Sound Transformations Based on the SMS High Level Attributes

Xavier Serra, Jordi Bonada
Audiovisual Institute, Pompeu Fabra University
Rambla 31, 08002 Barcelona, Spain
{xserra, jboni}@iua.upf.es http://www.iua.upf.es

Abstract

The basic Spectral Modeling Synthesis (SMS) technique models sounds as the sum of sinusoids plus a residual. Though this analysis/synthesis system has proved to be successful in transforming sounds, more powerful and intuitive musical transformations can be achieved by moving into the SMS high-level attribute plane. In this paper we describe how to extract high level sound attributes from the basic representation, modify them, and add them back before the synthesis stage. In this process new problems come up for which we propose some initial solutions.

1 Introduction

The goal of the research behind SMS has always been to get sound representations based on analysis that are musically intuitive and from which sound transformations can be obtained without generating any artifacts in the synthesized sound. Through this work it has become clear that it is impossible to obtain a single representation suitable for every sound and application. It has required to extend the original deterministic plus stochastic model [1] in order to include specific derivations for particular situations, while maintaining the overall analysis/synthesis framework as general as possible [2].

The most flexible representation can be achieved from the non real time analysis of well recorded sounds in a dry environment. These sounds should be pseudo-harmonic and monophonic, that is, melodies or single notes, played with a musical instrument. As these conditions are relaxed, one by one, other representations are possible that are not as flexible but that are powerful enough for many music and audio applications. In this article we concentrate on the most flexible representation possible, thus giving up some of the generality of SMS. However, it still shares the same analysis/synthesis framework with all other representations based on SMS, under which, representations can easily be found that are optimal for every sound and application.

We will start by describing the concept of high level attributes. We discuss the fact that in all natural sounds most of these attributes are interrelated, thus making the analysis harder. We also have to maintain these interrelations, or at least pay attention to them, when transformations are applied. Then, we describe the extraction of these attributes from the basic sinusoidal plus residual representation and discuss the issue of transformations based on this new representation.

2 High level attributes

The accomplishment of a meaningful parameterization for sound transformation applications is a difficult task. We want a parameterization that offers an intuitive control over the sound transformation process, with which we can access most of perceptual attributes of a sound, or even a group of related sounds, such as all the sounds produced by a given instrument.

The basic SMS analysis results in a simple parameterization appropriate for describing the microstructure of a sound. These parameters are the instantaneous frequency, amplitude and phase of each partial and the instantaneous spectral characteristics of the residual signal. There are other useful instantaneous attributes that give a higher level abstraction of the sound characteristics. For example we can describe fundamental frequency, amplitude and spectral shape of sinusoidal component, amplitude and spectral shape of residual component, and total amplitude. These attributes are easily calculated at each analysis frame from the output of the basic SMS analysis. At the same level, there are other attributes useful for the characterization of other aspects of the sound microstructure, like the degree of harmonicity, noisiness, spectral tilt, and spectral centroid.

A part from the instantaneous, or frame, values, it is also useful to have parameters that characterize the time evolution of the sound. These time changes can be described by the derivatives of each one of the instantaneous attributes.

Another important step towards a musically useful parameterization is the segmentation of a sound into regions that are homogeneous in terms of its sound attributes. Then we can identify and extract region attributes that will give higher level control over the sound. The simplest, and most general, segmentation process divides a melody into notes and silences and then each note into an attack, a steady state and a release regions. Global attributes that can characterize attacks and releases refer to the average variation of each of the instantaneous attributes, such as average fundamental frequency variation, average amplitude variation, or average spectral shape change. In the steady state regions it is meaningful to extract the average of each of the instantaneous attributes and measure other global attributes such as time-varying rate and depth of vibrato.

The concept of sound attributes can be taken a step further by considering the common attribute values of an entire instrument, i.e., for all the sounds produced by the instrument. The attributes of each note are compared and combined in order to group the characteristics that are common to the whole instrument, or some of its sounds, leaving each note only with the differences from the average values of the different attributes. This gives musical control at the instrument level without having to access each individual analyzed note.

3 Attribute correlations

In most cases the different attributes of a given sound are correlated and a change in one is accompanied by changes in others. Some of these correlations are due to the acoustic behavior of the actual vibrating system producing the sound, others to the way that the system is excited by the player. All musical instruments exhibit this property and we have to pay attention as to the way we calculate, extract and put back these attributes, in order to preserve the character of a given instrument and a specific performance of it. For musical reasons we might want to break these correlations in the synthesized sound, thus changing the natural behavior of the instrument and the perceptual relations between these attributes. But, even in this case, the results will be more interesting if we understand these correlations and define rules that generate "non-natural" correlations between the different perceptual attributes of a sound.

Some of the correlations are well known by musicians, and sound designers take them into account when they want to emulate the behavior of an acoustic instrument using a given synthesis technique. Examples of relevant correlations found in instrumental sounds are between fundamental frequency and spectral shape, between fundamental frequency and amplitude, or between amplitude and spectral shape. For example, in most instruments, as we go up the scale (higher fundamental frequency) the number of partials decreases, the spectral slope increases and, generally, the amplitude also increases. Also, as we play louder the resulting sound becomes brighter (flatter spectral slope).

