KEER2010, PARIS | MARCH 2-4 2010
INTERNATIONAL CONFERENCE ON KANSEI ENGINEERING AND EMOTION RESEARCH 2010

DESIGN OF AUDITORY PROCESSING AND SUBJECTIVE CLASSIFICATION FOR MUSIC

Masashi MURAKAMI[*a], Toshikazu KATO[a]

[a] *Chuo University, Japan*

ABSTRACT

We used Kansei engineering to analyze how humans relate to sound and music. We analyzed what features of sound humans paid attention to and how humans interpreted sound. At the physiological level of sound processing, sound is processed according to its auditory characteristics. At this level, humans don't interpret the image of the sound; there is no subjectivity. By using auditory characteristics, we investigated the features pertinent to the analysis of sound and music. We considered early processing in the auditory nervous system as extracting the changes in power, which are obtained from the segmentation of sound-signals by frequency band and time interval, and contrast, and the features obtained by that extraction. At the cognitive level, we analyzed the correlation of those features with words that subjects associated with the features. Through this modeling, we developed a method for retrieving sounds and music with similar associations, or images evoked by particular words.

Keywords: **Music, Retrieval, Contrast computing, Physiological filter**

---

[*] **Corresponding author**: 1-13-27, Kasuga, Bunkyo-ku, Tokyo 112-8551 Japan.

{masa_m, kato}@indsys.chuo-u.ac.jp

INTRODUCTION

Studies of music retrieval have been prompted by the increase in multimedia technology and content in recent years [1, 2]. However, retrieval to date is based on metadata, not on the music itself. As it cannot reveal musical similarities, listeners are not always satisfied with the music retrieved. Therefore, we attempted to analyze music by the attributes of auditory perception from the viewpoint of audiological psychology, which recognizes sound at a physiological level. With this approach, it should be possible to establish the features of music without the need for complicated models.

1. HIERARCHICAL KANSEI MODEL OF SENSITIVITY

We have found individual differences in the standard of interpretation revealed through the process of human perception, so we aimed to model this phenomenon [3-7]. Figure 1 shows our model of how humans process sound. The model has four levels: physical, physiological, psychological, and cognitive.

The physical level is based on the intrinsic features, such as the frequency, of a sound. This level comes before the perception of a sound.

The physiological level relies on the extraction of various features as represented by auditory perception and the nerve pathways.

The psychological level expresses or interprets the features of sound or music by classifying their weight. Features are obtained at the physiological level and by subsequent grouping.

Finally, the cognitive level interprets music by adopting an image word for each of the groups classified at the psychological level.

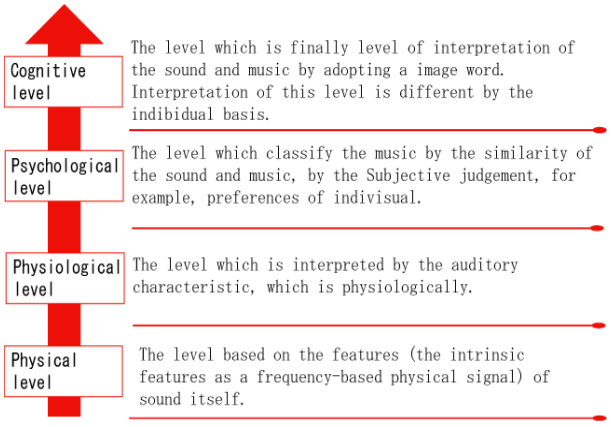

Figure 1: Hierarchical model of Kansei engineering for auditory information processing

2. SOUND AS A PHYSICAL SIGNAL

When conducting experiments with music, it is very important to consider the source of the sounds. Many computer synthesizers faithfully recreate the music written on scores [8, 9]. Yet we have a different impression of musical performances given by world-class musicians from that of the synthesizers. This difference can be explained by tempo fluctuations in musical performance [10].

3. AUDITORY FEATURE VALUES AT THE PHYSIOLOGICAL LEVEL

Humans notice the pitch, intensity, and tempo of music at the physiological level and evaluate them [11].

Pitch is not measured directly but is determined relatively. We used pitch as a feature value for classification. The intensity of sound is difficult to measure directly. Intensity also is measured relatively, on a logarithmic scale, by comparison of the intensity of two sounds. We used intensity as a feature value for classification. Humans also take notice of the time variation of sound. It is important to assess time variation in the study of auditory perception. We used it as a feature value also.

4. SIMILARITY RETRIEVAL AT THE PHYSIOLOGICAL LEVEL

For each feature value at the physiological level, we used a short-time Fourier transform. We conducted an experiment to retrieve similar music through the evaluation of feature values.

4.1. Short-time Fourier transform of music data

A power spectrum can be obtained by the use of a short-time Fourier transform. We divided the result of the short-time Fourier transform into several ranges by the following process to study the relationship between the time variation of sound and the attributes of auditory perception (Fig. 2).

(i) The frequency, which represents the vertical axis of the power spectrum, is divided into six bands (Table 1).

