Musical Instrument Recognizer "Instrogram" and Its Application to Music Retrieval based on Instrumentation Similarity

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
{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

Instrumentation is an important cue in retrieving musical content. Conventional methods for instrument recognition performing notewise require accurate estimation of the onset time and fundamental frequency (F0) for each note, which is not easy in polyphonic music. This paper presents a non-notewise method for instrument recognition in polyphonic musical audio signals. Instead of such notewise estimation, our method calculates the temporal trajectory of instrument existence probabilities for every F0 and visualizes it as a spectrogram-like graphical representation, called an instrogram. This method can avoid the influence by errors of onset detection and F0 estimation because it does not use them. We also present methods for MPEG-7-based instrument annotation and music information retrieval based on the similarity between instrograms. Experimental results with realistic music show the average accuracy of 76.2% for the instrument annotation and that the instrogram-based similarity measure represents the actual instrumentation similarity better than an MFCC-based one.

1. Introduction

The aim of our study is to enable users to retrieve musical pieces based on their instrumentation. When searching musical pieces, the type of instruments that are used is a important cue. In fact, the names of some musical forms are based on instrument names, such as "piano sonata" and "string quartet." There are two strategies for instrumentation-based music information retrieval (MIR). The first one allows users to specify musical instruments on which pieces that they want are played. This strategy is useful because specifying instruments does not require special knowledge unlike other musical elements such as chord progressions. The other one is the so-called Query-by-Example. In this strategy, once users specify musical pieces that they like, a system searches pieces that have similar instrumentation to the specified ones. This strategy is also useful particularly when automatically generating playlists for background music.

The key technology for achieving the above-mentioned MIR is to recognize musical instruments from audio signals. Whereas musical instrument recognition studies mainly dealt with solo musical sounds in the 1990s (e.g., [13]), the number of studies dealing with polyphonic music has been increasing in recent years. Kashino *et al.* [10] developed a computational music scene analysis architecture called OPTIMA, which recognizes musical notes and instruments based on the Bayesian probability network. They subsequently proposed a technique that identifies an instrument playing each musical note based on template matching with template adaptation [9]. Kinoshita *et al.* [11] improved the robustness of OPTIMA to the overlapping of frequency components, which occurs when multiple instruments play simultaneously, based on feature adaptation. Eggink *et al.* [2] tackled this overlapping problem with the missing feature theory. They subsequently dealt with the problem of identifying only the instrument playing the main melody on the assumption that the main melody's partials suffer less from other sounds occurring simultaneously [3]. Vincent *et al.* [17] formulated both music transcription and instrument identification as a single optimization based on independent subspace analysis. Essid *et al.* [4] achieved F0-estimation-less instrument recognition based on a priori knowledge about instrumentation of ensembles. Kitahara *et al.* [12] proposed an instrument identification method based on a mixed-sound template and musical context.

The common feature in most of these studies is that instrument identification is performed for each frame or each note. In the former case [2, 3], it is difficult to obtain a reasonable accuracy because temporal variations in spectra are important characteristics of musical instrument sounds. In the latter case [10, 9, 11, 12], the identification system has to first estimate the onset time and fundamental frequency (F0) of musical notes and then extract the harmonic struc-

Proceedings of the Eighth IEEE International Symposium on Multimedia (ISM'06)
0-7695-2746-9/06 $20.00 © 2006 IEEE

ture of each note based on the estimated onset time and F0. Therefore, the instrument identification suffers from errors of onset detection and F0 estimation. In the experiments reported in [9] and [12], in fact, correct data of the onset times and F0s were manually fed.

To cope with this vulnerability, we propose a new method that recognizes musical instruments in polyphonic musical audio signals without relying on onset detection nor F0 estimation. The key idea of this method is to visualize, as a spectrogram-like representation called an *instrogram*, the probability that the sound of each target instrument exists at each time and with each F0. Because this probability is calculated not for each note but for each point of the time-frequency plane, it can be calculated without using onset detection nor F0 estimation.

In addition, we provide methods for applying instrograms to MPEG-7 annotation and MIR based on instrumentation similarity. Although annotating musical content in a universal framework such as MPEG-7 is an important task for achieving sophisticated MIR, attempts on music annotation are less than those for visual media. Here, we introduce new MPEG-7 tags for describing instrograms. To achieve MIR based on instrumentation similarity, we introduce a new similarity measure between instrograms, which is calculated with the dynamic time warping (DTW). We have achieved a prototype system of MIR based on this similarity measure.

2. Instrogram

The instrogram is a spectrogram-like graphical representation of a musical audio signal, which is useful for finding which instruments are used in the signal. One image exists for each target instrument. Each image has horizontal and vertical axes representing time and frequency, and the intensity of the color of each point (t, f) shows the probability that the target instrument is used at time t and F0 f. An example is presented in Figure 1. This example is the result of analyzing an audio signal of "Auld Lang Syne" played on piano, violin, and flute. The target instruments of analysis were piano, violin, clarinet, and flute. If the instrogram is too detailed for some purposes, it can be simplified by dividing the whole frequency region into some subregions and by merging results within each subregion. A simplified version of Figure 1 is given in Figure 2. From the four images of the instrogram or from the simplified instrogram, we can see that this piece is played on flute, violin, and piano (no clarinet is played).

