INSTROGRAM: A NEW MUSICAL INSTRUMENT RECOGNITION TECHNIQUE WITHOUT USING ONSET DETECTION NOR F0 ESTIMATION

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

This paper describes a new technique for recognizing musical instruments in polyphonic music. Because the conventional framework for musical instrument recognition in polyphonic music had to estimate the onset time and fundamental frequency (F0) of each note, instrument recognition strictly suffered from errors of onset detection and F0 estimation. Unlike such a note-based processing framework, our technique calculates the temporal trajectory of *instrument existence probabilities* for every possible F0, and the results are visualized with a spectrogram-like graphical representation called *instrogram*. The instrument existence probability is defined as the product of a *nonspecific instrument existence probability* calculated using PreFEst and a *conditional instrument existence probability* calculated using the hidden Markov model. Experimental results show that the obtained instrograms reflect the actual instrumentations and facilitate instrument recognition.

1. INTRODUCTION

Musical instrument recognition is an important task for many applications including automatic music transcription, music information retrieval and computational auditory scene analysis. In particular, recent worldwide popularization of online music distribution services and portable digital music players makes musical instrument recognition more important. This is because musical pieces, especially classical music, are characterized by what instruments are used. In fact, the names of some music genres are based on instrument names, such as "piano sonata" and "string quartet." Musical instrument recognition can therefore be used when one wants to search for certain types of musical pieces, such as piano sonata or string quartet.

Whereas musical instrument recognition studies mainly dealt with solo musical sounds in 1990s (e.g., [1]), the number of those dealing with polyphonic music has been increasing in recent years. Kashino *et al.* [2] developed a computational music scene analysis architecture called OPTIMA, which recognizes musical notes and the instruments based on the Bayesian probability network. They subsequently proposed a method that identifies the instrument playing each musical note based on template matching with template adaptation [3]. Kinoshita *et al.* [4] 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.* [5] tackled this overlapping problem with the missing feature theory. They subsequently dealt with the problem of identifying only the instrument playing the main (the most predominant) melody on the assumption that the main melody's partials suffer less from other sounds occurring simultaneously [6]. Vincent *et al.* [7] proposed a new musical instrument identification method based on independent subspace analysis. Kitahara *et al.* [8] proposed techniques for feature weighting based on the robustness to the above-mentioned overlapping problem and for avoiding musically unnatural errors using musical context.

The common feature in the above-mentioned studies except for [7] is that instrument identification is performed for each frame or each note. In the former case [5, 6], it is difficult to obtain a reasonable accuracy because temporal variations of spectra are important characteristics of musical instrument sounds. In the latter case [2, 3, 4, 8], the identification system has to first estimate the onset time and fundamental frequency (F0) of musical notes and then extracts the harmonic structure of each note based on the estimated onset time and F0. Therefore, instrument identification suffers from errors of onset detection and F0 estimation. In the experiments reported in [3] and [8], in fact, correct data of the onset times and F0s were manually fed.

In this paper, we propose a new technique that recognizes musical instruments in polyphonic musical audio signals without relying on onset detection nor F0 estimation. The key idea of our technique is to visualize the probability that the sound of each of target instruments exists at each time and each frequency. The result of this analysis is a set of images, called *instrogram*, that are similar to a spectrogram except in that each point represents not the energy of the signal but the *instrument existence probability*. This probability is defined as the product of two kinds of probabilities, called *nonspecific instrument existence probability* and *conditional instrument existence probability*, and these are calculated using PreFEst [9] and the hidden Markov model (HMM), respectively. The advantage of our technique is that errors of calculating one probability do not influence the calculation of the other because the two probabilities can be calculated independently.

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. There exists one image for each target instrument, and each image, where the horizontal and vertical axes represent time and frequency respectively, shows the probability that the target instrument is used. An example is presented in **Fig. 1**. This example is the result of analyzing an audio signal of "Auld Lang Syne" with respect to piano, violin, clarinet, and flute. This signal, played on piano, violin, and flute, was prepared with the procedure described later, in Section 4. If the instrogram is too detailed for some purposes, it can be simplified by dividing the whole frequency region into some subregions and merging results within each subregion. The simplified version of **Fig. 1** is given in **Fig. 2**. From the four images of the instrogram or from

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Scientific Research (A), No.15200015, and Informatics Research Center for Development of Knowledge Society Infrastructure (COE program of MEXT, Japan).

