

## スペクトルデータを用いた音声クラスタリング合成法

小林良穂

慶應義塾大学 環境情報学部

t99515rk@sfc.keio.ac.jp

### 概要

本稿では、短時間フーリエ変換(STFT)によって得られる各瞬間のスペクトル・データを用いて、既存の音声に含まれる遷移の特徴を利用した、新しい音声合成法を提案する。本手法では、各瞬間の音声から得られるスペクトルを多変量データとみなし、ベクトル空間上に配置した上で、各ベクトル間の距離の数値化を実現する。その結果として、類似した音を音声データ中に発見することが可能となる。これらの類似音声をクラスタリングおよびラベリングすることで、音声の時間的変化の特徴を扱い易く表現する。これらの分析結果を利用するこことにより、他の音声に含まれる時間的変化の特徴を継承した新しい音声を合成する。

## Sound Clustering Synthesis Method Using Spectral Data

Ryoho Kobayashi

Faculty of Environmental Information Keio University

t99515rk@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 evaluation of distances between each vector is calculated. As a result, it is possible to detect the occurrences of similar sounds between analyzed sounds. Clustering and labeling these similar sounds, the features of a sound's transitions are represented in convenient form. Utilizing these analysis results, a new sound which inherits the transition features from a different sound will be synthesized.

## 1 はじめに

近年、音声・音楽の分野では、コンピュータの発達により、膨大な計算を伴う多くの興味深い試みがなされてきた。その中で、本手法の動機となっているのは、短時間フーリエ変換（STFT）[1,2]とアルゴリズミック・コンポジション[3]に関する研究である。

音声データに対してフーリエ変換を行うことにより、音声を各周波数成分に分解できることが知られていたが、コンピュータの普及と高速化により、これらの技術は現実的で身近なものになった。STFT[1,2]は、音声データを短時間のフレームに分解し、各瞬間の音声スペクトルを得るための技術である。これにより、周波数・音量・時間の3つの値を分解して考えることができるため、分析・合成等への様々な応用[4]が試みられており、多くの研究成果をあげている。

アルゴリズミック・コンポジション[3]は、既知の音楽理論や既存の楽曲を分析し、その結果を用いて新しい楽曲を作成するものである。非常に多くの手法が研究されており、現在では作曲手法として広く使われるまでになっている。

しかし、これまで研究してきたアルゴリズミック・コンポジションの手法は、楽曲を「ピッチの変化のシステム」と見なしたものがほとんどである。そのため、五線記譜法等の簡便な記譜法で表現されない音声や、音色変化を重視して表現された音声の特徴を抽出し、その分析結果から新しい音声を合成する試みは多くされていない。

提案手法は、音声を「音色変化のシステム」と見なし、STFT技術による分析結果を用いて、音声を合成する試みである。

## 2 音声クラスタリング合成手法の概要

本手法は、次の3行程により、既存の音声から得られる遷移の特徴を継承した音声を合成する。

- 1) STFT技術による、各瞬間における音声スペクトルの抽出。
- 2) 音声スペクトルを用いた類似音検出。
- 3) 分析音声の遷移特徴を利用した再合成。

利用者が音声データを入力すると、STFTを用いて各瞬間のスペクトルが抽出される。これにより、各瞬間・各周波数のマグニチードを得ることができる。

こうして得られた各瞬間のスペクトルを、それぞれ正規直交座標上に配置し、音声間の距離を「音種」「音量」の2つの側面から数値化する。以上の分析結果と利用者に与えられた指標に基づき、類似音声が発生している瞬間を見つけだすことができる。

入力音声中から抽出されたフレーム(A)を出力として選んだ場合、Aの類似フレームを入力音声中から探し出す。Aの類似フレーム(A',A'')が発見され、入力音声中に<A→B>、<A'→C>、<A''→D>という遷移が存在したとすると、Aの次の出力フレームとしてB、C、Dのいずれかが選ばれることになる。

この結果として得られる音声は、入力音声の遷移システムを継承していると考えることができる。

### 3 実現システム

本節では、提案手法の実現システムおよび定式化について示す。

#### 3.1 STFTによるスペクトルの抽出

入力音声  $I(t)(n)$  に対して(1)を適用することにより、各瞬間( $t$ )におけるフーリエ変換  $F(t)(k)$ を得ることができる。また、 $F(t)(k)$ に対し(2)を適用することにより、各瞬間・各周波数のマグニチュード  $M(t)(k)$  が得られる。STFTに関する詳細は文献[1,2]に述べられている。

For  $k = 0, 1, \dots, N-1$   $t = 0, 1, \dots$

$$F(t)(k) = \frac{1}{N} \sum_{n=0}^{N-1} W(n) I(t)(n) e^{-2\pi i kn/N} \quad (1)$$

$$M(t)(k) = \sqrt{F_{real}(t)(k)^2 + F_{imag}(t)(k)^2} \quad (2)$$

本手法では、人間の聴覚反応を考慮し、得られたマグニチュードを対数値に変換する。

$$S(t)(k) = \begin{cases} a \log \frac{M(t)(k)}{M_0} & M(t)(k) \geq M_0 \\ 0 & M(t)(k) < M_0 \end{cases} \quad (3)$$

