Toward Polyphonic Musical Instrument Identification using Example-based Sparse Representation

Mari Okamura, Masanori Takehara, Satoshi Tamura, and Satoru Hayamizu
Department of Information Science, Gifu University, Gifu, Japan
E-mail: (okamura,takehara)@asr.info.gifu-u.ac.jp, tamura@info.gifu-u.ac.jp, hayamizu@gifu-u.ac.jp

Abstract— **Musical instrument identification is one of the major topics in music signal processing. In this paper, we propose a musical instrument identification method based on sparse representation for polyphonic sounds. Such the identification has been still categorized into challenging tasks, since it needs high-performance signal processing techniques. The proposed scheme can be applied without any signal processing such as source separation. Sample feature vectors for various musical instruments are used for the base matrix of sparse representation. We conducted two experiments to evaluate the proposed method. First, the musical instrument identification is tested for monophonic sounds using five musical instruments. The average accuracy of 91.9% was obtained and it shows the effectiveness of the proposed method. Second, musical instrument composition of polyphonic sounds is examined, which contain two instruments. It is found that the estimated weight vector by sparse representation indicates the mixture ratio of two instruments.**

I. INTRODUCTION

The composition of musical instruments in music plays an important role to characterize music. It is useful to obtain the composition of musical instruments from music for music information retrieval such as music database retrieval [1]. However, identification and composition estimation of multi-instruments of polyphonic sound are still in a state of development.

A conventional method to identify multiple musical instruments from polyphonic sounds is based on a feature-extraction-based strategy; musical features are extracted from harmonic components using some signal and music processing techniques, for example, fundamental frequency extraction of a monophonic sound. In order to extract the features, monophonic sounds might be separated from the compound sounds. An instrument identification method for polyphonic music was proposed, which aims to improve robustness for overlapping frequency components [2].

In this paper, we introduce a Sparse Representation Classification (SRC) method for multiple musical instrument identification. SRC is used in the music signal processing domain, e.g. music genre classification [3]. It has typically used for signal compression and recovering, recently it has also shown success in face recognition [4]. The research of sparse decomposition of polyphonic music using Independent Component Analysis (ICA) which similar to SRC has been reported [5]. In the SRC method, weight coefficients of an input signal, which correspond to sample data including features of various musical instruments, are estimated. The coefficients are computed under the restriction where the input signal is represented as a linear combination of features in the sample data and the coefficient vector is sparse. For example, polyphonic sounds from piano and violin can be considered as the combination of a piano feature and a violin feature, therefore, it is expected that estimated weight components for piano and violin become large. Since the SRC method does not require any source separation technique, the method can robustly and correctly identify musical instruments. In order to evaluate the proposed method, we conduct two experiments. First, we investigate the effectiveness of the proposed method by comparing with Support Vector Machine (SVM) using monophonic sounds. Second, we apply the proposed method to polyphonic sounds, evaluating a mixture ratio of musical instruments.

Note that a similar approach has been investigated; Non-negative Matrix Factorization (NMF) is used to decompose a musical data matrix into two factors with non-negative entries. In the area of music signal processing, NMF has been applied to many tasks including music transcription [6]. NMF estimates two matrices so as to minimize the residual error between input data and the production of the matrices. However, NMF has disadvantage that bases in the matrix usually do not correspond to any instrument. In contrast, the base matrix is prepared by enumerating training vectors of every class in SRC, preventing from the disadvantage of NMF.

This paper is organized as follows: Section II introduces a theory of sparse representation classification. Our proposed musical instrument identification method is explained in Section III. Experiments using monophonic sounds are described in Section IV, subsequently experiments for polyphonic sounds are reported in Section V. Finally Section VI concludes this paper.

