Proceedings of the International Symposium on Musical Acoustics, March 31st to April 3rd 2004 (ISMA2004), Nara, Japan

3-S2-12

# Acoustical-similarity-based Musical Instrument Hierarchy and Its Application to Musical Instrument Identification

*Tetsuro Kitahara*[1], *Masataka Goto*[2], *Hiroshi G. Okuno*[1]

[1]Graduate School of Informatics, Kyoto University, Japan
[2]National Institute of Advanced Industrial Science and Technology, Japan
kitahara@kuis.kyoto-u.ac.jp m.goto@aist.go.jp okuno@i.kyoto-u.ac.jp

## Abstract

This paper describes a method of constructing a musical instrument hierarchy reflecting the similarity of acoustical features. Although this acoustical hierarchy is useful for various purposes as well as investigating the timbres of musical instruments, it has not been reported in the literature. The main issues in constructing such a hierarchy are what feature space is used and how to obtain the representative position of each instrument in the feature space. To solve the first issue, we use a feature space that facilitates identification of 6247 solo tones of 19 instruments with an accuracy of 80%. To solve the second issue, we approximate the distribution of each instrument using a large number of sounds. Experimental results using 6247 solo tones of 19 instruments show that the hierarchy obtained by our method is partially different from a conventional one. In addition, this paper reports category-level identification of non-registered instruments using our hierarchy.

## 1. Introduction

*Timbre*, which is also called *tone color* or *tone quality*, is one of the four basic psychoacoustical parameters of sounds: pitch, duration, intensity, and timbre. The timbre has been considered more complex than the other parameters and has never been fully definded. This is because it is difficult to put it on a physical scale, whereas the other parameters can easily be corresponded to physical values; the pitch, for example, can be corresponded to the fundamental frequency in usual.

One of the most useful approaches for understanding timbres of musical instruments is to construct a hierarchy (taxonomy) of musical instruments based on timbres. However, most of the existing hierarchies are not designed based on timbres. The hierarchy shown in *Table 1* is the most commonly used one, but the criterion for classifying musical instruments is not consistent. In fact, string, wind, and percussion are the material of instruments, the source that makes sounds, and the method of playing instruments, respectively. To solve this problem, Sachs [1] has proposed a new hierarchy based on the sounding mechanisms of musical instruments. However, since sounding mechanisms and timbres do not necessarily correspond, this hierarchy does not reflect the similarity of timbres.

*Table 1: A conventional hierarchy of musical instruments.*

| Higher level | Middle level | Lower level | Musical instruments* |
|---|---|---|---|
| Strings | — | Struck strings | PF |
| | | Plucked strings | CG, UK, AG |
| | | Bowed strings | VN, VL, VC |
| Winds | Wood winds | Air reeds | PC, FL, RC |
| | | Single reeds | SS, AS, TS, BS CL |
| | | Double reeds | OB, FG |
| | Brasses | (Rip reeds) | TR, TB |
| Percuss. | (omitted) | (omitted) | (omitted) |

*Notation of musical instruments is defined in *Table 3*.

There are two different strategies of constructing a timbre-based hierarchy. The first strategy is *perceptual classification* that constructs a hierarchy according to human-rated timbre similarity. In this case, the timbre means what humans feel about sounds. Although it was studied in the field of psychological acoustics [2]–[5], there were few reports of large-scale experiments because of the burden on human subjects. The second strategy is *acoustical classification* that automatically constructs a hierarchy on the basis of the similarity of acoustical features. In this case, the timbre means acoustical features of musical instruments. Although this strategy facilitates a large-scale hierarchy, it has not been reported because of the lack of large-scale musical sound databases.

In this paper, we adopt the second strategy and propose a method of constructing an acoustical hierarchy using a large-scale database, the RWC Music Database [6]. Our method uses a feature space where 6247 solo tones of 19 instruments can be identified with accuracy of 80% [7]; the method also approximates the distribution of each instrument using a large number of sounds. This feature space and approximatation make it possible to ob-

This research was partially supported by MEXT, Grant-in-Aid for Scientific Research (A), No.15200015, the Sound Technology Promotion Foundation, and COE program of MEXT, Japan.

tain an appropriate hierarchy that reflects the acoustical similarity of the timbre of musical instruments.