3. Algorithm for Calculating Instrogram

Let $\Omega = \{\omega_1, \cdots, \omega_m\}$ be the set of target instruments. Then, what needs to be solved is the calculation of the probability $p(\omega_i; t, f)$, called *instrument existence probability* (IEP), that a sound of the instrument ω_i with F0 of f exists at time t for every target instrument $\omega_i \in \Omega$. Here, we assume that more-than-one instruments are not played at the same time and with the same F0, that is, $\forall \omega_i, \omega_j \in \Omega$: $i \neq j \Longrightarrow p(\omega_i \cap \omega_j; t, f) = 0$, because separating simultaneous multiple sounds with the same F0 is too difficult with current technology. The IEPs satisfy $\sum_{\omega_i \in \Omega \cup \{\text{silence}\}} p(\omega_i; t, f) = 1$. By introducing the symbol "X", which stands for the existence of *some* instrument (i.e., $\mathrm{X} = \omega_1 \cup \cdots \cup \omega_m$), the IEP can be calculated as the product of two probabilities:

$$p(\omega_i; t, f) = p(\mathrm{X}; t, f)\, p(\omega_i | \mathrm{X}; t, f),$$

because $\omega_i \cap \mathrm{X} = \omega_i \cap (\omega_1 \cup \cdots \cup \omega_i \cup \cdots \omega_m) = \omega_i$. Above, $p(\mathrm{X}; t, f)$, called the *nonspecific instrument existence probability* (NIEP), is the probability that a sound of some instrument with F0 of f exists at time t, while $p(\omega_i | \mathrm{X}; t, f)$, called the *conditional instrument existence probability* (CIEP), is the conditional probability that, if a sound of some instrument with F0 of f exists at time t, the instrument is ω_i.

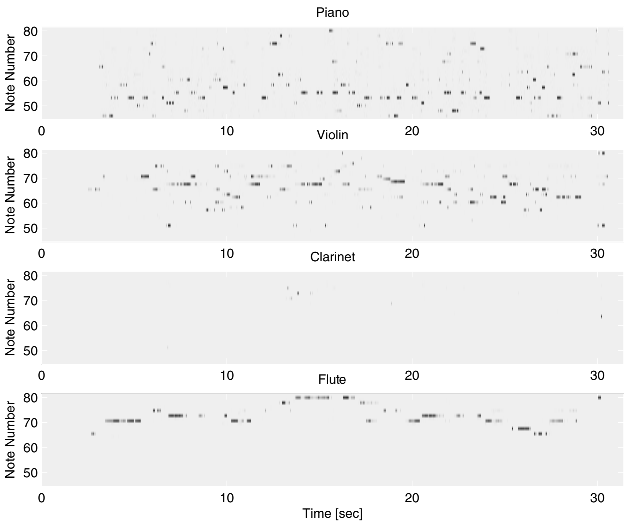


Figure 1. Example of instrograms. This is result of analyzing trio music, "Auld Lang Syne," played on piano, violin, and flute. Larger color version is available at: http://winnie.kuis.kyoto-u.ac.jp/˜kitahara/instrogram/ISM06/.

3.1. Overview

Figure 3 shows the overview of the algorithm for calculating an instrogram. Given an audio signal, the spectrogram is first calculated. In the current implementation, the short-time Fourier transform (STFT) shifted by 10 ms (441

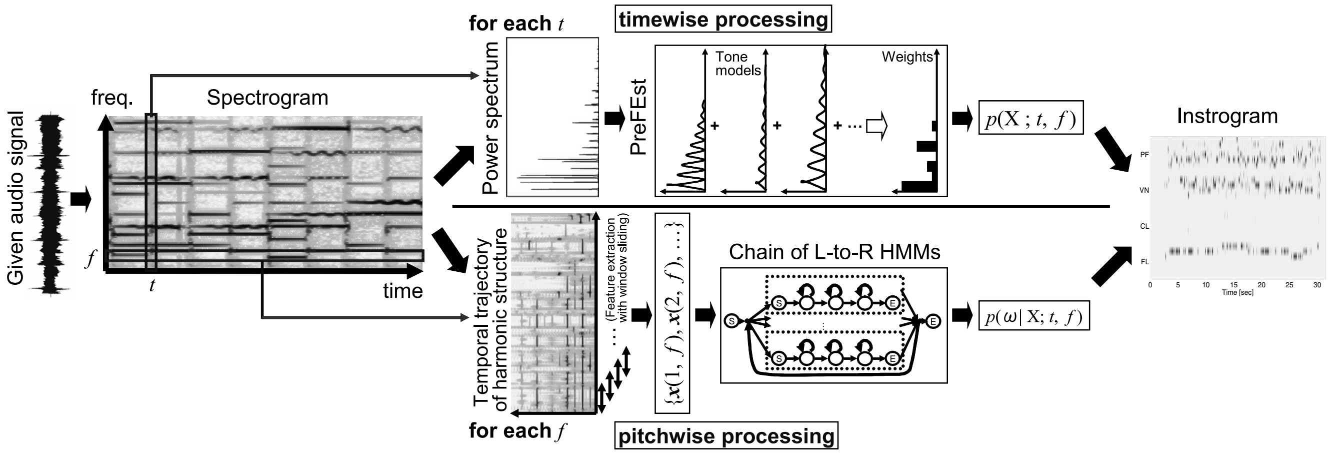


Figure 3. Overview of our technique for calculating instrogram.