II. SPARSE REPRESENTATION CLASSIFICATION

The fundamental issue of sparse representation is estimating weight coefficients (an unknown weight vector) so that an input signal can be represented as a combination of examples (a base matrix) and the weight vector. The outline of sparse representation is shown in Fig. 1. Let us denote a feature vector of an input signal by $\boldsymbol{y}$, a data matrix consisting of sample feature vectors for various musical instruments by $\boldsymbol{H}$, and a weight vector by $\boldsymbol{s}$. If we denote a feature matrix for a

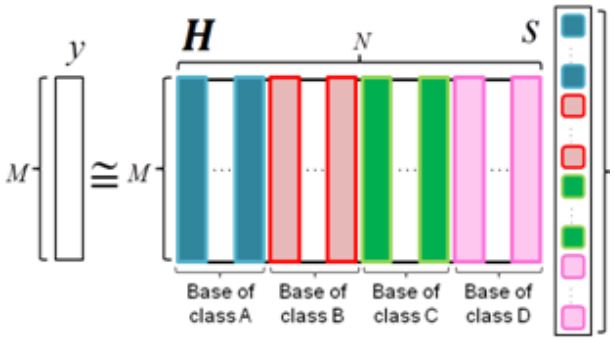

Fig. 1 Sparse representation of an input vector using a base matrix and a weight vector. N is the number of examples in the base matrix, M is a feature dimension.

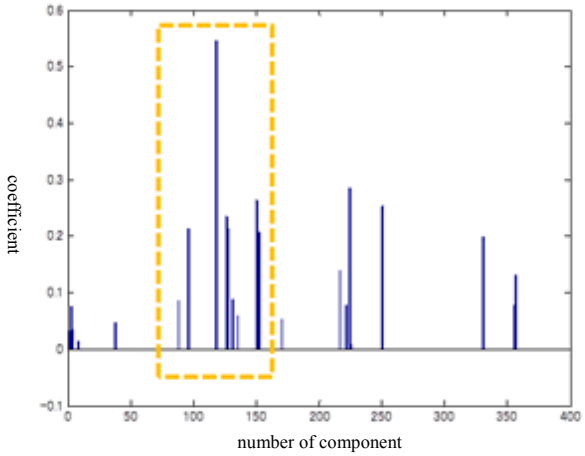

Fig. 2 An example of estimated weight vector. The vertical axis indicates weight coefficient values, and the horizontal axis indicates an index number of component of the vector. Weight coefficients for the correct class are shown in the dotted box. Large coefficients corresponding to the sub-matrix of the correct class are found.

class i by $\boldsymbol{H}_i$, the matrix $\boldsymbol{H}_i$ is defined by the following equation:

$$\boldsymbol{H}_i = [\boldsymbol{h}_1, \ \boldsymbol{h}_2, \ \cdots, \ \boldsymbol{h}_{N_i}] \tag{1}$$

where $\boldsymbol{h}_i$ is an i-th feature vector in the class. By concatenating all the feature matrices, the base matrix $\boldsymbol{H}$ can be formed. The weight vector $\boldsymbol{s}$ must be estimated so that the modeling error is minimized and the vector is sparse; given an input signal, the unknown weight vector can be obtained by solving a linear programming problem so as to minimize the 1-norm of the vector as:

$$\text{argmin}\|\boldsymbol{s}\|_1 \ \text{subj. to} \ \boldsymbol{y} = \boldsymbol{H}\boldsymbol{s} \tag{2}$$

The part of the vector $\boldsymbol{s}$ corresponding to $\boldsymbol{H}_i$ indicates the inclusion of data in the class i. For example, assume that an input signal is derived from a class B. It is then expected that the part corresponding to the class B of the vector $\boldsymbol{s}$ will have large values whereas the other parts become zero (see also Fig. 1). Similarly, if an input signal consists of a mixture of a class B and a class D, the components corresponding to the classes in the vector might have large values. The example of the estimated weight vector is shown in Fig. 2. The weight vector $\boldsymbol{s}$ can indicate how much an input signal contains the ingredient of each class. We can thus identify musical instruments in an input signal by estimating a weight vector $\boldsymbol{s}$, even from polyphonic sounds without any signal processing technique e.g. sound source separation.