In the context of SMS there are correlations that come up either in the analysis or the synthesis of a sound. Examples of these are between amplitude of sinusoids and amplitude of residual and between amplitude of partials, frequency of partials and spectral shape of sound. If we want to make transformations to a sound while maintaining its character we have to accompany the change of an attribute with the appropriate changes to the correlated attributes.

4 Attribute detection and extraction

From the basic sinusoidal plus residual representation it is quite simple to extract the attributes mentioned above. The critical issue is how to extract them in order to minimize interferences, thus obtaining, as much as possible, meaningful high level attributes free of correlations. We first extract instantaneous attributes and their derivatives, then we segment the sound, and finally we can extract region attributes.

4.1 Frame attributes

The basic frame, or instantaneous, attributes are: amplitude of sinusoidal and residual component, total amplitude, fundamental frequency, spectral shape of sinusoidal and residual component, harmonic distortion, noisiness, spectral centroid, and spectral tilt. These attributes are obtained at each frame using the information that results from the basic SMS analysis and not taking into account the data from previous or future frames. Some of them can be extracted from the frame data, leaving a normalized frame, others are information attributes that describe the characteristics of the frame and are not extracted from the original data.

The amplitude of the sinusoidal component is the sum of the amplitudes of all harmonics of the current frame expressed in dB,

$$AS_{total} = 20 \log_{10}\left(\sum_{i=1}^{I} a_i\right)$$

where a_i is the linear amplitude of the ith harmonic and I is the total number of harmonics found in the current frame.

The amplitude of the residual component is the sum of the absolute values of the residual of the current frame expressed in dB. This amplitude can also be computed by adding the frequency samples of the corresponding magnitude spectrum,

$$\begin{aligned} AR_{total} &= 20\ \log_{10}\left(\sum_{n=0}^{M-1}\left|x_R(n)\right|\right) \\ &= 20\ \log_{10}\left(\sum_{k=0}^{N-1}\left|X_R(k)\right|\right) \end{aligned}$$

where $x_R(n)$ is the residual sound, M is the size of the frame, $X_R(k)$ is the spectrum of the residual sound, and N is the size of the magnitude spectrum.

The total amplitude of the sound at the current frame is the sum of its absolute values expressed in dB. It can also be computed by summing the amplitudes of the sinusoidal and residual components,

$$\begin{aligned} A_{total} &= 20\ \log_{10}\left(\sum_{n=0}^{M-1}\left|x(n)\right|\right) = 20\ \log_{10}\left(\sum_{k=0}^{N-1}\left|X(k)\right|\right) \\ &= 20\ \log_{10}\left(\sum_{i=1}^{I}a_i + \sum_{k=0}^{N-1}\left|X_R(k)\right|\right) \end{aligned}$$

where $x(n)$ is the original sound and $X(k)$ is its spectrum.

The fundamental frequency is the frequency that best explains the harmonics of the current frame. This can be computed by taking the weighted average of all the normalized harmonic frequencies,

$$F_0 = \sum_{i=1}^{I}\frac{f_i}{i}\times\frac{a_i}{\sum_{i=1}^{I}a_i}$$

where f_i is the frequency of the ith harmonic. A more complete discussion on the issue of fundamental frequency in the context of SMS can be found in [3].

The spectral shape of the sinusoidal component is the envelope described by the amplitudes and frequencies of the harmonics, or its approximation,

$$Sshape = \{(f_1,a_1)(f_2,a_2)\ldots(f_I,a_I)\}$$

The spectral shape of the residual component is an approximation of the magnitude spectrum of the residual sound at the current frame. A simple function is computed as the line segment approximation of the spectrum,

$$Rshape = \{e_1, e_2, \ldots, e_q, \ldots, e_{N/M}\} = \max_k\left[\left|X_R(qM+k)\right|\right]$$

where $k = -M/2, -M/2+1, \ldots, M/2-1$, and M is the number of frequency samples used for each calculation of a local maximum. Other spectral approximation techniques can be considered depending on the type of residual and the application.

The harmonic distortion is a measure of the degree of deviation from perfect harmonic partials,

$$HarmDistortion = \sum_{i=1}^{I}\left|f_i - (F_0\times i)\right|\times\frac{a_i}{\sum_{i=1}^{I}a_i}$$

The noisiness is a measure of the amount of non sinusoidal information present in the frame. It is computed by taking the ratio of residual amplitude versus total amplitude,

$$Noisiness = \frac{\sum_{n=0}^{M-1}\left|x_R(n)\right|}{\sum_{n=0}^{M-1}\left|x(n)\right|}$$

Related noisiness measures result from studying the shape of each spectral peak and its deviation from the ideal sinusoidal peak.

