Unsounding Objects: Audio Feature Extraction for the Control of Sound Synthesis

Ian Hattwick
Center for Interdisciplinary Research in Music Media and Technology
Input Devices and Music Interaction Lab
McGill University
ian.hattwick@mail.mcgill.ca

Preston Beebe
Center for Interdisciplinary Research in Music Media and Technology
Digital Composition Studio
McGill University
preston.beebe@mail.mcgill.ca

Zachary Hale
Center for Interdisciplinary Research in Music Media and Technology
Percussion Performance
McGill University
zachary.hale@mail.mcgill.ca

Marcelo Wanderley
Center for Interdisciplinary Research in Music Media and Technology
Input Devices and Music Interaction Lab
McGill University
marcelo.wanderley@mcgill.ca

Phillippe Leroux
Center for Interdisciplinary Research in Music Media and Technology
Digital Composition Studio
McGill University
phillippe.leroux@mcgill.ca

Fabrice Marandola
Center for Interdisciplinary Research in Music Media and Technology
Percussion Performance
McGill University
fabrice.marandola@mcgill.ca

ABSTRACT

This paper presents results from the development of a digital musical instrument which uses audio feature extraction for the control of sound synthesis. Our implementation utilizes multi-band audio analysis to generate control signals. This technique is well-suited to instruments for which the gestural interface is intentionally weakly defined. We present a percussion instrument utilizing this technique in which the timbral characteristics of found objects are the primary source of audio for analysis.

Keywords

digital musical instrument, audio feature extraction, percussion, interdisciplinary

1. INTRODUCTION

Unsounding Objects is a project involving the creation of *digital musical instruments* [**?**] which use audio feature extraction for the control of sound synthesis. The instrument presented here, the *SpectraSurface*, is a percussion instrument which allows the use of found objects as input devices. Percussionists are accustomed to intimate control of timbre using a wide variety of performance techniques, and the *SpectraSurface* leverages this expert technique in order to allow for the intuitive control of a digital percussion instrument.

Most existing digital percussion controllers create control signals consisting of discrete events which are based upon the velocity of a hand or stick striking a membrane.[1] These instruments are meant to mimic traditional drum set and hand percussion performance practices and are well-suited to triggering audio samples; however, they are not well-suited to capturing percussionists nuanced control of timbre, do not allow for the incorporation of found objects into performance, and are not compatible with most extended performance techniques. The *Spectrasurface* was designed specifically to address these issues.

Unsounding Objects is an interdisciplinary research project whose team consists of a digital musical instrument designer, a composer, and a performer. An important part of this project has been the joint development of new musical instruments, performance practices, and compositional approaches, and especially how these three research areas inform each other. While one of the driving goals of our research is the application of scientific approaches to audio feature extraction and mapping strategies to the expansion of performance practice and composition for percussion instruments, we are aware that in order for the development of a new musical instrument to result in a real contribution to the musical community it must incorporate the contribution of the performers who play it and the composers who write for it. The technical details of the instrument design which is described in this paper was heavily influenced by the collaboration of the research team in regular workshops and our experience using the instrument in concert.

2. RELATED WORK

Audio feature extraction is commonly used within the Music Information Retrieval community for the automatic classification of music recordings [**?**]. It also has a long history of use for the generation of data structures for composition and music performance [**?**] [**?**]. Many different audio features are able to be extracted, such as brightness, spectral centroid, harmonic flux, noisiness, etc. For the purposes of this paper, we consider an audio feature to be a numerical representation of some perceptual aspect of the spectral content of an audio signal.

There are two main approaches taken to the use of audio features for control of synthesis. In the first approach audio signals are analyzed for specific characteristics such

[1] This includes commercial instruments such as the Roland V-Drum as well as instruments within the NIME community.

NIME'14, June 30 – July 03, 2014, Goldsmiths, University of London, UK.

Figure 1: The SpectraSurface showing the four surfaces as used in the composition *Unsounding Objects #1.*

as spectral centroid or harmonicity, which are then mapped directly to parameters of sound synthesis [**?**]. In the second approach, machine learning techniques are used to classify and correlate spectral information to pre-analyzed spectral templates [**?**, **?**]. Machine learning techniques are well-suited to analyzing large numbers of frequency components, such as those provided by FFTs; however, they also tend to be slower to respond to transients, and we found them to be less effective with complex spectra such as those which occur at note onsets. A typical use would be the classification of performance techniques as seen in Tindale et al [**?**].

