Sound Clustering Synthesis Using Spectral Data

Ryoho Kobayashi

Keio University Graduate School of Media and Governance
email: ryoho@sfc.keio.ac.jp

Abstract

This paper presents a new sound synthesis method utilizing the features of transitions contained in an existing sound, using spectral data obtained through Short-Time Fourier Transform (STFT) analysis. In this method, spectra obtained from each instantaneous sound are considered as multivariate data, and placed in a vector space, where an evaluation of distances between vectors is performed. As a result, it is possible to detect the occurrences of similarity between analyzed sounds. Clustering and labeling these similar sounds, the features of a sound's transitions are represented in a convenient form. Utilizing these analysis results, a new sound that inherits the transition features from an entirely different sound will be synthesized.

1 Introduction

Over the past few decades a considerable number of studies have been made with sound analysis and synthesis using computers. With the widespread deployment of these computer technologies, these studies have become familiar.

The sound clustering synthesis this paper presents is motivated by various applications of the Short-Time Fourier Transform (STFT) [1, 2] and Algorithmic Composition [3]. These are representative studies that have made rapid development as computer technology has advanced.

Algorithmic composition [3] is a method to compose, using an existing music or theory. Numerous attempts have been made on this subject, and it has been a common composition method.

However, these attempts have some limitations. Existing methods of algorithmic composition generally uses simple, schematic notations like staff notation, which are particularly unsuited to represent timbral transitions.

The purpose of the method this paper proposes is to analyze the timbral transitions of arbitrary music, including sundry ethnic music and noise music. To realize an analysis and synthesis system feasible for timbral transitions, this method this uses the STFT[3].

By using the STFT, magnitudes of each instantaneous frequency are obtained from analyzed sound data.

2 Steps to realize the analysis and resynthesis

The sound clustering synthesis method presented here works according to the following steps.

Step-1: Extracting spectrum from source sound data using STFT analysis.

Step-2: Detecting the similar sound frames of an elected frame.

Step-3: Resynthesizing a sound utilizing the analysis result.

By repeating Steps 2 and 3, a new sound can be synthesized that inherits the transition features from any number of analyzed source sounds.

3 Formulation

In this section, formulations to realize Sound Clustering Synthesis are presented.

3.1 Extracting spectra using STFT analysis

The STFT is essentially an overlapped and windowed Discrete Fourier Transform as described by formula (1). $F_{t,k}$ represents the collective spectra of the current input signal $I_{t,n}$.

$$For \quad k = 0,1,...,N-1 \quad t = 0,1,...$$

$$F_{t,k} = \frac{1}{N}\sum_{n=0}^{N-1} W_n I_{t,n} e^{-2\pi ikn/N} \qquad (1)$$

The index t is a frame index that corresponds to successive overlapped STFT spectra, and overlapped input signal segments.

After calculating the complex valued STFT spectra, a magnitude value, $M_{t,k}$, is calculated using formula (2) below.

$$M_{t,k} = (F_{real_{t,k}}^{2} + F_{imag_{t,k}}^{2})^{1/2} \quad (2)$$

The details of the STFT technique are presented in other articles [1, 2].

In this method, to consider the amplitude scaling properties of the human hearing mechanism, the magnitudes obtained above are mapped onto a logarithmic scale.

$$S_{t,k} = \begin{cases} n \log \dfrac{M_{t,k}}{M_0} & M_{t,k} \geq M_0 \\ 0 & M_{t,k} < M_0 \end{cases} \quad (3)$$

M_0 is minimum magnitude threshold used in the analysis; when magnitude $M_{t,k}$ is smaller than M_0, the magnitude is regarded as silence. n is a scalar used to produce suitable values for this analysis.

3.2 Detecting similar sounds

3.2.1 Positioning spectrum on rectangular coordinates

We position the vector $\mathbf{v}_t$, obtained through formula (4), on rectangular coordinates.

$$\mathbf{v}_t = S_{t,k} \quad k = 0,1,...,N-1 \quad (4)$$

3.2.2 Normalization

Taken as a vector, $\mathbf{v}_t$ includes both magnitude and direction. Here, magnitude corresponds to amplitude, and direction corresponds to timbre.

In this method, amplitude and timbre are analyzed separately.

The magnitude A_t of vector $\mathbf{v}_t$ is described by formula (5).

$$A_t = |\mathbf{v}_t| = (\sum_{n=0}^{N-1} S_{t,n}^{2})^{1/2} \quad (5)$$

Then, by obtaining a unit vector $\hat{\mathbf{v}}_t$, the element of amplitude is eliminated

$$\hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{A_t} \quad (6)$$

3.2.3 Expressing sound distances

Amplitude distance D is the simple distance between magnitudes A_t, as A_t is a scalar value.

$$D_{t_1,t_2} = |A_{t_1} - A_{t_2}| \quad (7)$$

Timbral distance is the measurement of the angle between vectors. Therefore, the inner product of unit vectors, namely, the cosine value of the angle between vectors is considered to be the Timbral Similarity Index K.

$$K_{t_1,t_2} = (\hat{\mathbf{v}}_{t_1}, \hat{\mathbf{v}}_{t_2}) \quad (8)$$

Here, the Timbral Similarity Index K ranges from 0 to 1, and when this value is large, the timbral distance is small.

3.3 Resynthesis

In this method, when the average of amplitude distances is smaller than D_0 and the average of the Timbral Similarity Index is larger than K_0, the frames are considered to be sounds of the same type.

$$C_{t_1} = \begin{cases} F_{t_2} \mid \dfrac{\sum_{n=0}^{f-1} D_{t_1+n,t_2+n}}{f} < D_0, \\ \dfrac{\sum_{n=0}^{f-1} K_{t_1+n,t_2+n}}{f} > K_0 \end{cases} \quad (9)$$

f is the number of analysis frames.

By utilizing the cluster obtained above, a new sound is synthesized.

When a particular frame $F_{output_{t_1-1}}$ is output, the next frame $F_{output_{t_1}}$ is decided by formula (10).

$$\begin{aligned} F_{output_{t_1}} &= F_{input_{t_2}}, \\ F_{input_{t_2-1}} &\in C_{output_{t_1-1}} \end{aligned} \quad (10)$$

Here, $C_{output_{t_1-1}}$ is a cluster of similar frames of output frame $F_{output_{t_1-1}}$, $F_{input_{t_2-1}}$ is selected from the cluster arbitrarily. Then, the next frame $F_{input_{t_2}}$ in input signal is output as the next output frame.

4 Conclusion

In this paper, a method for representing sound similarity presented. Utilizing this framework, a synthesis method to create a new sound containing the transition features of an existing sound is proposed.

This current method has some limitations. For example, transitions represented by the analysis are short. Therefore, this method does not currently have the generality to analyze long transitions such as sub-audio rate rhythms.

By applying the studies of cognitive science, we will be able to solve some problems this method has, and this timbral repesentation will be improved.

References

[1]Jont B. Allen, “Short Term Spectral Analysis, and Modification by Discrete Fourier Transform.” IEEE Transactions on Acoustics, Speech, and Processing, 25(3), pp. 235-238, 1977.

[2]F. Richard Moore, “Elements of Computer Music.” Prentice-Hall, 1988.

[3]Leach, J.a.J.F. "Nature, Music, and Algorithmic Composition." Computer Music Journal, 19(2), pp. 23-33, 1995.

[4]Christopher Penrose, “Extending Musical Mixing: Adaptive Composite Signal Processing” Proceedings of the International Computer Music Conference, Beijing, China, 1999.