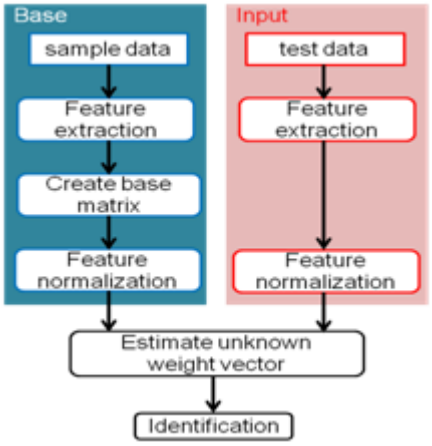

Fig. 3 An overview of our proposed method: "Base" is a flow to create a base matrix, and "Input" is a flow to compute an input vector.

TABLE I
THE LIST OF FEATURE(A) USED IN THIS PAPER.

	Feature Name	Number of dimensions
Spectral features	MFCC	39
	Spectral centroid	2
	Roll-off	1
	Flux	1
	Zero-crossing	1
	Low-energy ratio	1
	Spectral contrast	8
	Harmonic features	5
Temporal features	Time change of power	2

III. MUSICAL INSTRUMENT IDENTIFICATION

The overview of our proposed method is illustrated in Fig. 3. The rest of this section explains the proposed method in detail.

A. Feature extraction

In this paper, 60 features shown in Table I representing characteristics of musical instrument are extracted from spectral components and a waveform of an input signal. These features are prepared for each tone. In each feature, mean and variance parameters of all the frames in the signal are computed. As a result, 120-dimensional feature vector is obtained. The sampling rate is 44.1 kHz with 16 bits and the frame size and overlap are 50 msec and 20 msec, respectively. At first, spectrum components are computed, then spectral

features such as Mel Frequency Cepstral Coefficients (MFCCs) and a spectral centroid are extracted and finally time change of power coefficient, a ratio of a power at a first frame and the maximum power, as well as a gradient value of the envelope of input waveform are employed as temporal features [2]. These features are defined as feature(A). Spectral components divided by subband which contains 26 linear filters and 46 log filters in mel scale is defined as feature(B). The spectrum is used to apply polyphonic sounds because it has additibity of overlapping frequency components.

B. Feature normalization

After feature extraction, a feature vector is normalized. Let us denote an average and a standard deviation at an m–th feature in the sample matrix by μ_m and σ_m, respectively. The original feature vector $\boldsymbol{h} = (h_1, h_2, \cdots, h_M)^{\mathrm{T}}$ in the sample data matrix is then converted into a normalized vector $\boldsymbol{h}'$ as:

$$\boldsymbol{h}' = \left(\frac{h_1-\mu_1}{\sigma_1}, \frac{h_2-\mu_2}{\sigma_2}, \cdots, \frac{h_M-\mu_M}{\sigma_M}\right)^{\mathrm{T}} \quad (3)$$

An input feature vector is similarly normalized with μ_m and σ_m, then a normalized vector $\boldsymbol{y}'$ is obtained.

C. Estimation of unknown weight vector

A weight vector $\boldsymbol{s}$ is estimated using the base matrix and a feature vector of input signal. The weight vector is obtained by solving the 1-norm optimization problem shown by the following equation:

$$\mathrm{argmin}\|\boldsymbol{s}\|_1 \ \mathrm{subj.to}\ \boldsymbol{y}' = \boldsymbol{H}'\boldsymbol{s} \quad (4)$$

This optimization problem can be solved by LASSO [7]. The LASSO method reduces the estimation error by setting zero to the components that are not effective for the estimation of unknown weight vector. The estimated weight vector easily becomes sparse by using the 1-norm constraint.

