音色空間における距離の知覚
Perceptual Distance in Timbre Space

寺澤洋子 †, Malcolm Slaney†‡, Jonathan Berger†
Hiroko Terasawa†, Malcolm Slaney†‡, Jonathan Berger†

Center for Computer Research in Music and Acoustics†
Stanford University, Stanford, California, USA
IBM Almaden Research Center‡ San Jose, California, USA
{hiroko, malcolm, brg}@ccrma.stanford.edu

概 要

この研究では、定常音の音色の知覚空間を描写するために、知覚上の直交性を考慮した客観的尺度を定義し、音色の補間特性を測定した。３種類の音色の数値表現を比較し、心理測定結果との対応性を検討した。実験の結果、メルケプストラム（MFCC）による音色表現が知覚空間のモデル化に適することがわかった。

1　はじめに

音色は「ラウドネスとピッチが同じ二つの音を、等しい条件で提示した際に、違いが聴取されるための特質」と定義される [1]。本稿では、音の知覚において音色の役割を研究する際に重要である、音色知覚の多次元性について論じる。3 種類の音色推定手法と、音色知覚との関連性を比較検討した。

研究の動機として以下の二つが挙げられる。基礎研究としては、人間がどのように音と言語を知覚するのか理解し、色覚における三刺激値モデルと同等な音色知覚モデルを作りたい。そして応用のためには、よりよい音の分析のために最節約性を満たす基礎的な音色の数値表現を見つけたい。

この研究では音色の知覚について、これまでの研究とは異なったアプローチをとる。多次元尺度法を用いた音色の研究 [2, 3, 4] では、まず刺激音を準備し、知覚的距離を心理測定し、それを構成する座標系を見つける。我々の手法ではまず座標系を定め、それに基づいて音を合成してから、それぞれの音色の数値表現を使用した場合、どの程度まで音色間の距離を予測できるかを評価する。

本稿は以下の構成をとる。まず、音色の数値表現について説明を行う。そして、最後に音色の数値表現と知覚の対応を心理測定によって検討し、節約性と単純性を最も良く満たす数値表現を音色空間の最適モデルとする。

2　音色の数値表現

2.1　パラメタの設定

音の数値表現では、抽象化の度合いが多様である。たとえば、スペクトラムは可逆かつ完全な音の数値表現であるが、結果は複雑で、人間の知覚を予測するには更なる変換が必要である。

メルケプストラム（MFCC）は音声認識の分野で主に使用される。人間の聴覚を模したフィルタバンクを用い周波数軸をメル周波数に変換した後、各チャンネル出力の対数をとりラウドネス圧縮を行う。その後、離散コサイン変換（DCT）を行い、低次係数が MFCC とされる [5]。DCT はスペクトラムを滑らかにするだけでなく、係

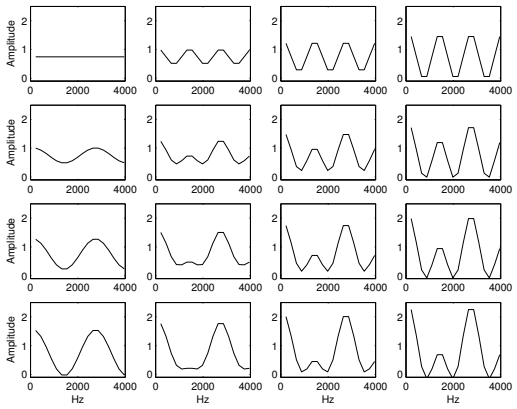

図 1: 2 次元の LFC パラメタから再構築されたスペクトラム。縦方向は C_3 が 0 から 0.75、横方向は C_6 が 0 から 0.75 をとる。図 2 と比べるとピークが等間隔に並んでいる。

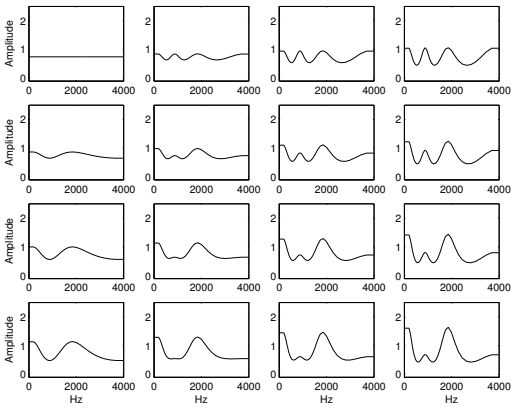

図 2: 2 次元の MFCC パラメタから再構築されたスペクトラム。縦方向は C_3 が 0 から 0.75、横方向は C_6 が 0 から 0.75 をとる。