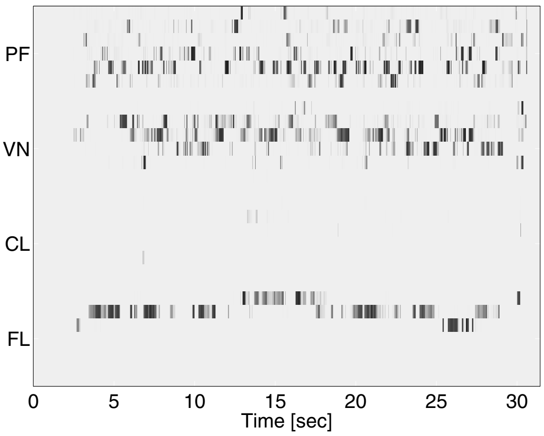


Figure 2. Simplified (summarized) instrogram of Figure 1.

points at 44.1 kHz sampling) with an 8192-point Hamming window is used. Next, the NIEPs and CIEPs are calculated. The NIEPs are calculated by analyzing the power spectrum at each frame (*timewise processing*) using PreFEst [6]. PreFEst models, at each frame, the spectrum of a signal containing multiple sounds as a weighted mixture of harmonic-structure tone models. The CIEPs are, on the other hand, calculated by analyzing the temporal trajectory of the harmonic structure with every F0 (*pitchwise processing*). The trajectory is analyzed with a framework similar to speech recognition, based on left-to-right hidden Markov models (HMMs) [15]. This HMM-based temporal modeling of harmonic structures is important because temporal variations in spectra characterize timbres well. This is the main difference from framewise recognition methodologies [2, 3]. Finally, the NIEPs and CIEPs are multiplied.

The advantage of this technique lies in that $p(\omega_i; t, f)$ can be estimated robustly because the two constituent probabilities are calculated independently and then integrated by multiplying them. In most previous studies, the onset time and F0 of each note were first estimated, and then the instrument of the note was identified by analyzing spectral components extracted based on the results of the note estimation. The upper limit of the instrument identification performance was therefore bound by the precedent note estimation, which is generally difficult and not robust for polyphonic music[1]. Unlike such a notewise symbolic approach, our non-symbolic and non-sequential approach is more robust for polyphonic music.

3.2. Nonspecific Instrument Existence Probability

By using the PreFEst, $p(\mathrm{X}; t, f)$ is estimated. The PreFEst models an observed power spectrum as a weighted mixture of tone models $p(x|F)$ of every possible F0 F. The tone model $p(x|F)$, where x is the log frequency, represents a typical spectrum of the harmonic structure, and the mixture density $p(x; \theta^{(t)})$ is defined as

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

where Fl and Fh denote the lower and upper limits of the possible F0 range, and $w^{(t)}(F)$ is the weight of a tone model $p(x|F)$ that satisfies $\int_{\mathrm{Fl}}^{\mathrm{Fh}} w^{(t)}(F) dF = 1$. If we can estimate the model parameter $\theta^{(t)}$ such that the observed spectrum is likely to have been generated from $p(x; \theta^{(t)})$, the spectrum can be considered to be decomposed into

[1] We tested the robustness to onset errors in identifying an instrument for every note using our previous method [12]. Giving onset times errors following the normal distribution with the standard deviation of e [s], we obtained the following results:

$e=0$	$e=0.05$	$e=0.10$	$e=0.15$	$e=0.20$
71.4%	69.2%	66.7%	62.5%	60.5%

Table 1. Overview of 28 features

Spectral features	
$\boxed{1}$	Spectral centroid
$\boxed{2}$	Relative power of fundamental component
$\boxed{3}$–$\boxed{10}$	Relative cumulative power from fundamental to i-th components ($i = 2, 3, \cdots, 9$)
$\boxed{11}$	Relative power in odd and even components
$\boxed{12}$–$\boxed{20}$	Number of components whose duration is p% longer than the longest duration ($p = 10, 20, \cdots, 90$)
Temporal features	
$\boxed{21}$	Gradient of straight line approximating power envelope
$\boxed{22}$–$\boxed{24}$	The temporal mean of differentials of power envelope from t to $t + iT/3$ ($i = 1, \cdots, 3$)
Modulation features	
$\boxed{25}$, $\boxed{26}$	Amplitude and Frequency of AM
$\boxed{27}$, $\boxed{28}$	Amplitude and Frequency of FM

harmonic-structure tone models, and $w^{(t)}(F)$ can be interpreted as the relative predominance of the tone model with F0 of F at time t. We therefore define the NIEP $p(\mathrm{X}; t, f)$ to be equal to $w^{(t)}(f)$. The weights can be estimated using the EM algorithm as described in [6].

