INSTRUMENT TIMBRE MODELS WITH NOISY PARTIAL INCLUSION

Conor O'Sullivan, Mikael Fernström

Interaction Design Centre
University Of Limerick
Ireland
conor.osullivan@ul.ie , mikael.fernstrom@ul.ie

ABSTRACT

This paper presents the results of work performed in the area of analysis, manipulation and re-synthesis of musical instrument sounds. The goal is an efficient method of musical instrument sound modelling.

Building on work previously carried out on analysis/re-synthesis methods, in addition to phase vocoding methods, the work here presented proffers an alternative method of harmonic partial analysis and filtration.

A new technique for sound model generation is presented and results from preliminary testing are discussed.

1. INTRODUCTION

Sound synthesis plays an important role in many areas of digital audio, multimedia and sonic-based interactive devices. Techniques such as Fourier-based additive synthesis, spectral modeling techniques, stochastic modeling, frequency modulation synthesis and analysis/re-synthesis timbre generation have all shown to be effective methods of modeling musical instrument sounds. Current devices with sonic generators such as PDA's, PC sound cards and mobile phones tend to use only basic synthesis algorithms such as those employing sampling, FM or square wave techniques.

The scope and potential for development of these devices through both high quality and perceptually-enhanced audio synthesis certainly exists [1][2][4]. Currently, the processing power of devices is one of the main restrictions on the quality of audio produced. This paper explores the possibility of reducing the steps involved in the synthesis of high quality digital musical instrument models, by performing an analysis on spectral content of those sounds.

The goal is to generate models using sufficient information to efficiently and accurately synthesize the sound.

2. OVERVIEW

The theory and relevance of analysis/re-synthesis based timbre generators are covered, as well as the specific introduction to the vocoding synthesis. The psychoacoustic basis for such an implementation is examined both in theory and by means of numerical analyses. Listening tests affirming the results of this implementation are carried out and the implications of the findings are discussed. These have a direct influence on the type of algorithm needed to perform an effective synthesis.

The above theory was tested and examined by performing analysis and re-synthesis. The goal here is to experiment with different complexities of the sound models, to determine a minimal level of partial information necessary to re-create a satisfactory version of an instrument timbre.

3. ANALYSIS AND RE-SYNTHESIS

3.1. Introduction

In order to perform an effective synthesis of any sound it is first necessary to perform an analysis on that sound. An intelligent and comprehensive assessment of the makeup of a sound needs to be constructed. This allows for its spectral qualities to be scrutinized and examined and a purer estimate for its synthesis to be determined [3]. This approach also allows for the potential of a wider and varied class of sounds to be synthesized, by changing the qualities of existing ones (for example phase or frequency envelopes). So in addition to the straightforward playback of an instrument sound, we are now open to a new range of effects on the original sound and a new class of hybrid timbres.

To merely playback a sample of a sound permits an extremely limiting number of options, such as duration or pitch, when it comes to synthesis and musical performance. Creating a model of that sound, however, sanctions a new level of audio synthesis [4], where the control of intonation and further synthesis opportunities are bounded only by creativity. Some examples of analysis/synthesis techniques are the phase vocoder, additive techniques and formant analysis/synthesis. The method employed in this study has been the phase vocoder, specifically using the short-time Fourier transform.

3.2. The Phase Vocoder

The phase vocoder [5] performs an analysis on a segment of digital audio and compiles spectral information about the sound. The analysis data is returned as a series of time frames, each separated into a number of bins. Each bin contains amplitude and frequency breakpoint values, which normally change on a frame-by-frame basis.

Such data can be taken by a phase vocoder module for plain re-synthesis, or can be altered in the meantime. This allows for effects such as timescale modification, whilst the sound still retains its short-time spectral characteristics. It is possible then to determine the movement of a sound's individual frequency

partial components over time and extract the envelope information that can be used in re-synthesis.

The equation for this type of synthesis of a sound sample $s_0(t)$ is commonly given by equations 1 and 2, below.

$$s_0(t) = \sum_{p=1}^{n} a_p(t)\cos(f_p(t)) \quad (1)$$

$$f_p(t) = f_p(0) + 2p\int_0^t f_p(u)du \quad (2)$$

4. SOUND MODEL GENERATION

4.1. Process

In order to perform an effective timbre generation, a suitable method for analysis, manipulation and re-synthesis was determined. Various methods were attempted, until the final process for generation of the models became:

- An instrument sound sample is analysed using a phase vocoder analysis.
- The output file produced by this analysis is imported into numerical analysis software.
- The content of the file is viewed numerically and graphically. This allows a determination to be made of the prominent spectral features of the sound, including relevant partial and harmonic content.
- A mathematical analysis is performed on the spectral data and a new binary file is written. This allows for the new sound model to be generated using prominent and/or noisy partials. The newly generated binary file can be read by the phase vocoding synthesis engine.
- The new sound can then be played to evaluate accuracy and quality.