数間の相関を取り除くためにも有効である。しかし、統計上の直交性は知覚上の直交性を意味しない。音声認識の経験則から、音声を時間の関数で表現するために 13 次までのベクトルが使用されることが多い。

LFC（Linear Frequency Coefficients）は比較のために提案されたモデルである。MFCC に似ているが、周波数軸も振幅も線形である。通常のスペクトラムに DCT を行い、13 次までの係数を使用してスペクトル包絡を表す。MFCC と LFC のどちらもスペクトル包絡を低次数で表し、係数を非相関とするが、違いは周波数と振幅の圧縮にある。

LFC と MFCC のどちらでも、定常音のスペクトル包絡が 13 次のベクトルで表される。ベクトル係数は C_0 から C_{12} と呼ばれ、最低次の C_0 はスペクトラムの平均振幅であり、今回の実験では定数とした。また高次の係数になるほど、スペクトル包絡のより細かな形状を表す。13 次の係数列を使ってどのように音合成を行うかは、次節において説明する。

Pollard による音色の三刺激値モデルは楽器音の音色表現によく使われる [6]。この方法では調波構造をもつ音が 2 次元平面上の点として表される。三刺激値係数は以下のように求められる。まず、Zwicker の方法に従って、各ハーモニクスのラウドネス N_i を計算する。次にハーモニクスは、$i = 1$、$i = 2 \ldots 4$ と、$i = 5 \ldots n$ の 3 つのグループに分類される。グループごとのラウドネスは Steven の法則に従って以下のように求められる。

$$N_i = 0.85 N_{max} + 0.15 \sum_j N_j \qquad (1)$$

ここで j はグループ i に分類されている調波であり、N_{max} はそのグループにおいて一番大きな振幅をもつハーモニクスのラウドネスである。最後に、各グループのラウドネスを、全てのグループのラウドネスの和で正規化する。$T1 = N_1/N$, $T2 = N_2^4/N$, $T3 = N_5^n/N$ となり、$N = N_1 + N_2^4 + N_5^n$ である。T1 が大きいほど強い基本周波数を意味し、T2 が大きければ中音域、T3 が大きければ高音域が強いことを示す。

2.2 スペクトラムの再構築

今回の実験では、13 次のベクトルをパラメタとして、LFC と MFCC の逆変換によりスペクトラムを再構築し、音合成を行った。

LFC の逆変換の場合、再構築されるスペクトラム $\tilde{S}(f)$ は LFC ベクトル C_i' の IDCT で与えられる。MFCC の逆変換の場合は、まず MFCC ベクトルの IDCT $\tilde{L}_i = \texttt{IDCT}(C_i)$ を求め、そこから 10 のべき乗数 $\tilde{F}_i = 10^{\tilde{L}_i}$ を求めると、そ

れがフィルタバンクのチャンネル i の出力となる。ここで、$\tilde{F}_i$ は各チャンネルの周波数中央値であると仮定し、線形補間をし、スペクトラム $\tilde{S}(f)$ を得る。

2.3 刺激音の種類

13 次元空間を全て実験するのは困難である。そこで、今回は MFCC と LFC からいくつかの２次元空間に限定し、心理測定を行った。測定対象として選ばれたのは、[C_3, C_6], [C_4, C_6], [C_3, C_4], [C_3, C_{12}], と [C_{11}, C_{12}] のペアからなる５種類の２次元空間である。

13 の係数のうち、２つが変数となるよう選ばれた。例えば、$[C_3, C_6]$ による空間の場合、パラメタベクトルは

$$[C_3, C_6] = [1, 0, 0, C_3, 0, 0, C_6, 0, 0, 0, 0, 0, 0]$$

となり、C_3 と C_6 が $C = [0, 0.25, 0.5, 0.75]$ と４段階に変化させ、他の係数は定数（１あるいは０）とした。このベクトルが LFC あるいは MFCC としてスペクトラム再構築に使用された。

