MUSICAL INSTRUMENT IDENTIFICATION BASED ON F0-DEPENDENT MULTIVARIATE NORMAL DISTRIBUTION

Tetsuro Kitahara,[†] *Masataka Goto,*[‡] and *Hiroshi G. Okuno*[†]

[†]Dept. of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

[‡]"Information and Human Activity", PRESTO,
JST / National Institute of Advanced
Industrial Science and Technology

kitahara@kuis.kyoto-u.ac.jp m.goto@aist.go.jp okuno@i.kyoto-u.ac.jp

ABSTRACT

The *pitch dependency* of timbres has not been fully exploited in musical instrument identification. In this paper, we present a method using an *F0-dependent multivariate normal distribution* of which mean is represented by a function of fundamental frequency (F0). This F0-dependent mean function represents the pitch dependency of each feature, while the F0-normalized covariance represents the non-pitch dependency. Musical instrument sounds are first analyzed by the F0-dependent multivariate normal distribution, and then identified by using the discriminant function based on the Bayes decision rule. Experimental results of identifying 6,247 solo tones of 19 musical instruments by 10-fold cross validation showed that the proposed method improved the recognition rate at individual-instrument level from 75.73% to 79.73%, and the recognition rate at category level from 88.20% to 90.65%.

1. INTRODUCTION

Musical instrument identification is an important subtask for many applications including auditory scene analysis and multimedia retrieval as well as for reducing ambiguities in automatic music transcription. The difficulties in musical instrument identification reside in the fact that some features depend on pitch and individual instruments. In particular, timbres of musical instruments are obviously affected by the pitch due to their wide range of pitch. For example, the pitch range of the piano covers over seven octaves.

To attain high performance of musical instrument identification, it is indispensable to cope with this *pitch dependency* of timbre. Most studies on musical instrument identification, however, have not dealt with the pitch dependency [1]–[6]. Martin used 31 features including spectral and temporal features with hierarchical classification and attained about 70% of identification by the benchmark of 1,023 solo tones of 14 instruments. He pointed out the importance of the pitch dependency, but left it as future work [1]. Eronen *et al.* used spectral and temporal features as well as cepstral coefficients used by Brown [2] and attained about 80% of identification by the benchmark of 1,498 solo tones of 30 instruments [3]. They treated the pitch as one element of feature vectors, but did not cope with the pitch dependency. Kashino *et al.* also treated the pitch similarly in their automatic music transcription system [4]. They also coped with the difference of individual instruments, but did not deal with the pitch dependency [5].

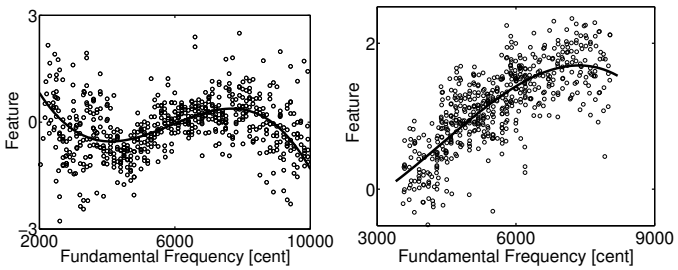
In this paper, to take into consideration the pitch dependency of timbre in musical instrument identification, each feature or basic vector of features is represented by an *F0-dependent multivariate normal distribution* of which mean is represented by a function of fundamental frequency (F0). This *F0-dependent mean function* represents the pitch dependency of each feature, while the *F0-normalized covariance* represents the non-pitch dependency. Musical instrument identification is performed both at individual-instrument level and at non-tree category level by a discriminant function based on the Bayes decision rule.

The rest of this paper is organized as follows: Section 2 proposes the F0-dependent multivariate normal distribution, and Section 3 describes the features and the discriminant function used in this paper. Sections 4 and 5 report the experimental results, and finally Section 6 concludes this paper.

