Feature-Based Synthesis for Sonification and Psychoacoustic Research

Matt Hoffman

Princeton University
Department of Computer Science
35 Olden Rd, Princeton, NJ 08540, USA
mdhoffma@cs.princeton.edu

Perry R. Cook

Princeton University
Departments of Computer Science and Music
35 Olden Rd, Princeton, NJ 08540, USA
prc@cs.princeton.edu

ABSTRACT

We present a general framework for synthesizing audio manifesting arbitrary sets of perceptually motivated, quantifiable acoustic features. Much work has been done recently on finding acoustic features that describe perceptually relevant aspects of sound. The ability to synthesize sounds defined by arbitrary feature values would allow perception researchers to more directly generate stimuli "to order," as well as providing an opportunity to directly test the perceptual relevance and characteristics of such features. The methods we describe also provide a straightforward way of approaching the problem of mapping from data to synthesis control parameters for sonification.

1. INTRODUCTION

Much of the work done in sound analysis focuses on applying machine learning techniques to descriptors consisting of vectors of perceptually motivated features derived from input audio signals. These features are statistics thought to correlate well with the sonic cues that humans use to perceive sound. We have implemented a general framework for synthesizing, given any set of such feature values, frames of audio matching them as closely as possible.

This approach differs from most existing sound synthesis techniques, whose control parameters are generally defined by underlying synthesis algorithms rather than by the desired output. Such approaches require that a human operator manually explore the space of sounds that such an algorithm is capable of producing to find a set of parameters that will produce the type of sound for which he or she is looking. The parameter spaces for flexible synthesis algorithms capable of producing a wide range of sounds can have dozens of interdependent dimensions. This can make it challenging for a user unfamiliar with the algorithm to synthesize a sound having specific predetermined characteristics, and nearly impossible for an untrained computer. By contrast, our feature-based synthesis system enables sounds characterized by any set of quantifiable features to be synthesized using any underlying synthesis algorithm, so long as there exists a set of synthesis parameters that maps to the desired feature values and can be found reasonably efficiently.

If one chooses a set of features for their perceptual relevance, the ability to use those features to drive synthesis becomes useful for a number of applications. The problem of how to map data onto synthesis parameters for sonification can be solved by directly mapping the data values onto synthesizable perceptual characteristics. Researchers designing experiments to quantify the relationships between various psychoacoustic and perceptual features of audio can synthesize tightly controlled audio stimuli that allow quantifiable relationships to be drawn between acoustic phenomena and perceptual impact. Similarly, researchers trying to model specific perceptual characteristics of sound for, e.g., music and audio information retrieval purposes can use feature-based synthesis to evaluate how well given feature sets meet their needs. By extracting a set of features from a sound and applying feature-based synthesis to resynthesize a sound matching those features, one can directly observe what information the chosen features do and do not capture, which may suggest new features capable of encoding relevant information. Finally, feature-based synthesis may be especially useful in experiments designed to explore and quantify the relationships between low-level acoustical features and the higher-level ecological features that people use to make sense of the world around them.

Our system frames the problem of feature synthesis in terms of minimizing an arbitrarily defined distance function between the target feature vector and the feature vector describing the synthesized sound over the set of underlying synthesis parameters. The mapping between the feature space and the parameter space can be highly nonlinear, complicating optimization. We present a modular framework that separates the tasks of feature extraction, feature comparison, sound synthesis, and parameter optimization, making it possible to mix and match various techniques in the search for an efficient and accurate solution to the broadly defined problem of synthesizing sounds manifesting arbitrary perceptual features.

2. PREVIOUS WORK

Previous techniques for automatically fitting synthesis parameters to some desired output have tended to fall into one of two categories. The first set of approaches revolves around finding synthesis parameter values that will produce a sound closely matching an existing sound. The second focuses on mapping some set of input parameters onto a set of synthesis parameters in a way that gives users straightforward control over the synthesized output without sacrificing the range of available sounds.

2.1. Matching Synthesis

This style of synthesis techniques is less applicable to the problems we hope to address, however we mention it due to parallels between such approaches and ours.