2.4 数値表現の比較

LFC および MFCC 空間におけるどの点も定常音を表す。図 1 と図 2 に、C_3 と C_6 を変化させ、C_0 は 1、それ以外は 0 に固定した場合のスペクトラムを示す。C_3 と C_6 が 0 で、$C_0 = 1$ のとき、スペクトラムは平坦である。図の下方に行くに従って、C_3 の値が増加し、直流と中音域の振幅が増加する。また、図の右方に行くにつれ、C_6 の値が増加、スペクトラム上の３つの隆起が大きくなる。

LFC と MFCC によって合成した音は三刺激値モデルで分析することが可能であり、２次元空間上の点として表現できる。[C_4,C_6] 空間の三刺激値モデルによるプロットを図 3 に示す。ここで、実験に使用したスペクトラムが T2-T3 空間に表した。MFCC のパラメタ空間においての直角格子は、三刺激値モデルでは非線形写像となる。

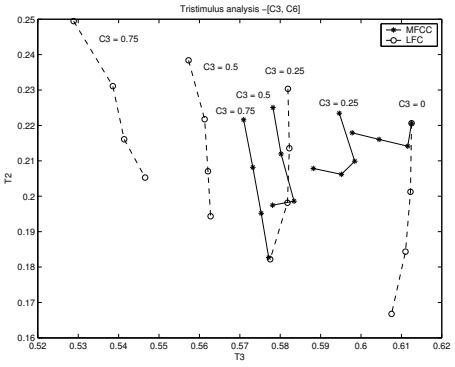

図 3: 刺激音の三刺激値モデルによるプロット。それぞれの線の上端において $C_6 = 0$、下端に行くに従って 0. ２５ずつ増加する。

2.5 FM 加算合成

実験に使われる音声を模した刺激音は、第 2.2 節で再構築されたスペクトラムから音源フィルタモデルを使用して合成された。音源は一定のピッチをもつインパルス列、フィルタリングには、加算合成を用いた。音源の基本周波数 f_0 は 220 Hz、ビブラートの周波数 v_0 は 6 Hz、周波数変調の振幅 V は 6 %とした。再構築されたスペクトラム包絡 $\tilde{S}(f)$ に基づいて、各ハーモニクスの振幅は重み付けされる。ハーモニクスの次数を n とすれば、合成音は下式で与えられる。

$$s = \sum_n \tilde{S}(n \cdot f_0) \cdot \sin(2\pi n f_0 t + V(1 - \cos 2\pi n v_0 t)) \tag{2}$$

3 実験の方法

実験では、被験者に二つの音を提示し、音色の違いを主観的に評価してもらい、音色のパラメタ間の距離を測定した。

一つの刺激につき、二つの音を連続して提示し、最初の音は常に基準音、次の音が評価対象の音である。基準音は全ての刺激において統一されており、C_0 以外が全て 0 の、平坦なスペクトラムをもつ（$[C_m, C_n] = [0, 0]$）。二つ目の評価音は各刺激で異なる。

刺激のグループ（LFC と MFCC で各５組、計 10 組の２次元空間）ごとに、５個の刺激を練習

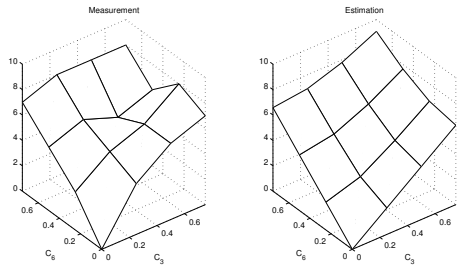

図 4: 被験者の一人による音色間の距離。(a) 測定結果 (b) フィッティング後の推定モデル

のために提示したのち、本番では、刺激を提示するたびに二つの音色間の距離を記録した。被験者は 1（全く同じ音色）から 10（最も異なる音色）の度数を用いて音色の類似度を評価した。各グループにつき 16 個の刺激がランダムな順序で提示された。10 人の学生（20 歳から 35 歳）が実験に参加し、刺激は静かなオフィス環境でヘッドフォンを使って提示された。

4 分析方法

実験結果の分析は二段階に分かれる。最初に、被験者ごとに音色の距離評価をユークリッド空間にフィッティングする。そして、距離評価とユークリッド距離の残差を被験者ごとに算定する。ここで各被験者の知覚において、音色の数値表現（LFC と MFCC）がどの程度ユークリッド距離の条件を満たすかがわかる。次の段階で、全ての被験者の残差から、残差平均と標準誤差を各刺激グループ（10 組の 2 次元空間）ごとに計算し、音色の数値表現を最終的に評価する。

