset>
    </note>
    <note>
     <pitch>
       <step>D</step>
       <octave>4</octave>
      </pitch>
     <onset>2.5</onset>
     <offset>3.0</offset>
    </note>
            . . . . . .
   </part>
   <part id = "P2">
    <note>
     <pitch>
      <step>D</step>
      <alter> +1 < /alter>
      <octave> 4 < /octave>
     </pitch>
     <onset>1.5</onset>
     <offset> 2.488541 < /offset>
    </note>
    <note>
     <pitch>
      <step>C</step>
      <alter> +1 < /alter>
      <octave> 4 < /octave>
     </pitch>
     <onset> 3.0 < /onset>
     <offset> 3.5 < /offset>
     </note>
            . . . . . .
   </part>
            . . . . . .
</score-partwise-simple>
```

Figure 8: Example of MusicXML annotation.

F0-estimation for a sound mixture is still a challenging problem, and the studies aimed at evaluating the performance of only their instrument identification methods. If the estimated F0s are used instead of the manually given correct F0s, the performance of instrument identification will decrease. In fact, Kinoshita et al. [11] reported that given random note patterns taken from three different instruments, the instrument identification performance was around 72–81% for correct F0s but decreased to around 66–75% for estimated F0s. Because multiple-F0 estimation has actively been studied [8, 31, 32], we plan to integrate and evaluate our instrument identification method with such a multiple-F0 estimation method in the future.

Second, most existing studies, including ours, used non-realistic music as test samples. For example, Kashino et al. [7] and Kinoshita et al. [11] tested their methods on polyphonic musical audio signals that were synthesized by mixing isolated monophonic sounds of every target instrument on an MIDI sampler. This was because information on the instrument for every note that was used as correct references in the evaluation was then easy to prepare. Strictly speaking, however, the acoustical characteristics of real music are different from those of such synthesized music. The performance of our method would decrease for real music because legato play sometimes causes overlapping successive notes with unclear onsets in a melody and because sound mixtures often involve reverberations. We plan to manually annotate the correct F0 information for real music and evaluate our method after integrating it with a multiple-F0 estimation method as mentioned above.

6. CONCLUSION

We have provided a new solution to an important problem of instrument identification in polyphonic music: the overlapping of partials (harmonic components). Our solution is to weight features based on their robustness to overlapping by collecting training data extracted from polyphonic sounds and applying LDA to them. Although the approach of collecting training data from polyphonic sounds is simple, no previous studies have attempted it. One possible reason may be that a tremendously large amount of data is required to prepare a thorough training data set containing all possible sound combinations. From our experiments, however, we found that a thorough training data set is not necessary and that a data set extracted from a few musical pieces is sufficient to improve the robustness of instrument identification in polyphonic music. Furthermore, we improved the performance of the instrument identification using musical context. Our method made it possible to avoid musically unnatural errors by taking the temporal continuity of melodies into consideration.

Because the F0 and onset time of each note were given in our experiments to check the performance of only the instrument identification, we plan to complete MusicXML annotation by integrating our method with a musical note estimation method. Our future work will also include the use of the description of musical instrument names identified using our method to build a music information retrieval system that enables users to search for polyphonic musical pieces by giving a query including musical instrument names.

REFERENCES

[1] M. Good, "MusicXML: an internet-friendly format for sheet music," in *Proceedings of the XML Conference & Exposition*, Orlando, Fla, USA, December 2001.

[2] P. Bellini and P. Nesi, "WEDELMUSIC format: an XML music notation format for emerging applications," in *Proceedings of the International Conference of Web Delivering of Music*, pp. 79–86, Florence, Italy, November 2001.

[3] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction of MPEG-7*, John Wiley & Sons, New York, NY, USA, 2002.

[4] T. Nagatsuka, N. Saiwaki, H. Katayose, and S. Inokuchi, "Automatic transcription system for ensemble music," in *Proceedings of the International Symposium of Musical Acoustics (ISMA '92)*, pp. 79–82, Tokyo, Japan, 1992.

[5] G. J. Brown and M. Cooke, "Perceptual grouping of musical sounds: a computational model," *Journal of New Music Research*, vol. 23, pp. 107–132, 1994.

[6] K. D. Martin, "Automatic transcription of simple polyphonic music," in *Proceedings of 3rd Joint meeting of the Acoustical Society of America and Japan*, Honolulu, Hawaii, USA, December 1996.

[7] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka, "Application of the Bayesian probability network to music scene analysis," in *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. G. Okuno, Eds., pp. 115–137, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1998.