Table 1: Frequency spectrum ranges

Frequency name	Range(Hz)
Treble	5000 - 10000
Soprano	2600 – 5000
Midrange	320 – 2600
Baritone	160 – 320
Bass	40 – 160
Sub-bus	20 – 40

(ii) The time period, which represents the horizontal axis of the power spectrum, is divided by *s* seconds.

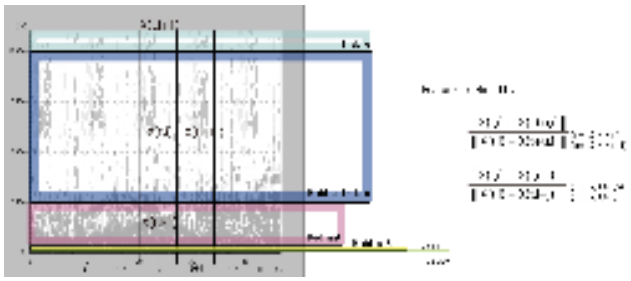

Figure 2: Division of spectrum by range and time

The value of *s* is established by equation (1) with the following conditions:

(A) The data sampling frequency is 22 050 Hz.
(B) The sampling frequency is halved in the process of the short-time Fourier transform.
(C) To implement the short-time Fourier transform, the number of data should be 2*n*:

$$s = \frac{2^n}{22050 \quad / 2} = \frac{2^{12}}{11025} \approx 0.37 \ (n = 12) \cdot \qquad (1)$$

At $n = 12$, the time period per divided range is equal to the time per note of a tune with a tempo of 160 beats per minute. We consider $n = 12$ to be the most appropriate value. The following data processing gives the statistics from the feature values for evaluation:

(a) According to the Weber–Fechner law, the amount of perception is proportional to the logarithm of the stimulus, so the power spectrum (PS) gives logPS.
(b) LogPS of each of the six ranges is divided by (1).
(c) LogPS is contrasted between the different ranges at time *t*.
(d) LogPS is contrasted between the different ranges at times *t* and *t* + 1.

To evaluate the significance of the feature values, we performed an experiment to retrieve similar music by the use of the statistics calculated from the values.

4.2. Summary of the experiment

Music samples were collected from the 576 tunes of the Sound Material Collection [12], excluding recordings of sound effects and voices. Fifty tunes were selected at random, and the experiment to retrieve similar music was performed using these tunes as the key tunes. The feature values were evaluated by seeking the average relevance ratio of the top 5, 10, 15, and 20 similar tunes. Further, to compare with the proposed feature values, the same experiment was performed using a power spectrum obtained from the fast Fourier transform as feature values.

4.3. Results

Table 2: Relevance ratios of the top 20 similar tunes

	→5th	→10th	→15th	→20th
Proposed feature values	72.8%	64.2%	56.4%	48.7%
Fast Fourier Transform	59.1%	53.2%	47.5%	42.3%

The relevance ratios (Table 2) were compared by Wilcoxon's rank-sum test, and the significance of differences was assessed. The relevance ratios differed at 1%.

5. SIMILARITY RETRIEVAL AT THE COGNITIVE LEVEL

To establish a Kansei model at the cognitive level, we conducted an experiment using "image words" to combine the subjects' responses and the physical features of the music.

5.1. Selection of image words and summary of experiment

A preliminary questionnaire was used to select image words frequently used for the evaluation of music. Table 3 lists the most commonly used words.

Table 3: Image words used in the experiment

Image words
vivid
fresh
tranquil
pleasant
monotonous
swinging

Next, we asked subjects to evaluate each word on a scale of 1–5 to determine how each word applied to 288 tunes selected at random from the collection of 576 tunes, and then created a model of auditory perception. We analyzed the physical feature values at a dimension of 216 without any modification and by the use of stepwise selection.

5.2. Results

When variable selection was applied to the image word "monotonous," it resulted in a low multiple correlation coefficient (Table 4). The other image words gave higher multiple correlation coefficients with variable selection. Wilcoxon's rank-sum test of the difference of the multiple correlation coefficients by variable selection showed no significance.

Table 4: Multiple correlation coefficients (adjusted for degrees of freedom) with and without variable selection

Image words	Variable selection performed	Variable selection not performed
vivid	0.5024	0.6534
fresh	0.6217	0.6516
tranquil	0.7997	0.8513
pleasure	0.5956	0.6842
monotonous	0.6661	0.5548
swinging	0.6817	0.7594

5.3. Evaluation of the Kansei model of auditory perception

We performed a retrieval experiment using the Kansei model of auditory perception that we established, and evaluated it by its retrieval precision. The experiment used 288 of the 576 tunes. The model was evaluated by determining the relevance ratio of the top 20 estimated values per image word. Further, we compared the multiple correlation coefficients by variable selection.

