USING A PERCEPTUALLY BASED TIMBRE METRIC FOR PARAMETER CONTROL ESTIMATION IN PHYSICAL MODELING SYNTHESIS

Hiroko Terasawa, Jonathan Berger, Julius O. Smith
Center for Computer Research in Music and Acoustics (CCRMA)
Department of Music, Stanford University

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. INTRODUCTION

Music instrument performance is a complex sensorimotor behavior. Through training with auditory timbre feedback, the production task becomes finely controllable and seemingly automated.Attainment of expertise can be described as the stage in which timbre is conceptualized and an integrated set of control parameters, some with a remarkable degree of subtlety of change, are set with seemingly effortless thought. For the composer and orchestrator expertise involves similar conceptualization of desired timbre and knowledge of the notational cues needed to suggest appropriate production to the performer.

In the domain of digital music synthesis the absence of efficient integration between conceptualized timbre and parameter controls creates an impediment to effective and efficient timbre manipulation. In order to create a better (i.e., efficient, intuitive and interactive) auditory feedback loop a composer/performer of a physical model would benefit from a perceptually informed algorithm that estimates timbre control parameters.

A primary goal of physical model (PM) synthesis is to create sound that convincingly and compellingly approximates the sound of a humanly performed instrument [1, 2]. Coupling a model of the performer's technique and the auditory feedback loop that is vital in the creation of the musical sound is a vital aspect toward achieving this goal.

Some work toward this goal has been done. These include studies to embed expressive nuance in MIDI scores using KTH rules [3] and in parameter estimation of PM synthesis. Diana Young and Stefania Serafin investigated the playability of the violin physical model controlled by bow force and bow position [4]. Caroline Traube et al. has estimated the plucking position of a guitar using the spectral centroid for the timbre evaluation [5]. The analysis-synthesis research group at IRCAM has worked on the estimation of the control parameter of trumpet by diverse approaches: by inverting the trumpet physical model [6], vector quantization on cepstral coefficients and derivative of cepstral coefficients [7], and minimization of two perceptual similarity criteria as a function of the control parameters [8], to which, we chose a very similar approach. In spite of the difficulty in controlling the non-linear system with delayed feed-back, these attempts have been quite successful. Guillemain et al. investigated the distribution of the clarinet timbre as a function of control parameters using classical timbre descriptors [9].

Our goal is to create a control system for PM synthesis, which produces: 1) Desired pitch, loudness and timbre at 2) Desired time so that this system could integrate with existing expressive nuance rendering systems in which pitch, loudness and time are the typical control parameters.

In addition to the primary PM synthesis inputs of pitch and time it is important to address issues of loudness and timbre. The Mel-Frequency cepstral coefficient (MFCC) is a perceptually valid metric to capture both loudness and timbre [10].

Our ideal purpose is to provide a means of coupling a conceptualization of a desired timbre to the production parameters of physical models of acoustic instruments.There are two primary motivations for our work:

1. to provide improved performance and practicality in music composition and performance of physical models,

2. as a preliminary step toward a model of the auditory feedback loop in music instrument performance.

Our research employs the STK clarinet PM [11] and focused on breath pressure and breath turbulence as variable control parameters. Recordings of acoustic clarinet tones are used as reference tones to imitate. In order to evaluate the perceptual timbre difference, the synthesized sounds and a reference sound are compared in terms of the mean

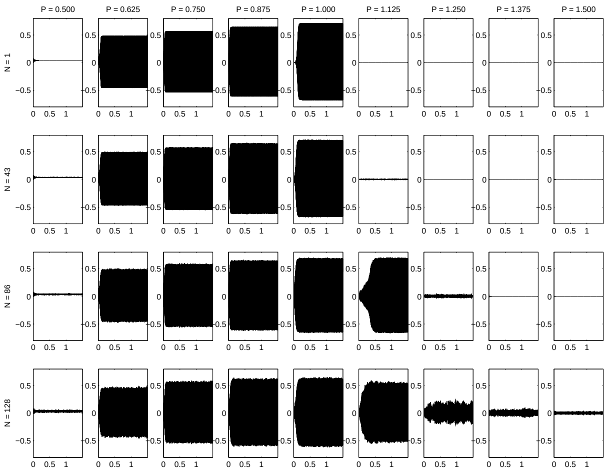

Figure 1. Waveforms of clarinet synthesis with increasing blowing pressure from left to right, and noise amplitude increasing from top to bottom. Note that both underblowing and over-blowing do not yield oscillation.

vector of Mel-Frequency Cepstral Coefficients (MFCC) $\bar{c}$ and the standard deviation $\bar{\sigma}$ of the MFCC over time frames. The $\bar{c}$ better captures the spectral envelope quality, while the $\bar{\sigma}$ better captures the breathiness and turbulence, which is important for naturalness in sound synthesis. We calculated the difference (sum of squared error) between $\bar{c}$ or $\bar{\sigma}$ of the reference and synthesis sounds. Finally, the control parameters which minimized the sum of squared error of $\bar{c}$ and/or $\bar{\sigma}$ are determined as the optimal parameters. The performances of comparisons using the MFCC, the standard deviation, and both are evaluated by informal listening.

