Spectral Modeling Synthesis

Xavier Serra and Julius O. Smith
Center for Computer Research in Music and Acoustics
Department of Music, Stanford University
Stanford, CA 94305

INTRODUCTION

In this paper a new analysis/synthesis method based on spectrum analysis is introduced, designed with the purpose of obtaining a musically useful representation for sound transformations. The underlying model considers a sound to be composed of a deterministic component plus a stochastic one. The deterministic component is represented as a series of sinusoids, where each one is defined by an amplitude and a frequency function. The stochastic component is represented as a series of magnitude-spectrum envelopes, understood as a time-varying filter excited by white noise. From this representation sounds can be synthesized that, in the absence of modifications, can behave as a perceptual identity, that is, they are perceptually equal to the original sound. At the same time the representation is easily modified to create a wide variety of new sounds.

This analysis/synthesis technique is based on the short-time Fourier transform (STFT). From the set of spectra returned by the STFT, the relevant peaks of each spectrum are detected and assigned to a number of frequency trajectories. The deterministic signal is obtained by synthesizing a sinusoid from each trajectory. Then, in order to obtain the stochastic component, a set spectra of the deterministic component are computed and subtracted from the corresponding spectra of the original sound. The resulting spectral residuals are approximated by a series of envelopes, and the stochastic signal generated by performing an inverse-STFT.

The result is a system that is appropriate for the manipulation of sounds. The intermediate representation is very flexible and offers unlimited possibilities for transformation.

In the next section the deterministic plus stochastic model is presented, followed by a description of the analysis/synthesis system.

THE DETERMINISTIC PLUS STOCHASTIC MODEL

A sound model assumes certain characteristics of the sound waveform or the sound generation mechanism. In general, every analysis/synthesis technique has an underlying model. The system presented is this paper assumes the input sound to be composed of a deterministic plus a stochastic component.

A deterministic signal is traditionally defined as anything that is not noise (i.e., a perfectly predictable part, predictable from measurements over any continuous interval). However in the present discussion the class of deterministic signals considered is restricted to sums of quasi-sinusoidal components (sines with piecewise linear amplitude and frequency variation). Each sinusoid models an actual sinusoidal component of the original sound and is described by an amplitude and a frequency function.

A stochastic, or noise, signal is fully described by its amplitude and its general frequency characteristics. When a signal is assumed stochastic it is not necessary to preserve either the instantaneous phase or the exact frequency information.

Therefore, the input sound $s(t)$ is the sum of a series of sinusoids plus a noise signal,

$$s(t) = \sum_{r=1}^{R} A_r(t) \cos[\theta_r(t)] + e(t) \tag{1}$$

where $A_r(t)$ and $\theta_r(t)$ are the instantaneous amplitude and phase of each sinusoid and $e(t)$ is the noise signal.

The model assumes that the sinusoids are stable partials of the sound and that each one can be characterized by its amplitude and frequency. The instantaneous phase is then taken to be the integral of the instantaneous frequency $\omega_r(t)$, and therefore satisfies

$$\theta_r(t) = \int_0^t \omega_r(t) dt \tag{2}$$

where $\omega(t)$ is the frequency in radians, and r is the sinusoid number.

By assuming that $e(t)$ is a stochastic signal, it can be described as filtered white noise,

$$e(t) = \int_0^t h(t-\tau) u(\tau) d\tau \tag{3}$$

where $u(t)$ is white noise and $h(t)$ is the impulse response of a slowly time varying filter. That is, the residual is modeled by the convolution of white noise with a frequency shaping filter.

The filtering of a noise signal corresponds to performing the inverse Fourier transform of the frequency response of the filter. This last approach is the one that will be taken to synthesize the stochastic signal.

DESCRIPTION OF THE SYSTEM

Fig. 1 and Fig. 2 show the block diagrams for the analysis and synthesis parts of the system. The first step is the derivation of a series of magnitude spectra of the waveform by computing the FFT of every windowed portion of the input signal (i.e., computing the STFT). Relevant issues for this system are the choice of the analysis window (the Kaiser window has shown to be very appropriate), its size (in order to get the best in the time-frequency trade-off), and the FFT-buffer size (a big zero-padding improves the performance of the technique). From the series of magnitude spectra the relevant peaks are detected by finding the highest peaks in each spectrum and performing a parabolic

interpolation to refine the peak-frequency and peak-amplitude calculations. These peaks are then organized into frequency trajectories by means of a peak continuation algorithm. The relevance of this algorithm is that it extracts the stable sinusoids present in the original sound (the deterministic component).

The stochastic part of the waveform is calculated by first computing the STFT of the deterministic component, in the same way that the STFT of the original waveform was obtained, and then subtracting each magnitude spectrum from the corresponding spectrum of the original waveform. The spectral envelope of the residual spectrum is then derived by performing a line-segment approximation. These envelopes represent the stochastic signal.

The generation of the deterministic part of the sound, the sinusoidal component, is done from the magnitude and frequency trajectories, or their transformation, by generating a sine wave for each trajectory.

The generation of the stochastic part is done by creating a complex spectrum (i.e., magnitude and phase spectra) for every spectral envelope of the residual, or its modification, and performing an inverse-STFT. The magnitude spectrum is the envelope itself, and the phase spectrum is generated by a random number generator. This process corresponds to the filtering of white noise by a filter with a frequency response equal to the spectral envelope.

CONCLUSIONS

Spectral modeling synthesis is a new analysis-based technique which is capable of capturing the perceptual characteristics of a wide variety of sounds. The analysis part is central to the system. It is a complex algorithm that requires the manual setting of a few control parameters. Further work may automate the analysis process, particularly if there is a specialization for a group of sounds. Also, some aspects of the analysis is open to further research, in particular the peak-continuation algorithm.