This paper also describes category-level identification of non-registered musical instruments (*i.e.*, instruments that are not included in the training data) as an application of our hierarchy. Although it is essential to deal with non-registered instruments in identifying musical instrument sounds, this has not been dealt with in previous studies [8]–[11]. Our method can identify the category name such as *strings* of a given sound even if it is not registered, as well as identifying the instrument name such as *violin* if it is registered.

## 2. Musical instrument hierarchy based on acoustical similarity

One of the most commonly used methods of constructing a hierarchy from feature vectors is *hierarchical clustering*. This first calculates distances between feature vectors in a feature space and then merges the closest pair of feature vectors (or clusters) into a single cluster recursively until all the feature vectors have been merged into a single cluster. This method is applicable for our purpose, but the following two problems make it difficult to obtain reasonable results:

**Problem 1** Clustering results depend on a feature space.

**Problem 2** If one sound is used as a representative of each musical instrument, the clustering results also depend on this sound. This is because features of musical instrument sounds depend on various factors including pitch and differences of individuals.

In this paper, to solve **Problem 1**, we use the feature space that we previously proposed [7], where 6247 solo tones of 19 instruments can be identified with accuracy of 80% [7]. To solve **Problem 2**, we perform hierarchical clustering on a multivariate normal distribution of each instrument, which is obtained from a large number of sounds. By using a multivariate normal distribution, instead of a single sound, for each instrument, we can obtain the appropriate representative position of the instrument in the feature space.

### 2.1. Details of the method

The hierarchy is constructed by the following three steps:

1. **Feature extraction**

The features proposed in our previous paper [7] are extracted. Specifically, the 129 features listed in *Table 2* are first extracted, and then the dimensionality of the 129-dimensional feature space is reduced by two successive processing steps: it is reduced to 79 dimensions by principal component analysis (PCA) with the proportion value of 99%, and then further reduced by linear discriminant analysis (LDA). The feature space is finally reduced to an 18-dimensional one because we deal with 19 instruments.

*Table 2: Overview of the 129 features used in [7].*

| | |
|---|---|
| (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 modulation of power, frequency, spectral centroid and MFCC |
| (4) | **Non-harmonic component features** (22 features) |
| | *e.g.*, temporal mean of kurtosis of spectral peaks of each harmonic component (their values decrease as sounds contain more non-harmonic components.) |

2. **Calculation of the Mahalanobis distances**

Once the distribution of each instrument $\omega_i$ in the feature space has been approximated by a multivariate normal distribution, the mean vector $\boldsymbol{\mu}_i$ and the covariance matrix $\Sigma_i$ of this distribution are calculated. The Mahalanobis distance $D_{\mathrm{M}}(\omega_i, \omega_j)$ of each instrument pair $(\omega_i, \omega_j)$ $(\omega_i \neq \omega_j)$ is calculated by the following equation:

$$D_{\mathrm{M}}(\omega_i, \omega_j) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' \Sigma_{i,j}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j),$$

where, $\Sigma_{i,j} = (\Sigma_i + \Sigma_j)/2$, and $'$ represents the transposition operator.

3. **Hierarchical clustering**

Hierarchical clustering is performed using the above Mahalanobis distances. We use the average-link clustering, which considers the distance between two clusters to be equal to the average distance from any member of one cluster to any member of the other.

### 2.2. Actual construction

We conducted experiments in constructing a hierarchy using a subset of a large-scale musical instrument sound database RWC-MDB-I-2001 [6]. This subset summarized in *Table 3* was selected by the quality of recorded sounds. It consists of 6247 solo tones of 19 orchestral instruments. All data were sampled at 44.1 kHz with 16 bits.

*Figure 1* shows the dendrogram obtained by our clustring method. We can obtain a hierarchy by merging instruments of which distances in *Figure 1* are less than a threshold into a single cluster. Higher, middle, lower level categories were obtained when the threshold was 30, 20, and 10, respectively.

The results are summarized below:

**Division into decayed and sustained instruments**

Our hierarchy divided all instruments into two categories: *decayed* and *sustained* instruments. This division matches reports on psychological acoustics [5] and manually constructed timbre-based hierarchies [8, 9]. This shows that our hierarchy approximately reflects the timbre similarity. This is one of the major differences between our hierarchy and a conventional one (*Table 1*).