D. Musical instrument identification

A musical instrument included in an input signal can be identified using a weight vector $\boldsymbol{s}$ estimated in the last subsection. A modeling vector is obtained as a product of the base matrix and a column vector $\boldsymbol{\varphi}_i(\boldsymbol{s})$ where all the values except the components for the class i are zero and the component for the i-th class are the same as those of the weight vector. An estimated class, that corresponds to a musical instrument, is finally obtained to minimize the residual error of the modeling vector and the input vector as:

$$\hat{i} = \mathrm{argmin}\left\|(\boldsymbol{y}' - \boldsymbol{H}'\boldsymbol{\varphi}_i(\boldsymbol{s}))\right\|_2 \quad (5)$$

IV. EXPERIMENT FOR MONOPHONIC SOUNDS

A. Experimental condition

For evaluation, we used the musical instrument sound database RWC-MDB-I-2001 [8]. Musical instrument sounds in the database were recorded in real environments. The summary of the data is shown in Table II. Both feature(A) and feature(B) are tested. The spectrum is used to apply polyphonic sounds because it has additibity of overlapping frequency components. The musical instruments used in the experiments were chosen from the category of various musical instruments. A weight vector used for identification is optimized by the linear programming toolbox cvx for MATLAB [9]. Identification accuracy rate is used to evaluate the method.

TABLE II
SPECIFICATIONS USED IN THIS PAPER.

Instrument names	Piano	Violin	Trumpet	Clarinet	Flute
Variation of manufacturer and performer	3	3	2	3	2
Articulation	Normal		Normal and Vibrato		
Intensity	Normal: Forte and Mezzo Vibrato: Mezzo				
Number of tones (Sample data)	271	256	70	160	74
Number of tones (Test data)	87	64	34	40	37

TABLE III
EXPERIMENTAL RESULT (IDENTIFICATION ACCURACY) FOR MONOPHONIC SOUND IDENTIFICATION USING FEATURE(A).

Identification method	Pf	Vn	Tr	Cl	Fl	average
SVM	100%	100%	97.1%	100%	78.4%	95.1%
The proposed method	95.4%	100%	88.2%	100%	75.7%	91.9%

TABLE IV
EXPERIMENTAL RESULT (IDENTIFICATION ACCURACY) FOR MONOPHONIC SOUND IDENTIFICATION USING FEATURE(B).

Identification method	Pf	Vn	Tr	Cl	Fl	average
SVM	100%	96.9%	41.2%	100%	83.8%	84.4%
The proposed method	100%	100%	61.8%	100%	73.0%	87.0%

B. Experimental result

The identification performance is summarized in Table III and IV. In the tables, the average accuracy of each class is shown. In the Table III, the average accuracy using SVM was 95.1%, while in the proposed method, the accuracy was 91.9%, a little lower than SVM. In contrast, in Table IV, the average accuracy rate of the proposed method obtained was higher than that of SVM. Since the proposed method achieved roughly 90% identification accuracy, and the performance is almost the same as SVM, it is found out that our proposed method for musical instrument identification is effective.

V. EXPERIMENT FOR POLYPHONIC SOUNDS

A. Experimental condition

Evaluation sounds were created by mixing two instrument sounds used in the previous experiment. In the experiment, feature(B) is used because the spectrum has additibity in a frequency domain. The polyphonic sounds consisted of two