[8] A. Klapuri, T. Virtanen, A. Eronen, and J. Seppanen, "Automatic transcription of musical recordings," in *Proceedings of Workshop on Consistent & Reliable Acoustic Cues (CRAC '01)*, Aalborg, Denmark, September 2001.

[9] Y. Sakuraba, T. Kitahara, and H. G. Okuno, "Comparing features for forming music streams in automatic music transcription," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 4, pp. 273–276, Montreal, Quebec, Canada, May 2004.

[10] K. Kashino and H. Murase, "Sound source identification system for ensemble music based on template adaptation and music stream extraction," *Speech Communication*, vol. 27, no. 3, pp. 337–349, 1999.

[11] T. Kinoshita, S. Sakai, and H. Tanaka, "Musical sound source identification based on frequency component adaptation," in *Proceedings of IJCAI Workshop on Computational Auditory Scene Analysis (IJCAI-CASA '99)*, pp. 18–24, Stockholm, Sweden, July-August 1999.

[12] J. Eggink and G. J. Brown, "A missing feature approach to instrument identification in polyphonic music," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 5, pp. 553–556, Hong Kong, April 2003.

[13] J. Eggink and G. J. Brown, "Application of missing feature theory to the recognition of musical instruments in polyphonic audio," in *Proceedings of International Symposium on Music Information Retrieval (ISMIR '03)*, Baltimore, Md, USA, October 2003.

[14] K. D. Martin, *Sound-source recognition: a theory and computational model*, Ph.D. thesis, MIT, Cambridge, Mass, USA, 1999.

[15] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," in

Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '00), vol. 2, pp. 753–756, Istanbul, Turkey, June 2000.

[16] A. Fraser and I. Fujinaga, "Toward real-time recognition of acoustic musical instruments," in *Proceedings of International Computer Music Conference (ICMC '99)*, pp. 175–177, Beijing, China, October 1999.

[17] I. Fujinaga and K. MacMillan, "Realtime recognition of orchestral instruments," in *Proceedings of International Computer Music Conference (ICMC '00)*, pp. 141–143, Berlin, Germany, August 2000.

[18] G. Agostini, M. Longari, and E. Pollastri, "Musical instrument timbres classification with spectral features," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 1, pp. 5–14, 2003.

[19] T. Kitahara, M. Goto, and H. G. Okuno, "Pitch-dependent identification of musical instrument sounds," *Applied Intelligence*, vol. 23, no. 3, pp. 267–275, 2005.

[20] J. Marques and P. J. Moreno, "A study of musical instrument classification using Gaussian mixture models and support vector machines," CRL Technical Report Series CRL/4, Cambridge Research Laboratory, Cambridge, Mass, USA, 1999.

[21] J. C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1933–1941, 1999.

[22] A. G. Krishna and T. V. Sreenivas, "Music instrument recognition: from isolated notes to solo phrases," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 4, pp. 265–268, Montreal, Quebec, Canada, May 2004.

[23] B. Kostek, "Musical instrument classification and duet analysis employing music information retrieval techniques," *Proceedings of the IEEE*, vol. 92, no. 4, pp. 712–729, 2004.

[24] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, vol. 3, pp. 245–248, Philadelphia, Pa, USA, March 2005.

[25] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument identification in polyphonic music: feature weighting with mixed sounds, pitch-dependent timbre modeling, and use of musical context," in *Proceedings of 6th International Conference on Music Information Retrieval (ISMIR '05)*, pp. 558–563, London, UK, September 2005.

[26] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, Cambridge, Mass, USA, 1990.

[27] D. Huron, "Tone and voice: a derivation of the rules of voice-leading from perceptual principles," *Music Perception*, vol. 19, no. 1, pp. 1–64, 2001.

[28] H. Kameoka, T. Nishimoto, and S. Sagayama, "Harmonic-temporal-structured clustering via deterministic annealing EM algorithm for audio feature extraction," in *Proceedings of 6th International Conference on Music Information Retrieval (ISMIR '05)*, pp. 115–122, London, UK, September 2005.

[29] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: popular, classical, and jazz music databases," in *Proceedings of 3rd International Conference on Music Information Retrieval (ISMIR '02)*, pp. 287–288, Paris, France, October 2002.

[30] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: music genre database and musical instrument sound database," in *Proceedings of 4th International Conference on Music Information Retrieval (ISMIR '03)*, pp. 229–230, Washington, DC, USA, October 2003.

