EMITTER International Journal of Engineering Technology
Vol. 5, No. 1, June 2017
ISSN: 2443-1168

A Time-Series Phrase Correlation Computing System With Acoustic Signal Processing For Music Media Creation

Keiichi Tsuneyama [†], Yasushi Kiyoki [‡]

[†] Faculty of Policy Management/Keio University
5322 Endo, Fujisawa, Kanagawa, Japan/+81-466-49-3404
E-mail: s13566kt@sfc.keio.ac.jp

[‡] Faculty of Environment and Information Studies/Keio University
5322 Endo, Fujisawa, Kanagawa, Japan/+81-466-49-3404
E-mail: kiyoki@sfc.keio.ac.jp

Abstract

This paper presents a system that analyzes the time-series impression change in the acoustic signal by a unit of music phrase. The aim is to support the music creation using a computer (computer music) by bringing out composers' potentially existing knowledge and skills. Our goal is to realize the cross-genre/cross-cultural music creation. Our system realizes the automatic extraction of musical features from acoustic signals by dividing and decomposing them into “phrases” and “three musical elements” (rhythm, melody, and harmony), which are meaningful for human recognition. By calculating the correlation between the target “target music piece” and the “typical phrase” in each musical genre, composers are able to grasp the time-series impression change of music media by the unit of music phrase. The system leads to a new creative and efficient environment for cross-genre/cross-cultural music creation based on the potentially existing knowledge on the music phrase and structure.

Keywords: Music Information Retrieval, Acoustic Signal Processing, Time-series Data Processing, Correlation Calculation

1. INTRODUCTION

In recent years, the environment surrounding the music creation using a computer (called *computer music*) has been greatly spreading in end-users. Everyone can start a music creation only with a music editing software and inexpensive musical instruments. In addition, W3C announced *Web Audio API* and *Web MIDI API* in 2013, which can process and synthesize audio and MIDI in Web applications. Through these efforts, more people will be able to access the world of computer music, and open music creation on the Web will spread to all over the world. The knowledge and skills to be needed for the music creation still depend on the personal inspiration and experience of

composers. Especially in the scene of computer music, this tendency seems to become stronger.

Based on this background, we propose a system for analyzing the time-series impression change in the acoustic signal by the unit of music *phrase* to support music media creation. The aim is to support the music creation by bringing out composers' potentially existing knowledge and skills. And our goal is to realize the cross-genre/cross-cultural music creation. This system extracts the musical feature from acoustic signals by dividing and decomposing the music piece to be a meaningful unit for human recognition – *phrase* and *three musical elements. Phrase* is a short-time music unit that gives some impression to human, such as bars and beats [1]. *Three musical elements* are melody, rhythm, and harmony, the basic structure of music [2]. By analyzing music with phrase and musical elements, composers are able to grasp the impression change of music media by the unit of music *phrase* and by each *musical element*.

2. RELATED WORKS

The following three research fields are included in related work of this research.

1. Acoustic Signal Processing
2. Impression-based Information Retrieval
3. Time-series Data Processing

In the field of Acoustic Signal Processing, classification of genre, automatic score creation, and auditory scene analysis have been studied as content-based music information retrieval [3][4][5]. In our system, the features extracted from acoustic signals by these methods are applied to calculate the correlation between phrases. In the field of Impression-based Information Retrieval, a "Semantic Computing" has been proposed [6][7][8], in which a computer interprets the semantics of multimedia as human. Our system is also positioned as a semantic computing system for multimedia in this sense. In the field of Time-series Data Processing, the processing and analyzing methods for communication signals, financial data, and so on [9][10]. Our system also treats acoustic signals as a time-series data.

3. APPROACH

The system calculates the correlations between two sets of acoustic signals by time-series and by a unit of phrase. The two sets of acoustic signals are extracted from a "target music piece (X)", which a composer expect to utilize as a reference for his/her music creation, and a "typical phrase (Y)" of each musical genre. The system analyzes the time-series impression change of a "target music piece (X)" by the unit of phrase, by the following steps. (Figure 1, Figure 2).

1. Dividing a "target music piece (X)" as a set of *phrases* based on the tempo of the music piece.
2. Decomposing a set of phrases into *three musical elements*: melody, rhythm, and harmony.
3. Calculating the correlation between each decomposed phrase of three musical elements and a "typical phrase (Y)" of each musical genre.

