音色知覚モデルを用いた物理モデル合成の制御パラメタ推定

寺澤 洋子, Jonathan Berger, Julius O. Smith
スタンフォード大学 CCRMA
{hiroko, brg, jos}@ccrma.stanford.edu

概要

物理モデル合成において、音色と音量は物理パラメタによって調整されるが、手動での調整は長時間の作業を要するうえ、パラメタ決定の過程は無計画であることが多い。本研究の目的は、最適パラメタの推定アルゴリズムを導入することで、物理モデルのパラメタ調整における偶然性を減らすことである。パラメタ推定には録音されたリファレンス音と合成音のセットとの間で、聴覚モデルを用いて音色間の距離を計算しそれを最小化する手法を基本とする。本稿では、定常状態の短時間 MFCC から連続した複数のフレームを抽出し、そのベクトル平均とその標準偏差を使用して音色を比較した。この手法を楽器演奏における聴覚フィードバックの基本モデルとして提案する。

Using A Perceptually Based Timbre Metric for Parameter Control Estimation in Physical Modeling Synthesis

Hiroko Terasawa, Jonathan Berger, Julius O. Smith
CCRMA, Stanford University
{hiroko, brg, jos}@ccrma.stanford.edu

Abstract

Manual adjustment of control parameters for physical modeling synthesis suffers from practical limitations of time-intensive and sometimes arbitrary and haphazard parameter tweaking. An efficient approach to automatic parameter estimation, the goal of this study, would potentially eliminate much of the hit or miss nature of parameter tuning by finding optimal control parameters for physical modeling synthesis. The method is based on psychoacoustically motivated timbre distance estimations between a recorded reference sound and a set of corresponding synthesized sounds. The timbre comparisons are based upon the sample mean and standard deviation between Mel-Frequency Cepstral Coefficients (MFCC) computed using several steady-state time frames from the reference and synthesized sounds. This framework serves as a preliminary model of the auditory feedback loop in music instrument performance.

1　はじめに

楽器の演奏は感覚と運動の統合によってなされる。音色の聴覚フィードバックを伴う訓練によって、楽音の生成は、細やかな制御が可能になり、ほぼ自動的に見えるほどまでに上達する。演奏技術の上達は、演奏者が複数の制御パラメタを微妙に変化させながら組み合わせ、生じる音色を概念化してゆく過程に他ならない。作曲とオーケストレーションの熟達には、同様に、楽音の音色を抽象化する訓練と、意図する音色を実現するために適切な記譜法を習得することが必要とされる。

音楽制作のためのデジタル音合成においては、音色の概念とパラメタによる音色の制御性との間に直感的なつながりがないことが、効率の良い音色制御の障害となっている。物理モデル合成を用いて作曲および演奏をする場合、楽器演奏におけるインタラクティブで直感的な聴覚フィードバックを再現するために、人間の聴覚をもとにした音色制御のためのパラメタ推定アルゴリズムが有益と考えられる。

物理モデル合成（PM 合成）の最大の目標は、楽器音の人間らしい演奏をリアルに再現することにある [1, 2]。その目標を達成するには、楽音を演奏するために重要な聴覚フィードバックと演奏技術のモデルを PM 合成システムとカップリングすることが必要不可欠である。

関連する研究としては、KTH ルールを用いて MIDI スコアに演奏表情を付与するもの [3] や、PM 合成のパラメタ推定をするものなどがある。Diana Young と Stefania Serafin は弓の圧力と位置によるバイオリン物理モデルの演奏性について報告した [4]。Caroline Traube らによる研究では、ギター演奏における撥弦位置をスペクトル重心を用いた音色評価によって推定している。Guillemain らはクラリネット物理モデルの音色分布を古典的な音色評価法によってモデル化している [6]。また IRCAM の分析合成グループではトランペットの制御パラメタ推定を様々な角度から行っており、物理モデルの逆関数をもとめる手法 [7]、ケプストラム係数とその微分係数をベクター量子化する手法 [8] がある。また音色の類似性を制御パラメタの関数として処理する研究 [9] があり、本研究もこれに非常に近いアプローチを取っている。これらは、トランペット物理モデルが遅延フィードバックを伴う非線形システムであるにも関わらずかなりの成功をおさめている。