3.3. Conditional Instrument Existence Probability

For every F0 f, the following steps are performed:

Step 1: Harmonic Structure Extraction

The temporal trajectory of the harmonic structure with F0 of f is extracted. This is represented as

$$H(t, f) = \{(F_i(t, f), A_i(t, f)) \mid i = 1, \cdots, h\},$$

where $F_i(t, f)$ and $A_i(t, f)$ are the frequency of amplitude of i-th partial of the sound with F0 of f at time t. $F_i(t, f)$ is basically equal to $i \cdot f$ but they are not exactly equal due to vibrato etc. We set h to 10.

Step 2: Feature Extraction

For every time t (every 10 ms in the implementation), we first excerpt a T-length bit of the harmonic-structure trajectory $H_t(\tau, f)$ $(t \le \tau < t + T)$ from the whole trajectory $H(t, f)$ and then extract a feature vector $\boldsymbol{x}(t, f)$ consisting of 28 features listed in Table 1 from $H_t(\tau, f)$. These features have been designed based on our previous studies [12]. Then, the dimensionality is reduced to 12 dimensions using the principal component analysis with the proportion value of 95%. T is 500 ms in the current implementation.

Step 3: Probability Calculation

We analyze the time series of feature vectors, $\{\boldsymbol{x}(t, f) | 0 \le t \le t_{\mathrm{end}}\}$, using $m+1$ left-to-right HMMs $M_1, \cdots, M_{m+1}$. The HMMs are basically same as those used in speech recognition. Each HMM M_i, consisting of 15 states, models sounds of each target instrument ω_i or silence, and those are chained as a Markov chain. Considering $\{\boldsymbol{x}(t, f)\}$ to be generated from this chain, we calculate the likelihood that $\boldsymbol{x}(t, f)$ is generated from each HMM M_i at each time t. This likelihood can be considered to be CIEP $p(\omega_i | \mathrm{X}; t, f)$ to be calculated here. Because features sometimes vary due to the influence of other simultaneous sounds, we use a mixed-sound template [12], in the training phase, which is a technique for building training data from polyphonic sounds.

3.4. Simplifying Instrograms

The instrogram calculates IEPs for every possible frequency, but some applications do not need such detailed results. If the instrogram is used for retrieving musical pieces including a certain instrument's sounds, for example, IEPs for rough frequency regions (e.g., high, middle and low) are sufficient. We therefore divide the whole frequency region into N subregions $I_1, \cdots, I_N$ and calculate the IEP $p(\omega_i; t, I_k)$ for k-th frequency region I_k. $p(\omega_i; t, I_k)$ is defined as $p(\omega_i; t, \bigcup_{f \in I_k} f)$, which can be obtained by iteratively calculating the following equation because the frequency axis is practically discrete.

$$\begin{aligned} & p(\omega_i; t, f_1 \cup \cdots \cup f_i \cup f_{i+1}) \\ & \quad = \; p(\omega_i; t, f_1 \cup \cdots \cup f_i) \; + \; p(\omega_i; t, f_{i+1}) \\ & \qquad - \; p(\omega_i; t, f_1 \cup \cdots \cup f_i)\, p(\omega_i; t, f_{i+1}), \end{aligned}$$

where $I_k = \{f_1, \cdots, f_i, f_{i+1}, \cdots, f_{n_k}\}$.

4. MPEG-7-based Instrogram Annotation

Describing multimedia content including musical one in a universal framework is an important task for content-based multimedia retrieval. In fact, a universal framework for multimedia description, MPEG-7, has been established. Here, we discuss music description based on our instrogram analysis in the context of the MPEG-7 standard.

There are two choices for transforming instrograms to MPEG-7 annotations. First, we can simply represent the IEPs as a time series of vectors. If one aims at the Query-by-Example such as the one discussed in the next section, this annotation method should be used. Because the MPEG-7 standard has no tag for the instrogram annotation, we added several original tags as shown in Figure 4. This example shows the time series of the 8-dimensional IEPs for the piano (line 16) with the 10ms time resolution (line 6). Each dimension corresponds to a different frequency region, which is defined by dividing the entire range from 65.5 Hz to 1048 Hz (line 3) by 1/2 octave (line 4).

Second, we can transform instrograms into a symbolic (event-oriented) representation. If one aims at the Query-

```
 1:<AudioDescriptor
 2:    xsi:type="AudioInstrogramType"
 3:    loEdge="65.5" hiEdge="1048"
 4:    octaveResolution="1/2">
 5:  <SeriesOfVector totalNumOfSamples="5982"
 6:      vectorSize="8" hopSize="PT10N1000F">
 7:    <Raw mpeg7:dim="5982 8">
 8:      0.0 0.0 0.0 0.0 0.718 0.017 0.051 0.0
 9:      0.0 0.0 0.0 0.0 0.724 0.000 0.085 0.0
10:      0.0 0.0 0.0 0.0 0.702 0.013 0.089 0.0
11:      0.0 0.0 0.0 0.0 0.661 0.017 0.063 0.0
12:                  ......
13:    </Raw>
14:  </SeriesOfVector>
15:  <SoundModel
16:      SoundModelRef="IDInstrument:Piano"/>
17:</AudioDescriptor>
```