2. F0-DEPENDENT MULTIVARIATE NORMAL DISTRIBUTION

The distribution of tone features in the feature space is represented by an *F0-dependent multivariate normal distribution* with two parameters: the *F0-dependent mean function* and *F0-normalized covariance*. The reason why the mean of the distribution is approximated as a function of F0, that is an *F0-dependent mean function*, is that tone features at different pitches have different positions (means) of distributions in the feature space. In this paper, the F0-dependent

This research was partially supported by MEXT, Grant-in-Aid for Scientific Research (B), No.12480090, and Informatics Research Center for Development of Knowledge Society Infrastructure (COE program of MEXT, Japan)



(a) Piano's 4th basic vector of features. (b) Cello's first basic vector of features.

Fig. 1. Examples of F0-dependent mean functions.

mean function for each musical instrument ω_i, $\boldsymbol{\mu}_i(f)$, is approximated as a cubic polynomial by using the least squares method. For example, piano's fourth basic vector of features and cello's first basic vector are depicted in **Fig. 1** (a) and (b), respectively.

On the other hand, the non-pitch dependency of each feature is represented by the *F0-normalized covariance*. Since the F0-dependent mean function represents the mean of features, the covariance obtained by subtracting the mean from each feature eliminates the pitch dependency of features. For each musical instrument ω_i, the F0-normalized covariance Σ_i is defined as follows:

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

where $'$ is the transposition operator, χ_i and n_i are the set of the training data of the instrument ω_i and its total number, respectively. $f_{\boldsymbol{x}}$ denotes the F0 of the data $\boldsymbol{x}$.

3. FEATURES AND A DISCRIMINANT FUNCTION

3.1. Features for Musical Instrument Identification

We used spectral, temporal, and modulation features as well as non-harmonic component features resulting in 129 features in total listed in **Table 1**. The features except the non-harmonic component features are determined by consulting the literatures [1, 3, 4]. The non-harmonic component features are original and have not been used in the literature. We incorporated features as many as possible, since the feature space is transformed to a lower-dimensional space.

Each musical instrument sound sampled by 44.1 kHz with 16 bits are first analyzed by STFT (short time Fourier transform) with Hanning windows (4096 points) for every 10 ms, and spectral peaks are extracted from the power spectrum. Then, the F0 and the harmonic structure is obtained from these peaks.

The number of dimensions of the feature space is reduced by principal component analysis (PCA): the 129-dimensional space is reduced to a 79-dimensional space with the proportion value of 99%. It is further reduced to the minimum dimension by linear discriminant analysis (LDA).

Table 1. Overview of 129 features.

(1)	**Spectral features** (40 features)
	e.g., Spectral centroid, Relative power of the fundamental component, Relative power in odd and even components
(2)	**Temporal features** (35 features)
	e.g., Gradient of a straight line approximating power envelope, Average differential of power envelope during onset
(3)	**Modulation features** (32 features)
	e.g., Amplitude and frequency of AM, FM, modulation of spectral centroid and modulation of MFCC
(4)	**Non-harmonic component features** (22 features)
	e.g., Temporal mean of kurtosis of spectral peaks of each harmonic component (Their values become lower as sounds contain more non-harmonic components.)

In this paper, the space is reduced to an 18-dimensional space, since we deal with 19 instruments.

3.2. A Discriminant Function for the F0-dependent Multivariate Normal Distribution

Once parameters of the F0-dependent multivariate normal distribution are estimated, the Bayes decision rule is applied to identify the musical instrument or category of instruments. The discriminant function $g_i(\boldsymbol{x}; f)$ for the musical instrument ω_i is defined by

$$g_i(\boldsymbol{x}; f) = \log p(\boldsymbol{x}|\omega_i; f) + \log p(\omega_i; f), \qquad (1)$$

where $\boldsymbol{x}$ is an input data, $p(\boldsymbol{x}|\omega_i; f)$ is a probability density function (PDF) of this distribution and $p(\omega_i; f)$ is a priori probability of the instrument ω_i.