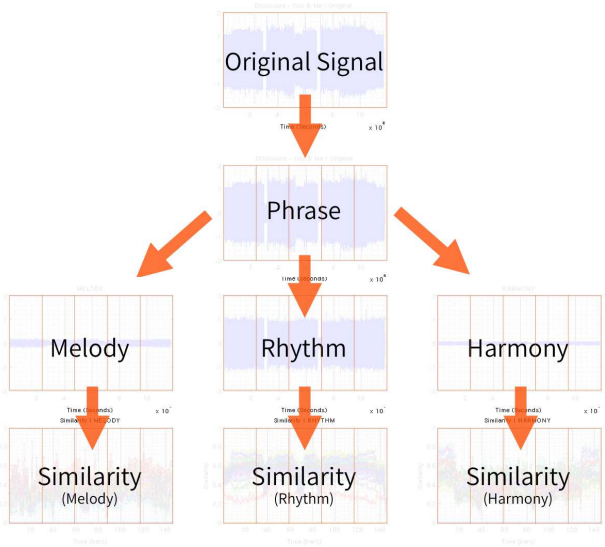

Figure 1. System overview

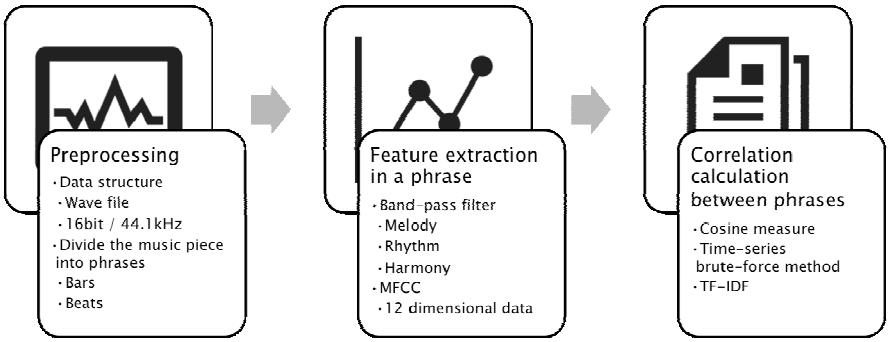

Figure 2. Implementation flow

The concrete use case of the system is as follows. When a composer selects a "target music piece (X)" as a query, the system calculates the correlations between the query and a large number of "typical phrases (Y)" stored in a "phrase database", and plots the analyzed results of time-series impression change in a graph. Also, the composer can play the "typical phrases (Y)" of analyzed results by time-series and by three musical elements on the Web application. Through this process, the composer is able to grasp the music structure of a "target music piece (X)". For example, the composer is able to understand the time-series impression change intuitively such as

"the rhythm of music piece P has Rock taste in the chorus, but Jazz taste in the verse" or "the melody of music piece Q has strong Pops elements in total, but Jazz tastes in the bridge", and so on.

3.1. Preprocessing

3.1.1. Data Structure

In the system, before the correlation calculation, the system convert acoustic signals into WAVE file (16bit/44.1kHz) format because it is needed to make the dimension of data consistent. This format is corresponding to the music data in general audio CD. In the step of correlation calculation, a "target music piece (X)" and "typical phrases (Y)" of each genre are transformed into [$n \times 44100$] matrixes with Fast Fourier Transform (FFT). The each line of this matrix represents a frequency spectrum (1Hz…44,100Hz) from *Phrase 1* to *Phrase n*, in which n depends on the length (the number of phrases) of the music piece. (Figure 3, Figure 4).

$$X, Y := \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ \vdots \\ P_n \end{array} \begin{array}{c} \begin{array}{cccccc} f_1 & f_2 & f_3 & \cdots & f_{44100} \end{array} \\ \begin{pmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \end{pmatrix} \end{array}$$

Figure 3. Data structure of target music data

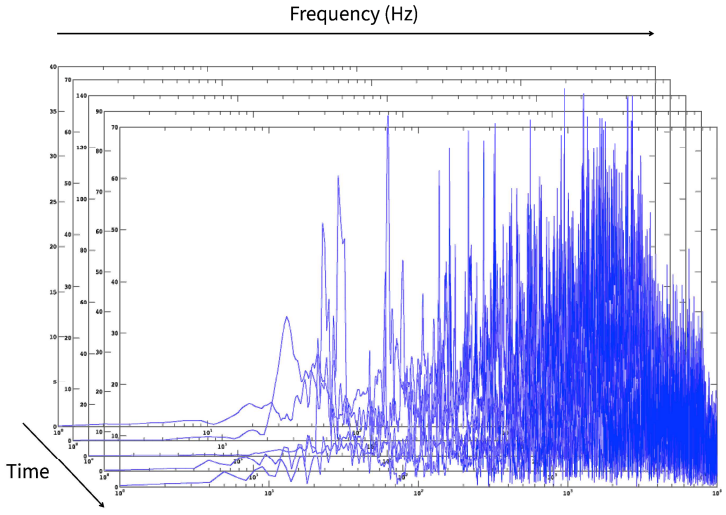

Figure 4. Time-series data matrixes (on a set of frequency spectrum)

3.1.2. Music Audio Tempo Estimation

In this step, the system analyzes the tempo of a music piece with a measure called *BPM (Beats Per Minute)*. BPM means the number of beats in a minute. BPM is calculated based on the gap of the amplitude between two adjoining frames in the acoustic signal. A frame is a divided signal by a very short period. By analyzing the tempo of a music piece, it is possible to divide the music piece into a meaningful phrase such as beat(s) and bar(s), which are recognizable by a human being. In our system, 4beats = 1bar is defined as a basic unit of phrase, assuming that all the input data are music in quadruple time.