この研究の目的は、PM 合成の制御システムを構築することであり、以下の二条件が必要である。（1）意図通りの音高、ラウドネス、音色をもつ音を（2）意図通りの時間に合成する。このようなシステムがあれば、現存する演奏表情付与システム（インタフェースは音高、ラウドネス、時間である場合が多い）と組み合わせることができる。

音高と時間は殆どの PM 合成システムの基本的な入力であるので、音色とラウドネスの制御が問題となる。MFCC（Mel-Frequency Cepstral Coefficient）は主に音声認識などの分野で使われるスペクトル分析法であるが、著者らのこれまでの研究から、MFCC は聴覚的にも有意な音色とラウドネスの評価法とわかっている [10]。そこで、本研究では MFCC を用いてパラメタ推定を行う。

このシステムの最終的な目標は、意図される音色の概念化と、楽器の物理モデルのパラメタ生成を結合することにある。その動機として、

1. 物理モデルによる作曲および音楽演奏の実用性を改善すること
2. 楽器演奏における聴覚フィードバックのモデル化へ向けた試み

があげられる。

本研究では STK（The Synthesis ToolKit）[11] に含まれるクラリネット物理モデルを使用し、吹鳴圧と吹鳴ゆらぎの二つの制御パラメタを変数とした。また、実際のクラリネットの録音を模倣のためのサンプル音とした。合成音とサンプル音の STFT（短時間フーリエ変換）を行い、定常状態から複数フレームを抽出し、複数フレームにまたがる MFCC のベクトル平均 $\bar{c}$ とその標準偏差 $\bar{\sigma}$ を用いて、合成音とサンプル音の音色を比較した。MFCC 平均ベクトル $\bar{c}$ はスペクトルエンベロープの特徴量であるのに対し、標準偏差 $\bar{\sigma}$ は合成音が自然に聞こえるために不可欠なスペクトルゆらぎの特徴量である。合成音とサンプル音の間で、これら二つの特徴量について残差平方和を求め、残差平方和を最小にする制御パラメタを最適パラメタとして決定した。残差平方和を求める際に、$\bar{c}$ と $\bar{\sigma}$ のそれぞれ、あるいは両方を用いた方法を提案し、それらの結果を

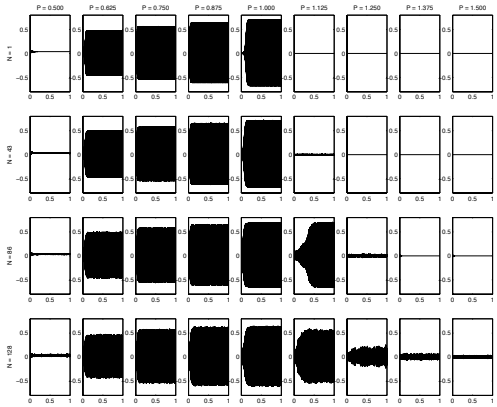

図 1: クラリネット合成音の波形。吹鳴圧は左から右へ、ノイズ振幅は上から下へと大きい値となる。吹鳴圧が低すぎても高すぎても安定した発振は得られない。

公式な聴取実験ではないものの、リスニングによって比較評価した。

2　クラリネット PM 合成による音色レンダリング

今回、クラリネット PM の吹鳴圧と、ゆらぎを加えるためのノイズ振幅を可変とし、他の全てのパラメタは固定とした。合成音、サンプル音ともに基本周波数は 440 Hz に揃えた。

図 1 は吹鳴圧と雑音振幅を変化させた場合の合成音の波形を表にしたものである。吹鳴圧と信号の振幅、そしてノイズ振幅と気息音の間にはっきりとした相関がある。加えて、強すぎる吹鳴圧がかかった場合にはリードが閉まり発振がえられないために、吹鳴圧が高すぎると合成音は無音になる。

推定のためには、吹鳴圧は２０段階、ノイズ振幅は１０段階に設定し、合成音を計２００個作成、サンプル音と比較した。

サンプル音として利用したクラリネットの録音は米国アイオワ大学から提供されている楽器音サンプル [13] を利用した。音高は A4 でダイナミクスは *pp, mf, ff* のサンプルを B 管および Es 管クラリネットの録音から採用した。