Horner, Beauchamp, and Haken [1] used genetic algorithms to find the parameters for single modulator, multiple carriers FM that minimize the difference between the resulting signal's spectrum and that of an arbitrary instrumental tone. Horner, Cheung, and Beauchamp [2] also applied similar techniques to other synthesis algorithms, all

to good effect. More recently, Horner, Beauchamp, and So [3] and Wun, Horner, and Ayers [4] explored other error metrics, including perceptually influenced metrics, and search methods, respectively.

Although the approaches taken by Horner et al. over the years have, like our system, used various algorithms to find synthesis parameters that minimize the error of the audio generated by several synthesis functions with respect to some target sound, our primary goal is not the accurate resynthesis of existing sounds but the synthesis of new sounds for auditory display and psychoacoustic research.

2.2. User Input to Synthesis Parameter Mapping

Much of the work on facilitating the expressive use of real-time controllers for music has focused on the problem of mapping the input parameters from such controllers onto the parameters controlling a synthesis algorithm. This problem is directly analogous to the problem of sonifying multivariate data. Whereas the first problem is concerned with translating the musical intention of a performer (as represented by a number of inputs varying through time) into sound that makes that intention perceptible to others, sonification is concerned with translating the salient trends and features present in a (frequently multidimensional) data set perceptible to those interested in analyzing that data set.

A system that underscores these parallels particularly strongly is Joseph and Lodha's MUSART system [5], which allows users to map data values to traditional musical parameters such as melody, rhythm, and harmony. However, MUSART provides no mechanism for the continuous modulation of timbre, whereas most of the attempts at developing techniques for mapping musical controller parameters onto synthesis parameters have focused intently on controlling timbre.

The problem of mapping input parameters into a timbre space, which [6], among others, points to as a potentially powerful technique for sonifying data, has been explored on several occasions in the context of musical controllers. Lee and Wessel [7] explored the use of neural networks to map from an input space into a user-generated timbre space. Taking a more straightforward, mathematical approach, Bowler et al. [8] developed a method for efficiently mapping N control parameters to M synthesizer parameters, where the mapping was again defined in terms of a user-generated timbre space of sorts filled out by interpolating between control points. Although such techniques may be effective for specific instruments and synthesis algorithms, they demand that the user define an explicit mapping from synthesis parameter space to an ersatz timbre space. This must be repeated any time a new synthesis algorithm is used, which limits the generalizability of such techniques.

2.3 Feature-Specific Synthesis

Some work has been done in synthesizing sounds to fit values for specific features. Lidy, Pölzlbauer, and Rauber [9], for example, developed a technique to synthesize audio from their Rhythm Patterns feature descriptor having very specific beat patterns in various frequency bands, and Lakatos, Cook, and Scavone [10] have synthesized noises with specific spectral centroids for their studies of human perception. But, although it is frequently possible or even trivial to find a closed-form solution to the problem of synthesizing sounds having specific values for one or two feature values, finding parameters to synthesize a sound exactly fitting many feature values very quickly grows intractable.

3. MOTIVATION

Our approach's flexibility permits its application to a number of problems. We give three examples – generation of stimuli for psychoacoustic and perceptual research, mapping of multidimensional data for sonification, and mapping sounds into human-generated timbre spaces.

3.1. Psychoacoustic Stimulus Generation

Studies of selective attention such as those described in [10] and [11] study the impact of various characteristics of sounds on listeners' abilities to perceive or attend to those sounds in the presence of distractions. These studies have tended to either focus on relatively simple acoustic features such as pitch and spectral centroid or on ecological characteristics of the objects producing the sounds being used as stimuli. Using feature-based synthesis, such studies could be conducted on more complex features that may be difficult to synthesize directly such as spectral flux (the difference between the power spectra of successive frames), dissonance (as measured using techniques such as that described in [12]), or spectral sparseness (the number of frequency bins in the power spectrum with power below the average bin). Feature-based synthesis offers the additional ability to vary one feature while keeping other features constant, allowing exploration of the differences in the perceptual effects of one feature in the presence of other features in differing amounts.