3.1.3. Frequency Filtering

In this step, we apply *band-pass filtering* for acoustic signals to decompose a music piece into the three musical elements: melody, rhythm, and harmony. (Table 1, Figure 5). We select the frequency range to filter out to emphasize vocal sound for melody, high-hats and cymbals sound for rhythm, and bass sound for harmony [11].

Table 1. Band-pass filter ranges of each musical element

Musical Elements	Band-pass filter range
Harmony	0Hz-300Hz
Melody	200Hz-4000Hz
Rhythm	3540-10,000Hz

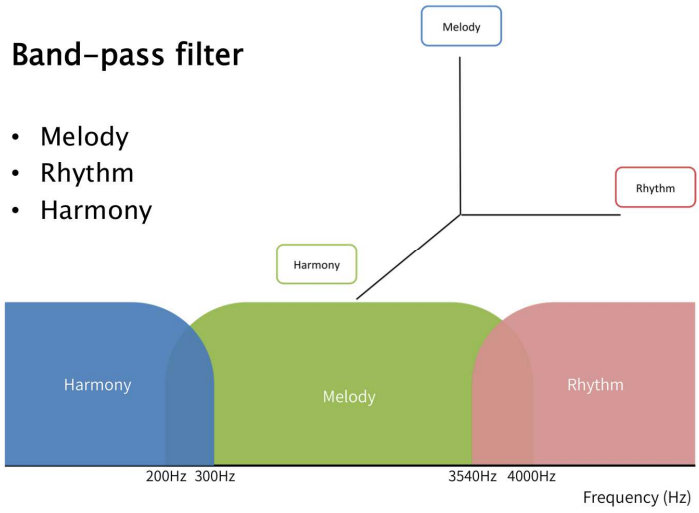

Figure 5. The conceptual image of band-pass filtering

3.2. Feature Extraction

3.2.1. Mel-Frequency Cepstrum Coefficients (MFCC)

The system extracts audio feature called *Mel-Frequency Cepstrum Coefficients (MFCC)* from each music phrase after preprocessing. MFCC has been widely used in the field of not only speech recognition but also music retrieval. [12] [13]. The advantages of MFCC are that 1) MFCC approximates the human auditory system's response; 2) MFCC can reduce the dimension of audio data. In this paper, we apply the following process to get MFCC feature and make the 12-dimensional matrix. (Figure 6)

1. Emphasize high frequency of the audio signal using pre-emphasis filter.
2. Take the Fast Fourier Transform (FFT) of the audio signal to make the power spectrums.
3. Compress the power spectrums by the Mel-filter bank.
4. Take the discrete cosine transform of compressed data.

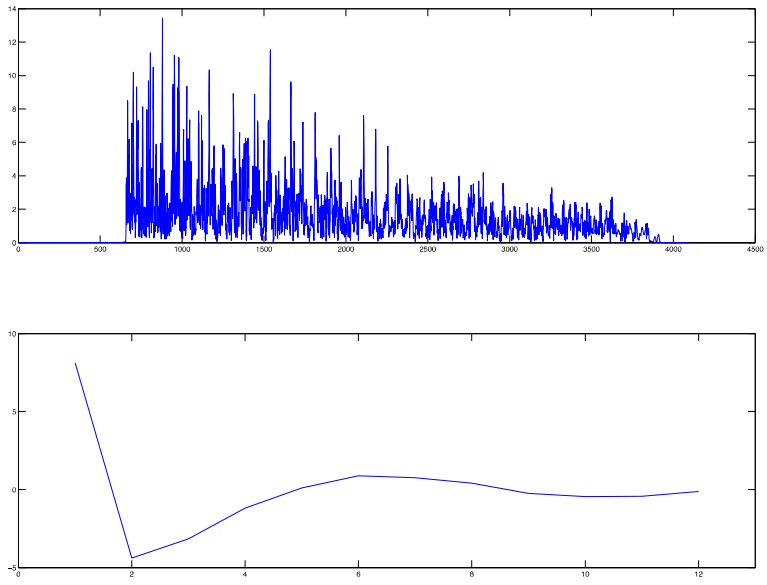

Figure 6. Frequency spectrum and 12-dimensional MFCC matrix

Correlation Calculation

3.2.2. Cosine Measure and Brute-force Method

The system calculates the correlation between each decomposed *phrase* in *three musical elements* of a “target music piece (X)” and a “typical phrase (Y)” of each musical genre by the following steps.