ここで、 $M_0$ は分析に使用する最低音量を示すもので、これより小さい値は無音として扱う。また、 $a$ は分析に適したスケールに変換するために与えられる値である。

#### 3.2 類似音検出法

類似音声の検出手法を以下に示す。

##### Step-1 スペクトルの直交座標への配置

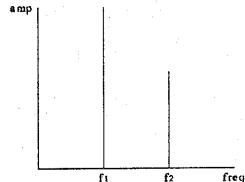
(4)で得られるベクトル  $v(t)$  を直交ベクトル空間上に配置する。

$$v(t) = S(t)(k) \quad k = 0, 1, \dots, N-1 \quad (4)$$

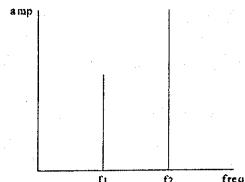
ここで、例として 2 つの周波数( $f_1, f_2$ )からなる音声  $\alpha, \beta, \gamma$  を考える。

$$\alpha = (5, 3), \quad \beta = (3, 5), \quad \gamma = (2, 1)$$

$\alpha$ )



$\beta$ )



$\gamma$ )

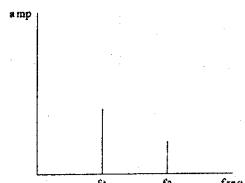


図 1:  $\alpha, \beta, \gamma$  の各周波数成分

ベクトル  $\alpha, \beta, \gamma$  を 2 次元の直交座標上に配置する。これを、図 2 のように表すことができる。

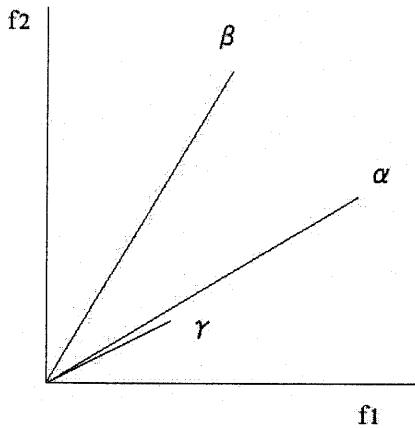


図 2:  $\alpha$ 、 $\beta$ 、 $\gamma$ の直交座標への配置

### Step-2 正規化

図 2 より、 $\alpha$ と $\beta$ 、 $\gamma$ との距離を計測すると、 $\alpha$   $\beta$ 間が $\alpha$   $\gamma$ 間よりも近いことが分かる。しかし、実際にこれらの音を聴いた場合、 $\alpha$ 、 $\gamma$ は音量の違う「似た音」だと感じられることが予想できる。

このような問題が起きる原因是、この座標上に「音種」と「音量」の2つの要素が、同時に表されてしまっている点にある。音声が各周波数を含むバランス（本稿では「音種」と呼ぶ）は、人間が音を似ていると感じるための、非常に重要な要素である。それに対し、音量のみから音声の類似性を感じるのは、その音量が非常に大きい、あるいは小さい場合など、特殊な状況に限られる。

そこで、これら「音種」と「音量」の2つの要素を、それぞれ独立に分析することが有用であると考えられる。

「音種」は、座標上に配置されたベクトルの向きに表される。また、「音量」はベクトルの大きさに表される。

ベクトル  $\mathbf{v}(t)$  の音量  $A(t)$  の数値化は、(5)によってベクトルの大きさを求めることで達成される。

$$A(t) = \|\mathbf{v}(t)\| = \sqrt{\sum_{n=0}^{N-1} S(t)(n)^2} \quad (5)$$

次に、単位ベクトル  $\mathbf{u}(t)$ を得ることで、音量の要素を排除し、直交座標上に各ベクトルを再配置する。

$$\mathbf{u}(t) = \frac{\mathbf{v}(t)}{A(t)} \quad (6)$$

これにより、前に例示されたベクトル  $\alpha$ 、 $\beta$ 、 $\gamma$ は、図 3 のように再配置される。

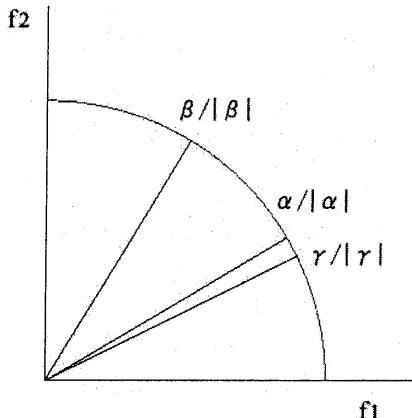


図 3:  $\alpha$ 、 $\beta$ 、 $\gamma$ の単位ベクトル

### Step-3 音声間の距離の数値化

「音量」「音種」のそれぞれの要素について、音声間の距離を数値化する。

- ・ 音量差の数値化

音量に関しては、音量  $A(t)$  がスカラー値で与えられるため、この差を音量差  $D$  とする。

$$D(t_1, t_2) = |A(t_1) - A(t_2)| \quad (7)$$

## ・ 音種類似度の数値化

音種差はベクトルの向きの差として表されるため、提案手法では単位ベクトルの内積、すなわち、2つのベクトル間の角度から得られるコサイン値を音種類似度  $K$  とする。