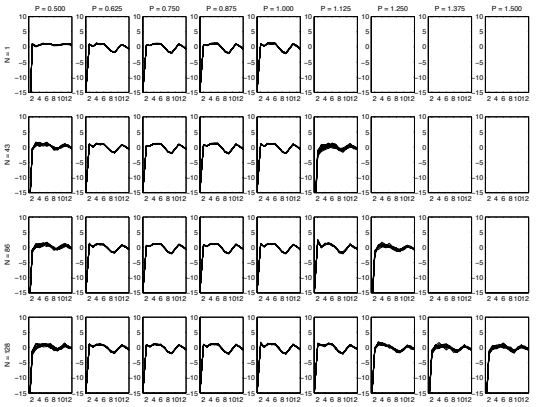

図 2: 図 1 に示したクラリネット合成音の MFCC。１０フレームを重ねてプロットした。音のゆらぎが大きくなるほど MFCC にもばらつきが見られる。

3　MFCC を用いた音色評価

定常状態における音色をあらわすために、MFCC は聴覚的に有意な測定法である。MFCC の計算は以下のように行われる。まず、聴覚フィルタバンクを使い、スペクトラムから臨界帯域あたりの総エネルギーを求める。MFCC は、その IDCT の係数であり、本研究では低次（１３次まで）の係数を使用して対数スペクトルエンベロープを表した。実際の計算には MATLAB の Auditory Toolbox [14] を利用した。

図 2、図 3 に定常状態における合成音とサンプル音の MFCC を示す。それぞれの図には MFCC が１０フレームずつ重ねてプロットされている（フレーム長 5.8 ms、ステップ幅 2.3 ms。）ゆらぎが大きいときは MFCC のばらつきが大きくなることが見てとれる。そこで MFCC のベクトル平均 $\bar{c}$（連続した８０フレーム分、時間にしておよそ 0.2 秒）と

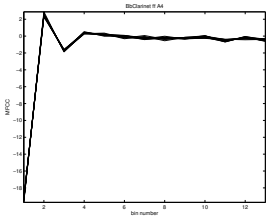

図 3: B 管クラリネットサンプル音（ A4, *ff* ）の MFCC。１０フレームを重ねてプロットした。

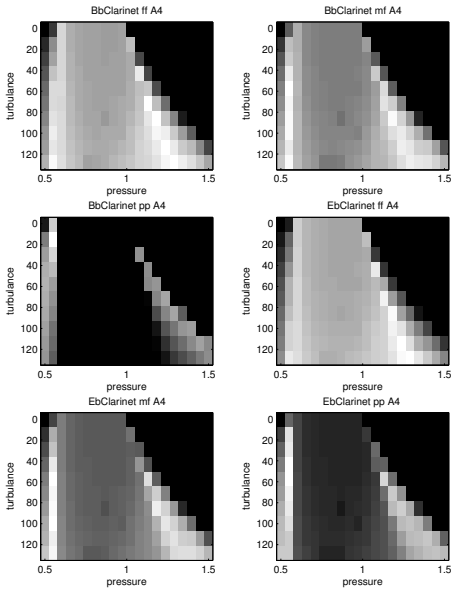

図 4: 比較法（１）による音色類似度分布（MFCC 平均ベクトルのみを使用。）サンプル音は B 管と Es 管、ダイナミクスが *ff, mf, pp* である。色が明るい部分ほどサンプル音と合成音の類似度が高い。類似度の分布はサンプル音によって変化することがわかる。

MFCC 標準偏差ベクトル $\bar{\sigma}$ を音色比較のために使用した。これら $\bar{c}$ と $\bar{\sigma}$ の両方とも１３次のベクトルである。

4　パラメタ推定の手法

4.1　音色の比較

合成音とサンプル音の音色を、以下に示す残差平方和 $D_{\bar{c}}$、$D_{\bar{\sigma}}$ と D_{norm} を用いて比較した。

$$D_{\bar{c}} = \sum \mid \bar{c}_{synth} - \bar{c}_{ref} \mid^2 \qquad (1)$$

ここで $\bar{c}_{synth}$ と $\bar{c}_{ref}$ はそれぞれ合成音とサンプル音（リファレンス）の MFCC ベクトル平均である。

$$D_{\bar{\sigma}} = \sum \mid \bar{\sigma}_{synth} - \bar{\sigma}_{ref} \mid^2 \qquad (2)$$

ここで $\bar{\sigma}_{synth}$ と $\bar{\sigma}_{ref}$ はそれぞれ合成音とサンプル音（リファレンス）の MFCC 標準偏差ベクトルである。

$$D_{norm} = \frac{D_{\bar{c}}}{\max(D_{\bar{c}})} + \frac{D_{\bar{\sigma}}}{\max(D_{\bar{\sigma}})} \qquad (3)$$

