Musical Sound Modeling with Sinusoids plus Noise

Xavier Serra
Audiovisual Institute, Pompeu Fabra University
Rambla 31, 08002 Barcelona, Spain
URL: http://www.iua.upf.es
email: xserra@iua.upf.es

[published in C. Roads, S. Pope, A. Picialli, G. De Poli, editors. 1997. "Musical Signal Processing". Swets & Zeitlinger Publishers]

1. Introduction

When generating musical sound on a digital computer, it is important to have a good model whose parameters provide a rich source of meaningful sound transformations. Three basic model types are in prevalent use today for musical sound generation: instrument models, spectrum models, and abstract models. Instrument models attempt to parametrize a sound at its source, such as a violin, clarinet, or vocal tract. Spectrum models attempt to parametrize a sound at the basilar membrane of the ear, discarding whatever information the ear seems to discard in the spectrum. Abstract models, such as FM, attempt to provide musically useful parameters in an abstract formula.

This article addresses the second category of synthesis techniques: spectrum modeling. The main advantage of this group of techniques is the existence of analysis procedures that extract the synthesis parameters out of real sounds, thus being able to reproduce and modify actual sounds. Our particular approach is based on modeling sounds as stable sinusoids (partials) plus noise (residual component), therefore analyzing sounds with this model and generating new sounds from the analyzed data. The analysis procedure detects partials by studying the time-varying spectral characteristics of a sound and represents them with time-varying sinusoids. These partials are then subtracted from the original sound and the remaining "residual" is represented as a time-varying filtered white noise component. The synthesis procedure is a combination of additive synthesis for the sinusoidal part, and subtractive synthesis for the noise part.

This analysis/synthesis strategy can be used for either generating sounds (synthesis) or transforming pre-existing ones (sound processing). To synthesize sounds we generally want to model an entire timbre family, i.e., an instrument, and that can be done by analyzing single tones and isolated note transitions performed on an instrument, and building a data base that characterizes the whole instrument or any desired timbre family, from which new sounds are synthesized. In the case of the sound processing application the goal is to manipulate any given sound, that is, not being restricted to isolated tones and not requiring a previously built data-base of analyzed data.

Some of the intermediate results from this analysis/synthesis scheme, and some of the techniques developed for it, can also be applied to other music related problems, e.g., sound compression, sound source separation, musical acoustics, music perception, performance analysis,... but a discussion of these topics is beyond the current presentation.

2. Background

Additive synthesis is the original spectrum modeling technique. It is rooted in Fourier's theorem which states that any periodic waveform can be modeled as a sum of sinusoids at various amplitudes and harmonic frequencies. Additive synthesis was among the first synthesis techniques in computer music. In fact, it was described extensively in the very first article of the very first issue of the Computer Music Journal (Moorer, 1977).

In the early 1970s, Andy Moorer developed a series of analysis programs to support additive synthesis. He first used the "heterodyne filter" to measure the instantaneous amplitude and frequency of individual sinusoids (Moorer, 1973). The heterodyne filter implements a single frequency bin of the Discrete Fourier Transform (DFT), using the rectangular window. The magnitude and phase derivative of the complex numbers produced by the sliding DFT bin provided instantaneous amplitude and frequency estimates. The next implementation (Moorer, 1978) was based on the Digital Phase Vocoder (Portnoff, 1976). In this system, the fast Fourier transform (FFT) was used to provide, effectively, a heterodyne filter at each harmonic of the fundamental frequency. The use of a non rectangular window gave better isolation among the spectral components.

The main problem with the phase vocoder was that inharmonic sounds, or sounds with time-varying frequency characteristics, were difficult to analyze. The FFT can be regarded as a fixed filter bank or "graphic equalizer": If the size of the FFT is N, then there are N narrow bandpass filters, slightly overlapping, equally spaced between 0 Hz and the sampling rate. In the phase vocoder, the instantaneous amplitude and frequency are computed only for each "channel filter" or "bin." A consequence of using a fixed-frequency filter bank is that the frequency of each sinusoid is not normally allowed to vary outside the bandwidth of its channel, unless one is willing to combine channels in some fashion which requires extra work. (The channel bandwidth is nominally the sampling rate divided by the FFT size.) Also, the analysis system was really set up for harmonic signals—you could analyze a piano if you had to, but the progressive sharpening of the partials meant that there would be frequencies where a sinusoid would be in the crack between two adjacent FFT bins. This was not an insurmountable condition (the adjacent bins could be combined intelligently to provide accurate amplitude and frequency envelopes), but it was inconvenient and outside the original scope of the analysis framework of the phase vocoder.