The PDF of this distribution is defined by

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

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

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

Substituting equation (2) into equation (1), thus, generates the discriminant function $g_i(\boldsymbol{x}; f)$ as follows:

$$g_i(\boldsymbol{x}; f) = -\frac{1}{2} D^2(\boldsymbol{x}, \boldsymbol{\mu}_i(f)) - \frac{1}{2} \log |\Sigma_i| - \frac{d}{2} \log 2\pi + \log p(\omega_i; f).$$

The name of the instrument that maximizes this function, that is ω_k satisfying $k = \mathrm{argmax}_i\, g_i(\boldsymbol{x}; f)$, is determined as the result of musical instrument identification.

The a priori probability $p(\omega_i; f)$ represents whether the pitch range of the instrument ω_i includes f, that is,

$$p(\omega_i; f) = \begin{cases} 1/c & (\text{if } f \in R_i) \\ 0 & (\text{if } f \notin R_i) \end{cases}$$

where R_i is the pitch range of the instrument ω_i, and c is the normalizing factor to satisfy $\sum_i p(\omega_i; f) = 1$.

Table 2. Contents of the database used in this paper.

Instrument names	Piano (PF), Classical Guitar (CG), Ukulele (UK), Acoustic Guitar (AG), Violin (VN) , Viola (VL), Cello (VC), Trumpet (TR), Trombone (TB), Soprano Sax (SS), Alto Sax (AS), Tenor Sax (TS), Baritone Sax (BS) , Oboe (OB), Fagotto (FG), Clarinet (CL), Piccolo (PC), Flute (FL), Recorder (RC)
Individuals	3 individuals except TR, OB, FL. TR, OB, FL: 2 individuals.
Intensity	Forte, normal, piano.
Articulation	Normal articulation style only.
Number of tones	PF: 508, CG: 696, UK: 295, AG: 666, VN: 528, VC: 558, TR: 151, TB: 262, SS: 169, AS: 282, TS: 153, BS: 215, OB: 151, FG: 312, CL: 263, PC: 245, FL: 134, RC: 160.

Table 3. Categorization of 19 instruments.

Categories	Instruments
Piano	Piano
Guitars	Classical Guitar, Ukulele, Acoustic Guitar
Strings	Violin, Viola, Cello
Brasses	Trumpet, Trombone
Saxophones	Soprano Sax, Alto Sax, Tenor Sax, Baritone Sax
Double Reeds	Oboe, Faggoto
Clarinet	Clarinet
Air Reeds	Piccolo, Flute, Recorder

4. EXPERIMENTS AND RESULTS

4.1. Experimental Conditions

Musical instrument identification is performed not only at individual-instrument level but also at category level to evaluate the improvement of recognition rates by the proposed method based on the F0-dependent multivariate normal distribution. The recognition rate was obtained by 10-fold cross validation. We compared the results by the method using usual multivariate normal distribution (called *baseline*) with those by the method using the proposed F0-dependent multivariate normal distribution (called *proposed*).

The benchmark used for evaluation is a subset of the large musical instrument sound database RWC-MDB-I-2001 developed by Goto *et al.* [7, 8]. This subset summarized in **Table 2** was selected by the quality of recorded sounds and consists of 6,247 solo tones of 19 orchestral instruments. All data are sampled by 44.1 kHz with 16 bits.

The categories of musical instruments summarized in **Table 3** are determined based on the sounding mechanism of instruments and existing studies [1, 3]. The category of instruments is useful for some applications including music retrieval. For example, when a user wants to find a piece of piano solo on a music retrieval system, the system can reject pieces containing instruments of different categories, which can be judged without identifying individual instrument names.

4.2. Results of Musical Instrument Identification

Table 4 summarizes the recognition rates by both the *baseline* and *proposed* methods. The proposed F0-dependent method improved the recognition rate at individual-instrument level from 75.73% to 79.73% and reduced recognition errors by 16.48% in average. At category level, the proposed method improved the recognition rate from 88.20% to 90.65% and reduced recognition errors by 20.67%. The observation of these experimental results is summarized below:

Improvement by the pitch dependency

The recognition rates of six instruments (PF, TR, TB, SS, BS, and FG) were improved by more than 7%. In particular, the recognition rate for pianos was improved by 9.06%, and its recognition errors were reduced by 35.13%. This big improvement was attained, since their pitch dependency is salient due to their wide range of pitch.