1. Cosine measure
2. Brute-force method (Figure 7)

Cosine measure has a merit that the calculation results are normalized, and Brute-force method has a merit that the calculation results does not depend on the length of each music piece. This kind of calculation method

applying to time-series data has been proposed in the field of signal processing and music retrieval. [9][10]

When a "target music piece (X)" has *phrase 1* to *phrase u* ($P_1 \ldots P_u$) and a "typical phrase (Y)" has *phrase 1* to *phrase v* ($P_1 \ldots P_v$) under a condition ($u>v$), Cosine Measure S_t for *phrase t* ($1 \le t < u-v$) is represented by a formula shown (1). $X.P_t$ and $Y.P_t$ mean a 12-dimensional MFCC data of a "target music piece (X)" and that of a "typical phrase (Y)" respectively. The value of Cosine measure S_t ranges $0.0 \le S_t \le 1.0$, and when the value becomes close to 1.0, it means that the two phrases are similar.

$$S_t = \frac{1}{v} \sum_{i=t}^{t+v-1} \frac{\overrightarrow{X.P_i} \cdot \overrightarrow{Y.P_{i-t+1}}}{\left|\overrightarrow{X.P_i}\right| \left|\overrightarrow{Y.P_{i-t+1}}\right|} \quad (u > v, 1 \le t < u - v) \tag{1}$$

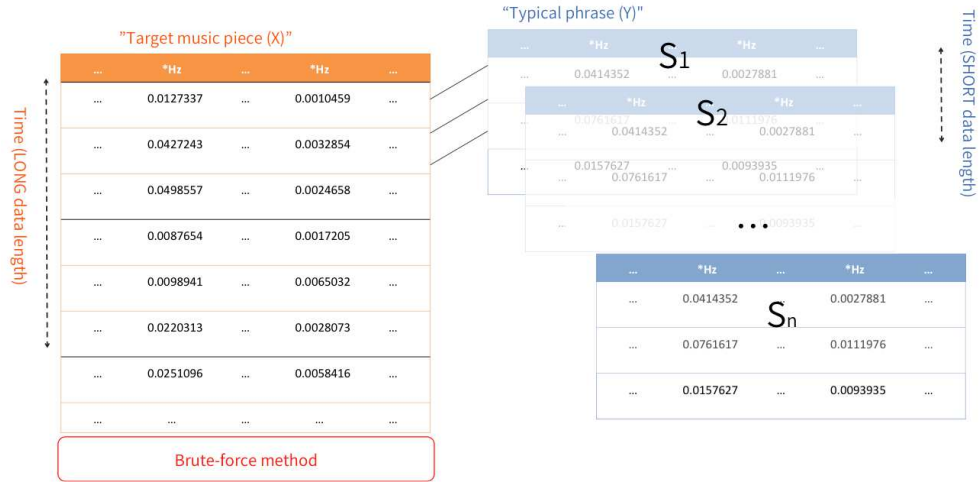

Figure 7. Correlation calculations with brute-force method

3.2.3. TF-IDF

In this step, we apply TF-IDF (Term Frequency*Inversed Document Frequency) method, which is widely used in text processing, to search the most characteristic phrase for the "target music piece (X)". In our system, we regard a "typical phrase (Y)" in a genre as a "word", and a "target music piece (X)" as "document". Then the system calculates the TF-IDF value for each phrase. For example, a music piece consisting of 70 phrases is processed as a document consisting of 70 words.

3.3. Visualization

3.3.1. Time-series Impression Change in a Graph

The system visualizes the result of correlation calculation in 2 rows (Figure 8). The 1st row shows the waveform of the "target music piece (X)" and the 2nd row shows the time-series impression change of "target music piece (X)". In the 1st row, the vertical axis represents the amplitude of the waveform, and the horizontal axis represents time on the scale of phrase

(4beats = 1bar, mentioned in *Music audio tempo estimation* section). In the 2nd row, the values of cosine measure *St* between "target music piece (X)" and each "typical phrase (Y)" are plotted in time-series by phrase. The vertical axis represents the range of S_t ($0.0 \leq S_t \leq 1.0$), and the horizontal axis represents time corresponding to the 1st row of the graph.

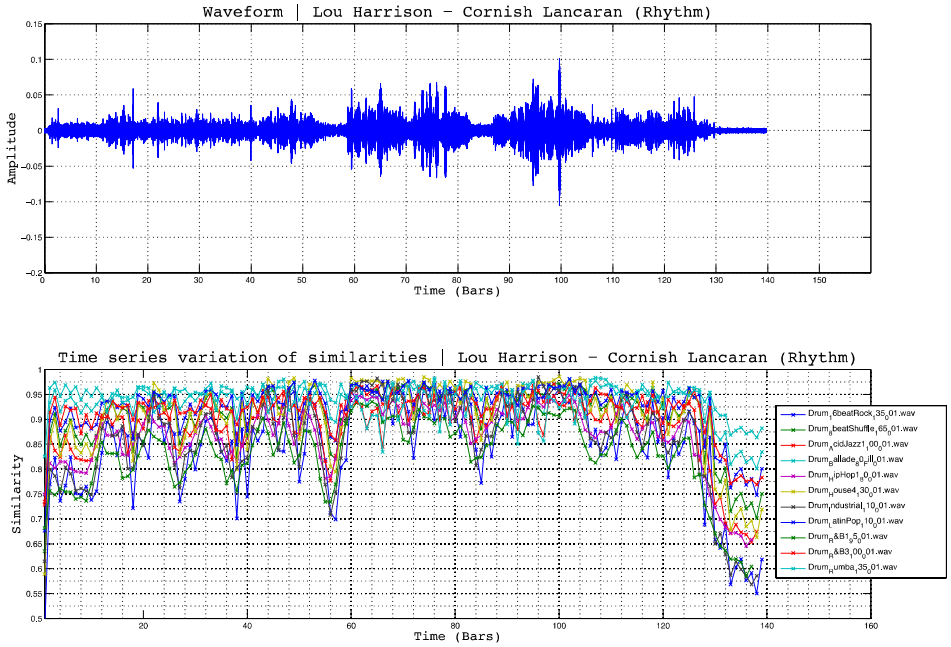

Figure 8. Time-series impression change in "target music piece (X)" expressed in a graph

3.3.2. Web Interface

Figure 9 shows the Web interface with 4 columns. The descriptions of each column are below.