In the mid eighties Julius Smith developed the program PARSHL for the purpose of supporting inharmonic and pitch-changing sounds (Smith and Serra, 1987). PARSHL was a simple application of FFT peak-tracking technology commonly used in the Navy signal processing community (General Electric, 1977; Wolcin 1980a; 1980b; Smith and Friedlander, 1984). As in the phase vocoder, a series of FFT frames is computed by PARSHL. However, instead of writing out the magnitude and phase derivative of each bin, the FFT is searched for peaks, and the largest peaks are "tracked" from frame to frame. The principal difference in the analysis is the replacement of the phase derivative in each FFT bin by interpolated magnitude peaks across FFT bins. This approach is better suited for analysis of inharmonic sounds and pseudo-harmonic sounds with important frequency variation in time.

Independently at about the same time, Quatieri and McAulay developed a technique similar to PARSHL for analyzing speech (McAulay and Quatieri, 1984; 1986). Both systems were built on top of the short-time Fourier transform (Allen, 1977).

The PARSHL program worked well for most sounds created by simple physical vibrations or driven periodic oscillations. It went beyond the phase vocoder to support spectral modeling of inharmonic sounds. A problem with PARSHL, however, is that it was unwieldy to represent noise-like signals such as the attack of many instrumental sounds. Using sinusoids to simulate noise is extremely expensive because, in principle, noise consists of sinusoids at every frequency within the band limits. Also, modeling noise with sinusoids does not yield a flexible sound representation useful for music applications. Therefore the next natural step to take in spectral modeling of musical sounds was to represent sinusoids and noise as two separate components (Serra, 1989; Serra and Smith, 1990).

3. 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 system has an underlying model. The sounds produced by musical instruments, or by any physical system, can be modeled as the sum of a set of sinusoids plus a noise residual. The sinusoidal, or deterministic, component normally corresponds to the main modes of vibration of the system. The residual comprises the energy produced by the excitation mechanism which is not transformed by the system into stationary vibrations plus any other energy component that is not sinusoidal in nature. For example, in the sound of wind-driven instruments, the deterministic signal is the result of the self-sustained oscillations produced inside the bore and the residual is a noise signal that is generated by the turbulent streaming that takes place when the air from the player passes through the narrow slit. In the case of bowed strings the stable sinusoids are the result of the main modes of vibration of the strings and the noise is generated by the sliding of the bow against the string, plus by other non-linear behavior of the bow-string-resonator system. This type of separation can also be applied to vocal sounds, percussion instruments and even to non-musical sounds produced in nature.

A deterministic signal is traditionally defined as anything that is not noise (i.e., an analytic signal, or 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 (sinusoids with slowly varying amplitude and frequency). Each sinusoid models a narrowband 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 power spectral density which gives the expected signal power versus frequency. When a signal is assumed stochastic, it is not necessary to preserve either the instantaneous phase or the exact magnitude details of individual FFT frames.

Therefore, the input sound $s(t)$ is modeled by,

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

where $A_r(t)$ and $\theta_r(t)$ are the instantaneous amplitude and phase of the r^{th} sinusoid, respectively, and $e(t)$ is the noise component at time t (in seconds).

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

$$\theta_r(t) = \int_0^t \omega_r(\tau)d\tau$$

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,

$$e(t) = \int_0^t h(t,\tau)u(\tau)d\tau$$

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

This model has problems with sounds that include “noisy partials” (for example, produced by a modulation). We have found this type of component, which is in between a deterministic and a stochastic signal, in the higher partials of vocal sounds, in some string sounds, specially when they have vibrato, and in the sound of metal plates, like a crashed cymbal. Due to these problems, the assumed separation between deterministic and stochastic components of a sound is rarely a clear one and the implementation of this process should be flexible enough to give the user some control over how it is done.