5.4. Results

Table 5: Relevance ratios of the top 20 tunes

Image words	Variable selection performed	Variable selection not performed
vivid	70.0%	80.0%
fresh	75.0%	75.0%
tranquil	80.0%	95.0%
pleasure	70.0%	80.0%
monotonous	80.0%	85.0%
swinging	80.0%	90.0%

The relevance ratio was higher in all image words when variable selection was performed (Table 5). Wilcoxon's rank-sum test of the difference of relevance ratios showed a significance of 1%. In addition, variable selection gave differences. Table 7 illustrates the high accuracy of the Kansei model based on the proposed feature values with variable selection.

6. IMPROVEMENT OF FEATURE VALUES

As the number of dimensions of the feature values decreased by only 1/10 when variable selection was performed, we consider that the proposed feature values are redundant. To seek the feature values necessary for each image word, we calculated the standard partial regression coefficient by multiple regression of the model, using the five most important feature values. Table 6 shows the example of "vivid". As the contrasts between various ranges of sound are ranked high on the axis of the "whole of midrange," both features are important over a wide range, from brief melody to the whole of the tune.

Table 7 shows the results for "fresh". Similar features to "vivid" are shown. As the 3rd and 4th terms show the whole of the midrange, all of the music is important. Table 8 shows the results for "tranquil". The sub-bass range appears and the treble range disappears. Therefore, the bass range of the whole tune is important.

Table 6: Top 5 terms in importance of "vivid"

Image terms	"Vivid"
1st	Whole of midrange
2nd	Contrast between soprano and midrange
3rd	Contrast between treble and bass
4th	Contrast between baritone and treble
5th	Contrast between soprano and treble

Table 7: Top 5 terms in importance of "fresh"

Image terms	"Fresh"
1st	Contrast between treble and bass
2nd	Contrast between soprano and midrange
3rd	Whole of midrange
4th	Whole of midrange
5th	Contrast between midrange and treble

Table 8: Top 5 terms in importance of “tranquil”

Image terms	“Tranquil”
1st	Contrast between bass and midrange
2nd	Whole of sub-bass
3rd	Contrast between midrange and soprano
4th	Contrast between midrange and soprano
5th	Contrast between bass and baritone

7. ACKNOWLEDGMENTS

In this study, we propose feature values of sound from the viewpoint of human attributes of auditory perception at the physiological level. The results show the possibility of classification by the use of the proposed feature values, independent of the name of the tune or the composer. We will work to improve the statistics, the divisions of the sound range, and experimental data. Through the use of this auditory perception model, we are attempting the quantification and classification of musical performances given by virtuosi and synthesizers.

REFERENCES

1. Tsuyoshi, T., Yosuke, M., Tomotada, I., Tetsunori, K.: A Conversational Robot utilizing Facial and Body Expressions. In: 2000 IEEE International Conference on System, Man and Cybernetics, pp. 858–863. IEEE Press, New York ,2000
2. Yuya, I., Tetsuji, T., Satoru, H.: Recommendation System of Musical Composition by the Use of the Grouping of Impression Words. In: National Convention of Japanese Society for Artificial Intelligence (20th)
3. Syunichi, K., Katsumasa S.: Research and Development of Sensitivity Agent and Human Media Data Base—Sensitivity Work Shop. In: System/Information/Control “Vol. 42, No. 5, pp. 253–259 ,1998
4. Kato, T.: Trans-category Retrieval Based on Subjective Perception Process Models. In: Proc.IEEE Multimedia and Expo ICME 2004, TP9–5 ,2004
5. Kouta, H., Tosikazu K.: Design Approach of the Retrieval System of Sensitivity; Sensitivity System Modeling. In: Journal of the Information Processing Society of Japan; Data Base; Vol. 46, No. SIG 19 (TOD 29) ,2005
6. Masahiro, T., Tosikazu K.: Modeling of the Sensitivity of Auditory Perception by the Use of Hierarchical Classification and Application for the Retrieval of Similar Image. In: Journal of the Information Processing Society of Japan Data Vol 44, No. SIG8 (TOD 18), pp. 37–45 ,2003
7. Yasuhiko, T., Tosikazu, K.: Feature Analysis of the Range of Similar Images and the Modeling of the Sensitivity of Auditory Perception. In: Journal of the Institute of Electronics, Information and Communication Engineers; D-II, Vol. J87-D-II, No. 10, pp. 1983–1995 ,2004

8. Yasuhiko, O., Kenichi, S., Yutaka, M., Yosinobu, K., Yasuo N.: Establishment of an Automatic Playing System for Piano, adding the Information of Musicians/Extraction of the Features of Performance in Topical Parts? by Neutral Network//. In: Study Report, Information Processing Society of Japan
9. Takayuki, H., Susumu H.: Automatic Playing of Music for Piano by use of Standard Performance Data. In: Study Report, Information Processing Society of Japan
10. Tetsuya, O.: Base of the Style of Musical Performance. In: Shunjusha Co., Ltd. pp 158–165, 1998
11. B.C.J.Moor, Kengo, O., (translation supervisor): Survey of Audiological Psychology. In: Seisinshobo Co., Ltd ,1994
12. On-Man-Tan DX. In: E-Frontier Co., Ltd