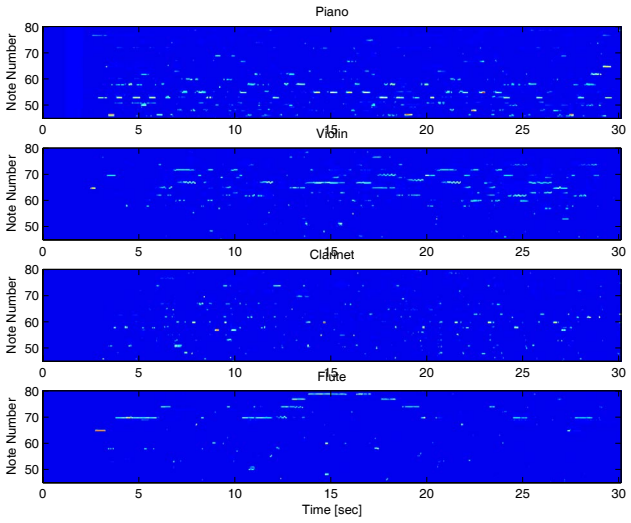


Fig. 1. An example of instrograms. This is the result of analyzing trio music, "Auld Lang Syne," played on piano, violin, and flute.

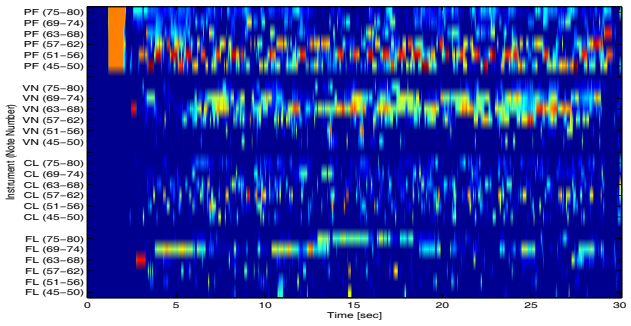


Fig. 2. The simplified (summarized) instrogram of **Fig. 1**.

the simplified instrogram, we can see that the melodies in the high (approx. note numbers 70–80), middle (60–75), and low (45–60) pitch regions are played on flute, violin, and piano.

3. ALGORITHM FOR CALCULATING INSTROGRAM

Each image of the instrogram is a plane with horizontal and vertical axes representing time and frequency. The intensity of the color of each point (t, f) in the image represents the probability $p(\omega_i; t, f)$ that a sound of the target instrument ω_i exists at time t and frequency f. We call this *instrument existence probability*. It can be calculated as the product of two probabilities:

$$p(\omega_i; t, f) = p(\text{exist}; t, f)\; p(\omega_i|\text{exist}; t, f),$$

where $p(\text{exist}; t, f)$ called *nonspecific instrument existence probability* is the probability that a sound of a certain instrument exists at time t and frequency f, while $p(\omega_i|\text{exist}; t, f)$ called *conditional instrument existence probability* is the conditional probability that, if a sound of a certain instrument exists at time t and frequency f, the instrument is ω_i.

The nonspecific instrument existence probability is calculated using PreFEst [9]. PreFEst models, at each frame, an observed spectrum of the input signal containing multiple musical instrument sounds as a weighted mixture of harmonic-structure tone models with every possible F0. The weight of each tone model represents how relatively predominant its tone model is. We define $p(\text{exist}; t, f)$ as this weight because this weight can be interpreted as the probability that there exists a certain sound at its F0.

The conditional instrument existence probability, on the other hand, is calculated by using HMMs because temporal characteristics of an instrument sound are important in recognizing its instrument. In each possible frequency f, the temporal trajectory $H(t, f)$ of the harmonic structure with F0 of f can be considered to be generated from a Markov chain of $m + 1$ models of possible instruments $\omega_1, \cdots, \omega_m$ and silence. Each model is an HMM that consists of multiple states. Then, $p(\omega_i|\text{exist}; t, f)$ can be calculated from the likelihoods of paths in the chain.

The instrument existence probability $p(\omega_i; t, f)$ can be estimated robustly because the two constituent probabilities corresponding to the note estimation and instrument identification 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 result 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. Unlike such note-based symbolic approach, our technique is based on a non-symbolic and non-sequential approach that is more robust for polyphonic music.