D_{norm} は MFCC のベクトル平均と標準偏差ベクトルを最大値で正規化して和をとったものである。

残差平方和を２００の合成音全てと、ひとつのサンプル音（例えば B 管の *ff*）の間で計算し、残差を最小にする吹鳴圧とノイズ振幅の組み合わせを、合成に使われたパラメタセットの中で最適なパラメタとして採用する。ここで、3 つの比較法が検討された。（１） $D_{\bar{c}}$ を最小にする、（２）$D_{\bar{\sigma}}$ を最小にする、（３）D_{norm} を最小にする。図 4 は比較法（１）を用いた時の、６つのサンプル音との類似度分布をプロットしたものである。

4.2　２次補間を用いた推定

合成音のデータベースから類似度の高いものを選び出すだけでは、あくまでも合成に使われたパラメタセット（この場合では２００組）しか考慮されない。真に最適なパラメタセットは離散的な段階（吹鳴圧は２０段階、ノイズ振幅は１０段階）の間にある可能性が高い。そこで隣り合う 3 点を使って 2 次補間を行い最適パラメタを求めた。例えば、比較法（１）で $D_{\bar{c}}$ を最小にするパラメタと、その隣り合う 2 点を使って 2 次曲線フィッティングをし、その２次曲線の極を最適パラメタとする。比較に使う量は比較法（２）と（３）では $D_{\bar{\sigma}}$ および D_{norm} となる。ピーク周辺における 2 次補間は文献 [15] に詳細が述べられている。

表 1: パラメタ推定結果

	Method 1		Method 2		Method 3	
Ref.	p_{est}	n_{est}	p_{est}	n_{est}	p_{est}	n_{est}
B♭ *ff*	1.17	91.4	1.06	128.0	1.11	102.6
B♭ *mf*	0.56	59.8	0.63	75.0	0.61	92.6
B♭ *pp*	0.54	15.8	0.54	1.0	0.53	9.7
E♭ *ff*	1.27	128.0	0.89	103.8	0.63	43.9
E♭ *mf*	0.55	60.5	1.06	64.5	0.57	96.0
E♭ *pf*	0.54	51.6	1.07	98.2	0.55	53.8

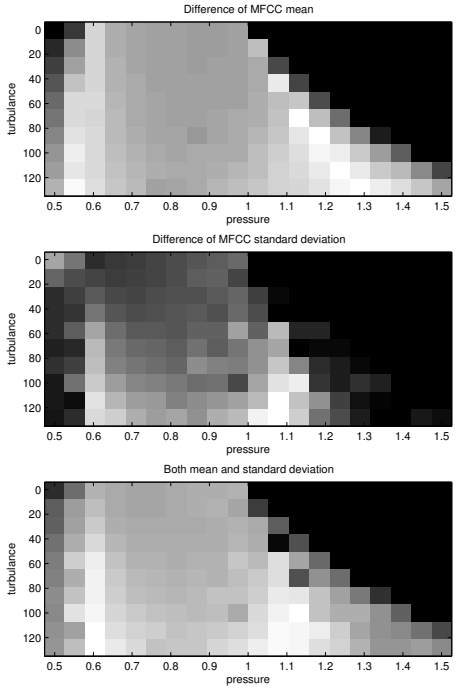

図 5: B 管クラリネット（ *ff* ）との音色類似度分布の変化。 色が明るい部分ほど類似度が高い。上：MFCC ベクトル平均、中：MFCC 標準偏差ベクトル、下：平均と標準偏差を用いて比較。

今回は２つのパラメタを取り扱っているので、吹鳴圧とノイズ振幅の二つの値が最小残差の検索の結果として与えられる。ここで、最小残差点付近で曲率がゼロでないと仮定し、２次補間は吹鳴圧とノイズ振幅の両方において行われる。