2.1 Audio Feature Extraction in DMI Design

Audio feature extraction has been applied to digital musical instrument design in many ways. Jehan proposed a timbre model which uses pitch, loudness, and brightness to control sound synthesis; this model was then embedded into the design of the Hyperviolin [**?**]. Shiraishi presented a timbre tracking model in which an input timbre space is constructed which is a representation of the actual range of timbres which an acoustic instrument can create. In performance, an input sample frame is analyzed and then its location within this timbre space is determined. The input timbre space is then projected onto a target timbre space which controls sound synthesis [**?**]. O'Modhrain and Essl's Pebblebox uses onset detection, amplitude measurement, and spectral content to allow the manipulation of pebbles to control granular synthesis [**?**]. Their approach is noteworthy for the correlation of selected audio features and synthesis method. The impulses generated by large quantities of pebbles striking each other, for example, bear a direct relationship to the triggering of clouds of audio events in granular synthesis. Further related works are Settel and Lippe's implementations of FFT analysis and re-synthesis [**?**] and Puckette's use of spectral envelopes for parameter mapping [**?**].

Two related instruments have also been recently developed at IRCAM. Lorenzo Pagliei's *Geecos* use machine learning techniques to identify specific performance techniques as well as using perceptual audio features to directly control synthesis parameters.[2] Bruno Zamberlin's MoGees[3] utilize piezo sensors to detect audio signals which are then processed using a mobile phone application which appears to be based on his earlier work implementing gesture following techniques using Hidden Markov Models [**?**].

[2] `www.youtube.com/watch?v=6Si2Y9Sm4AE`, accessed January 29, 2013.

[3] `www.brunozamborlin.com/mogees/`, accessed February 4, 2014.

2.2 Relation to Direct and Indirect Gesture Acquisition

Direct acquisition of performer gesture through sensor data is the most common approach to digital musical instrument design [**?**]. Indirect gesture acquisition uses audio analysis to extract information about a performer's gestures, which is then used to control sound synthesis. Both of these methods depend upon the explicit definition of correct performance gestures, and are relatively intolerant of gestures which do not fit this definition. While you can strum the strings of a Yamaha Disklavier directly, you cannot use this gesture for the control of sound synthesis.

Audio feature extraction has the benefit that any sound created on the instrument will be analyzed, regardless of performer gesture, as long as an emphasis on predefined models isn't built into its algorithm. In practice, the algorithms, mappings, and synthesis techniques will be chosen based on the instrument's expected use, thus limiting the expressive potential of arbitrary gestures. However, there is still the potential for extended techniques in the sense of gestures outside the expected norm.

3. DESIGN GOALS FOR THE SPECTRASURFACE

In creating the SpectraSurface our primary goal was the creation of a digital percussion instrument whose interface can be any found object, played with any performance technique. We took the approach of using surfaces upon which arbitrary objects can be placed. A piezo contact microphone is attached to this surface. When a performer interacts with objects placed upon the surface the sound produced is captured by the microphone and sent to a computer for analysis. The implementation described here consists of four surfaces thus equipped, mounted into a suitcase. A dedicated 4-channel preamp was created whose output is sent directly to the line inputs on a multi-channel audio interface. A fixed-gain preamp (powered by a DC adaptor) and ADC were found to be preferable in order to assure consistent audio input levels.

Different surfaces including metal pans, metal sheets, foam board, cardboard, corrugated plastic sheets, solid plastic sheets, and wooden boards were tested for frequency response. While we had initial concerns regarding high-frequency damping in plastic materials and uneven frequency responses overall, in practice we found that any sufficiently rigid and hard surface performed adequately. The surfaces used in the implementation pictured are a metal pan, plastic turntable lid, and plywood panels.

Audio analysis and sound synthesis takes place in Max/MSP. The Zsa.Descriptors library for Max/MSP was used to calculate Bark coefficients for one of the analyses [**?**]; the remainder of the analyses and sound syntheses were programmed by the authors.

3.1 Performance Techniques

Two compositions for the SpectraSurface have been written, focusing on two different performance techniques. The first composition utilizes different sized metal bowls which are placed on the four surfaces. For the beginning of the composition marbles are thrown into the bowls in such a way that they spin around the bowls' diameter. The spinning of the marbles generates a self-sustaining process which drives the sound synthesis. Later in the composition quantities of rice are placed in the bowl, either gradually or suddenly,

which the performer then manipulates with his hands.

In the second composition metal pots and cymbals are placed on the surfaces. A light pair of wooden sticks are used to strike these objects as well as the surfaces themselves.

4. SOFTWARE IMPLEMENTATION

Two analysis algorithms were implemented, each with their own approach both to analysis and also to mapping to sound synthesis. These algorithms are described below while details regarding sound synthesis will be provided in future publications.

4.1 Fuzzy

Fuzzy was designed to have a fast transient response and analyzes only four wide frequency bands. The center frequency of each band is set to 200, 500, 1000, and 2000 Hz. The first derivative of the amplitude of each frequency band is taken and run through an averaging filter and then an envelope follower with a fast attack and medium decay. The four output signals are then made available to mapping to synthesis parameters. An additional set of signals are created by generating four bands of spectral tilt through comparing the amplitude of neighbouring bands as well as the amplitudes of the first and fourth band.