4.2. Partial Analysis and Manipulation

A numerical analysis and manipulation was performed on the frequency/amplitude breakpoints. A cut-off point is chosen so that, upon analysis, partials with a given number of frequency points with amplitude greater than the cut-off would be included in the new sound model, where a specific number of partials would only be present. The process for this type of partial inclusion in a sound model, here referred to as the Greatest Amplitude model, can be described mathematically as follows. The new model, $s_1(t)$ is formed when the partials chosen in $s_0(t)$ satisfy Condition 1, below.

4.2.1. Condition 1

Partial *p* has at least *x* pairs $(a_p(t), f_p(t))$ such that $a_p(t) > C$, where *x* and *C* are pre-defined or derived integers, for all *t* in $s_0(t)$ (3)

An experimental algorithm for the inclusion of envelope partials with 'noisy' frequency components was also addressed in this study.

The inclusion of noisy information in a musical instrument sound is a key factor in determining the recognisability of that sound [6]. This algorithm analyses the partial envelope and performs a differentiation on the data so that the most active envelopes can be ascertained. It is foreseen that these envelopes will contain data that can be interpreted as noise. This area of the study is not intended as a replacement for existing work in area of sinusoid/noise modeling, but rather a lower-cost alternative and an exploration into other methods of partial selection.

Using this process, a cut-off point is chosen so that, upon analysis, partials with a given number of frequency points with amplitude greater than the cut-off would be included in the new sound model. In addition to this, partials with the greatest number of upward sloping curvatures, found by taking the second derivative of the partial envelope, are included. The total number of envelopes shall not exceed a specific (given) number. The intersection of these two groups is included only once. The process for this type of partial inclusion in a sound model, here referred to as the Greatest Amplitude Plus noise model, can be described mathematically as follows. A new model $s_2(t)$ is formed when the partials chosen in $s_0(t)$ satisfy either condition 1 above or condition 2, below.

4.2.2. Condition 2:

Partial *p* has at least *y* pairs $(a_p(t), f_p(t))$ such that $f''(a_p(t)) > 0$, for all t in $s_0(t)$, where *y* is a pre-defined or derived integer. (4)

4.3. Sound Model Representation

Figures 1 and 2 below offer two representations of a selection of bins. There are 128 bins in total, over 206 time frames. Figure 1 shows an amplitude-frame representation, and figure 2 gives an amplitude-frequency display (Figure 2 can be thought of as the end-view of the elevation in Figure 1). The distinct lines track the movement of a bin and give us information about the significant partials of the sound. Notice the amount of noisy information along the x-axis, or frequency -axis of Figure 2.

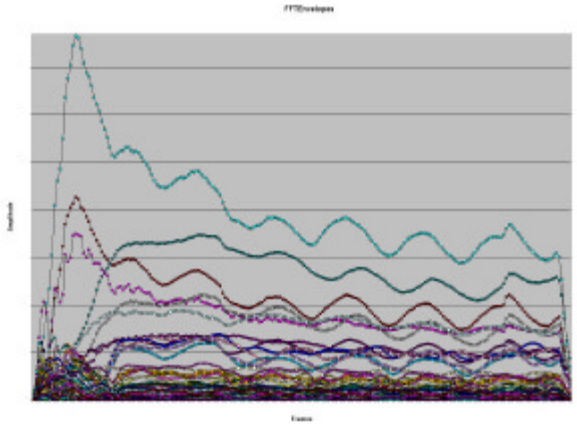

Figure 1 Frequency partial amplitudes over time

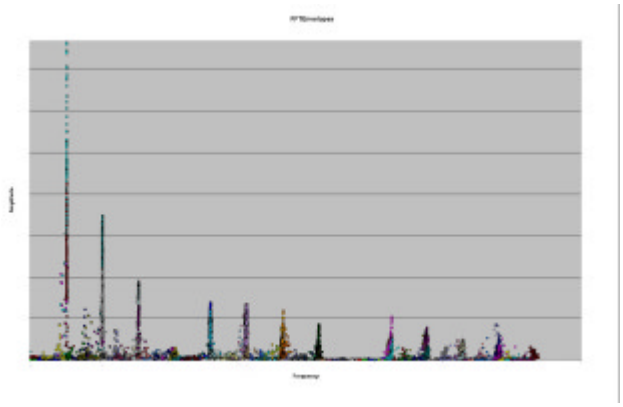

Figure 2 The amplitude-frequency spectrum

This graphical representation allows the prominent frequency partial information to be determined. This information is important because it shows the constitution of the sinusoidal components that extend to make the sound when re-synthesised. Therefore it is possible to cut out some of the superfluous noise, which is represented by random sinusoids; and to re-generate a phase vocoder file that can be used by the synthesis software.