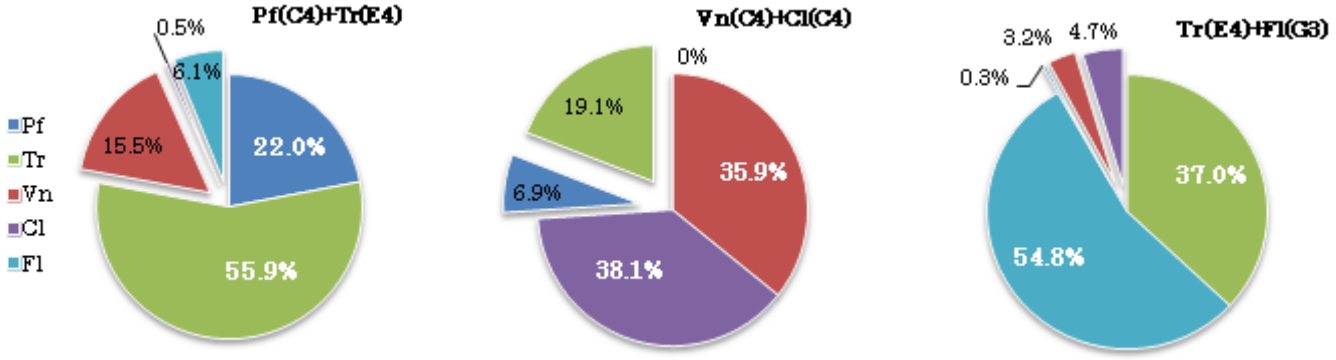

Fig. 4 A mixture ratio for polyphonic sounds consist two instruments. The input (correct) classes are written in white.

pitches (C4,E4,G4). 90 patterns of the polyphonic sounds were created by combination of each pitch and instrument. The accuracy is evaluated by a mixture ratio of each instrument.

B. Experimental result

46 combinations were estimated correctly in 90 patterns and the accuracy was 51.1%. One of the two instruments was identified correctly in 87 patterns.

Fig. 4 depicts examples of mixture ratios for correctly identified cases. It is observed that the correct classes had high percentages compared to the others. These results show a possibility that the proposed method would be applicable to polyphonic sounds. On the other hand, application of SVM for polyphonic sound is difficult since every SVMs for possible combinations of instruments should be prepared.

The precondition of sparse representation is that an input signal can be represented as a linear combination of feature components in the base matrix. If a feature cannot be theoretically added to another, an SRC technique may not be applied, or the classification performance is damaged. In order to robustly identify musical instruments for polyphonic sounds, further investigation of the feature that can be simply added to another and achieve high performance is essential.

VI. Conclusions

In this paper we propose a musical instrument identification method based on sparse representation. Compared with conventional methods, our proposed approach has an advantage that sound source separation is not required and sound overlapping can be treated as a combination of features obtained from musical instruments contained in the polyphonic sounds. At first, we evaluated the effectiveness of the proposed method for monophonic sounds, then the average accuracy rate of 91.9% was obtained and it was roughly as same as SVM. And in the case of using the spectrum divided by subband as the features, the accuracy of the proposed method higher than of SVM. It is found that the proposed method is thus effective. Secondly, we evaluated the performance for polyphonic sounds mixed with two instrument sounds. 46 samples were correctly estimated and one of the instrument was correctly identified in 87 samples among 90 patterns. A possibility that our method could be applicable for polyphonic sounds is found.

Our future work includes further investigation of robust features in order to accomplish the higher identification rate, evaluation of the method for the other instruments, and experiments for polyphonic sounds containing more than three instrument sounds. Also it is necessary to compare the proposed method to other methods for polyphonic sounds.

References

[1] T. Kitahara et al., "*Musical instrument identification based on F0-dependent multivariate normal distribution,*" *Proc. ICASSP 2003*, vol.5, pp.421–424, 2003.

[2] T. Kitahara et al., *identification in polyphonic music: feature weighting with mixed sounds, pitch dependent timbre modeling, and use of musical context,*" *Proc. ISMIR2005,* pp.558–563, 2005.

[3] Y. Panagakis et al., "*Music genre classification via sparse representations of auditory temporal modulation,*" *Proc. EUSIPCO,* pp.1-5, 2009.

[4] J. Wright et al., "*Robust face recognition via sparse representation,*" *IEEE Trans. On Pattern Analysis and Machine Intelligence,* vol.31, pp.210–227, 2009.

[5] M D. Plumbley et al., "*Sparse Representations of Polyphonic Music,*" *Signal Processing,* ELSEVIER, vol.86, pp.417-431, 2009.

[6] G. Grindlay et al., "*A probabilistic subspace model for multi-instrument polyphonic transcription,*" *Proc. ISMIR2010,* pp.21–26, 2010.

[7] R. Tibshirani, "*Regression shrinkage and selection via the lasso,*" *Journal of the Royal Statistical Society.* Series B, vol.58, no.1, pp.267-288, 1996.

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

[9] M. Grant, et al., "*CVX: Matlab Software for Disciplined Convex Programming,*"