A Parametric Model for Spectral Sound Synthesis of Musical Sounds

Cornelia Kreutzer
University of Limerick
ECE Department
Limerick, Ireland
cornelia.kreutzer@ul.ie

Jacqueline Walker
University of Limerick
ECE Department
Limerick, Ireland
jacqueline.walker@ul.ie

Michael O'Neill
University College Dublin
School of Computer
Science and Informatics
m.oneill@ucd.ie

Abstract

We introduce a reduced parameter synthesis model for the spectral synthesis of musical sounds, which preserves the timbre and the naturalness of the musical sound. It also provides large flexibility for the user and reduces the number of synthesis parameters compared to traditional analysis/re-synthesis methods. The proposed model is almost completely independent from a previous spectral analysis. We present a frequency estimation method using a random walk to keep the naturalness of the sound without using a separate noise model. Three different approaches have been tested to estimate the amplitude values for the synthesis, namely, local optimization, the use of a lowpass filter and polynomial fitting. All of these approaches give good results, especially for the sustain part of the signal.

1. Introduction

Unlike physical modeling, where a set of algorithms and equations is used to simulate the different parts of the sound source [8], spectral sound synthesis uses the spectral representation of the sound itself.

This spectral approach has been applied to model speech signals based on their sinusoidal representation [10], before it was adapted to musical sounds for the Spectral Modeling Synthesis (SMS) framework [14,17]. The SMS framework provides separate models for the harmonic and the residual parts of the sound. This separation allows for a flexible transformation and synthesis framework. However, due to the characteristics of musical sounds, especially the complexity of the musical timbre [5,7], spectral modifications can lead to sound artifacts [6].

Several methods have been suggested to improve the sound analysis, such as more accurate partial tracking by using linear prediction [9] or Hidden Markov Models [2]. Concerning the synthesis model, some approaches have been proposed to reduce the number of synthesis parameters [3,16]. Unlike the standard SMS, where the synthesis is based on the frequency, amplitude and phase parameters of the sound, these methods focus more on high level attributes of the musical sound.

We propose a synthesis model, the Reduced Parameter Synthesis Model (RPSM), that is almost completely independent from a previous spectral analysis without using high level sound attributes. The method is based on a frequency and an amplitude model with a reduced number of synthesis parameters compared to the standard SMS. The model also allows the synthesis of musical sounds outside the range of a particular instrument by preserving the timbre of the instrument and the *naturalness* of the sound.

2. Spectral Modeling Synthesis

Spectral Modeling Synthesis (SMS) is a framework for spectral analysis, synthesis and transformations of musical sounds introduced by Serra [14]. The basic principle is to analyse the spectral content of a given sound sample in order to perform a spectral synthesis using the analysis results. Therefore, the SMS framework consists of a deterministic and a stochastic model.

The deterministic model is used for the sinusoidal parts of the sound. Once the sound spectrum is obtained by means of the STFT, the prominent spectral peaks are detected and tracked using a peak tracking algorithm. The objective of this algorithm is to detect magnitude, frequency and phase parameters of the sinusoidal partials. In case the sound is pseudo-harmonic a pitch detection method can be used to improve this process.

Subsequently the sinusoidal part of the sound is subtracted from the overall signal to obtain the sound residual. This residual part of the sound – sometimes also referred to as the noise part - is modeled using a stochastic model, e.g., using a time varying filter [15]. The deterministic and the stochastic model are independent from each other, which allows a flexible analysis and re-synthesis process. Results obtained throughout the analysis/synthesis process can also be used for other music related applications like sound source separation or sound transformations.

Since its introduction, the original framework has been further developed [15,1], and a number of extensions have been proposed, like additional models for transient parts of the sound [18] or feature based sound transformation methods [16].

3. Parametric Synthesis Model

3.1. Frequency estimation

In order to determine the frequency values within the synthesis model we use a flexible model that is not based on a preceding spectral analysis but, on the basic knowledge about the sound. The fundamental frequency - or pitch - as well as the number of harmonic partials are user defined values. This is particularly important if the synthesised sound lies outside the range of the instrument the model is supposed to mimic. Also, within the range of an instrument there is no restriction of the pitch value or the number of harmonics that can be chosen, since both values are entirely user defined. Consequently we can model whole tones, semitones or quarter tones of an instrument as well as other *notes* whose pitch values is anywhere in between or outside these tones.