3.2. Sonification

In recent years, many general-purpose toolkits and frameworks of various flavors have been developed to facilitate the mapping of data to some set of synthesis parameters. The MUSART system mentioned above [5] chooses parameters derived from music theory; the Sonification Sandbox [13] offers more contextual cues, but less flexible synthesis options than MUSART; SonART [14] leverages the flexibility of the Synthesis ToolKit to allow a broad range of parameters; and Pauletto and Hunt's [15] Interactive Sonification Toolkit offers a number of synthesizers, mostly using traditional analog synthesizer designs. Of these, it is worth noting that only SonART and the Interactive Sonification Toolkit offer the ability to manipulate timbre in a continuous fashion, and in any case the most perceptually appropriate mapping of data values onto synthesis parameters is not necessarily obvious.

Feature-based synthesis provides a straightforward paradigm for mapping between data values of interest and synthesis parameters. Data values can be mapped directly onto a set of perceptually salient features, with the system handling the details of choosing low-level synthesis parameters.

3.3. Mapping into Human-Generated Timbre Spaces

Studies such as [10] [16] [17] have investigated the human ability to perceive various physical attributes of sound sources. As Ottaviani and Rocchesso [18] point out, listeners can easily attend to these attributes seperately, making them good candidates for use in auditory display. We suggest that feature-based synthesis could be of use in studying the low-level acoustical properties that human listeners use to deduce the more complex physical attributes of a sound's source. We can generate sounds defined over a set of features we expect to correlate with listeners'

perceptions of, e.g., size, material, or shape, and then use techniques like those described in [19] to determine how those sounds map to the ecological features we wish to study. From the data points obtained in this way, we may be able to discover consistent relationships between acoustical and human-generated features that can be used to predict how a sound manifesting certain acoustic feature values will be perceived. This information can then be used to synthesize sounds with arbitrary perceived physical attributes.

4. IMPLEMENTATION

Our architecture focuses on four main modular components: feature evaluators, parametric synthesizers, distance metrics, and parameter optimizers. Feature evaluators take a frame of audio as input and output an n-dimensional vector of real-valued features. Parametric synthesizers take an m-dimensional vector of real-valued inputs and output a frame of audio. Distance metrics define some arbitrarily complex function that compares how "similar" two n-dimensional feature vectors are. Finally, parameter optimizers take as input a feature evaluator **F**, a parametric synthesizer **S**, a distance metric **D**, and an n-dimensional feature vector **v** generated by **F** (which **D** can compare to another such feature vector). The parameter optimizer **P** outputs a new m-dimensional synthesis parameter vector **u'**, a new n-dimensional feature vector **v'**, and a frame of audio representing the output of **S** when given **u'**. This frame of audio produces **v'** when given as input to **F**. **v'** represents the feature vector as close to **v** (where distance is defined by **D**) as **P** was able to find in the parameter space of **S**.

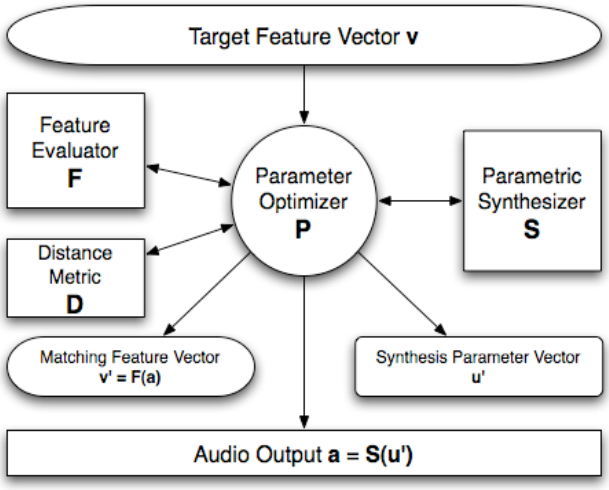

Figure 1. *Overview of the architecture of the framework. Given a target feature vector* **v**, **P** *searches through the parameter space of* **S** *to find a set of synthesis parameters* **u'** *that will minimize* **D**(**v**, **F**(**S**(**u'**))).