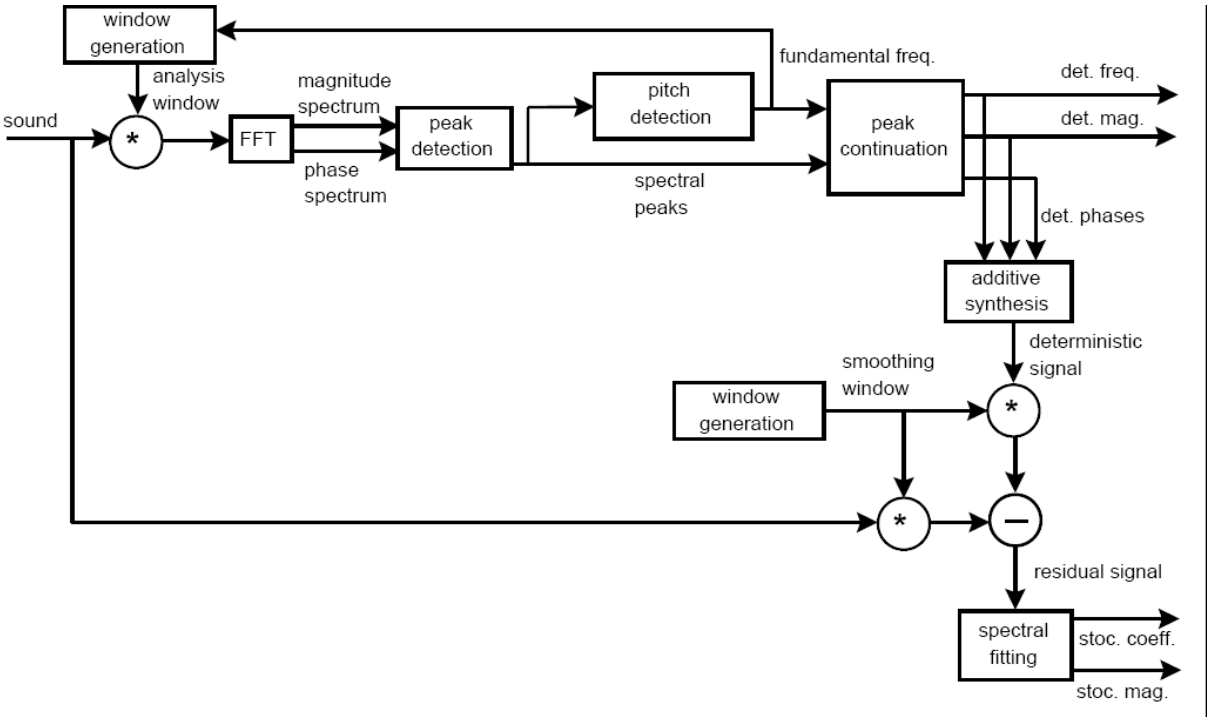

Figure 1: *Block diagram of the analysis process.*

4. General Diagram of the Analysis/Synthesis Process

The deterministic plus stochastic model has many possible implementations and we will present a general one while giving indications on variations that have been proposed. Both the analysis and synthesis are frame based processes with the computation done one frame at a time. Throughout this description we will consider that we have already processed a few frames of the sound and we are ready to compute the next one.

Fig. 1 shows the block diagram for the analysis. First, we prepare the next section of the sound to be analyzed by multiplying it with an appropriate analysis window. Its spectrum is obtained by the Fast Fourier Transform (FFT) and the prominent spectral peaks are detected and incorporated into the existing partial trajectories by means of a peak continuation algorithm. The relevance of this algorithm is that it detects the magnitude, frequency and phase of the partials present in the original sound (the deterministic component). When the sound is pseudo-harmonic, a pitch detection step can improve the analysis by using the fundamental frequency information in the peak continuation algorithm and in choosing the size of the analysis window (pitch-synchronous analysis).

The stochastic component of the current frame is calculated by first generating the deterministic signal with additive synthesis, and then subtracting it from the original waveform in the time domain. This is possible because the phases of the original sound are matched and therefore the shape of the time-domain waveform preserved. The stochastic representation is then obtained by performing a spectral fitting of the residual signal.