4.1 データフィッティング

2 次元パラメタの場合、音色の知覚距離 d はユークリッドモデルで以下のように推定する。

$$d^2 = ax^2 + by^2 \tag{3}$$

x は係数のうちの一つ（例えば C_3）で y はもう一つの係数（例えば C_6）である。この式で d^2、x^2、y^2 は既知の値である。多次元線形回帰分析によって心理測定された距離がユークリッドモデルに適合するかを評価した。

回帰分析における推定は最小自乗法を使用した。疑似逆行列は線形推定において最少誤差を保証する。線形モデルからの残差を以下に示す。

$$d_{res} = \frac{1}{16} \sum_{x,\ y} \Big| \ d - \hat{d} \ \Big| \tag{4}$$

ここで、$\hat{d}$ は線形回帰分析による推定モデルである。図 4 に心理測定による距離と、推定モデルを示した。

4.2 データの統合

各被験者のモデル残差がでたところで、数値表現ごとに残差平均が求められた。

$$\bar{d}_{res} = \frac{1}{N} \sum_{i=1}^{N} d_{res,i} \tag{5}$$

ここで N は被験者数である。標準誤差 σ_{Mean} は以下のように求める。

$$\sigma_{Mean} = \frac{\sqrt{\sum_{i=1}^{N} |d_{res,i} - \bar{d}_{res}|^2}}{N} \tag{6}$$

数値表現ごとに、残差平均と標準誤差を比較し、どの数値表現が人間の音色の知覚のモデルとして適切か判断する。

5 結果

図 5 では、5 種類の 2 次元パラメタで実験を行った際の、LFC と MFCC による音色の知覚空間を比較した。どちらの数値表現においても、心理測定値を 10 度数のうち 1 度数の標準誤差で予測することができる。全てのパラメタセットにおいて、単純化された LFC よりも、MFCC のほうが音色空間の推定に適したモデルである。つまり、今回実験したほかの数値表現に比べて、MFCC のほうが正確に音色の補間ができる上、パラメタ軸が知覚空間の直交座標軸に近い。

殆どの 2 次元パラメタ空間においては残差平均がほぼ一定であり、試験した 2 次元パラメタが音色の直交知覚空間を成すことを示唆する。

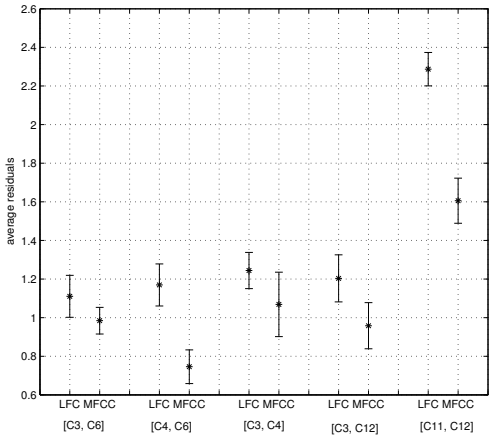

図 5: ５種類の２次元パラメタ空間における、LFC と MFCC の残差平均。標準誤差をエラーバーで示した。

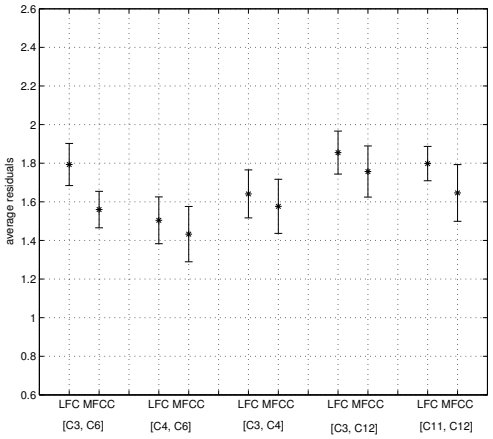

図 6: 三刺激値でモデル化した場合の、１０種類の刺激セットの残差平均。標準誤差をエラーバーで示した。

これは２次元パラメタの係数が C_3 と C_4 のように近い場合でも、C_3 と C_{12} のように遠い場合でも共通である。ここで、C_{11} と C_{12} の２次元空間では残差平均が大幅に増加することは注目に値する。C_3 と C_{12} の場合は残差は少ないので、高次係数の場合でも知覚空間は線形であることを示す。しかし C_{11} と C_{12} が組み合わされた場合の残差増加は、この二つの次元の組は他の次元の組合わせほど直交でないことを意味する。