Furthermore, we apply a random walk to several frequency partials in order to reconstruct the naturalness of the sound. Figure 1 shows a representative result of the SMS partial tracking algorithm; in this particular case the result for a flute note (A4, played forte, non Vibrato). As illustrated there, some of the partials, especially the upper ones, show a certain amount of variation or *noisiness*. Due to this noisiness a reconstruction of the sinusoidal parts of the sound does keep the sound characteristics of the original recording, although the residual part of the signal is neglected for the reconstruction. Because of this observation we incorporate this noisiness into the sinusoidal partials of our synthesis model rather than defining a separate noise model.

This is achieved by the use of a one-dimensional random walk [4] to determine the frequencies of the upper harmonics. A one-dimensional random walk can be described as a path starting from a certain point, and then taking successive steps on a one-dimensional grid. The step size is constant and the direction of each step is chosen randomly with all directions being equally likely.

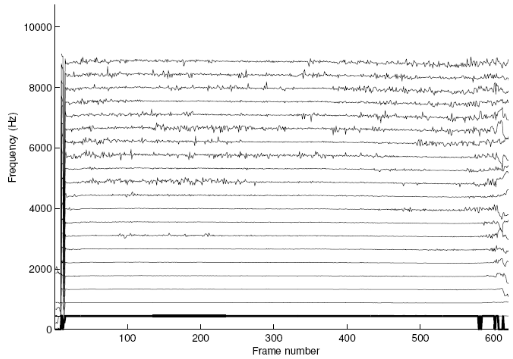

Figure 1. SMS Frequency analysis result (flute, A4, played forte, non vibrato)

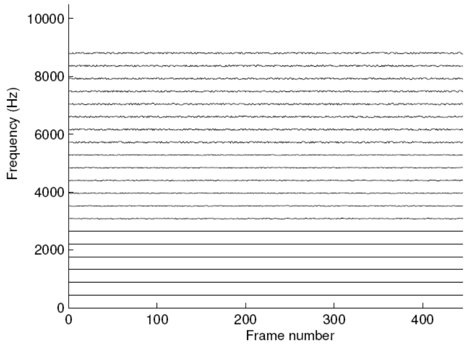

Figure 2. Estimated frequency tracks for 20 harmonics (f0: 440 Hz)

For the purpose of our synthesis model random walks are applied to certain harmonic partials in the following way. First, the harmonic partials are divided into three groups, where each group represents a third of the overall number of harmonics. This follows from the results of the SMS analysis, which shows different levels of variations for the lower, the middle and the upper harmonics. Concerning the lowest third of the harmonic partials - starting from the fundamental frequency - no random walk is applied as the analysis of these lower partials shows very little variation. For the middle and the upper harmonic partials a random walk is applied, where the starting point of the random walk is determined by the basic frequency of the harmonic partial. Basic frequency in that case means the integer multiple of the fundamental frequency. Again, from the analysis result it can be seen that the upper harmonics show more variation than the middle ones. Due to that, and after testing several levels of noisiness, the step size of the random walk is set to 30 Hz for the upper harmonics and to 15 Hz for the middle ones. Figure 2 shows the estimated frequency tracks for the synthesis model with the same conditions as the flute sound in Figure 1 (440 Hz fundamental frequency and 20 sinusoidal partials).

3.2. Amplitude estimation

In contrast to the frequency estimation, which is not directly taken from the sound analysis results, we use SMS analysis results as a basis for estimating the amplitude values of the harmonic partials. However, we reduce the number of parameters to provide a flexible synthesis model that is mostly independent from the preceding sound analysis process. Additionally, our main concern is to keep the quality and naturalness of the musical sound after the synthesis process in order to mimic real instruments. Therefore, different methods have been applied to the analysis amplitude data. We have carried out amplitude estimation by means of local optimization, lowpass filter estimation and polynomial fitting. A detailed discussion of all these methods will be provided in the following sections after a short description of the applied analysis procedure.

