Feature-Based Synthesis: A Tool for Evaluating, Designing, and Interacting with Music IR Systems

Matt Hoffman
Princeton University
Computer Science Department
35 Olden Street
Princeton, NJ 08544
mdhoffma@cs.princeton.edu

Perry R. Cook
Princeton University
Computer Science & Music Departments
35 Olden Street
Princeton, NJ 08544
prc@cs.princeton.edu

Abstract

We present a general framework for performing feature-based synthesis – that is, for producing audio characterized by arbitrarily specified sets of perceptually motivated, quantifiable acoustic features of the sort used in many music information retrieval systems.

1. Introduction

We have implemented a general framework for performing feature-based synthesis, which attempts to synthesize, given any set of feature values, audio that matches those feature values as closely as possible. Depending on how one chooses the values to synthesize, feature-based synthesis can be used to evaluate the usefulness of a given set of features for a particular audio IR domain, to diagnose why a system is not performing as well as expected, as a tool for gaining insight into what information a set of features is encoding, and to generate stimuli for use in studies of human perception.

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

2. Motivation

2.1 Feature Evaluation and Selection

Feature-based synthesis can be used to address this question of what relevant qualities, if they were encoded by one's feature set, might enable better performance on a problem. As Lidy, Pölzlbauer, and Rauber [1] observe, one way of qualitatively evaluating the meaningfulness of a feature set is through an analysis-by-synthesis process where one extracts the features in question from multiple sounds from the target domain, synthesizes new sounds matching the extracted features, and compares the original and resynthesized versions. If the resynthesized version of a sound file lacks some quality relevant to the problem at hand, then it is likely that adding a feature representing that quality to the feature set will improve performance.

2.2 Feature Exploration

Our system provides an interface for synthesizing audio manifesting feature values specified in real-time, which can be used to gain a more intuitive understanding of how the various features one is using map to actual sounds. Attempting to generate sounds with specific perceptual characteristics in this way can stimulate insights into how much descriptive power a feature set has.

2.3 Perceptual Study Stimulus Generation

Studies such as [2] [3] [4] have investigated the human ability to perceive various physical attributes of sound sources. 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 [5] 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.

2.4 Classification System Evaluation

We can also treat the confidence outputs of entire classification systems as features to match, enabling us to gain insights into what sorts of audio a system strongly believes fit into one category or another, as well as what sorts of audio it finds difficult to classify.

3. Related Work

Our system's approach of synthesizing audio to fit quantifiable targets is not entirely without precedent. For example, a presentation at the 2004 ISMIR graduate school [6] mirrors many of the ideas behind our system. Another approach seeks synthesis parameter values that will produce a sound closely matching the spectrum of an existing sound (e.g. [7, 8, 9]). Some work has been done on synthesizing audio manifesting a limited number of specific feature values (e.g. [1]). Finally, relatively recent work in concatenative synthesis and audio mosaicing, although sometimes working to somewhat different ends, faces some of the same challenges as feature-based synthesis. An overview of current work in concatenative synthesis and audio mosaicing can be found in [10].

4. Implementation

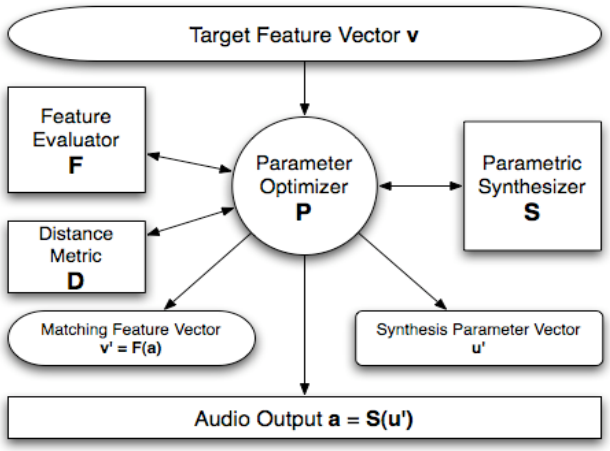

Figure 1. Overview of the architecture of the framework. Given a target feature vector v**,** P **searches through the parameter space of** S **to find a set of synthesis parameters** u' **that will minimize** D(v, F(S(u'))).

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**.

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.

5. Future Work

The next phase of the project will involve implementing more feature evaluators, synthesizers, optimizers, and distance metrics, and evaluating the system's performance more rigorously in a variety of domains. Additionally, we will extend the framework to more directly handle time-domain features and interpolate between parameters smoothly to produce smoother, more natural output.

References

[1] 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.

[2] 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.

[3] 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*, 2003.

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

[5] 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

[6] S. Le Groux, "Extraction of Relevant Controllers for the 'Analysis by Synthesis' of Musical Sounds," research proposal presentation given at ISMIR 2004 Fifth Int. Conf. On Music Inf. Retr., graduate school, available at http://www.iua.upf.es/mtg/ismir2004/graduateschool/

[7] 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.

[8] 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.

[9] 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.

[10] D. Schwarz, "Current Research in Concatenative Sound Synthesis," in *Proceedings of the International Computer Music Conference,* 2005, pp. 802-805.05, pp. 802-805.