1. Embed audio data of "target music piece (X)"
   - Can play the audio or video via web services such as YouTube, SoundCloud, and so on
   - Rotated 90 degrees to the right. When the user plays the sound, the time bar goes down to the bottom.
2. Rhythm element of "typical phrase (Y)" in highest St value by phrase
3. Melody element of "typical phrase (Y)" in highest St value by phrase
4. Harmony element of "typical phrase (Y)" in highest St value by phrase

On the Web interface, composers can play both "target music piece (X)" and "typical phrase (Y)" at the same time by phrase and three musical elements. Therefore, this Web interface helps the composer to understand

the time-series impression change of "target music piece (X)" more intuitively.

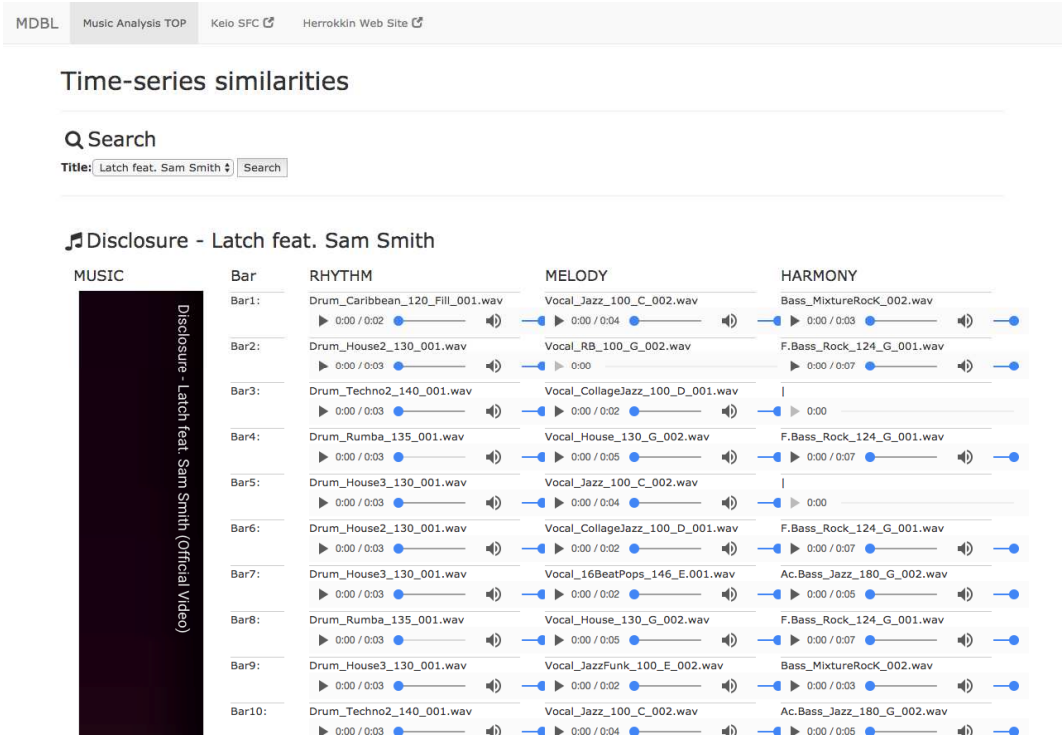

Figure 9. Web interface for composers

4. EXPERIMENTS

We performed two kinds of experiments below.

1. Evaluation as a musical genre classifier
2. Feasibility experiment using specific music pieces

In the 1st experiments, we evaluate our system as a musical genre classifier with the *average accuracy* [14]. In the 2nd experiments, we set the following two specific music pieces as "target music piece (X)" and evaluated the feasibility of our system.

- Eminem – "Lose Yourself"
- Gershwin – "Rhapsody In Blue"
- Donald Fagen – "I.G.Y."

The evaluation was performed by two parts: (1) plotting of time-series impression change, and (2) extraction of characteristic phrases of a "target music piece (X)" by TF-IDF.

4.1. Dataset for Experiments

The datasets shown in Table 2 were used for this evaluation experiment.