In order to obtain the basic amplitude parameters a standard SMS analysis has been carried out, as described in [1]. The STFT is performed using a sampling rate of 44.1 kHz and a Blackman-Harris window with a window size of 1024 points and a hop size of 256 points. Zero-padding is applied in the time domain - using a zero-padding factor of 2 - to increase the number of spectral samples per Hz and, improve the accuracy of the peak detection process. From the resulting frequency spectrum after the Fourier analysis, 100 spectral peaks per frame are detected and subsequently used to track the harmonic partials of the sound. The number of partials to be tracked was set to 20. This analysis has been applied to sound samples taken from the RWC database [12]. In particular to all notes over the range of a flute, a violin and a piano. Given the amplitude tracking results one representative note for each instrument has been chosen to provide the basis for the amplitude values of the presented synthesis model. Figure 3 shows an example for the obtained SMS analysis results for a flute note.

3.2.1. Local optimization. The SMS analysis provides one amplitude value for each harmonic partial and for each frame of a given sound signal.

We reduce that parameter size by determining the local maxima of each amplitude track. This reduces the number of parameters to about a third of the SMS analysis result. For example, for the flute note (A4, played forte, non Vibrato) the SMS analysis consists of 12680 amplitude values. This is reduced to 3015 values, which represent all the local maxima of the 20 harmonic partials. To determine the shape of each amplitude track, which is necessary for the synthesis process, we then perform a simple one-dimensional linear interpolation between the local maxima of the track. Figure 3 illustrates an estimation result. As can be seen the shape of the tracks are close to the SMS analysis result. However, this is not the case for the attack and the release part of the sound.

3.2.2. Lowpass filter estimation. The second curve fitting method we chose to estimate the overall amplitude envelope of each harmonic partial uses a lowpass filter. Therefore we apply a 3rd order Butterworth lowpass filter to the analysis data. We perform zero-phase digital filtering by processing the input data in both the forward and reverse directions. After filtering in the forward direction, the filtered sequence is reversed and runs back through the filter. The resulting sequence has precisely zero-phase distortion and double the filter order. Figure 4 shows an example for amplitude tracks of a flute note estimated with the lowpass filter.

As with the local optimization the shape of the estimated amplitude tracks is very close to the SMS analysis result. However, the filter method takes significantly longer to be performed and is also not able to provide a sufficient estimate for the synthesis of the attack and the release part of the sound signal.

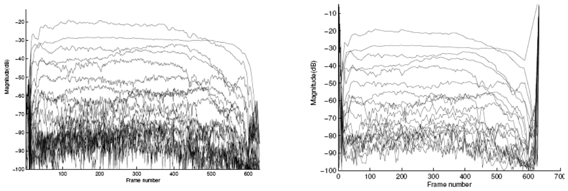

Figure 3. Amplitude tracks for a flute note, A4, forte, non vibrato (SMS analysis result (left) and estimated tracks using local maxima estimation (right))

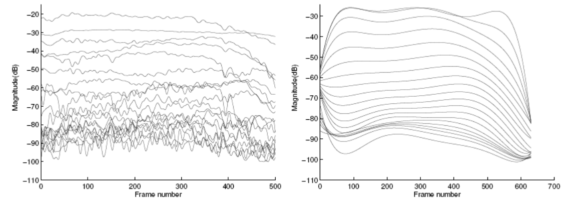

Figure 4. Estimated amplitude tracks for a flute note, A4, forte, non vibrato (using an LP filter (left) and using a polynomial fit (right))

3.2.3. Polynomial interpolation. Additionally we performed polynomial fitting to obtain an estimate for the several amplitude tracks. For each amplitude envelope the coefficients of a polynomial of degree 6 are computed to fit the data - in our case the analysis result - in a least squares sense. This computation is performed using a Vandermonde matrix [11]

$$V = \begin{bmatrix} 1 & \alpha_1 & \alpha_1 & \dots & \alpha_1^{n-1} \\ 1 & \alpha_2 & \alpha_2 & \dots & \alpha_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_m & \alpha_m & \dots & \alpha_m^{n-1} \end{bmatrix} \qquad (1)$$