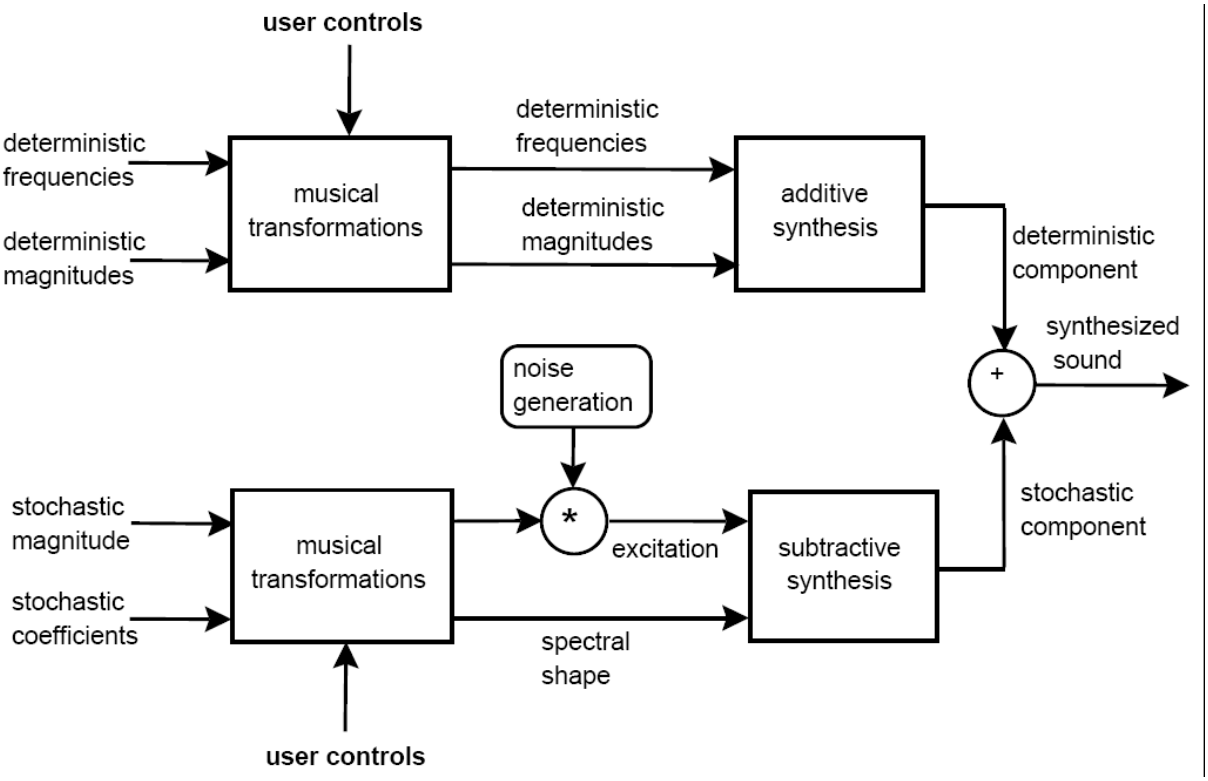

Figure 2: *Block diagram of the synthesis process.*

In Fig. 2 the block diagram of the synthesis process is shown. The deterministic signal, i.e., the sinusoidal component, results from the magnitude and frequency trajectories, or their transformation, by generating a sine wave for each trajectory (i.e., additive synthesis). This can either be implemented in the time domain with the traditional oscillator bank method or in the frequency domain using the inverse-FFT approach.

The synthesized stochastic signal is the result of generating a noise signal with the time-varying spectral shape obtained in the analysis (i.e., subtractive synthesis). As with the deterministic synthesis, it can be implemented in the time domain by a convolution or in the frequency domain by creating a complex spectrum (i.e., magnitude and phase spectra) for every spectral envelope of the residual and performing an inverse-FFT.

5. Magnitude and Phase Spectra Computation

The computation of the magnitude and phase spectra of the current frame is the first step in the analysis. It is in these spectra that the sinusoids are tracked and the decision takes place as to whether a part of the signal is considered deterministic or noise. The computation of the spectra is carried out by the short-time Fourier transform, STFT, technique (Allen, 1977; Serra, 1989).

The control parameters for the STFT (window-size, window-type, FFT-size, and frame-rate) have to be set in accordance with the sound to be processed. First of all, a good resolution of the spectrum is needed since the process that tracks the partials has to be able to identify the peaks which correspond to the deterministic component. Also the phase information is particularly important for subtracting the deterministic component to find the residual; we should use an odd-length analysis window and the windowed data should be centered in the FFT-buffer at the origin in order to obtain the phase spectrum free of the linear phase trend induced by the window (“zero-phase” windowing). A discussion on windows is beyond this article, Harris (Harris, 1978) gives a good introduction on this topic.