Difference between accuracy at two levels

The recognition rates of the four types of saxophones at individual-instrument level (47–73%) were lower than those at category level (77–92%). This is because sounds of those saxophones were quite similar. In fact, Martin reported that sounds of various saxophones are very difficult for the human to discriminate [1].

Instrument-dependent difficulty of identification

Since we adopt the flat (non-hierarchical) categorization, the recognition rates at category level depend on the category. The recognition rates of guitars and strings at category level were more than 94%, while those of brasses, saxophones, double reeds, clarinet and air reeds were about 70–90%. This is because instruments of these categories have similar sounding mechanism: these categories are subcategories of "wind instruments" in conventional hierarchical categorization.

5. EVALUATION OF THE BAYES DECISION RULE

The effect of the Bayes decision rule in musical instrument identification was evaluated by comparing with the 3-NN rule (3-nearest neighbor rule) with/without LDA. Three variations of the dimension reduction are examined:

(a) reduction to 79 dimension by PCA,
(b) reduction to 18 dimension by PCA, and
(c) reduction to 18 dimension by PCA and LDA.

The last one is adopted in the proposed method.

The experimental results listed in **Table 5** showed that the Bayes decision rule performed better in average than the 3-NN rule. Some observation are as follows:

Table 4. Accuracy by usual distribution (baseline) and F0-dependent distribution (proposed).

	Individual-instrument level			Category level		
	Usual	F0-dpt	diff.	Usual	F0-dpt	diff.
PF	74.21%	83.27%	+9.06%	74.21%	83.27%	+9.06%
CG	90.23%	90.23%	±0.00%	97.27%	97.13%	−0.14%
UK	97.97%	97.97%	±0.00%	97.97%	98.31%	+0.34%
AG	81.23%	83.93%	+2.70%	94.89%	95.65%	+0.76%
VN	69.70%	73.67%	+3.97%	98.86%	99.05%	+0.19%
VL	73.94%	76.27%	+2.33%	93.22%	94.92%	+1.70%
VC	73.48%	78.67%	+5.19%	95.16%	96.24%	+1.08%
TR	73.51%	82.12%	+8.61%	76.82%	85.43%	+8.61%
TB	76.72%	84.35%	+7.63%	85.50%	89.69%	+4.19%
SS	56.80%	65.89%	+9.09%	73.96%	80.47%	+6.51%
AS	41.49%	47.87%	+6.38%	73.76%	77.66%	+3.90%
TS	64.71%	66.01%	+1.30%	90.20%	92.16%	+1.96%
BS	66.05%	73.95%	+7.90%	81.40%	86.05%	+4.65%
OB	71.52%	72.19%	+0.67%	75.50%	74.83%	−0.67%
FG	59.61%	68.59%	+8.98%	64.74%	71.15%	+6.41%
CL	90.69%	92.07%	+1.38%	90.69%	92.07%	+1.38%
PC	77.56%	81.63%	+4.07%	89.39%	90.20%	+0.81%
FL	81.34%	85.07%	+3.73%	82.09%	85.82%	+3.73%
RC	91.88%	91.25%	−0.63%	92.50%	91.25%	−1.25%
Ave.	75.73%	79.73%	+4.00%	88.20%	90.65%	+2.45%

Usual: Usual (F0-independent) distribution (baseline)
F0-dpt: F0-dependent distribution (proposed)

Table 5. Accuracy by 3-NN rule and the Bayes decision rule.