3.1. Nonspecific Instrument Existence Probability

By using the PreFEst, the nonspecific instrument existence probability $p(\text{exist}; t, f)$ is estimated on the basis of the maximum likelihood estimation without assuming the number of sound sources in a mixture. The PreFEst, which was originally developed for estimating F0s of melody and bass lines, consists of three processes: *PreFEst-front-end* for frequency analysis, *PreFEst-core* for estimating the relative dominance of every possible F0, and *PreFEst-back-end* for evaluating the temporal continuity of the F0. Because the problem to be solved here is not to estimate the predominant F0s as melody and bass lines but to calculate $p(\text{exist}; t, f)$ of every possible F0 f, we only use the PreFEst-core.

The PreFEst-core 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

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

$$\theta^{(t)} = \{w^{(t)}(F) | \text{Fl} \leq F \leq \text{Fh}\},$$

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 $\int_{\text{Fl}}^{\text{Fh}} w^{(t)}(F) dF = 1$. 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 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 nonspecific instrument existence probability $p(\text{exist}; t, f)$ to be equal to $w^{(t)}(f)$. The weights can be estimated using the *Expectation-Maximization* (EM) algorithm as described in [9].

3.2. Conditional Instrument Existence Probability

The conditional instrument existence probability $p(\omega_i|\text{exist}; t, f)$ is calculated through the following steps.

3.2.1. Short-time Fourier transform

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

Table 1. Overview of 28 features

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

3.2.2. Harmonic structure extraction

In each possible frequency f, the temporal trajectory $H(t, f)$ of the harmonic structure (10 harmonics) whose F0 is f is extracted.

3.2.3. Feature extraction

Designing effective features is an important issue in musical instrument recognition. In the field of speech recognition studies, mel-frequency cepstrum coefficents (MFCCs) and Delta MFCCs are commonly used. Although these features may be considered possible to be applied to musical instrument recognition, we designed features optimized for musical instrument sounds because musical instrument sounds have more complicated temporal variations (e.g., amplitude and frequency modulations). We adopt 28 features listed in **Table 1**, which are partly modified from those we previously proposed [8]. For every time t (every 10 ms in the implementation), we first excerpt a T-length bit of the harmonic-structure trajectory $H_t(\tau, f)$ $(t \leq \tau < t + T)$ from the whole trajectory $H(t, f)$ and then extract a feature vector $\boldsymbol{x}(t, f)$ consisting of 28 features from $H_t(\tau, f)$. T is 50 ms in the current implementation.

3.2.4. Probability calculation

We train HMMs, each consisting of 15 states, for target instruments $\omega_1, \cdots, \omega_m$ and silence, and then consider the time series $\boldsymbol{x}(t, f)$ of feature vectors to be generated from a Markov chain of those $m + 1$ HMMs. The problem to be solved here is thus to calculate, for each HMM M_i $(i = 1, \cdots, m,$ silence), the probability $p(\boldsymbol{x}(t, f)|M_i; t)$ that the feature vector $\boldsymbol{x}(t, f)$ is generated from M_i at time t. Because the Markov chain assumes that the model at time t only depends on that at the previous time $t - 1$, this probability can be calculated by

$$p(\boldsymbol{x}(t, f)|M_i; t) = \sum_j p(\boldsymbol{x}(t, f)|M_i)p(M_i|M_j),$$

where $p(\boldsymbol{x}(t, f)|M_i)$ is the probability that the model M_i generates the feature vector $\boldsymbol{x}(t, f)$ at any time and $p(M_i|M_j)$ is the transition probability from M_j to M_i. Here $p(\boldsymbol{x}(t, f)|M_i)$ is calculated by the trained HMM M_i. The conditional instrument existence probability $p(\omega_i|\text{exist}; t, f)$ can thus defined to be equal to $p(\boldsymbol{x}(t, f)|M_i; t)$.

3.3. Simplifying Instrograms

The instrogram calculates instrument existence probabilities 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, instrument existence probabilities 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 instrument existence probability $p(\omega_i; t, I_k)$ 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.