2. TIMBRE RENDERING IN CLARINET PM SYNTHESIS

In this study, the blowing pressure and noise amplitude each have one degree of freedom, and all the other control parameters are fixed. The fundamental frequency is set to 440 Hz for all the synthesis and recording sounds.

Figure 1 shows the waveforms of the resulting synthesis with the blowing pressure and noise amplitude parameters varied. There is a clear correspondence between the amplitude and the blowing pressure, and between breathiness and the noise amplitude. In addition, over-blowing the instrument does not provide a steady oscillation. Therefore when the amplitude pressure grows too large, it finally mutes the synthesis (reed blown shut).

For the estimation, synthetic sounds were compared to the reference sound, with the blowing pressure varied with 20 levels and the noise amplitude varied with 10 levels.

The recordings of clarinet sounds from the University of Iowa Musical Instrument Samples are used as the reference sounds to imitate [13]. We picked A4 sounds with the dynamics *pp*, *mf*, *ff* from both B♭ and E♭ clarinets.

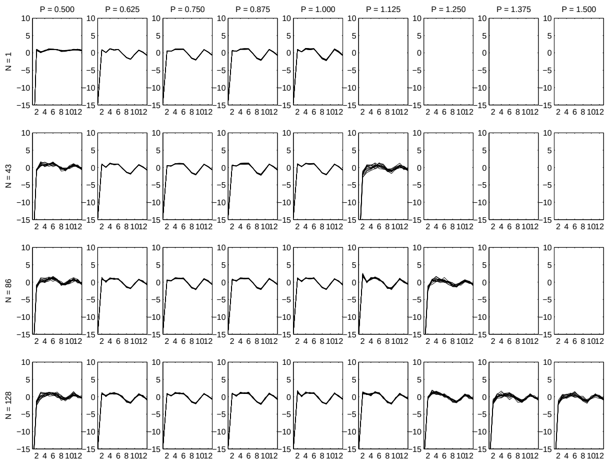

Figure 2. MFCC of clarinet synthesis shown in Figure 1. Ten frames are overlaid and the deviation of the MFCC becomes larger as the turbulence grows.

3. TIMBRE EVALUATION BY MFCC

The MFCC is a perceptually valid metric for a timbre description of static sounds. It is calculated as follows: A spectrum is converted into total energy per critical band using an auditory filterbank. Then the first 13 coefficients from the inverse DCT of the filterbank output represent the log-magnitude spectral envelope. The Auditory Toolbox is used for the actual computation [14].

Figures 2 and 3 show MFCCs for the steady-state portions of recorded and resynthesized clarinet sounds. In each subfigure, MFCCs from ten successive time frames are overlaid (frame length = 5.8 ms, frame step size = 2.3 ms). When the turbulence is larger, we see more deviation in the MFCC. Therefore, the sample mean MFCC vector $\bar{c}$ (the average of MFCC vectors across 80 successive time frames), and the sample standard deviation MFCC vector $\bar{\sigma}$ are used as the basis for the timbre comparison.

Both $\bar{c}$ and $\bar{\sigma}$ are vectors of length 13.

4. ESTIMATION METHOD

4.1. Comparison between Timbres

The timbre of synthesis sound is compared to the timbre of the reference sound, in terms of the sum of squared errors

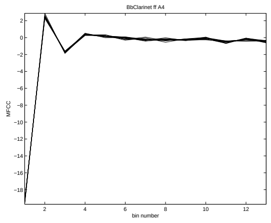

Figure 3. Overlay of the MFCC of ten successive frames from the recorded sound: A4, *ff*, B-flat Clarinet.

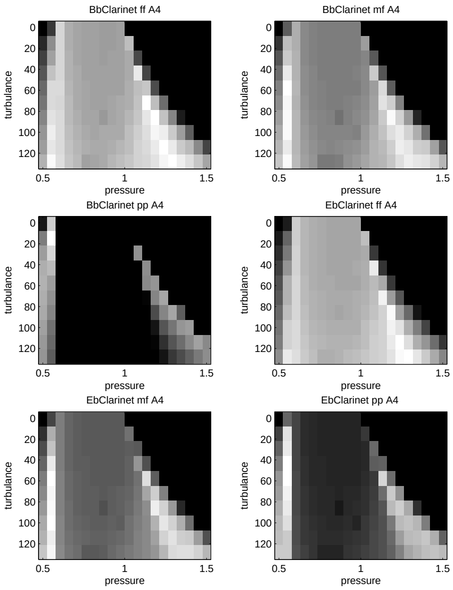

Figure 4. Method 1: Timbre difference $D_{\bar{c}}$ between reference and synthesis sounds. Reference sounds are from a B♭ and E♭ Clarinet at dynamic levels *ff*, *mf* and *pp*. The brighter cells correspond to more similarity between reference and synthesis sounds. Note how the distribution of similarity varies with reference sound.