	3-NN rule			Bayes decision rule		
	(a)	(b)	(c)	(a)	(b)	(c)
PF	53.94%	46.46%	63.39%	55.91%	59.06%	83.27%
CG	79.74%	77.16%	75.72%	98.28%	97.27%	90.23%
UK	94.58%	92.54%	97.63%	67.12%	80.00%	97.97%
AG	95.05%	92.79%	97.00%	19.97%	44.14%	83.93%
VN	47.73%	46.02%	45.83%	89.58%	84.47%	73.67%
VL	55.93%	54.24%	61.86%	71.19%	79.24%	76.27%
VC	86.20%	85.84%	84.23%	45.16%	30.82%	78.67%
TR	36.42%	38.41%	47.02%	41.72%	72.85%	82.12%
TB	70.99%	54.58%	77.86%	75.19%	78.24%	84.35%
SS	23.08%	14.20%	24.85%	48.52%	66.86%	65.89%
AS	37.59%	29.79%	40.43%	72.70%	41.84%	47.84%
TS	62.09%	66.01%	68.63%	30.07%	61.44%	66.01%
BS	68.84%	67.91%	66.98%	55.35%	54.42%	73.95%
OB	47.68%	48.34%	49.01%	43.71%	81.46%	72.19%
FG	64.10%	65.06%	74.36%	40.38%	30.12%	68.59%
CL	93.45%	87.93%	93.10%	95.51%	93.45%	92.07%
PC	84.08%	84.90%	84.08%	63.27%	58.37%	81.63%
FL	88.06%	72.39%	94.03%	35.82%	84.33%	85.07%
RC	97.50%	93.75%	97.50%	85.00%	96.25%	91.25%
Ave.	70.27%	66.98%	72.53%	62.11%	66.50%	79.73%

(a) Dimensionality reduction to 79 dim. using PCA only
(b) Dimensionality reduction to 18 dim. using PCA only
(c) Dimensionality reduction to 18 dim. using both PCA and LDA

(1) The Bayes decision rule with 79-dimension showed poor performance for AG, TR, SS, TS, OB and FL, since the number of their training data is not enough for estimating parameters of a 79-dimensional normal distribution. For such small training sets with 79-dimension, 3-NN is superior to the Bayes decision rule.

(2) LDA with the Bayes decision rule improved the accuracy of musical instrument identification from 66.50% to 79.73% in average. Although it seemed that PCA with 79-dimension performed better than LDA for CG, VN and AS, the cumulative performance of LDA for the categories of strings and saxophones is better than that of PCA.

6. CONCLUSIONS

In this paper, we presented a method for musical instrument identification using the *F0-dependent multivariate normal distribution* which takes into consideration the pitch dependency of timbre. The method improved the recognition rates at individual-instrument level from 75.73% to 79.73%, and at category level from 88.20% to 90.65% in average, respectively. The Bayes decision rule with dimension reduction by PCA and LDA also performed better than the 3-NN method.

Future works include evaluation of the method with different styles of playing, evaluation of the robustness of each feature against mixture of sounds, and automatic music transcription.

Acknowledgments: We thank everyone who has contributed to building and distributing the RWC Music Database (Musical Instrument Sound: RWC-MDB-I-2001) [7, 8]. We also thank Kazuhiro Nakadai and Hideki Asoh for their valuable comments.

7. REFERENCES

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

[2] J. C. Brown, "Computer Identification of Musical Instruments Using Pattern Recognition with Cepstral Coefficients as Features," *J. Acoust. Soc. Am.*, **103**, 3, pp.1933–1941, 1999.

[3] A. Eronen and A. Klapuri, "Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features," *Proc. of ICASSP*, pp.753–756, 2000.

[4] K. Kashino, K. Nakadai, T. Kinoshita and H. Tanaka, "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.

[5] 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.

[6] I. Fujinaga and K. MacMillan, "Realtime Recognition of Orchestral Instruments," *Proc. of ICMC*, 2000.

[7] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," *IPSJ SIG Notes*, 2002-MUS-45, pp.19–26, 2002. *(in Japanese)*

[8] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, "RWC Music Database: Popular, Classical, and Jazz Music Databases," *Proc. of ISMIR 2002*, pp.287–288, 2002.