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

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

3.4. Application to Musical Instrument Identification

The simplest way for applying instrograms to musical instrument identification is to obtain the instrument maximizing $p(\omega_i; t, f)$ for each time t and frequency f. However, the results obtained in this way include some errors because it does not take the temporal continuity into consideration. Here, we consider the time series of the instruments maximizing $p(\omega_i; t, f)$ for each frequency to be outputs of a Markov chain whose states are the instruments $\omega_1, \cdots, \omega_m$ and silence. In the Markov 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. Once the most likely path in the chain is obtained, the times when the instrument ω_i begins and stops playing can be estimated from the times of transitions from the silence state to the ω_i state and vice versa, respectively.

4. EXPERIMENTS

We conducted experiments on obtaining instrograms from audio signals. We used trio music of "Auld Lang Syne" that Kashino *et al.* used [2]. Audio data of Auld Lang Syne were generated by mixing audio data of RWC-MDB-I-2001 [10] on a computer according to a standard MIDI file that we input using a MIDI sequencer based on Kashino's score. Target instruments were piano (PF), violin (VN), clarinet (CL), and flute (FL). The time resolution was 10 ms from the beginning to the end of the musical piece, and the resolution was every 20 cent from A2 to A5. The width of each frequency region I_k was 600 cent. We used Fujihara's implementation [11] for PreFEst and HTK 3.0 for HMMs.

The results of generating instrograms for four different instrumentations are shown in **Fig. 3**. Note that **Fig. 3** shows only simplified instrograms due to limited space. Comparing (a) and (b), we can see that (a) has high existence probabilities for the flute in high frequency regions while (b) has very low (almost zero) probabilities. In addtion, (c) has the highest existence probabilities for the violin and the secondarily highest existence probabilities for the piano, whereas it has lower probabilities for the other instruments. Similarly, (d) has high existence probabilities for the piano, whereas it has lower probabilities for the other instruments.

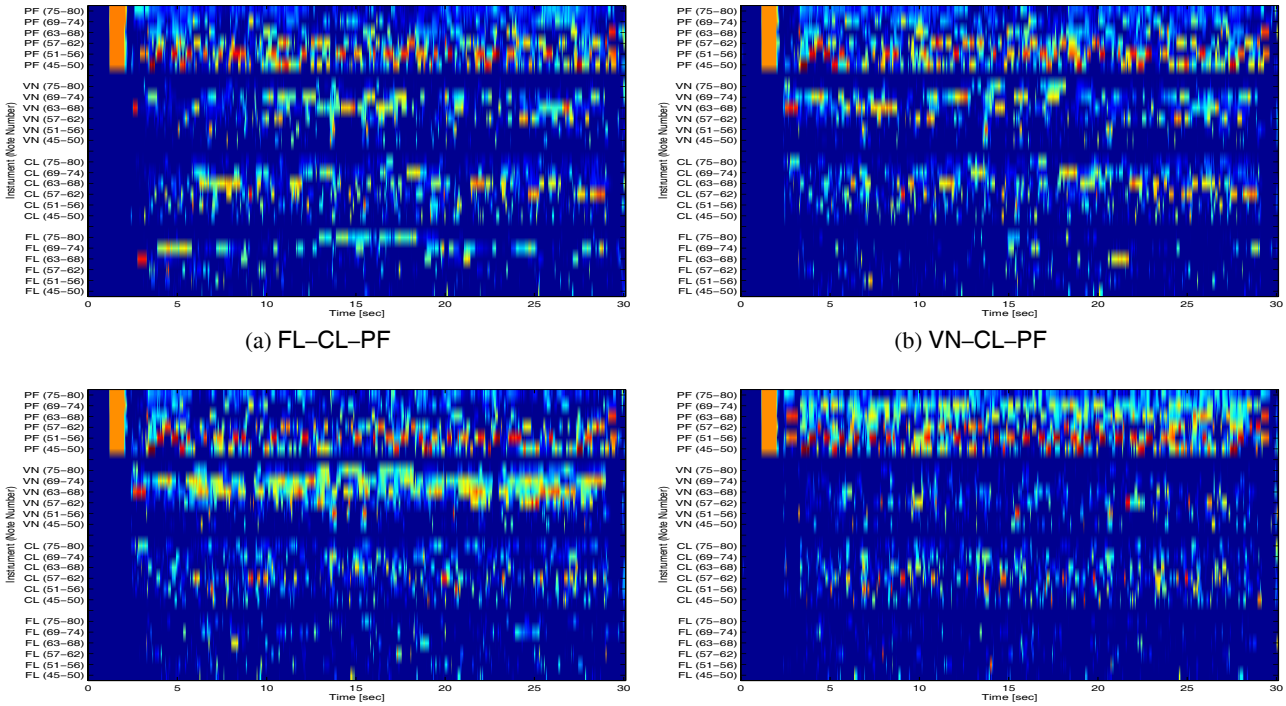
Next, we conducted experiments on musical instrument recognition using the obtained instrograms based on the method in Section 3.4. The accuracy of the recognition is evaluated with

$$\text{(Recognition rate)} = \frac{\text{(Num of correctly recognized frames)}}{\text{(Num of the whole frames)}}.$$

The results are listed in **Table 2**. From the table, we can see that the instrogram was useful for musical instrument recognition.