$$K(t_1, t_2) = \mathbf{u}(t_1) \cdot \mathbf{u}(t_2) \quad (8)$$

ここで、音種類似度  $K$  の取り得る値は、0 から 1 の間の実数値であり、この値が大きいものほど音種差が小さいと考えられる。

## 3.3 再合成法

分析結果を利用し、音声の遷移システムを利用した再合成法を示す。

本手法では、音量差の平均が基準値  $D_0$  より小さく、音種類似度の平均が基準値  $K_0$  より大きい音を同類の音と見なしクラスタリングする。

$$C(t_1) = \left\{ F(t_2) \mid \sum_{n=0}^{f-1} D(t_1 - n, t_2 - n) < fD_0, \right. \\ \left. \sum_{n=0}^{f-1} K(t_1 - n, t_2 - n) > fK_0 \right\} \quad (9)$$

ここで、 $f$  は分析フレーム数を示し、現フレームから  $f-1$  フレーム前までを分析対象とする。

次に、こうして得られたクラスターを利用し、音声を再合成する。

フレーム  $F_{output}(t_1-1)$  が出力されたとき、次に出力されるフレーム  $F_{output}(t_1)$  は以下によって決定される。

$$\begin{aligned} F_{output}(t_1) &= F_{input}(t_2), \\ F_{input}(t_2-1) &\in C(t_1-1) \end{aligned} \quad (10)$$

ここで、 $C(t_1-1)$  は、出力フレーム  $F_{output}(t_1-1)$  の類似音声として、入力音声から得られたフレームの集合(クラスター)であり、 $F_{input}(t_2-1)$  は、そのクラスターからランダムに選ばれたフレームである。そして、入力音声における  $F_{input}(t_2-1)$  の次のフレーム  $F_{input}(t_2)$  を出力する。

本手法では、類似音声の集合  $C$  は出力フレーム毎に与えられるため、独立性を持たない。しかしながら、 $C$  に含まれるすべてのフレームは同類の音声と見なされるため、擬似的な遷移確率に基づいた再合成が実現していると考えることができる。

図 4 は、クラスター  $C_0$  に含まれるフレームからの遷移を矢印で示している。遷移の特徴から、仮想クラスター  $C_1$ 、 $C_2$  への遷移システムを見出すことができる。

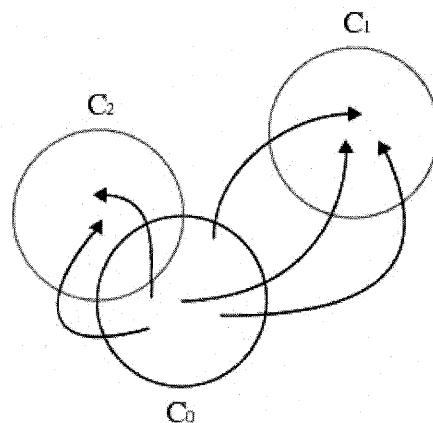


図 4: 仮想クラスターへの遷移のイメージ

## 4まとめ

本稿では、類似音声の検出手法を示した上で、既存音声の遷移システムを継承した音声の合成手法を提案した。音楽分野において、音声スペクトルを用いた遷移システムの分析・再合成法について充分に研究されていない現状で、提案手法は価値のあるものだと考えられる。

しかし、本手法には多くの問題点がある。

分析過程で指摘されるべき最も大きな問題は、本手法の示す「類似音声」と実際の人間の聴覚反応との相違である。本手法では、周波数成分の間隔を一定で得ることになる。また、人間の聴覚は、同量の刺激であっても周波数によって反応の大きさが違う。これらの問題により、本手法では高音成分の影響を必要以上に大きく受ける傾向がある。一般的な音声データは高音成分をあまり多く含まず、含まれる高音成分のほとんどが、より低音の音声と強く関係している。そのため、この問題の影響は大きなものにならないと考えられるが、今後改善を要する部分である。

再合成過程は、フレーム単位の短時間の遷移システムを継承するものになっている。そのため、音声に含まれるリズム等の比較的長時間の音声から構成される要素を分析・継承することが難しい。こうした長時間におけるパターンは、音声の特徴を決定付ける大きな要素である。そのため、このような要素を取り入れることも必要であると考えられる。

## 参考文献

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.