Table 2. Dataset for experiments

Group	Dataset name	Description
Target music piece (X)	GTZAN Genre Collection [15]	Music pieces classified by 10 musical genres. Each musical genre includes 100 music pieces.
Target music piece (X)	Eminem - "Lose Yourself"	Artist: Eminem Song title: Lose Yourself Genre: Rap Rock
Target music piece (X)	Gershwin - "Rhapsody In Blue"	Artist: Gershwin Song title: Rhapsody In Blue Genre: Symphonic Jazz
Target music piece (X)	Donald Fagen – "I.G.Y."	Artist: Donald Fagen Song title: I.G.Y. Genre: Jazz-Rock
Typical phrase (Y)	Cakewalk Music Creator 5	Powered by Roland. Includes over 3000 music phrases such as drum, bass, and vocal by musical genres

Experimental Results

4.1.1. Evaluation as a musical genre classifier

This experiment has been performed to evaluate whether the "target music pieces (X)" in GTZAN Genre Collection are classified into the correct genre labeled by the dataset. In this experiment, we set Hip-Hop, Jazz, Pops and Rock for the musical genre, and rhythm for the musical element for a band-pass filtering. The dataset for the "target music piece (X)" is GTZAN Genre Collection and for the "typical phrase (Y)" is Cakewalk Music Creator 5. We perform the experiment by the following step and calculate *average accuracy* [14] value.

1. Calculate the time-series similarities *St* between 400 music pieces (4 genres * 100 music pieces) in the GTZAN Genre Collection and 128 typical rhythm phrases (4 genres * 32 phrases) in the Cakewalk Music Creator 5.
2. Select the "typical phrase (Y)" in highest *St* value by each phrase of "target music piece (X)" and set them as predicted class.
3. Make the *confusion matrix* based on the result of step 2 to get *True Positive*, *True Negative*, *False Positive*, and *False Negative* values by each four genres.
4. Calculate *average accuracy* value by the following formula (2).

$$Average\ Accuracy = \frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l} \quad (2)$$

Where…
tp_i: *True Positive* for class C_i
tn_i: *True Negative* for class C_i
fp_i: *False Positive* for class C_i
fp_i: *False Negative* for class C_i
l: The number of classes

Table 3. The number of *TP*, *TN*, *FP*, and *FN* values by each four genres (Musical element: Rhythm)

Musical genre	True Positive	True Negative	False Positive	False Negative
Hip-Hop	1571	847	3818	207
Jazz	12	4514	412	1505
Pop	40	4743	110	1550
Rock	19	4424	461	1539

Table 3 shows the number of *True Positive*, *True Negative*, *False Positive* and *False Negative* values by each four genres. Based on Table 3 and Formula (2), the *average accuracy* is calculated at *0.627*. Though there is a tendency that rhythm phrases of Hip-Hop are calculated higher *St* values than that of other musical genres, the system has the basic capability to classify the rhythm element of four musical genres above correctly.

4.2. Feasibility experiment using specific music pieces

4.2.1. Eminem – Lose Yourself

4.2.1.1. Time-series Impression Change

Figure 10 shows the time-series impression change of the song in rhythm element. In the first half of the song in which the impression of rhythm is weak, the difference in St value by genre is small. On the other hands, after the middle of the song in which the impression of rhythm becomes stronger, the value of St varies depending on genre.

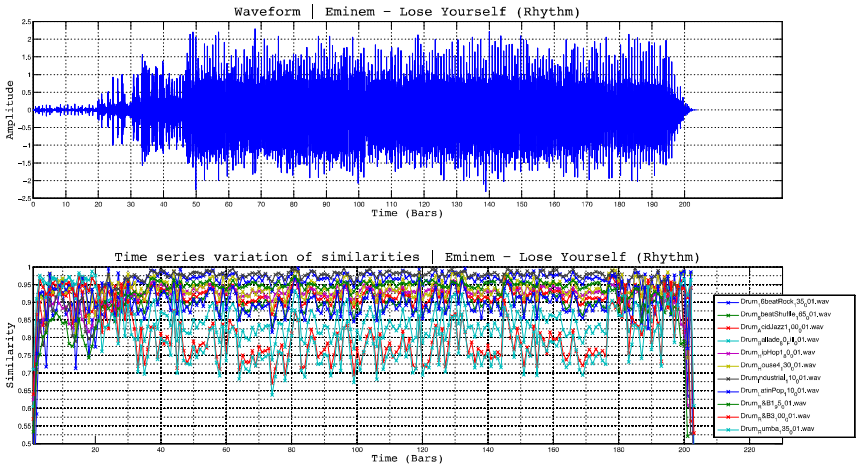

Figure 10. Time-series impression change of "Eminem – Lose Yourself" (Rhythm)

4.2.1.2. Most Characteristic Phrase in the "Target music piece (X)"

Table 4 shows that this music piece includes a characteristic rhythm phrase of "Drum_16beatRock_135_Fill_003.wav". This means that the system is able to extract the feature of Rock taste from the song that is known for "Rap Rock" genre.

Table 4. Characteristic phrases in "Eminem – Lose Yourself" (Rhythm)

Typical phrase (Y) name	TF-IDF Value
Drum_8beatPop3_145_003.wav	0.068
Drum_16beatRock_135_Fill_003.wav	0.037
Drum_SwingJazz2_110_Fill_002.wav	0.027

4.2.2. Gershwin - Rhapsody In Blue

4.2.2.1. Time-series Impression Change

Figure 11 shows that the first half of the music (1st – 20th bar), which gives impression calm, has low similarity to any typical phrase of the rhythm element. On the other hands, the last half of the music (after 21st bar), which gives impression dynamic/active, has a high similarity to the typical phrases of rhythm element.