These analysis signals are mapped to synthesis using a dual-stage preset system. The first stage consists of making direct mappings between synthesis parameters and control signals. The second stage consists of interpolating between multiple mapping presets in a two-dimensional space. The standard Max/MSP object *Nodes* was used to implement this second stage. This system allows for dramatic changes in timbre as the frequency content of the input signal changes, as well as allowing for the positioning of mapping presets within a specified timbral range.

4.2 Cat's Eye

For Cat's Eye, the incoming audio was processed with an initial high-shelf filter giving a wide 16db boost at 500hz. A hipass filter at 25hz was also applied. The audio signal was then analyzed using the zsa.energy and zsa.bark externals in order to get a list of 25 Bark coefficients. The coefficients are then analyzed to find the local maximum; the other coefficients are divided by this maximum to generate a list of relative amplitudes. The square root of each value of this list is taken in order to compress the overall amplitude range. The value of each item in the list is then fed through an envelope follower whose attack and decay times are exposed for mapping. The resulting list of values contains the relative amplitude of 25 consecutive frequency bands.

5. OBSERVATIONS

In this section we will share our observations regarding the use of audio feature extraction for control of sound synthesis as well as integration with direct sensing of performer gesture.

5.1 Characteristics of Audio Feature Extraction

The audio feature extraction algorithms described above share certain characteristics with implications for the control of sound synthesis.

5.1.1 Spectral Tilt

The two analysis methods described above produce values which represent spectral tilt, or relative amplitudes between spectral bands. Spectral tilt is theoretically amplitude independent – however, the fact that frequency spectrum of an instrumental sound is highly amplitude and time dependent causes spectral tilt to generally follow amplitude changes. For an instrument which is capable of generating sustained timbres spectral tilt can be a relatively constant value. One of the challenges of analyzing percussion instruments, however, is that their amplitude and therefore spectra changes quickly over time. The spectra produced is also complex, inharmonic, and highly variable – especially during transients. As the values from any two spectral bands change, their relative amplitudes can change to an even greater degree. This makes it difficult to derive discrete, static, or slow moving control signals – all of the values for spectral tilt tend to look like amplitude envelopes. A lowpass filter on each filter bands' amplitude can help to mitigate sudden jumps but will have an impact on responsivity.

It is necessary therefore to find a way to react quickly to these changes while minimizing the jitter resulting from these natural fluctuations in spectral content. Care can be taken during control signal processing and mapping to make a clear difference between the results of actual amplitude envelope following and the spectral information, especially since it is natural to map the amplitude envelope from the control signal to the amplitude of the synthesized audio. Another approach is to use envelope followers with different attack and decay speeds for amplitude envelope versus spectral tilt values.

5.1.2 Frequency Band Selection

The selection of frequency bands plays a large part on the responsivity of an instrument based on multiband audio feature extraction. Two approaches were described above, the first of which uses four unequally spaced frequency bands which were determined heuristically while the second used 25 frequency bands which were based on Bark coefficients. Since Bark coefficients are based on human perceptual characteristics utilizing them for audio feature extraction has the benefit of making an intuitive correlation between the perception of a sound and the control signals generated by the sound. However, there is no reason why the frequency content of a sound will map in an optimal way to the frequency bands generated by Bark coefficients.

In our first approach we found that as few as four frequency bands provide a reasonable amount of variation between the amplitude of consecutive frequency bands. To facilitate frequency band selection for different sound sources it may help to implement a learning stage in which an adaptive algorithm is used to determine frequency bands with relatively equal amplitude responses.

5.1.3 Secondary Sound Interference

On the hardware side, our use of a surface equipped with contact mics to pick up the audio from objects placed on top of the surface has several shortcomings. The primary one is that the surface acts to attenuate the audio coming from the object as it moves towards the microphone, requiring significant preamplification which increases the noise floor. This also causes sympathetic vibrations induced in the surface to be relatively prominent in relation to the sound from the objects. In addition, the audio resulting from the coupling of objects to the surface they are placed upon is often very different from the sound perceived by the performer through the air. This can cause the performer's perception of the control signal to be different from the actual control signal.

One constant of audio feature extraction is the inability to distinguish intentional and unintentional sound input.

The use of direct sensing to supplement audio feature extraction may help in, for example, wind instruments where the audio input to the algorithm is muted when the mouth is taken away from the mouthpiece. However, it may be that sympathetic resonances and secondary sounds can be seen as either tolerable or even desirable. As on an acoustic instrument expert performance and programming may mitigate the problem of background sound to a sufficient extent. Certainly this also opens the possibility of performance techniques which may not have been foreseen by the instrument designer.