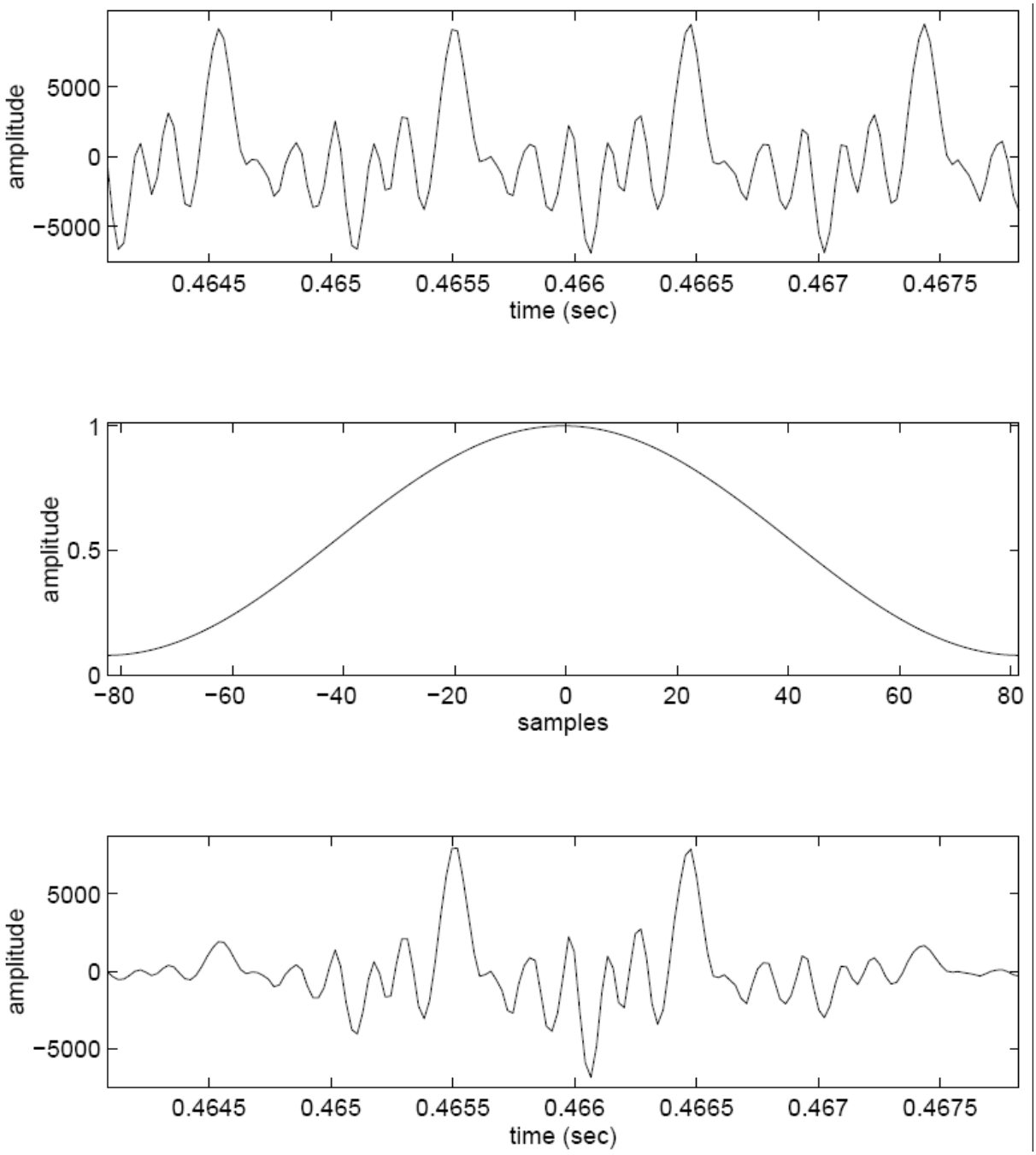

Figure 3: *Sound selection and windowing. a. Portion of a violin sound to be used in the analysis of the current frame. b. Hamming window. c. Windowed sound.*