An amplitude cut-off point is chosen here, such that any bin with at least 3 partial points with amplitude greater than the cut-off is included in the new filtered file. All other bins are discarded and their amplitude points set to zero. This ensures that anomalous elements of sinusoidal components that appeared perhaps because of unwanted noise are removed. The resulting timbre models, here illustrated by Figures 3 and 4 are a reduced version of the original sounds.

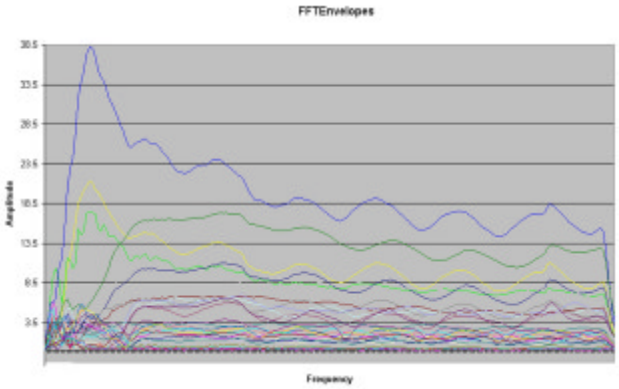

Figure 3 Filtered model over time

The reduction is quite significant, as a timbre file can be reduced from, say, 128 partials to 30 or even less. This is evident visually, above, and also numerically in terms of the generation algorithm. It then remains to be determined how a significantly reduced file fares in terms of quality for the listener. Listening tests were performed to this end

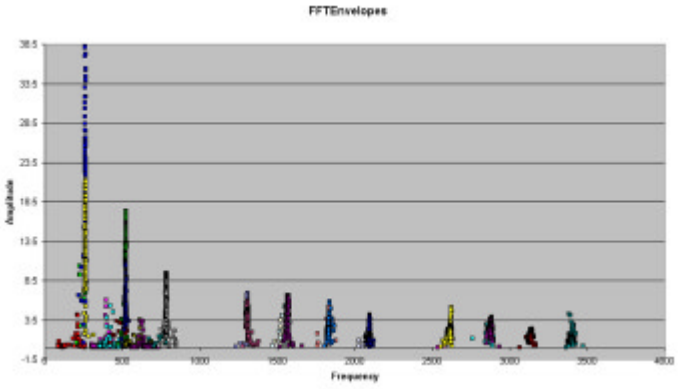

Figure 4 Filtered model's frequency spectrum

A visual display of the filtration process provided by the Greatest Amplitude Plus noise model is also given.

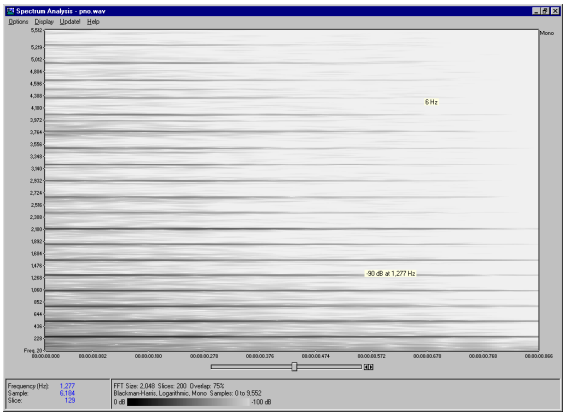

Figure 5 Original spectrum

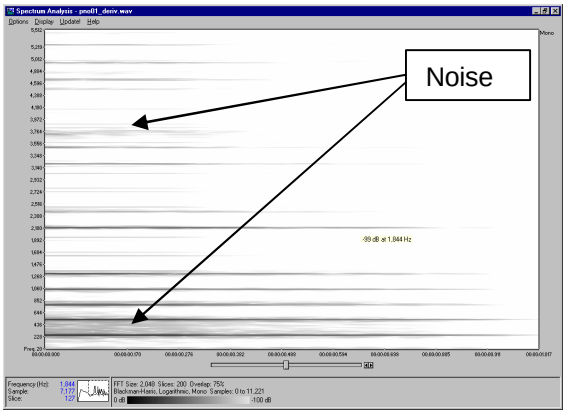

Figure 6 Greatest Amplitude Plus noise model

The sonogram representations given in Figures 5 and 6 also display the consequences of the algorithms. Figure 5 is a frequency domain representation of a piano sample used in the sound model generations. After analysis, filtration and re-synthesis was performed, Figure 6 shows the spectral content of the resulting sound. This particular sound model was generated using 12 of the most amplitude-significant partials, plus 12 of the noisiest partials. All other irrelevant sinusoidal information was dumped, as described above. Note how, even visually, the significant partial information has been included. The algorithm has also captured some of the noise (light grey) information, appearing here in the upper and lower registers of the frequency spectrum.