since solving the system of linear equations $Vu = y$ for u with V being an $n\ x\ n$ Vandermonde matrix is equivalent to finding the coefficients u_j of the polynomial

$$P(x) = \sum_{j=0}^{n-1} u_j x_j \qquad (2)$$

of degree <= *n-1* with the values y_i at α_i [11]. An example for the estimation results is shown in Figure 4. In contrast to the two other methods being used, the results are very smooth amplitude envelopes missing all the small variation that can be seen in the SMS analysis result. Nevertheless, the synthesized sounds preserve the timbre of the particular instrument and the sound quality of the original recordings. Regarding the flute and the violin the polynomial estimation also gives a sufficient estimate for the attack and the release part of the sound signal.

3.3. Spectral synthesis

For the synthesis we use an implementation of additive synthesis based on the inverse FFT [1,13]. Compared to the traditional use of oscillator banks for additive synthesis, this is a more efficient and faster approach. The method takes advantage of the fact that a sinusoid in the frequency domain is a sinc-type function, using the transform of the window, and not all samples in these functions have the same weight [1]. So we only need to calculate the main lobe samples of the window transform with the specific amplitude, frequency and phase values to generate a sinusoid in the frequency domain. All the main lobes of the sinusoids we want to compute are then placed into an FFT buffer and by performing an inverse FFT we obtain the synthesized time-domain signal. Applying an overlap-add method then gives the time varying characteristics of the sound.

3.4. Empirical evaluation

The presented model has been tested for notes covering the whole range of a flute (37 notes), a violin (64 notes) and a piano (88 notes). An SMS analysis has been carried out for all these notes using recorded samples from the RWC database [12]. The analysis was performed to find a representative note for the presented amplitude model and to compare the RPSM synthesis results with the standard SMS results. As mentioned before, the amplitude analysis results of only one note per instrument have been used as a basis for the synthesis model. This way the amplitude shape stays the same for all notes of an instrument and only the frequencies are changed to obtain the presented synthesis results. However, the model also allows to modify the amplitude data if desired. For example, different amplitude *templates* can be used for different parts of the range of an instrument.

The frequency estimation works well and allows a large flexibility when choosing the fundamental frequency. Due to the random walk that is applied to higher frequency partials the synthesised sound keeps the natural noisiness of the real instrument recording without the need for a separate noise model. From the three different amplitude estimation methods the polynomial fit gives overall the best results. The estimation is fast and although the resulting envelopes are very smooth, the synthesised sound is of high quality. This is also the only method that gives a satisfactory estimate for the attack and the release parts of the signal. The local optimization is also fast and performs well for the sustain part of the signal, but does not give a satisfactory estimate of the shape of the attack and the release part. Applying a lowpass filter to

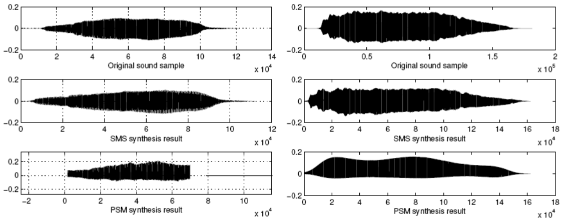

Figure 5. Violin: original signal, SMS result and sustain part of RPSM result with local maxima estimation (left), Flute: original sound, SMS result with polynomial fit (right) - time domain plots

estimate the amplitude tracks performs rather poorly compared to the other methods. The estimation results can be compared with the local optimization but the computation is significantly slower.

Figure 5 shows comparisons of SMS and RPSM synthesis results in the time domain for different instruments and different RPSM amplitude estimation techniques. When modeling the amplitudes with local optimization, the sustain part of the synthesised sound is very close to the original. Attack and release are not shown here, as the amplitude values are very high. Using a polynomial fit the resulting signal has a highly smoothed envelope, but gives good results for the attack and the release part.