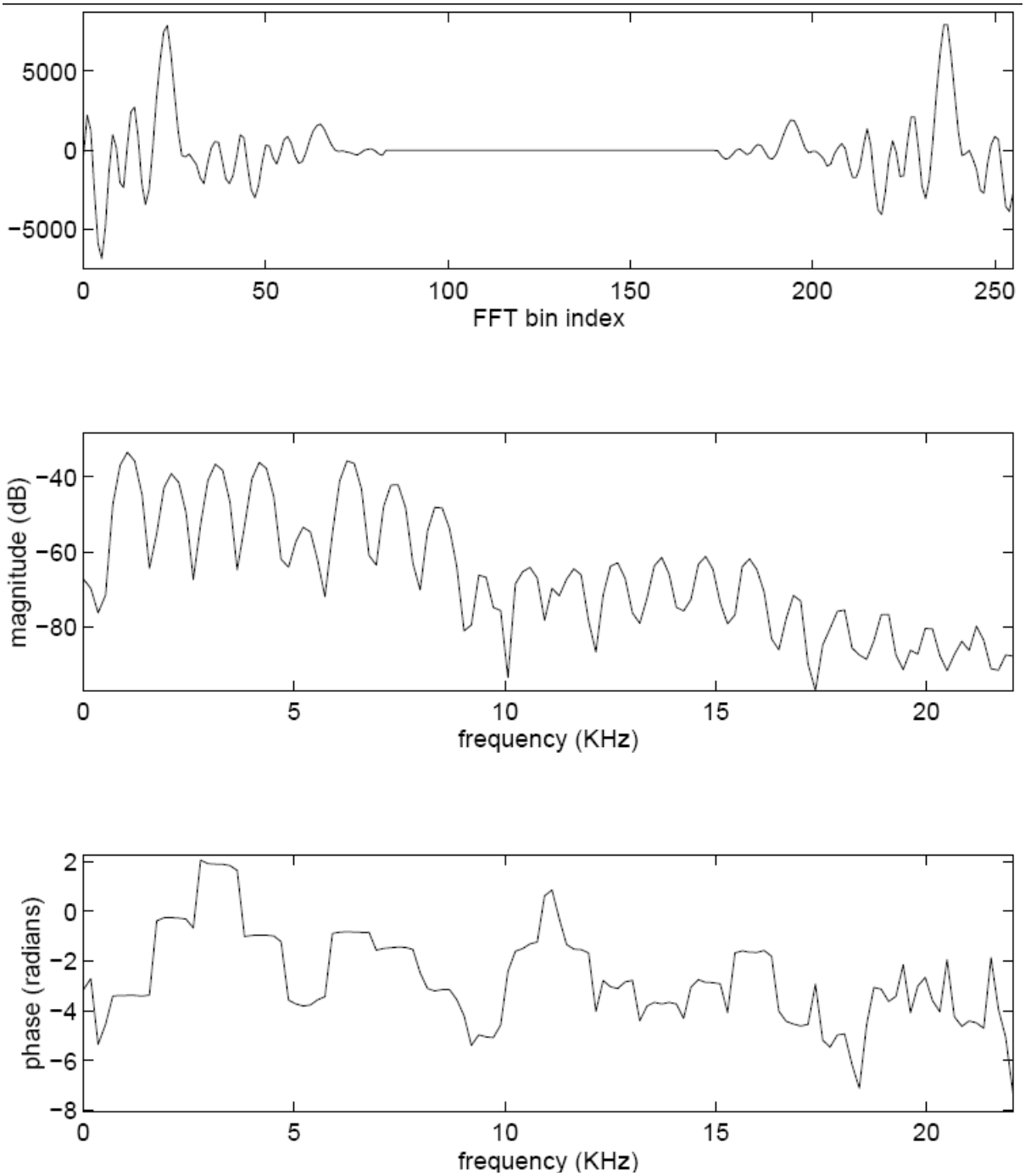

Figure 4: *Computing the FFT. a. packing of the sound into the FFT buffer for a zero phase spectrum. b. Magnitude spectrum. c. Phase spectrum.*

Since the synthesis process is completely independent from the analysis, the restrictions imposed by the STFT when the inverse transform is also performed, i.e. that the analysis windows add to a constant, are unnecessary here. The STFT parameters are more flexible, and we can vary them during the course of the analysis, if that is required to improve the detection of partials.

The time-frequency compromise of the STFT has to be well understood. For the deterministic analysis it is important to have enough frequency resolution to resolve the partials of the sound. For the stochastic analysis the frequency resolution is not that important, since we are not interested in particular frequency components, and we are more concerned with a good time resolution. This can be accomplished by using different parameters for the deterministic and the stochastic analysis.

In stable sounds we should use long windows (several periods) with a good side-lobe rejection (for example, Blackman-Harris 92dB) for the deterministic analysis. This gives a good frequency resolution, therefore a good measure of the frequencies of the partials, but most sounds cannot afford these settings and a compromise is required. In the case of harmonic sounds the actual size of the window will change as pitch changes, in order to assure a constant time-frequency trade-off for the whole sound. In the case of inharmonic sounds we should set the window-size depending on the minimum frequency difference that exists between partials.

6. Peak Detection

Once the spectrum of the current frame is computed, the next step is to detect its prominent magnitude peaks. Theoretically, a sinusoid that is stable both in amplitude and in frequency (a partial) has a well defined frequency representation: the transform of the analysis window used to compute the Fourier transform. It should be possible to take advantage of this characteristic to distinguish partials from other frequency components. However, in practice this is rarely the case since most natural sounds are not perfectly periodic and do not have nicely spaced and clearly defined peaks in the frequency domain. There are interactions between the different components, and the shapes of the spectral peaks cannot be detected without tolerating some mismatch. Only some instrumental sounds (e.g., the steady-state part of an oboe sound) are periodic enough and sufficiently free from prominent noise components that the frequency representation of a stable sinusoid can be recognized easily in a single spectrum. A practical solution is to detect as many peaks as possible and delay the decision of what is a deterministic, or “well behaved” partial, to the next step in the analysis: the peak continuation algorithm.

A “peak” is defined as a local maximum in the magnitude spectrum, and the only practical constraints to be made in the peak search are to have a frequency range and a magnitude threshold. In fact, we should detect more than what we hear and get as many sample bits as possible from the original sound, ideally more than 16. The measurement of very soft partials, sometimes more than 80dB below maximum amplitude, will be hard and they will have little resolution. These peak measurements are very sensitive to transformations because as soon as modifications are applied to the analysis data, parts of the sound that could not be heard in the original can become audible. The original sound should be as clean as possible and have the maximum dynamic range, then the magnitude threshold can be set to the amplitude of the background noise floor. To get a better resolution in higher frequencies, preemphasis can be applied before the analysis, which is then undone during the synthesis.