Figure 4. Excerpt of example of instrogram annotation.

```
 1:<MultimediaContent xsi:type="AudioType">
 2:  <Audio xsi:type="AudioSegmentType">
 3:    <MediaTime>
 4:      <MediaTimePoint>T00:00:06:850N1000
 5:                          </MediaTimePoint>
 6:      <MediaDuration>PT0S200N1000
 7:                          </MediaDuration>
 8:    </MediaTime>
 9:    <AudioDescriptor xsi:type="SoundSource"
10:        loEdge="92" hiEdge="130">
11:     <SoundModel
12:      SoundModelRef="IDInstrument:Piano"/>
13:    </AudioDescriptor>
14:  </Audio>
                    ......
```

Figure 5. Excerpt of example of symbolic annotation.

by-Instrument (i.e., retrieving pieces by specifying instruments by a user), this annotation method is more useful than the first one. We also added several original tags as shown in Figure 5. This example shows that an event of the piano (line 12) at a pitch between 92 and 130 Hz (line 10) occurs at 6.850 s (line 4) and continues during 0.200 s (line 6). To obtain this symbolic representation, we have to estimate the event occurrence and its duration within every frequency region I_k. We therefore obtain the time series of the instrument maximizing $p(\omega_i; t, I_k)$ and then consider this time series to be an output of a Markov chain whose states are the instruments $\omega_1, \cdots, \omega_m$ and silence. In the chain, the transition probabilities from a state to the same state, from a non-silence state to the silence state, and from the silence state to a non-silence state are more than zero, and the other probabilities are zero. After obtaining the most likely path in the chain, we can estimate the occurrence and duration of an instrument ω_i from the transitions between the silence state and the state ω_i.

5. Instrumentation-similarity-based MIR

One of the advantages of the instrogram which is a non-symbolic representation is to provide a new instrumentation-based similarity measure. The similarity between two instrograms enables the MIR based on instrumentation similarity. As we pointed out in Introduction, this key technology is important for automatic playlist generation and content-based music recommendation. Here, instead of calculating the similarity, we calculate the distance (dissimilarity) between instrograms by using dynamic time warping (DTW) [14] as follows:

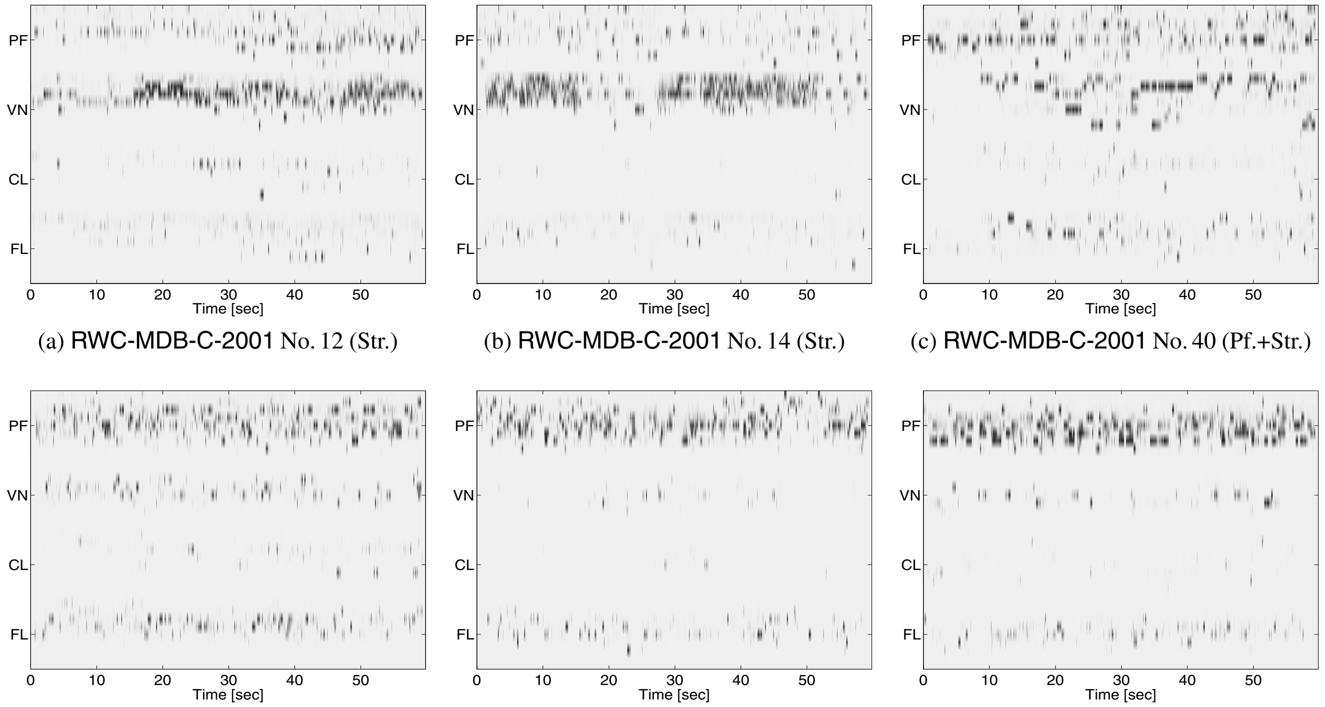
1. A vector $\boldsymbol{p}_t$ for every time t is obtained by concatenating the IEPs of all instruments:
$$\boldsymbol{p}_t = (p(\omega_1; t, I_1), p(\omega_1; t, I_2), \cdots, p(\omega_m; t, I_N))',$$
where $'$ is the transposition operator.