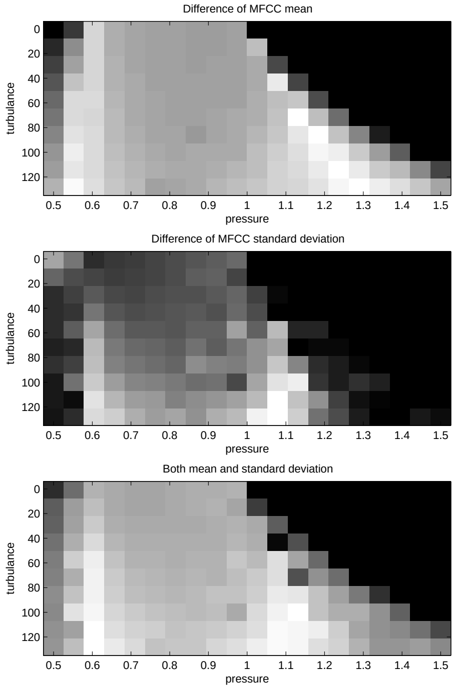

Figure 5. Timbre similarity distribution for the B♭ Clarinet *ff* sound measured by, Top: Mean MFCC, Middle: MFCC standard deviation, Bottom: Both.

$D_{\bar{c}}$, $D_{\bar{\sigma}}$ and D_{norm} which are defined as follows:

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

where $\bar{c}_{synth}$ and $\bar{c}_{ref}$ are the mean MFCC for the synthesis and reference sounds, respectively.

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

where $\bar{\sigma}_{synth}$ and $\bar{\sigma}_{ref}$ are the standard deviation of the MFCCs for synthesis and reference sounds, respectively.

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

We calculate the sum square errors for all 200 synthesis sounds and one of the reference sounds, e.g., for the B♭ Clarinet at *ff*, and we determine the pair of blowing pressure and noise amplitude which provide the minimum error as the best control parameters among the parameters used in the synthesis. Three methods of judgment are investigated: (1) Minimize $D_{\bar{c}}$. (2) Minimize $D_{\bar{\sigma}}$. (3) Minimize D_{norm}. Figure 4 shows the distribution of similarity for six reference sounds measured by Method 1.

Table 1. Estimated parameters

	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

4.2. Estimation by Quadratic Interpolation

Finding minimum points from the distribution of similarity, we are only able to know the best pair of parameters from the set of discrete parameters. The optimal parameter may fall into the gap between the discrete levels. In order to estimate such parameters, we used quadratic interpolation based on three adjacent points: For example, in Method 1, the point which minimizes $D_{\bar{c}}$ and two adjacent points in the parameter space are picked for the quadratic curve fitting. The metric could be either $D_{\bar{\sigma}}$ or D_{norm} for Methods 2 and 3. Quadratic interpolation of peaks is described in, e.g., [15].

In our parameter estimation, the two parameters, the blowing pressure and noise amplitude, are given as the result of minimum error search in the distribution of similarity. We assume nonzero curvature in the neighborhood of the minimum error point, and quadratic interpolations is performed along both blowing pressure and noise amplitude.

5. RESULTS

The estimated parameters are given in Table 1. The synthesized sounds using these estimated parameters are provided online at http://ccrma.stanford.edu/˜hiroko/ICMC05/.

Figure 5 shows that the distribution of similarity is different among the three comparison metrics. Therefore, each comparison method provides a different estimate. We have not organized formal psychoacoustic listening tests for evaluating the comparison methods. Based on informal listening, we believe Methods 1 and 3 provided better estimates than Method 2. Method 3 seems to be a good way to incorporate timbre, loudness and breathiness. However, it should be possible to improve Method 3 (Eqn. 3) by optimizing weighting coefficients applied to $D_{\bar{c}}$ and $D_{\bar{\sigma}}$.

6. CONCLUSIONS AND FUTURE WORK

Minimizing short-time MFCC and MFCC standard deviation differences between a reference and synthesized sound was found to be effective for control parameter estimation in physical modeling synthesis. This framework is expected to be similarly effective for other kinds of sound synthesis models. Further work on the timbre comparison metric should enable improved estimation. We consider this to be a preliminary step toward an integrated model of a performer's interaction with a physical model of a musical instrument.

7. ACKNOWLEDGMENT

We appreciate Malcolm Slaney's helpful tutorials on the auditory models and timbre metrics.

8. REFERENCES

[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] 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.

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

[8] 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.

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

[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/.