5. CONCLUSIONS

We have described a new *instrogram* representation obtained by a new musical instrument recognition technique that does not rely



(a) FL–CL–PF

(b) VN–CL–PF

(c) VN–VN–PF

(d) PF–PF–PF

Fig. 3. Instrograms of "Auld Lang Syne" with four different instrument combinations. "FL–CL–PF" means that the treble, middle, and bass parts are played on flute, clarinet, and piano, respectively.

Table 2. Recognition rate for each instrument combination.

PF–PF–PF	97.1%	VN–CL–PF	73.2%
PF–VN–PF	84.7%	FL–PF–PF	82.8%
PF–CL–PF	86.9%	FL–VN–PF	78.5%
VN–PF–PF	81.1%	FL–CL–PF	78.7%
VN–VN–PF	86.7%		

on both onset detection and F0 estimation. Whereas most previous studies first estimated the onset time and F0 of each note and then identified the instrument of each note, our technique calculates the instrument existence probability for each target instrument in each point of the time-frequency plane. This non-symbolic approach made it possible to avoid bad influences caused by errors of the onset detection and F0 estimation. In addition, by introducing a Markov chain whose states correspond to target instruments and silence for every possible F0, we achieved the identification of musical instruments (i.e., to symbolize the instrogram analysis) for polyphonic music.

Although we have applied instrograms to musical instrument identification in this paper, the instrograms have a wider range of potentials. If the similarity between the instrograms of two musical pieces can be calculated, for example, it will make instrumentation-similarity-based music information retrieval possible. Future work will include the development of such useful applications of instrograms as well as the improvement of the accuracy of calculating the instrument existence probabilities.

Acknowledgments: We thank everyone who has contributed to building and distributing the RWC Music Database (RWC-MDB-I-2001) [10]. We also thank Mr. Hiromasa Fujihara for giving us permission to use his program.

6. REFERENCES

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

[2] K. Kashino *et al.*, "Application of the Bayesian Probability Network to Music Scene Analysis," *Computational Auditory Scene Analysis*, D. F. Rosenthal and H. G. Okuno (eds.), Lawrence Erlbaum Associates, pp.115–137, 1998.

[3] K. Kashino and H. Murase, "A Sound Source Identification System for Ensemble Music based on Template Adaptation and Music Stream Extraction," *Speech Communication*, **27**, pp.337–349, 1999.

[4] T. Kinoshita *et al.*, "Musical Sound Source Identification based on Frequency Component Adaptation," *Proc. IJCAI CASA Workshop*, pp.18–24, 1999.

[5] J. Eggink and G. J. Brown, "A Missing Feature Approach to Instrument Identification in Polyphonic Music," *Proc. ICASSP*, **V**, pp.553–556, 2003.

[6] J. Eggink and G. J. Brown, "Instrument Recognition in Accompanied Sonatas and Concertos," *Proc. ICASSP*, **IV**, pp.217–220, 2004.

[7] E. Vincent and X. Rodet, "Instrument Identification in Solo and Ensemble Music using Independent Subspace Analysis," *Proc. ISMIR*, pp.576–581, 2004.

[8] T. Kitahara *et al.*, "Instrument Identification in Polyphonic Music: Feature Weighting with Mixed Sounds, Pitch-dependent Timbre Modeling, and Use of Musical Context," *Proc. ISMIR*, pp.558–563, 2005.

[9] 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*, **43**, pp.311–329, 2004.

[10] M. Goto *et al.*, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," *Proc. ISMIR*, pp.229–230, 2003.

[11] H. Fujihara *et al.*, "Singer Identification based on Accompaniment Sound Reduction and Reliable Frame Selection," *Proc. ISMIR*, pp.329–336, 2005.