2. The distance between two vectors, $\boldsymbol{p}$ and $\boldsymbol{q}$, is defined as the cosine distance:
$$\mathrm{dist}(\boldsymbol{p}, \boldsymbol{q}) = 1 - (\boldsymbol{p}, \boldsymbol{q})/||\boldsymbol{p}||\cdot||\boldsymbol{q}||,$$
where $(\boldsymbol{p}, \boldsymbol{q}) = \boldsymbol{p}'R\boldsymbol{q}$, and $||\boldsymbol{p}|| = \sqrt{(\boldsymbol{p}, \boldsymbol{p})}$. $R = (r_{ij})$ is a positive definite symmetric matrix that gives the relationship between elements. One may want to give a high similarity to pieces where the same instrument is played at different pitch regions (e.g., $p(\omega_1; t, I_1)$ vs. $p(\omega_1; t, I_2)$) or pieces where different instruments within the same instrument family (e.g., violin vs. viola) are played. They can reflect such relations in the distance measure by setting r_{ij} for the corresponding elements to a value more than zero. When R is the unit matrix, $(\boldsymbol{p}, \boldsymbol{q})$ and $||\boldsymbol{p}||$ are equivalent to the standard innerproduct and norm, respectively.

3. The distance (dissimilarity) between $\{\boldsymbol{p}_t\}$ and $\{\boldsymbol{q}_t\}$ is calculated by applying DTW with the above-mentioned distance measure.

Also in previous MIR-related studies [16, 1], the timbral similarity was used. The timbral similarity was calculated on the basis of spectral features, such as mel-frequency cepstrum coefficients (MFCCs), directly extracted from complex mixtures of sounds. Such features sometimes do not clearly reflect actual instrumentation, as will be implied in the next section, because they are influenced from not only instrument timbres but also arrangements including the voicing of chords. Because instrograms directly represent instrumentation, on the other hand, they will facilitate the appropriate calculation of the similarity of instrumentation. Moreover, instrograms have the following advantages:

Intuitiveness The musical meaning is intuitively clear.



(a) RWC-MDB-C-2001 No. 12 (Str.)

(b) RWC-MDB-C-2001 No. 14 (Str.)

(c) RWC-MDB-C-2001 No. 40 (Pf.+Str.)

(d) RWC-MDB-C-2001 No. 43 (Pf.+Fl.)

(e) RWC-MDB-J-2001 No. 1 (Pf.)

(f) RWC-MDB-J-2001 No. 2 (Pf.)

Figure 6. Results of calculating instrograms from real-performance audio signals. Color versions are available at: http://winnie.kuis.kyoto-u.ac.jp/˜kitahara/instrogram/ISM06/.

Table 2. Musical pieces used and their instrumentation.

Classical	(i) No. 12, 14, 21, 38	Strings
	(ii) No. 19, 40	Piano+Strings
	(iii) No. 43	Piano+Flute
Jazz	(iv) No. 1, 2, 3	Piano solo

Controllability By appropriately setting R, users can ignore or make little of the difference among pitch regions within the same instrument and/or the difference among instruments within the same instrument family.

6. Experiments

We conducted experiments on obtaining instrograms from audio signals. We used 10 recordings of real performances of classical and jazz music taken from the RWC Music Database [7]. The instrumentation of every piece is listed in Table 2. The target instruments were piano (PF), violin (VN), clarinet (CL), and flute (FL). Therefore, the IEPs for violin should also be high when string instruments other than violin are played, and the IEPs for clarinet should always be low. Training data of these four instruments were taken from both RWC-MDB-I-2001 [8] and NTTMSA-P1 (a non-public musical sound database). The time resolution was 10 ms, and the frequency resolution was every 100 cent from C2 to C6. The width of each frequency region was 600 cent. We used HTK 3.0 for HMMs.

The results are shown in Figure 6. We can see that (a) and (b) have high IEPs for violin while (e) and (f) have high IEPs for piano. For (c), the IEPs for violin increase after 10 sec, whereas those for piano are high from the beginning. It reflects the actual performances of these instruments. When (d) is compared to (e) and (f), the former has slightly higher IEPs for flute than the latter, though the difference is unclear. This unclear difference is because the acoustic characteristics of real performances have high variety. It can be improved by adding appropriate training data.

Based on the instrograms obtained, we conducted experiments on symbolic annotation using the method described in Section 4. The results were evaluated by

$$\frac{\sum_i \sum_k \text{\# frames correctly annotated as } \omega_i \text{ at } I_k}{\sum_i \sum_k \text{\# frames annotated as } \omega_i \text{ at } I_k},$$

The results are shown in Figure 7, where C12 for example stands for Piece No. <u>12</u> included in RWC-MDB-<u>C</u>-2001. The average of the accuracies was 76.2%, and the accuracies for eight of the ten pieces were over 70%.