These four components together make up a complete system for synthesizing frames of audio characterized by arbitrary feature vectors. Any implementation of one of these components is valid, so long as it adheres to the appropriate interface. The design of the framework makes it possible to use any parameter space search algorithm to seek parameters for any synthesis algorithm that produce audio characterized by any set of automatically extractable features and any distance function.

5. PRELIMINARY RESULTS

Preliminary results using relatively crude optimizers and a simple Euclidean L2 distance function with stationary sinusoids of various frequencies and spectrally shaped noise (with noise shape controlled by Mel-Frequency Cepstral Coefficients as implemented by Slaney [20]) to match spectral centroids, spectral rolloffs, multiple pitch histogram data [21] and spectral sparseness have been good. Figures 2 and 3 present an example of a sound file synthesized using our system to have its 33% spectral rolloff (i.e., the frequency in the spectrum below which 33% of the total spectral energy resides) stationary at 30% of the Nyquist frequency (in this case 22050 * 0.3 = 6615 Hz), and its 67% spectral rolloff move from 50% of the Nyquist frequency to 80% of the Nyquist frequency. In this example we use 13 MFCC coefficients to control the shaping of noise as our underlying synthesis algorithm. For the most part we are able to closely match the feature values we attempt to synthesize, although some error is inevitable.

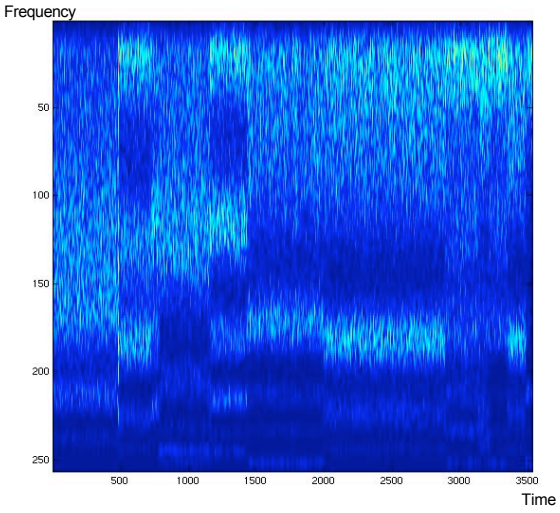

Figure 2. *Time versus frequency spectrogram of sound synthesized by our system to match two spectral rolloff values simultaneously.*

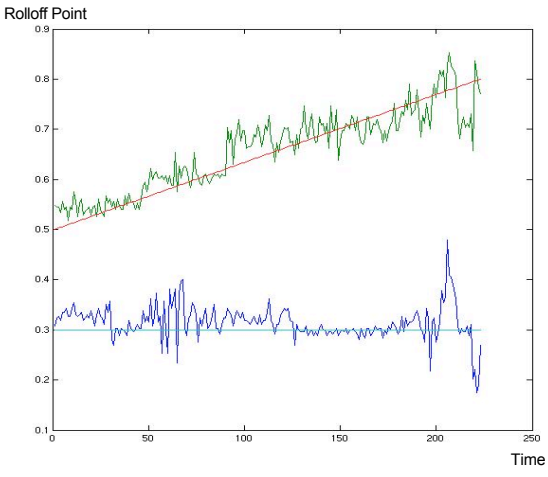

Figure 3. *Plot of the values of the target and synthesized 33% rolloff values (below) and the target and synthesized 67% rolloff values (above). The x-axis is time, and the y-axis is spectral rolloff.*

6. CONCLUSION AND FUTURE WORK

We have implemented a framework for synthesizing frames of audio described by arbitrary sets of well-defined feature values. The techniques implemented by our system can be used to a wide variety of ends such as synthesizing stimuli for studies of human perception, sonification, and timbre space modeling.

The next stage consists of implementing more feature evaluators, synthesizers, optimizers, and distance metrics, and evaluating the system's performance more rigorously in a variety of domains. Finding ways to improve search times will be particularly valuable in preparing the framework for use in real-time situations. We also plan to implement an interactive interface to facilitate testing and development.