*Table 3: 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 for each instrument except for 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. |

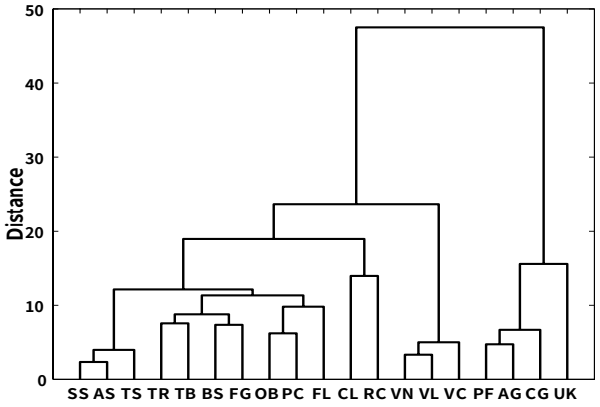


*Figure 1: The dendrogram obtained by hierarchical clustering.*

**Categories that consist of only one instrument**

Three instruments, the ukulele, the clarinet and the recorder, each formed a category singly at the lower level. The reason why the ukulele and the clarinet did so is that the Mahalanobis distances between them and others are large due to their peculiar characteristics. Ukuleles decay the fastest of the four decayed instruments. Clarinets have small powers of even-ordered harmonic components, especially 2nd one. On the other hand, the reason why the recorder did so is that the variance of the recorder's distribution is small. In recorders, the flow is fixed by the form of the narrow windway, while in flutes it is fixed by the form of the player's lips. Therefore, sounds of recorders do not vary much from player to player. This is why the variance of the recorder's distribution was small.

**Influence of pitch range**

In classifying wind instruments, instruments that have a similar pitch range tended to be placed into the same category. This result means that the features of musical instrument sounds depend on not only the sounding mechanisms but also the pitch. This matches the literature on psychological acoustics [5].

*Table 4: Our musical instrument hierarchy based on the acoustical similarity.*

| Higher level | Middle level | Lower level | Musical Instruments |
|---|---|---|---|
| Decayed | —— | Ukulele | UK |
| | | Others | PF, CG, AG |
| Sustained | Strings | —— | VN, VL, VC |
| | Winds | Saxophones | SS, AS, TS |
| | | Clarinet | CL |
| | | Recorder | RC |
| | | Brasses, etc. | TR, TB, BS, FG |
| | | Others | OB, PC, FL |

**Saxophones and Clarinets**

Although saxophones and clarinets have single reeds, our results show that their sounds are not similar. This is because clarinets are cylindrical while saxophones are conical. This shape difference causes spectral differences, especially powers of even-ordered harmonic components. While conventional hierarchies such as *Table 1* do not take these timbre differences into consideration, our hierarchy does.

## 3. Application to identification of non-registered instruments

In this section, as an application of our musical instrument hierarchy, we perform identification of instruments that are not included in the training data (called *non-registered instruments*).

Most studies on musical instrument identification [8]–[11] have used training data including a limited number of musical instruments and have assumed that all the instruments used in an input were included in the training data. Because there are numerous kinds of musical instruments in the world, it is impossible to prepare training data that covers all of them. Moreover, the recent development of digital audio technology has made it possible to create infinite kinds of original musical sounds (ranging from sounds similar to natural instruments to sounds of instruments that do not exist actually). It is therefore essential to deal with non-registered instruments when identifying musical instrument sounds.

To solve this problem, we propose category-level identification of non-registered instruments. For example, a sound that is similar to a violin and a viola but not the same (for example, a sound made from the two instruments using a synthesizer) is identified as "strings." When humans listen to this sound for the first time, they will think "I do not know this instrument, but it must be one of the strings." This study aims to achieve such human-like recognition on a computer.

*Table 5: Musical instrument sounds used for identifying non-registered musical instruments.*

| | |
|---|---|
| Sound names | Electric Piano (ElecPf),<br>Synth Strings (SynStr),<br>Synth Brass (SynBrs) |
| Variations | 2 variations for each sound name |
| Velocity | 100 |
| Pitch range | C3–C5 (A4=440Hz) |

In this experiment, electric sounds played by a MIDI tone generator (MU2000, Yamaha), listed in *Table 5*, were used as non-registered instruments, and sounds of natural instruments listed in *Table 3* were used as training data and test data of registered instruments. We divided all the data of *Table 3* into two groups at random and assigned one to training data and the other to test data of registered instruments.