Due to the sampled nature of the spectra returned by the FFT, each peak is accurate only to within half a sample. A spectral sample represents a frequency interval of f_s / N Hz, where f_s is the sampling rate and N is the FFT size. Zero-padding in the time domain increases the number of spectral samples per Hz and thus increases the accuracy of the simple peak detection. However, to obtain frequency accuracy on the level of 0.1% of the distance from the top of an ideal peak to its first zero crossing (in the case of a Rectangular window), the zero-padding factor required is 1000. A more efficient spectral interpolation scheme is to zero-pad only enough so that quadratic (or other simple) spectral interpolation, using only samples immediately surrounding the maximum-magnitude sample, suffices to refine the estimate to $.1\%$ accuracy.

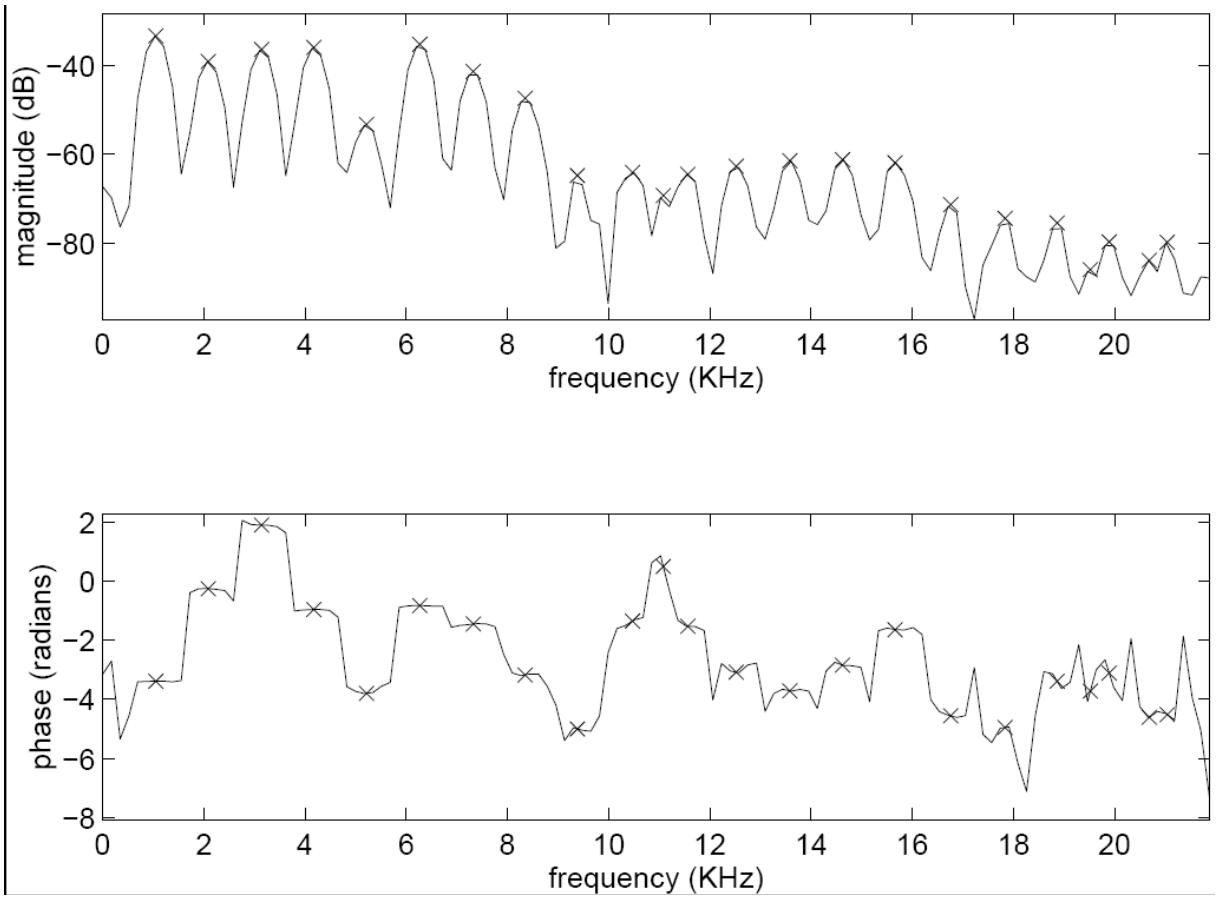

Figure 5: *Peak detection. a. Peaks in the magnitude spectrum. b. Peaks in the phase spectrum.*

Although we cannot rely on the exact shape of the peak to decide whether it is a partial or not, it is sometimes useful to have a measure of how close its shape is to the ideal sinusoidal peak. This measure can be obtained by calculating the difference from the samples of the measured peak to the samples of the analysis window transform centered at the measured frequency and scaled to the measured magnitude. This information, plus the frequency, magnitude, and phase of the peak, can help in the peak continuation process.