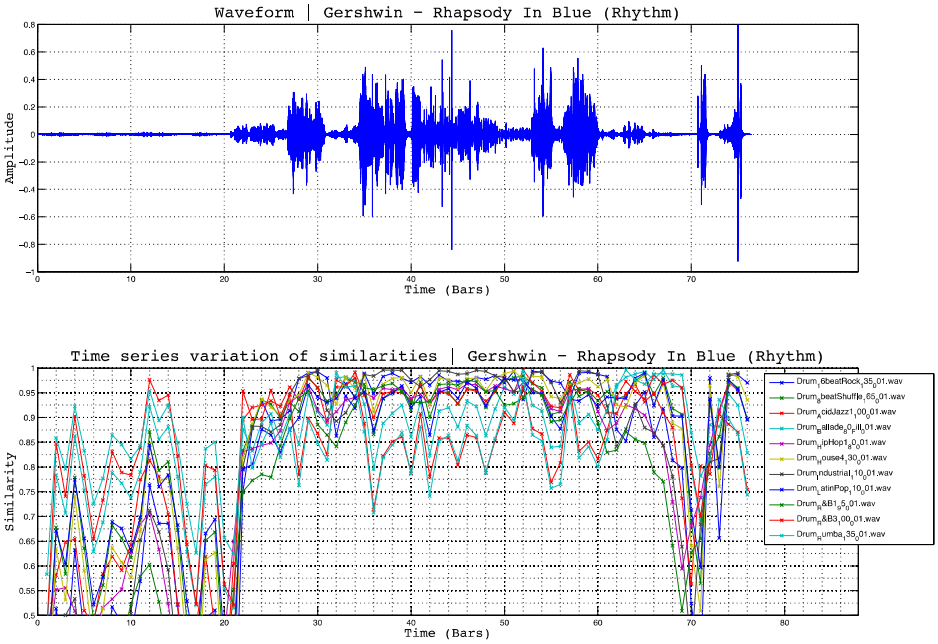

Figure 11. Time-series impression change of "Gershwin – Rhapsody In Blue" (Rhythm)

4.2.2.2. Most Characteristic Phrase in the "Target music piece (X)"

Table 5 shows that this music piece includes a characteristic rhythm phrase of "Drum_AcidJazz1_100_001.wav" by fitting in with the reality of that this music piece is known as a fusion style of Classic and Jazz, which is called "Symphonic Jazz".

Table 5. Characteristic phrases in "Gershwin – Rhapsody In Blue" (Rhythm)

Typical phrase (Y) name	TF-IDF Value
Drum_16beatRock_135_001.wav	0.030
Drum_8beatPop3_145_003.wav	0.027
Drum_AcidJazz1_100_001.wav	0.021

4.2.3. Donald Fagen - I.G.Y.

4.2.3.1. Time-series Impression Change

Figure 12 shows that a typical phrase of Rock rhythm "Drum_16beatRock_135_004.wav" has a high similarity with more than 0.98 in the 57th bar, and a typical phrase of Jazz rhythm "Drum_SwingJazz2_110_Fill_002.wav" has a high similarity in the 60th bar. This fit in with the reality that this music has both Jazz and Rock taste and it is generally classified in a genre called "Jazz-Rock".

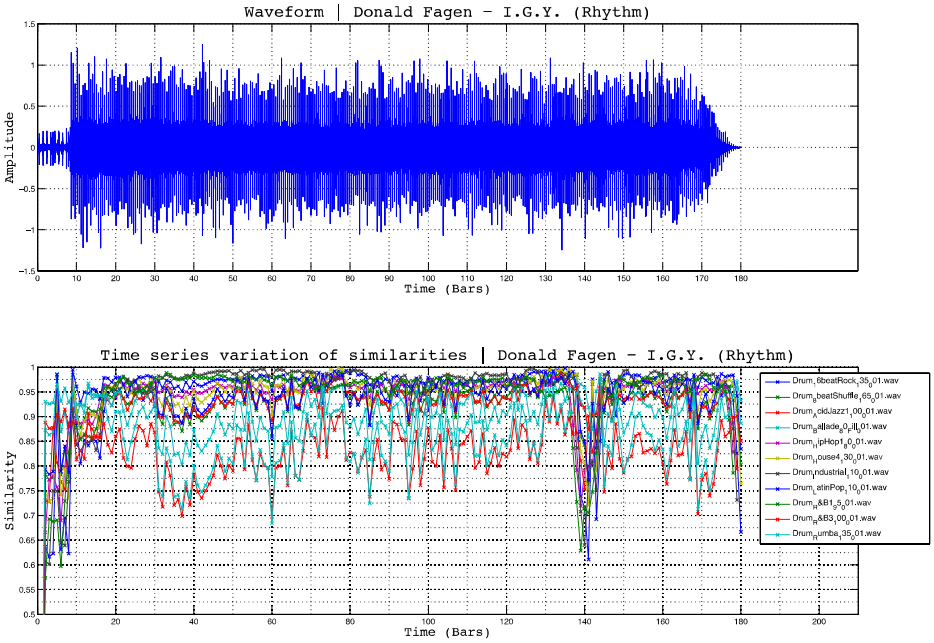

Figure 12. Time-series impression change of "Donald Fagen – I.G.Y." (Rhythm)