The spectral centroid is the midpoint of the energy distribution of the magnitude spectrum of the current frame. One might also think of it as the "balance point" of the spectrum,

$$Centroid = \sum_{k=0}^{N-1}\frac{k}{N}f_s\times\frac{\left|X(k)\right|}{\sum_{k=0}^{N-1}\left|X(k)\right|}$$

The spectral tilt of the sinusoidal component is the slope of the linear regression of the data points (f_i, a_i),

$$Stilt = \frac{1}{\sum_{i=1}^{I}t_i^2}\sum_{i=0}^{I}\frac{t_i a_i}{\sigma_i}$$

where $t_i = \frac{1}{\sigma_i}\left(f_i - \frac{\sum_{i=0}^{I} f_i/\sigma_i^2}{\sum_{i=0}^{I} 1/\sigma_i^2}\right)$, and σ_i is a weight factor for each data point that we have found useful to set it proportional to the amplitude value,

$$\sigma_i = \frac{a_i}{\sum_{i=1}^{I}a_i}.$$

For completeness we could also compute the spectral tilt of the residual component but we have not considered a relevant attribute for our purposes.

4.2 Frame variation attributes

The frame to frame variation of each attribute is a useful measure of its time evolution, thus an indication of changes in the sound. It is computed in the same way for each attribute,

$$\Delta = \frac{Val(l) - Val(l-1)}{M}$$

where $Val(l)$ is the attribute value for the current frame and $Val(l-1)$ is the attribute value for the previous one.

4.3 Segmentation

Sound segmentation has proved important in automatic speech recognition and music transcription algorithms. For our purposes it is very valuable as a way to apply region dependent transformations. For example, a time stretching algorithm would be able to transform the steady state regions, leaving the rest unmodified.

The simplest, and most general, segmentation process divides a melody into notes and silences and then each note into an attack, a steady state and a release regions. Attack and release regions are identified by the way the instantaneous attributes change in time and the steady state regions are detected by the stability of these same attributes.

The techniques originally developed for speech [4], based on Pattern-Recognition, Knowledge-Based or Neural Network methodologies, start to be used in music segmentation applications [5]. Most of the approaches apply classification methods that start from sounds features, such as the ones described in this paper, and are able to group sequences of frames into predefined categories. No reliable and general purpose technique has been found. Our experience is that they require narrowing the problem to a specific type of musical signal or including a user intervention stage to help direct the segmentation process.

4.4 Region attributes

Once a given sound has been segmented into regions we can study and extract the attributes that describe each one. Most of the interesting attributes are simply the mean and variance of each of the frame attributes for the whole region. For example, we can compute the mean and variance for the amplitude of sinusoidal and residual components, the fundamental frequency, the spectral shape of sinusoidal and residual components, or the spectral tilt.

Vibrato is a specific attribute present in many steady state regions of sustained instrumental sounds that requires a special treatment. There is another article from our research group that describes this issue in detail [6].

Region attributes can be extracted from the frame attributes in the same way that the frame attributes were extracted from the low level SMS data. The result of the extraction of the frame and region attributes is a hierarchical multi-level data structure where each level represents a different sound abstraction.

5 Attribute transformations

The hierarchical data structure that includes a complete description of a given sound offers many possibilities for sound transformations. Most musically meaningful transformations are done by modifying several attributes at the same time and at different abstraction levels.

Higher level transformations can refer to aspects like sound character, articulation or expressive phrasing. These ideas lead to the development of front ends such as graphical interfaces [7] or knowledge-based systems [8] that are able to deal with the complexity of this sound representation.

6 Conclusion

In this article we have presented an overview of the work being carried out at the Audiovisual Institute in the direction of extending the SMS sound representation towards higher level abstractions. It opens new research problems that will lead to interesting and exciting music applications.

References

[1] X. Serra. "Musical Sound Modeling with Sinusoids plus Noise". G. D. Poli and others (eds.), *Musical Signal Processing*, Swets & Zeitlinger Publishers, 1997.

[2] X. Serra and others. "Integrating Complementary Spectral Models in the Design of a Musical Synthesizer". *Proceedings of the ICMC*, 1997.

[3] P. Cano. "Fundamental Frequency Estimation in the SMS Analysis". *Proceedings of the*

Digital Audio Effects Workshop (DAFX98), 1998.

[4] E. Vidal and A. Marzal. "A Review and New Approaches for Automatic Segmentation of Speech Signals". L. Torres and others (eds.), *Signal Processing V: Theories and Applications*, Elsevier Science Publishers, 1990.

[5] S. Rossignol and others. "Feature Extraction and Temporal Segmentation of Acoustic Signals". *Proceedings of the ICMC*, 1998.

[6] P. Herrera. "Vibrato Extraction and Parameterization in the Spectral Modeling Synthesis Framework". *Proceedings of the Digital Audio Effects Workshop (DAFX98)*, 1998.

[7] A. Loscos and E. Resina. "SmsPerformer: A Real-Time Synthesis Interface for SMS". *Proceedings of the Digital Audio Effects Workshop (DAFX98)*, 1998.

[8] J. L. Arcos and others. "Saxex: a Case-Based Reasoning System for Generating Expressive Musical Performances". *Proceedings of the ICMC,* 1997.