7. REFERENCES

[1] A. Horner, J. Beauchamp, and L. Haken, "Machine Tongues XVI: Genetic algorithms and their application to FM matching synthesis," *Computer Music Journal* 17(3), pp. 17-29, 1993.

[2] A. Horner, N. Cheung, and J. Beauchamp, "Genetic algorithm optimization of additive synthesis envelope breakpoints and group synthesis parameters," in *Proceedings of the International Computer Music Conference* 1995*,* pp. 215-222.

[3] A. Horner, J. Beauchamp, and R. So, 2004, "A search for best error metrics to predict discrimination of original and spectrally altered musical instrument sounds," in *Proceedings of the International Computer Music Conference* 2004*,* pp. 9-16.

[4] S. Wun, A. Horner, and L. Ayers, "A comparison between local search and genetic algorithm methods for wavetable matching," in *Proceedings of the International Computer Music Conference,* 2004, pp. 386-389.

[5] A.J. Joseph and S.K. Lodha, "MUSART: Musical Audio Transfer Function Real-Time Toolkit," in *Proceedings of the International Conference on Auditory Display*, Kyoto, Japan, 2002.

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

[7] M. Lee, and D. Wessel, "Connectionist models for real-time control of synthesis and compositional algorithms," in *Proceedings of the International Computer Music Conference* 1992 pp. 277-280.

[8] I. Bowler, A. Purvis, P. Manning, and N. Bailey, "On mapping N articulation onto M synthesizer-control parameters," in *Proceedings of the International Computer Music Conference* 1990, pp. 181-184.

[9] T. Lidy, G. Pölzlbauer, and A. Rauber, "Sound re-synthesis from rhythm pattern features – audible insight into a music feature extraction process," in *Proceedings of the International Computer Music Conference* 2005*,* pp. 93-96.

[10] S. Lakatos, P. Cook, and G. Scavone, "Selective attention to the parameters of a physically informed sonic model," *Acoustics Research Letters Online*, Acoustical Society of America, March 2000.

[11] Scharf, B. "Auditory attention: The psychoacoustical approach," in *Attention*, edited by H. Pashler et al. Psychology Press, Hove, U.K., 1998.

[12] A. Kameoka, and M. Kuriyagawa, "Consonance theory, Part II: Consonance of complex tones and its calculation method," *Journal of the Acoustical Society of America*, 45, 1460–1469, 1969.

[13] B. N. Walker and J. T. Cothran, "Sonification Sandbox: A graphical toolkit for auditory graphs," in *Proceedings of the International Conference on Auditory Display*, Boston, MA, 2003.

[14] O. Ben-Tal, J. Berger, B. Cook, M. Daniels, G. Scavone, and P. Cook, "SONART: the sonification application research toolbox," in *Proceedings of the Int. Conf. On Auditory Display*, 2002.

[15] S. Pauletto and Andy Hunt, "A toolkit for interactive sonification," in *Proceedings of the Int. Conf. on Auditory Display*, 2004.

[16] P. Cook and S. Lakatos, "Using DSP-based parametric synthesis models to study human perception," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, NY, 2003.

[17] D. Rocchesso, "Acoustic cues for 3-D shape information," in *Proceedings of the 2001 International Conference on Auditory Display*, Espoo, Finland, 2001.

[18] L. Ottaviani and D. Rocchesso, "Auditory perception of 3D size: experiments with synthetic resonators," *ACM Transactions on Applied Perceptions*, Vol. 1, No. 2, Pages 118–129, October 2004.

[19] Scavone, G., Lakatos, S., Cook, P., and Harbke, C., "Perceptual spaces for sound effects obtained with an interactive similarity rating program," in *Proceedings of the International Symposium on Musical Acoustics,* 2001

[20] M. Slaney, "Auditory toolbox," Technical Report # 1998-010. Interval Research Corporation, Palo Alto, CA, 1998.

[21] G. Tzanetakis, A. Ermolinskyi, P. Cook, "Pitch histograms in audio and symbolic music information retrieval," in *Proceedings of ISMIR*, 2002.