5　結果

パラメタ推定の結果を表 1 に示す。この表で、p_{est} は吹鳴圧、n_{est} はノイズ振幅である。このパラメタを用いて合成された音はオンラインで公開されている。(http://ccrma.stanford.edu/~hiroko/ICMC05/) 図 5 を見ると三種類の比較法によって音色の類似度分布が異なることがわかる。そこで比較法によって異なるパラメタ推定結果が得られる。今回、推定結果を評価するための心理音響実験は行わなかったが、非公式な聴取によって、比較法（１)、(３）は（２）よりよい合成音をもたらす印象を受けた。特に比較法（３）は音のゆらぎとラウドネスのバランスがうまくとられるように考えられる。比較法（３）においては式（３）における $D_{\bar{c}}$ と $D_{\bar{\sigma}}$ の重み係数を調整することでより良い結果が得られる可能性が高い。

6　おわりに

本稿では、物理モデルのパラメタ推定のために音色の知覚モデルを結合させる手法について述べた。クラリネットの物理モデルに短時間 MFCC のベクトル平均と標準偏差ベクトルによる音色評価を使用することで、２次元パラメタ空間での最適パラメタを求めた。この手法は他の様々な楽器の線形物理モデルに応用可能である。これからの課題としては、音色評価法を改善することでより良いパラメタ推定が可能になるであろう。また、本研究は聴覚フィードバックによる楽器演奏者と楽器のインタラクションの基本モデルであり、今後さらに研究を進めたい。

7　謝辞

本研究を進めるにあたり、聴覚モデルについてご教示頂いた IBM 研究所の Malcolm Slaney 氏に感謝いたします。

参考文献

[1] Smith, J. O. “Virtual Acoustic Musical Instruments: Review and Update”, *Journal of New Music Research*, vol. 33, no. 3, pp. 283–304, 2004.

[2] Smith, J. O. *Physical Audio Signal Processing: Digital Waveguide Modeling of Musical Instruments and Audio Effects, August 2004 Draft*, Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, 2004. Web published at http://ccrma.stanford.edu/~jos/pasp04/.

[3] Sundberg, J., Friberg, A., and Bresin, R. "Attempts to reproduce a pianist's expressive timing with Director Musices performance rules", *Journal of New Music Research*, 32:3, 317-325, 2003.

[4] Young, D., Serafin, S. "Playability Evaluation of a Virtual Bowed String Instrument", *Proceedings of International Conference on New Interfaces for Musical Expression*, Montreal, Canada, 2003.

[5] Traube, C., Dapalle, P., Wanderley, M. "Indirect Acquisition of Instrumental Gesture Based on Signal, Physical and Perceptual Information", *Proceedings of International Conference on New Interfaces for Musical Expression*, Montreal, Canada, 2003.

[6] Guillemain, P., Helland, R. Kronelend-Martinet, R. Ystad, S. "The clarinet timbre as an attribute of expressiveness", *Computer Music Modeling and Retrieval (CMMR2004)* pp. 246-259, Springer, 2004.

[7] Vergez, C., Rodet, X. "Trumpet and Trumpet Player: Model and Simulation in a Musical Context", *Proceedings of the International Computer Music Conference*, Havana, Cuba, 2001.

[8] Dubnov, S., Rodet, X. "Timbre Recognition with Combined Stationary and Temporal Features", *Proceedings of the International Computer Music Conference*, Ann Arbor, USA, 1998.

[9] D'haes, W. Rodet, X. "A New Estimation Technique for Determining the Control Parameters of a Physical Model of a Trumpet", *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)* London, UK, 2003.

[10] Terasawa, H., Slaney, M., Berger, J. "Perceptual Distance in Timbre Space", *Proceedings of the International Conference on Auditory Display (ICAD05)* Limerick, Ireland, 2005.

[11] Cook, P., Scavone, G. "The Synthesis ToolKit in C++ (STK)", available at http://ccrma.stanford.edu/software/stk/.

[12] Aoki, N., Ifukube, T. "Analysis and perception of spectral 1/f characteristics of amplitude and period fluctuations in normal sustained vowels", *J. Acoust. Soc. Am.* 106(1), July, 1999.

[13] Fritts, L. "University of Iowa Musical Instrument Samples", available at http://theremin.music.uiowa.edu/.

[14] Slaney, M., "Auditory Toolbox Ver. 2", Technical Report #1998-010, Interval Research Corporation. Available at http://www.slaney.org/malcolm/pubs.html.

[15] Smith, J. O., "PARSHL: A Program for the Analysis/Synthesis of Inharmonic Sounds Based on a Sinusoidal Representation", *Proceedings of the International Computer Music Conference*, Champaign-Urbana, USA, 1987. Extended version online at http://ccrma.stanford.edu/~jos/parshl/.