4. Conclusion and future work

We introduced a flexible parametric synthesis model for the spectral synthesis of musical sounds. Unlike traditional spectral analysis/synthesis methods, the model is largely independent from a previous analysis of a recorded sound. The model has been tested for notes covering the whole range of three different instruments. The timbre and the perceptual quality of the sound is preserved even for notes at the upper end of the instrument range and for sounds that are outside the range of the instrument. This is not always the case for traditional analysis/re-synthesis approaches, mostly due to the quality of the recorded sound samples and the complex analysis procedure. The synthesis of sounds outside the instrument range by means of an analysis/re-synthesis method also requires additional transformations after the analysis, which can lead to artifacts in the synthesised sound too. Future work will be focused on defining a sufficient model for the attack and release part of the sound signals and on carrying out listening tests to gain more detailed results for a comparison between the recorded sounds, the SMS synthesis results and the RPSM results.

5. Acknowledgment

This work was supported by the Science Foundation Ireland (SFI) under the National Development Plan (NDP) and Strategy for Science Technology & Innovation (SSTI) 2006-2013.

10. References

[1] X. Amatriain, J. Bonada, A. Loscos, and X. Serra. *Spectral Processing* in *DAFx – Digital Audio Effects*, chapter 10, pages 373 – 439. edited by Udo Zoelzer. John Wiley & Sons, 2002.

[2] P. Depalle, G. Garcia, and X. Rodet. *Tracking of Partials for Additive Synthesis using Hidden Markov Models.* pages 225 – 228. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1993. Minneapolis, USA.

[3] M. Desainte Catherine, and S. Marchand. *Structured Additive Synthesis: Towards a Model of Sound Timbre and Electroacoustic Music Forms.* pages 260 – 263. International Computer Music Conference (ICMC), 1999. Beijing, China.

[4] W. Fellner. *Introduction to Probability Theory and its Applications.* Wiley series in probability and mathematical statistics. John Wiley & Sons, 3rd edition, 1968.

[5] J. M. Grey. *Multidimensional Perceptual Scaling of Musical Timbre.* Journal of the Acoustical Society of America, 61(5):1270 – 1277, 1977.

[6] J. M. Grey and J. W. Gordon. *Perceptual Effects of Spectral Modifications on Musical Timbres.* Journal of the Acoustical Society of America, 63(5):1493 – 1500, 1978.

[7] K. Jensen. *Timbre Models of Musical Sounds.* PhD thesis, University of Copenhagen, Copenhagen, Denmark, 1999.

[8] K. Karplus and A. Strong. *Digital Synthesis of Plucked String and Drum Timbres.* Computer Music Journal, 7(2):43 – 55, 1983.

[9] M. Lagrange, S. Marchand, M. Raspaund, and J. B. Rault. *Enhanced Partial Tracking using Linear Prediction.* pages 1 - 6. 6th International Conference on Digital Audio Effects (DAFx-03), 2003. London, UK.

[10] R. McAuley and T. Quatieri. *Speech Analysis/Synthesis Based on a Sinusoidal Representation.* 34:744 - 754, 1986. IEEE Transactions on Acoustics, Speech and Signal Processing.

[11] C. Meyer. *Matrix Analysis and Applied Linear Algebra,* chapter 4. SIAM, Philadelphia, PA, 2000.

[12] R. Music Database. *Musical Instrument Sound.* RWC-MDB-I-2001 No. 01 - 50. Tokyo, Japan, 2001.

[13] X. Rodet and P. Depalle. *Spectral Envelopes and Inverse FFT Synthesis.* 93rd AES Convention, San Francisco, AES Preprint No. 3393(H-3), 1992.

[14] X. Serra. *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition,* PhD thesis, Stanford University 1989.

[15] X. Serra. *Musical Sound Modeling with Sinusoids plus Noise,* chapter 3, pages 91 – 123. Musical Signal Processing. Swets & Zeitlinger, Lisse, The Neatherlands, 1997.

[16] X. Serra and J. Bonada. *Sound Transformations Based on the SMS High Level Attributes.* Digital Audio Effects (DAFx) Workshop, 1998. Barcelona, Spain.

[17] X. Serra and J. Smith. *Spectral Modeling Synthesis: A Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition.* Computer Music Journal, 14(4):12 – 24, 1990.

[18] T. S. Verma and T. H. Y. Meng. *Extending Spectral Modeling Synthesis with Transient Modeling Synthesis.* Computer Music Journal, 24(2):47 – 59, 2000.