The synthesis from the deterministic plus stochastic representation is simple and can be performed in real-time with current technology. A real-time implementation of this system would allow the use of this technique in performance. The representation would be precomputed and stored, and the performer would modify the sounds interactively.

For a detailed discussion of the technique see Serra's dissertation (Serra, 1989).

REFERENCES

Serra, Xavier. 1989. *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition,* Ph.D. Dissertation, Stanford University, Stanford, California.

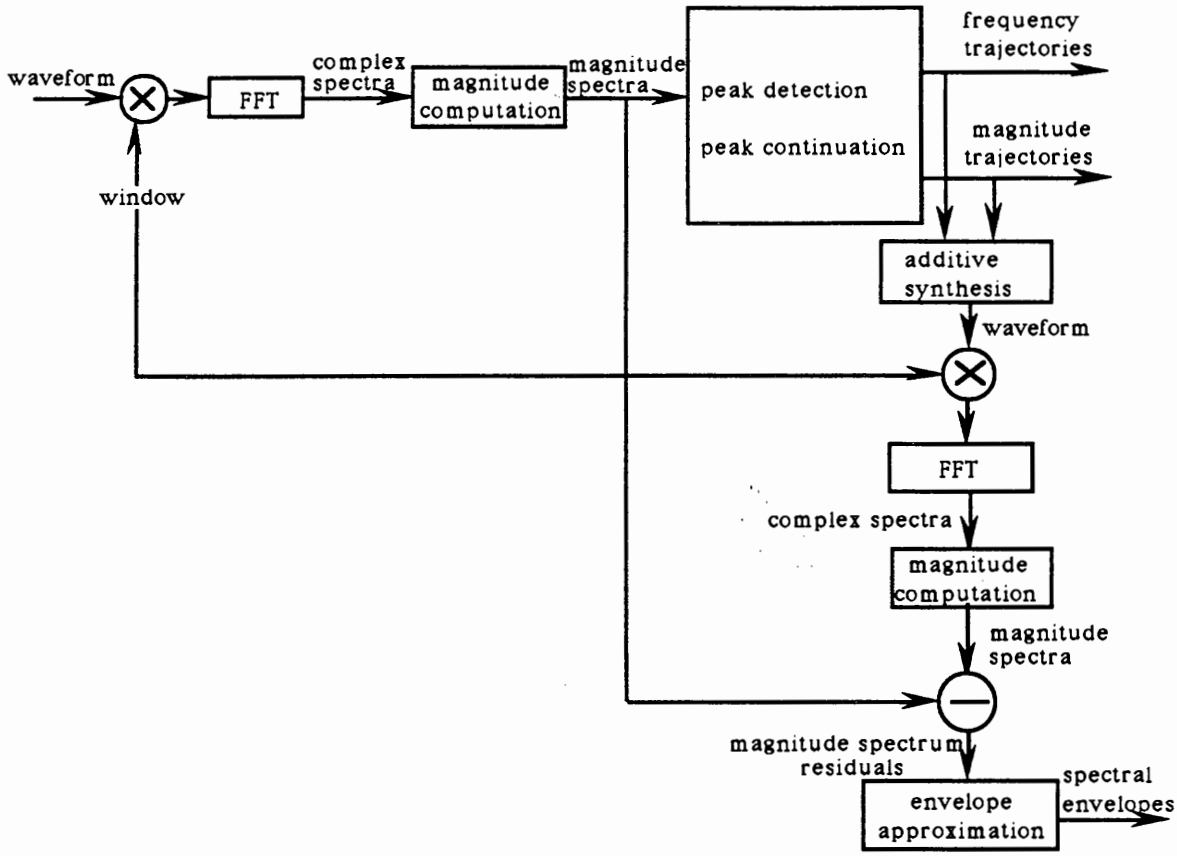

Figure 1: Block diagram of the analysis part of the system.

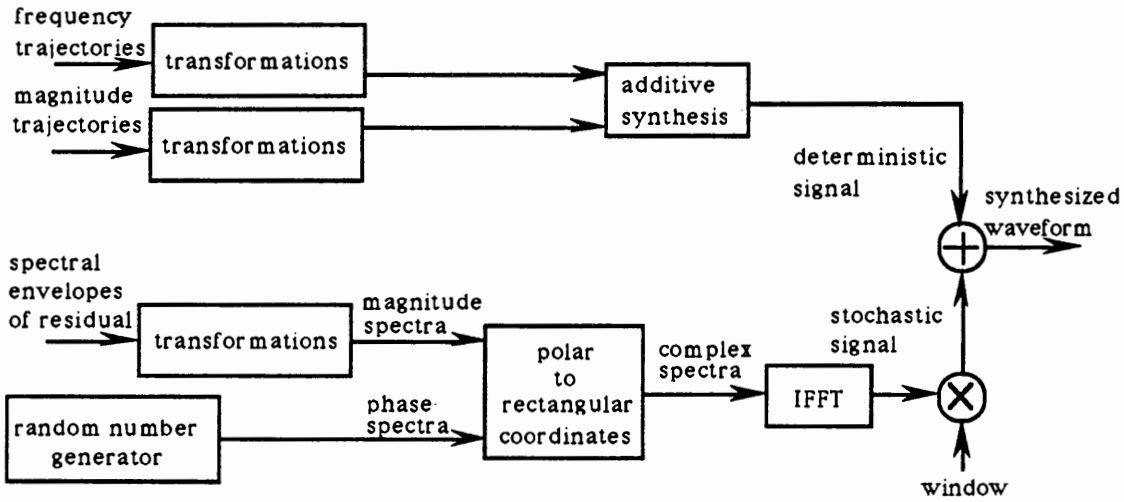

Figure 2: Block diagram of the synthesis part of the system.