残差の分散は LFC の場合で 6.8、MFCC の場合で 3.9（ともに１０度数で測定）、どちらの場合でも心理測定の結果の 66%までを予測することができ, ユークリッドモデルは音色知覚の推定に優れているといえる。

図 6 には三刺激値モデルを使った場合の残差平均を示した。この分析では、LFC と MFCC の実験に使用した刺激音について三刺激値を求め、そのベクトル間ユークリッド距離と心理測定値の結果を線形回帰分析によってモデル化した。LFC と MFCC を使用した際（図 5）に比べて、三刺激値による数値表現の方が残差平均が大きくなる（図 6）。

推定の精度が落ちる原因として、使用された基準音が三刺激値空間において原点 $(0, 0)$ にないことが挙げられる。この問題を補正するために、オフセットを導入したユークリッドモデル（$d^2 = ax^2 + by^2 + c$）を使用したが、残差平均は減少したものの、MFCC モデルによる残差平均よりも大きい値であった。

6　まとめ

本稿では、音色空間を評価するための基準について述べ、３種類の音色数値表現を紹介し、音色の距離評価を心理測定した後、MFCC による音色のモデル化で 66%まで心理評価を推定出来ることを示した。

提案された評価基準は、音色のクオリティを表現するために望まれる客観的なものであり、定常音の音色表現に関して、他のモデルよりも MFCC がその基準を満たすことが明らかである。これまでの研究では MFCC とその他の DCT をもとにしたモデルが統計的に独立な数値表現であることがわかっている。今回の結果は、人間の聴覚がこの統計的な独立性と一致し、MFCC が聴覚的に直交空間をなすことを示唆する。本稿で述べた手法は音色空間の問題に関して閉じた形の解ではなく、あくまでも数値表現が心理測定について再倹約性を満たすかを観察するものであり、今後さらなる研究が必要である。

三刺激値モデルによる数値表現では、実験に使った音は合成に使ったパラメタに関係なく表

現されるので、図 6 によって直接 LFC と MFCC の比較はできない。また、全てのモデルにおいて局地的に線形で、音色空間の端では推定精度が落ちることも考えられる。三刺激値によるプロット（図 6）を見ると、LFC 逆変換で合成した音は、MFCC によって合成された音よりも範囲が大きい。推定精度の違いは、この音色の範囲の違いが原因とも考えられる。今回の LFC と MFCC の比較では、それぞれ LFC と MFCC パラメタ空間から合成された刺激音を基にしているため、推定結果の違いが刺激音セットの違いによって生じる可能性は否定出来ない。

また、今回の実験に使用した数値表現では定常音しか描写できない。これまでの研究で、立ち上がり時間などの時変パラメタが音色の知覚に大きく影響することがわかっている。しかし、本稿で述べた線形性、直交性などの基準も、音色空間を特徴づけるために重要な要素である。

最後に、文脈による音色の知覚の違いについてはこれからの研究が必要であるが、本研究により音色の数値表現における基本的な問題を述べた。

謝辞

この研究は２００４年の Telluride Neuromorphic Workshop において始められた。また、Shihab Shamma, Stephen McAdams, Dan Ellis, Tom Rossing の各氏の御助言に感謝申し上げます。

参考文献

[1] B.C.J.Moore. *An introduction to the psychology of hearing, fifth ed.* Academic Press, 2003.

[2] J.Grey. “Multidimensional Scaling of Musical Timbres.” *Journal of the Acoustical Society of America* 61(5): pp. 1270–1277, 1976.

[3] S.McAdams, W.Winsberg, S. Donnadieu, G.De Soete, and J.Krimphoff. “Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes.” *Psychological Research*, 58, pp. 177–192, 1995.

[4] S.Lakatos. “A common perceptual space for harmonic and percussive timbres” *Perception & Psychophysics*, 62 (7), pp. 1426–1439, 2000.

[5] J.F.Blinn. “Jim Blinn’s Corner: What’s the Deal with the DCT?” *IEEE Computer Graphics & Applications (July 1993)*, pp. 78–83, 1993.

[6] H.F.Pollard, E.V.Jansson. “A Tristimulus Method for the Specification of Musical Timbre” *Acustica*, 51, pp. 162–171, 1982.

[7] D.C.Dennett. “Quining Qualia.” *Consciousness in Modern Science* Eds. A.Marcel, and E.Bisiach, Oxford University Press, Oxford, 1988.