5. MODEL TESTING

5.1. Method

Listening tests were performed with eight subjects, out of a possible response pool of ten. The eleven musical instrument sounds were presented in a random order, adjusted for smooth dispersion so that no two 'similar' instrument sounds were adjacent. The object of the test was for the listener to perform an identification of timbre and an assessment of quality on each instrument sound. The listener was asked to

- Play the sound.
- Choose which instrument the sound was most similar to from a list.
- Rate the quality of that sound and perceived closeness to the instrument.
- Give any additional comments, optionally.

The list of instruments from which selection was made included the intended instrument, as well as a broad range of others, spanning brass, percussion, string, wind and electronic categories. An 'other' option was also included to allow the listener associate their own choice of instrument. The scale of judgement was between 1 and 10, with 10 being the best possible sound model, or as close as perceivable to the listener's knowledge of the acoustic instrument. The listener was asked to assign a rating, giving any additional comments as thought necessary. The order in which the sound models were judged was not absolute, and there was no time limit specified for the tests. The tests were run using headphones or speakers, as the listener desired.

5.2. Sound Models

The timbre models that were used consisted of various spectral forms of three different acoustic instrument sounds, specifically, piano, violin and trumpet. The sounds were analysed and re-synthesised, according to methods outlined above, so that a range of partial combinations would be used to generate their timbre. It was decided that, for each instrument, a timbre model would be generated containing 12, 20 and 30 of the most significant partials. In the piano and violin sounds, an extra timbre model was generated to incorporate an algorithm to generate a timbre using additive techniques that includes 'noisy' partials.

5.3. Results

The results returned by the testing were consistent with expectation on the whole. Generally speaking, the greater the number of partials included in the spectral information of the timbre model, the more the sound became recognisable and associated with its intended instrument. The notable exception to this is the Piano sound with 24 partials, found using the Greatest Amplitude Plus noise model, which scored either better than or just as well as the piano model with 30 partials, found using the Greatest Amplitude model. The answers to the second part of the test, which is based on instrument recognition, were fairly consistent among subjects.

5.4. Partial Information Inclusion

Overall, the results of the listening tests indicate a necessity of inclusion of partial information on the order of 20 to 30. Clearly, as the results support, the more partials that are included, the more recognisable the sound becomes, with the quality also judged as having improved. The instrument becomes fairly recognisable at 20 partials and almost universally recognisable at 30, so the level of partial inclusion would therefore depend on any computational restrictions and be bound by the level of audio quality desired. One result that also shines through from the listening tests is that the inclusion of only 12 partials in a timbre model can be ruled if a fairly high level of audio quality is being aimed for. This is endorsed by the pattern that emerged from the tests in instrument association and accuracy ratings. Those sounds were generally considered 'poor', given lower ratings, and indeed were not correctly assigned to the intended targets.

As the number of partials increase in the sound, timbral information encoded in the noisier, residual part of the sound also enters the spectrum. It is a result of the stochastic nature of the physical instruments and irregularities like breath and

bowing sounds. It is true to say, therefore, that if this information could be included at synthesis stage, the overall quality of the representation would increase, but also that the number of partials required would decrease.

6. CONCLUSIONS

The results found to date are certainly promising. The approach to synthesis has been one of investigation with optimal algorithm design central. A short-time Fourier transform-based phase vocoder method is here favoured and its advantages shown.

A visual and mathematical analysis of original instrument samples was undertaken and a re-synthesis method was defined. The synthesis method was tested with a group of subjects and favourable results were recorded after partial reduction and filtration.

7. REFERENCES

[1] W. Gaver, "Auditory interfaces," In Helander, M. G., Landauer, T. K., & Prabhu, P. V. (Eds.), *Handbook of human-computer interaction*, Amsterdam, The Netherlands: North-Holland, 1997.

[2] C. O'Sullivan, M. Fernström, "A Formulation of Holophonic Display", in *Proc. ICAD*, Kyoto, Japan, 2002.

[3] K. Jensen, "Timbre Models of Musical Sounds", PhD thesis, University Of Copenhagen, 1999

[4] D. Rocchesso, R. Bresin and M. Fernström, "Sounding Objects", paper submitted to *IEEE Multimedia* (http://www.soundobject.org/articles.html) 2002.

[5] M. Dolson, "The Phase Vocoder: A Tutorial". *Computer Music Journal* 10(4):14-27, 1986.

[6] X. Serra, "Musical Sound Modeling With Sinusoids Plus Noise". Audiovisual Institute, Pompeu Fabra University (http://www.iua.upf.es/~xserra/articles/msm/) September 2002.