[31] M. Goto, "A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[32] H. Kameoka, T. Nishimoto, and S. Sagayama, "Audio stream segregation of multi-pitch music signal based on time-space clustering using Gaussian Kernel 2-dimensional model," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, vol. 3, pp. 5–8, Philadelphia, Pa, USA, March 2005.

Tetsuro Kitahara received the B.S. degree from Tokyo University of Science in 2002 and the M.S. degree from Kyoto University in 2004. He is currently a Ph.D. Course Student at Graduate School of Informatics, Kyoto University. Since 2005, he has been a Research Fellow of the Japan Society for the Promotion of Science. His research interests include music informatics. He received five awards, including TELECOM System Technology Award for Student in 2004 and IPSJ 67th National Convention Best Paper Award for Young Researcher in 2005. He is a Student Member of IPSJ, IEICE, JSAI, ASJ, JSMPC, and IEEE.

Masataka Goto received the Doctor of Engineering degree in electronics, information, and communication engineering from Waseda University, Japan, in 1998. He then joined the Electrotechnical Laboratory (ETL), which was reorganized as the National Institute of Advanced Industrial Science and Technology (AIST) in 2001, where he has been a Senior Research Scientist since 2005. He served concurrently as a Researcher in Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Corporation (JST) from 2000 to 2003, and as an Associate Professor of the Department of Intelligent Interaction Technologies, Graduate School of Systems and Information Engineering, University of Tsukuba, since 2005. His research interests include music information processing and spoken language processing. He has received 18 awards, including the Information Processing Society of Japan (IPSJ) Best Paper Award and IPSJ Yamashita SIG Research Awards (special interest group on music and computer, and spoken language processing) from the IPSJ, the Awaya Prize for Outstanding Presentation and Award for Outstanding Poster Presentation from the Acoustical Society of Japan (ASJ), Award for Best Presentation from the Japanese Society for Music Perception and Cognition (JSMPC), WISS 2000 Best Paper Award and Best Presentation Award, and Interaction 2003 Best Paper Award.

Kazunori Komatani is an Assistant Professor at the Graduate School of Informatics, Kyoto University, Japan. He received a B.S. degree in 1998, an M.S. degree in Informatics in 2000, and a Ph.D. degree in 2002, all from Kyoto University. He received the 2002 FIT Young Researcher Award and the 2004 IPSJ Yamashita SIG Research Award, both from the Information Processing Society of Japan.

Tetsuya Ogata received the B.S., M.S., and Ph.D. degrees of Engineering in mechanical engineering in 1993, 1995, and 2000, respectively, from Waseda University. From 1999 to 2001, he was a Research Associate in Waseda University. From 2001 to 2003, he was a Research Scientist in Brain Science Institute, RIKEN. Since 2003, he has been a faculty member in Graduate School of Informatics, Kyoto University, where he is currently an Associate Professor. Since 2005, he has been a Visiting Associate Professor of the Humanoid Robotics Institute of Waseda University. His research interests include human-robot interaction, dynamics of human-robot mutual adaptation, and intersensory translation in robot system. He received the JSME Medal for Outstanding Paper from the Japan Society of Mechanical Engineers in 2000.

Hiroshi G. Okuno received B.A. and Ph.D. degrees from the University of Tokyo in 1972 and 1996, respectively. He worked for Nippon Telegraph and Telephone, JST Kitano Symbiotic Systems Project, and Tokyo University of Science. He is currently a Professor of Department of Intelligence Technology and Science, Graduate School of Informatics, Kyoto University. He was a Visiting Scholar at Stanford University, and Visiting Associate Professor at the University of Tokyo. He has done research in programming languages, parallel processing, and reasoning mechanism in AI, and he is currently engaged in computational auditory scene analysis, music scene analysis, and robot audition. He received various awards including the 1990 Best Paper Award of JSAI, the Best Paper Award of IEA/AIE-2001 and 2005, and IEEE/RSJ Nakamura Award for IROS-2001 Best Paper Nomination Finalist. He was also awarded 2003 Funai Information Science Achievement Award. He edited with David Rosenthal "*Computational Auditory Scene Analysis*" from Lawrence Erlbaum Associates in 1998 and with Taiichi Yuasa "*Advanced Lisp Technology*" from Taylor and Francis Inc. in 2002. He is a Member of the IPSJ, JSAI, JSSST, JCCS, RSJ, ACM, IEEE, AAAI, ASA, and ISCA.