Next, we tested the calculation of the dissimilarities between instrograms. We used the unit matrix as R. The re-

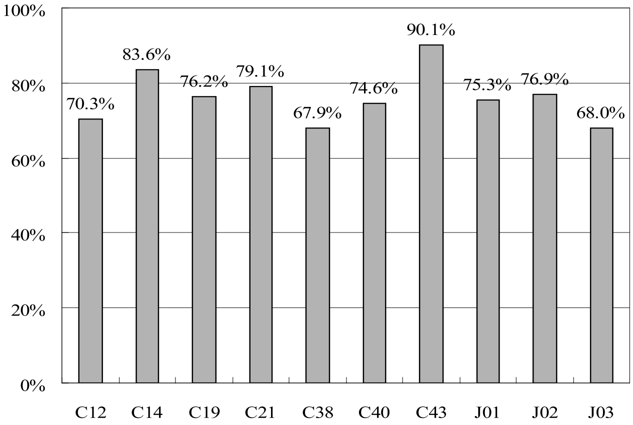


Figure 7. Accuracy for symbolic annotation from instrogram. C and J represent genres and following numbers represent piece numbers described in Table 2.

sults, listed in Table 3 (a), can be summarized as follows:

- The dissimilarities within each group were mostly less than 7000 (except for Group (ii)).
- Those between Groups (i) (played on strings) and (iv) (piano) were mostly more than 9000, and some were more than 10000.
- Those between Groups (i) and (iii) (piano+flute) were also around 9000.
- Those between Groups (i) and (ii) (piano+strings), (ii) and (iii), and (ii) and (iv) were around 8000. In these pairs, one instrument is commonly used, so these dissimilarities were reasonable.
- Those between Groups (iii) and (iv) were around 7000. Because the difference between these groups is only the presence of flute, these similarities were also reasonable.

For comparison, Table 3 (b) shows the results using MFCCs and Table 4 shows the 3-best-similarity pieces from each of the ten pieces by using both methods. Comparing the results of the two methods, we can see the following differences:

- The dissimilarities within Group (i) and the dissimilarities between Group (i) and others in the case of IEPs were more different than those in the case of MFCCs. In fact, all of the 3-best-similarity pieces from the pieces in Group (i) belonged to the same Group (i) in the case of IEPs, while those in the case of MFCCs contained pieces out of Group (i).
- All of the 3-best-similarity pieces from the four pieces without strings (Groups (iii) and (iv)) also did not contain strings in the case of IEPs, while those in the case of MFCCs contained pieces with strings (C14, C21).

Table 3. Dissimilarity of instrograms. (i)–(iv) represent categories defined in Table 2.

(a) Using IEPs (instrograms)

	(i)				(ii)		(iii)	(iv)		
	C12	C14	C21	C38	C19	C40	C43	J01	J02	J03
C12	0									
C14	6429	0								
C21	5756	5734	0							
C38	7073	6553	6411	0						
C19	7320	8181	7274	7993	0					
C40	8650	8353	8430	8290	8430	0				
C43	8910	9635	9495	9729	8148	8235	0			
J01	9711	10226	10252	10324	8305	8214	6934	0		
J02	9856	10125	10033	10610	8228	8139	7216	6397	0	
J03	9134	9136	8894	9376	8058	8327	7480	6911	7223	0

(b) Using MFCCs

	(i)				(ii)		(iii)	(iv)		
	C12	C14	C21	C38	C19	C40	C43	J01	J02	J03
C12	0									
C14	17733	0								
C21	17194	18134	0							
C38	18500	18426	18061	0						
C19	17510	18759	18222	19009	0					
C40	17417	19011	18189	19099	18100	0				
C43	18338	17459	17728	18098	18746	18456	0			
J01	17657	17791	17284	17834	18133	17983	16762	0		
J02	17484	17776	17359	18009	17415	17524	17585	15870	0	
J03	17799	18063	17591	18135	17814	18038	17792	16828	16987	0

We also developed a prototype system that enables a user to retrieve pieces having instrumentation similar to a piece specified by the user (Figure 8). After the user selects a musical piece as a query, the system calculates the (dis)similarity between the selected piece and each of the pieces in a collection using the method described in Section 5 and then shows the list of musical pieces in order of similarity. When the user selects a piece from the list, the system plays back its piece with audio-synchronized visualization of its IEPs: it shows bar graphs of IEPs in real time like those of the power spectrum display on digital music players. The demonstration of our MIR system is available at:

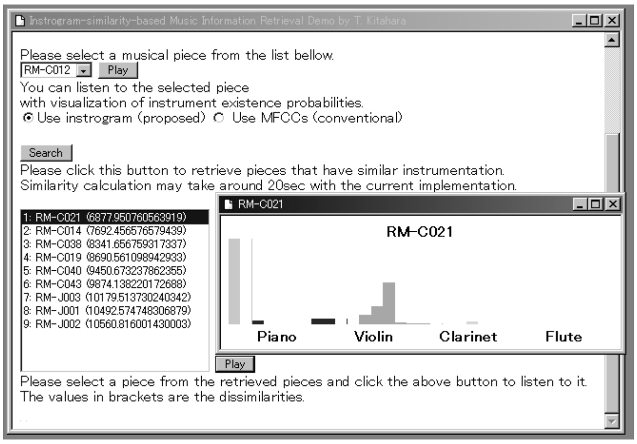
http://winnie.kuis.kyoto-u.ac.jp/~kitahara/instrogram/ISM06/