5.2 Integration with Direct Sensing

Finally, as noted above, the supplementation of audio feature extraction with direct sensing of performer gesture may create a best-of-both-worlds situation, in which the intuitive control of timbre provided by audio feature extraction is augmented with the ability for control of discrete events provided by direct sensing. One problem with this approach is the relative difficulty of implementing audio feature extraction on the microcontroller platforms which are frequently used in direct sensing. One characteristic of the algorithms described above which may make implementation easier is the fact that the audio and analysis resolution may be relatively low-quality. For example, if a large part of the spectral content of a sound is below 1k there is no need for typical audio sampling frequencies. In addition, relatively small FFT sizes may be used as accurate reproduction of the source signal is not required.

6. FUTURE WORK

The work described above represents the first implementation of *Unsounding Objects*. Implementations utilizing breath controllers as well as vibrating strings for audio input have been created in order to explore the generalizability of the algorithms described above.

A constant concern has been CPU load of the analysis algorithms, which can be significant even on modern PCs. As described above we plan on implementing audio feature extraction using low-fidelity ADC conversion. We also plan on creating implementations directly on microcontrollers, leading to possibilities for the use of large arrays of audio signals as well as the use of direct gesture acquisition combined with audio feature extraction within the same instrument.

7. CONCLUSIONS

The research presented in this paper demonstrates an approach utilizing multi-band audio feature extraction for the control of sound synthesis in a digital percussion instrument. While the control signals produced by audio feature extraction are more difficult to work with than those produced by direct sensing of performance gesture, audio feature extraction is more tolerant of extended performance techniques and may not require the performer to learn tightly defined performance gestures. These qualities make it an effective alternative and complement to direct sensing for the control of sound synthesis.

8. ACKNOWLEDGMENTS

This research was supported by a Student Research Grant from the Center for Interdisciplinary Research in Music Media and Technology, and also by the CIRMMT Director's Interdisciplinary Excellence Prize.

9. REFERENCES

[1] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guedy, and N. Rasamimanana. Continuous Realtime Gesture Following and Recognition. In S. Kopp and I. Wachsmuth, editors, *Gesture in Embodied Communication and Human-Computer Interaction*, pages 73–84. Springer-Verlag, Heidelberg, 2010.

[2] T. Jehan. *Perceptual Synthesis Engine : An Audio-Driven Timbre Generator Perceptual Synthesis Engine : An Audio-Driven Timbre Generator*. PhD thesis, Massachusetts Institute of Technology, 2001.

[3] J.-m. Jot and O. Warusfel. Spat ˜ : A Spatial Processor for Musicians and Sound Engineers. In *Proceedings of the International Conference on Acoustic and Musical Research*, 1995.

[4] M. Malt and E. Jourdan. Zsa.Descriptors : a library for real-time descriptors analysis. In *Proceedings of the Sound and Music Computing Conference*, 2009.

[5] E. R. Miranda and M. M. Wanderley. *New Digital Musical Instruments: Control And Interaction Beyond the Keyboard*. A-R Editions, Middleton, WI, 2006.

[6] S. O'Modhrain and G. Essl. PebbleBox and CrumbleBag : Tactile Interfaces for Granular Synthesis. In *Proceedings of the 2004 conference on New interfaces for musical expression*, 2004.

[7] M. Puckette. Low-dimensional parameter mapping using spectral envelopes. In *Proceedings of the International Computer Music Conference*, 2004.

[8] S. Rossignol, X. Rodet, J. Soumagne, J. Collette, and P. Depalle. Automatic characterisation of musical signals: feature extraction and temporal segmentation. *Journal of New Music Research*, 28(4):281–295, 1999.

[9] Z. Settel and C. Lippe. Real-Time Musical Applications using FFT-based Resynthesis. In *Proceedings of International Computer Music Conference*, 1994.

[10] S. Shiraishi. *A Real-Time Timbre Tracking Model Based on Similarity*. PhD thesis, Royal Conservatory, The Hauge, 2006.

[11] A. R. Tindale, A. Kapur, W. A. Schloss, G. Tzanetakis, A. Kapur, W. A. Schloss, and G. Tzatenakis. Indirect Acquisition of Percussion Gestures Using Timbre Recognition. pages 10–12, 2005.

[12] C. Traube, P. Depalle, and M. Wanderley. Indirect Acquisition of Instrumental Gesture Based on Signal, Physical and Perceptual Information. In *Proceedings of Conference on New Interfaces for Musical Expression*, pages 42–47, 2003.

[13] D. L. Wessel. Timbre Space as a Musical Control Structure. *Computer Music Journal*, 3(2):45–52, 1979.