Identification was performed in the following steps:

1. Identify a musical instrument sound at the instrument-name level;
2. Calculate the Mahalanobis distance from the sound to the distribution of the above result;
3. Judge it to be *registered* if the distance is less than a threshold or *non-registered* if the distance is not;
4. Output the instrument name as a result if the sound is judged to be registered, or the category name after re-identifying at the category level if the sound is judged to be non-registered.

Details of the identification method in step 1 are given in [7]. To calculate the Mahalanobis distances, we used a 23-dimensional feature space obtained by PCA. In step 4, the lower-level categories were used.

The experimental results listed in *Table 6* show that 77% of non-registered instrument sounds were correctly identified at the category level while distinguishing them from registered sounds. The recognition rate for ElecPf A was poor, because it was not recognized as a non-registered instrument but as a registered one. Actually, it sounds like a real piano to human listeners.

## 4. Conclusion

We constructed a musial instrument hierarchy according to the acoustical similarity and observed differences between it and the conventional hierarchy. Comparing these hierarchies with the literature on psychological acoustics showed that our hierarchy reflects the timbre similarity better than the conventional one. Because the literature we used was based on smaller databases than ours, however, there is still room for improvement in the way of this comparison. Future work will therefore include a psychoacoustical experiment using our large database and the comparison between its results and our hierarchy in more detail.

*Table 6: Experimental results on flexible musical instrument identification.*

| *Registered* | PF | CG | UK | AG | VN | VL | VC |
|---|---|---|---|---|---|---|---|
| Correct I | 69% | 83% | 97% | 68% | 62% | 69% | 70% |
| Correct II | 17% | 12% | 0% | 14% | 14% | 11% | 10% |
| Incorrect | 14% | 5% | 3% | 17% | 24% | 20% | 20% |
| | TR | TB | SS | AS | TS | BS | OB |
| Correct I | 64% | 63% | 47% | 40% | 30% | 49% | 48% |
| Correct II | 15% | 17% | 11% | 17% | 26% | 20% | 19% |
| Incorrect | 22% | 20% | 42% | 43% | 44% | 31% | 33% |
| | FG | CL | PC | FC | RC | Av. | |
| Correct I | 56% | 91% | 66% | 45% | 89% | 67% | |
| Correct II | 16% | 0% | 17% | 20% | 0% | 13% | |
| Incorrect | 27% | 9% | 17% | 35% | 11% | 20% | |

| *Non* | ElecPf | | SynStr | | SynBrs | | Av. |
|---|---|---|---|---|---|---|---|
| *registered* | A | B | A | B | A | B | |
| Correct II | 44% | 76% | 88% | 100% | 60% | 96% | 77% |
| Incorrect | 56% | 24% | 12% | 0% | 40% | 4% | 33% |

Correct I: Correct at instrument name level.
Correct II: Correct at category level while rejecting instrument-name-level results.


## Acknowledgments

We thank everyone who has contributed to building and distributing the RWC Music Database (Musical Instrument Sound: RWC-MDB-I-2001) [6]. We also thank Professor Haruhiro Katayose, Dr. Kunio Kashino and Dr. Kazuhiro Nakadai for their valuable comments.


## References


[1] C. Sachs, "*The History of Musical Instruments*," WW Norton & Co., 1940.

[2] L. Wedin and G. Goude, "Dimension Analysis of the Perception of Instrument Timbre," *Scand. J. Psychol.*, **13**, pp.228–240, 1972.

[3] J. M. Grey: "Multidimensional Perceptual Scaling of Musical Timbres," *JASA*, **61**, 5, pp.1270–1277, 1977.

[4] P. Toiviainen, *et al.*, "Musical Timbre: Similarity Ratings Correlate with Computational Feature Space Distances," *J. New Music Research*, **24**, pp.292–298, 1995.

[5] J. Marozeau, *et al.*, "The Dependency of Timbre on Fundamental Frequency," *JASA*, **114**, 5, pp.2946–2957, 2003.

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

[7] T. Kitahara, *et al.*, "Musical Instrument Identification based on F0-dependent Multivariate Normal Distribution," *Proc. ICASSP*, **V**, pp.421–424, 2003.

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

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

[10] 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.

[11] I. Fujinaga and K. MacMillan, "Realtime Recognition of Orchestral Instruments," *Proc. ICMC*, pp.141–143, 2000.