7. Conclusions

We proposed a non-notewise musical instrument recognition method based on *instrogram*, the time-frequency representation of instrument existence probabilities (IEPs). Whereas most previous methods first estimated the onset time and F0 of each note and then identified the instrument of each note, our method calculates the IEP for each target instrument at each point of the time-frequency plane and hence does not rely on either onset detection nor F0 estimation. We also presented methods for applying instrograms to MPEG-7 annotation and MIR based on instrumentation similarity. The experimental results with ten pieces of realistic music were promising. In the future, we plan to extend our method to deal with pieces containing drum sounds

Table 4. 3-best-similarity pieces from each of ten pieces.

		Using IEPs	Using MFCCs
(i)	C12	C21, C14, C38	C21, C40, J02
	C14	C21, C12, C38	C43, C12, J02
	C21	C14, C12, C38	C12, J01, J02
	C38	C21, C14, C38	J01, J02, C21
(ii)	C19	C21, C12, C38	J02, C12, J03
	C40	J02, J01, C43	C12, J02, J01
(iii)	C43	J01, J02, J03	J01, C14, J02
(iv)	J01	J02, J03, C43	J02, C43, J03
	J02	J01, C43, J03	J01, J03, C21
	J03	J01, J02, C43	J01, J02, C21


Instrogram-similarity-based Music Information Retrieval Demo by T. Kitahara
RM-C012
Play
Search
RM-C021
RM-C021
Piano
Violin
Clarinet
Flute
Play

Please select a musical piece from the list bellow.

You can listen to the selected piece
with visualization of instrument existence probabilities.

● Use instrogram (proposed) ○ Use MFCCs (conventional)

Please click this button to retrieve pieces that have similar instrumentation.
Similarity calculation may take around 20sec with the current implementation.

1: RM-C021 (6877.950760563919)
2: RM-C014 (7692.456576579439)
3: RM-C038 (8341.656759317337)
4: RM-C019 (8690.561098942933)
5: RM-C040 (9450.673237862355)
6: RM-C043 (9874.138220172688)
7: RM-J003 (10179.513730240342)
8: RM-J001 (10492.574748306879)
9: RM-J002 (10560.816001430003)

Please select a piece from the retrieved pieces and click the above button to listen to it.
The values in brackets are the dissimilarities.

Figure 8. Demonstration of our MIR prototype.

by incorporating a drum sound recognition method. We also plan to compare our instrumentation similarity measure with a perceptual one through listening tests.

This study is based on the standpoint that transcribing music as a score is not the essence of music understanding [5]. People can enjoy listening to music without mental score-like transcription, but most previous studies dealt with not such human-like music understanding but score-based music transcription. We therefore plan to establish a computational model of music understanding by integrating the instrogram technique with models for recognizing other musical elements.

Acknowledgments: This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Scientific Research 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 and NTT Communication Science Laboratories for giving us permission to use NTTMSA-P1. We would also like to thank Dr. Shinji Watanabe for his valuable comments.

References

[1] J.-J. Aucouturier and F. Pachet. Music similarity measure: What's the use? In *Proc. ISMIR*, pages 157–163, 2002.

[2] J. Eggink and G. J. Brown. Application of missing feature theory to the recognition of musical instruments in polyphonic audio. In *Proc. ISMIR*, 2003.

[3] J. Eggink and G. J. Brown. Extracting melody lines from complex audio. In *Proc. ISMIR*, pages 84–91, 2004.

[4] S. Essid, G. Richard, and B. David. Instrument recognition in polyphonic music based on automatic taxonomies. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(1):68–80, 2006.

[5] M. Goto. Music scene description project: Toward audio-based real-time music understanding. In *Proc. ISMIR*, pages 231–232, 2003.

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

[7] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical, and jazz music databases. In *Proc. ISMIR*, pages 287–288, 2002.

[8] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Music genre database and musical instrument sound database. In *Proc. ISMIR*, pages 229–230, 2003.

[9] K. Kashino and H. Murase. A sound source identification system for ensemble music based on template adaptation and music stream extraction. *Speech Comm.*, 27:337–349, 1999.

[10] K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka. Application of the Bayesian probability network to music scene analysis. In D. F. Rosenthal and H. G. Okuno, editors, *Computational Auditory Scene Analysis*, pages 115–137. Lawrence Erlbaum Associates, 1998.

[11] T. Kinoshita, S. Sakai, and H. Tanaka. Musical sound source identification based on frequency component adaptation. In *Proc. IJCAI CASA Workshop*, pages 18–24, 1999.

[12] 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 *Proc. ISMIR*, pages 558–563, 2005.

[13] K. D. Martin. *Sound-Source Recognition: A Theory and Computational Model*. PhD thesis, MIT, 1999.

[14] C. S. Myers and L. R. Rabiner. A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell Syst. Tech. J.*, 60(7):1389–1409, 1981.

[15] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

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

[17] E. Vincent and X. Rodet. Instrument identification in solo and ensemble music using independent subspace analysis. In *Proc. ISMIR*, pages 576–581, 2004.