4.2.3.2. Most Characteristic Phrase in the "Target music piece (X)"

Table 6 shows that this music piece includes a characteristic rhythm phrase of "Drum_HipHop1_80_Fill_001.wav" and "Drum_8beatPop3_145_003.wav". Based on the results above, our system is able to extract not only the feature of the popular genre but also the feature of the unpopular genre by the unit of phrase in a heuristic way.

Table 6. Characteristic phrases in "Donald Fagen – I.G.Y." (Rhythm)

Typical phrase (Y) name	TF-IDF Value
Drum_HipHop1_80_Fill_001.wav	0.094
Drum_8beatPop3_145_003.wav	0.081
Drum_16beatRock_135_004.wav	0.025

5. CONCLUSION AND FUTURE WORK

In this paper, we presented *"A time-series phrase correlation computing system with acoustic signal processing for music media creation"*. Our system realizes the automatic extraction of musical features from acoustic signals by dividing and decomposing them into phrases and three musical elements, which are meaningful for human recognition. In the experiment, we indicated the feasibility of our system when the selected musical element is "rhythm". These results have shown that our system leads to a new creative and efficient environment for cross-genre/cross-cultural music creation based on the potentially existing knowledge on the music phrase and structure. As our future work, we focus on the improvement of the human interface and the application of machine learning method. The former is to implement the application with intuitive interface [16] and the latter is to apply some machine learning methods to extract the feature of "typical phrases (Y)".

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Shiori Sasaki from Keio University and Dr. Ali Ridho Barakbah from EEPIS who are the program chairs of KCIC 2016. This work was supported by a Research Grant 2016 from Keio SFC Academic Society.

REFERENCES

[1] Schoenberg, Arnold, and Leonard Stein, **Fundamentals of musical composition**, Ed. Gerald Strang, Faber & Faber, 1970.
[2] Dennis DeSantis, **Making Music: 74 Creative Strategies for Electronic Music Producers**, Ableton AG, 2015.
[3] Casey, Michael, et al, **Content-based music information retrieval: Current directions and future challenges,** Proceedings of the IEEE 96.4 (2008): 668-696.
[4] George Tzanetakis, and Emiru Tsunoo, **Frontiers of Music Information Processing Technologies: Audio-based Classification of Musical Signals,** Information Processing 50.8 (2009): 746-750. (in Japanese)

[5] Tsunoo, Emiru, Nobutaka Ono, and Shigeki Sagayama, **Musical Bass-Line Pattern Clustering and Its Application to Audio Genre Classification,** ISMIR, 2009.
[6] Yasushi Kiyoki, **Database System for Calculating Kansei and Semantics: Memory System of Human and Information System,** Keio SFC journal 13.2 (2013): 19-26. (in Japanese)
[7] Takashi Kitagawa, Yasushi Kiyoki, and Youichi Hitomi, **A Mathematical Model of Meaning and Its Implementation Method,** DE 93.56 (1993): 25-32 (in Japanese)
[8] Chalisa VEESOMMAI, and Yasushi KIYOKI, **The rSPA Processes of River Water-quality Analysis System for Critical Contaminate Detection, Classification Multiple-water-quality-parameter Values and Real-time Notification**, EMITTER International Journal of Engineering Technology, 2016
[9] Kashino, Kunio, Gavin A. Smith, and Hiroshi Murase, **A quick search algorithm for acoustic signals using histogram features–time‐series active search,** Electronics and Communications in Japan (Part III: Fundamental Electronic Science) 84.12 (2001): 40-47.
[10] Kosugi, Naoko, Yasushi Sakurai, and Masashi Morimoto, **SoundCompass: a practical query-by-humming system; normalization of scalable and shiftable time-series data and effective subsequence generation,** Proceedings of the 2004 ACM SIGMOD international conference on Management of data. ACM, 2004.
[11] Davida Rochman, **How to Read a Microphone Frequency Response Chart**, Shure Blog, http://blog.shure.com/how-to-read-a-microphone-frequency-response-chart/
[12] Sahidullah, Md, and Goutam Saha, **Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition,** Speech Communication 54.4 (2012): 543-565.
[13] Logan, Beth, **Mel Frequency Cepstral Coefficients for Music Modeling,** ISMIR, 2000.
[14] Sokolova, Marina, and Guy Lapalme, **A systematic analysis of performance measures for classification tasks,** Information Processing & Management 45.4 (2009): 427-437.
[15] Marsyas, **GTZAN Genre Collection**, http://marsyasweb.appspot.com/download/data_sets/
[16] YAMAHA, **KITTAR**, http://ses.yamaha.com/kittar/ (in Japanese)