7. Pitch Detection

Before continuing a set of peak trajectories through the current frame it is useful to search for a possible fundamental frequency, that is, for periodicity. If it exists, we will have more information to work with, and it will simplify and improve the tracking of partials. This fundamental frequency can also be used to set the size of the analysis window, in order to maintain constant the number of periods to be analyzed at each frame and to get the best time-frequency trade-off possible. This is called a pitch-synchronous analysis.

Given a set of spectral peaks, with magnitude and frequency values for each one, there are many possible fundamental detection strategies (Piszczalski and Galler, 1979; Terhardt, 1982; Hess, 1983; Doval and Rodet, 1993; Maher and Beauchamp, 1994). For this presentation we restrict ourselves to single-source sounds and assume that a fundamental peak or one of its first few partials exists. With these two constraints, plus the fact that there is some number of buffered frames, the algorithm can be quite simple.

The fundamental frequency can be defined as the common divisor of the harmonic series that best explains the spectral peaks found in the current frame. The first step is to find the possible candidates

inside a given range. This can be done by stepping through the range by small increments, or by only considering as candidates the frequencies of the measured spectral peaks and frequencies related to them by simple integer ratios (e.g., $1/2, 1/3, 1/4$) that lie inside the range. This last approach simplifies our search enormously.

Once the possible candidates have been chosen we need a way to measure the "goodness" of the resulting harmonic series compared with the actual spectral peaks. A suitable error measure (Maher and Beauchamp, 1994) is based on the weighted differences between the measured peaks and the ideal harmonic series (predicted peaks).

The predicted to measured error is defined as:

$$Err_{p\to m} = \sum_{n=1}^{N} E_{\omega}(\Delta f_n, f_n, a_n, A_{\max}) = \sum_{n=1}^{N} \Delta f_n \cdot (f_n)^{-p} + (\frac{a_n}{A_{\max}}) \times \left[q\Delta f_n \cdot (f_n)^{-p} - r \right]$$

where Δf_n is the difference between a predicted and its closest measured peak, f_n and a_n are the frequency and magnitude of the predicted peaks, and A_{max} is maximum peak magnitude.

The measured to predicted error is defined as:

$$Err_{m\to p} = \sum_{k=1}^{K} E_{w}(\Delta f_k, f_k, a_k, A_{\max}) = \sum_{k=1}^{K} \Delta f_k (f_k)^{-p} + (\frac{a_k}{A_{\max}}) \times \left[q\Delta f_k \cdot (f_k)^{-p} - r \right]$$

where Δf_k is the difference between a measured and its closest predicted peak, f_k and a_k are the frequency and magnitude of the measured peaks, and A_{max} is maximum peak magnitude.

The total error is:

$$Err_{total} = Err_{p\to m}/N + \rho Err_{m\to p}/K$$

Maher and Beauchamp propose to use $p = 0.5$, $q = 1.4$, $r = 0.5$ and $\rho = 0.33$. The harmonic series with the smallest error is chosen, and since there is a certain frame memory, the preferred fundamental in a given frame will be compared with the fundamentals found in the previous ones. If the new fundamental is very different from the preceding ones it is possible that something is wrong. Either we are in the middle of a note transition or the new fundamental is not a real one. When this new fundamental value does not prevail for a few frames it will not be accepted and the frame will be considered as containing only stochastic data. After a few frames, once it is clear that we were in a note transition and that we are now in a new note, we can re-analyze the previous frames by setting the window-size according to the fundamental of the new note. This will improve the measurements in the preceding frames and their time-frequency trade-off. We can also set the window-size for the next frame according to it.

8. Peak Continuation

Once the spectral peaks of the current frame have been detected, the peak continuation algorithm adds them to the incoming peak trajectories. The schemes used in PARSHL (Smith and Serra, 1987) and in the sinusoidal model (McAulay and Quatieri, 1984; 1986) find peak trajectories both in the noise and deterministic parts of a waveform, thus obtaining a sinusoidal representation for the whole sound.

These schemes are unsuitable when we want the trajectories to follow just the partials. For example, when the partials change in frequency substantially from one frame to the next, these algorithms easily switch from the partial that they were tracking to another one which at that point is closer.

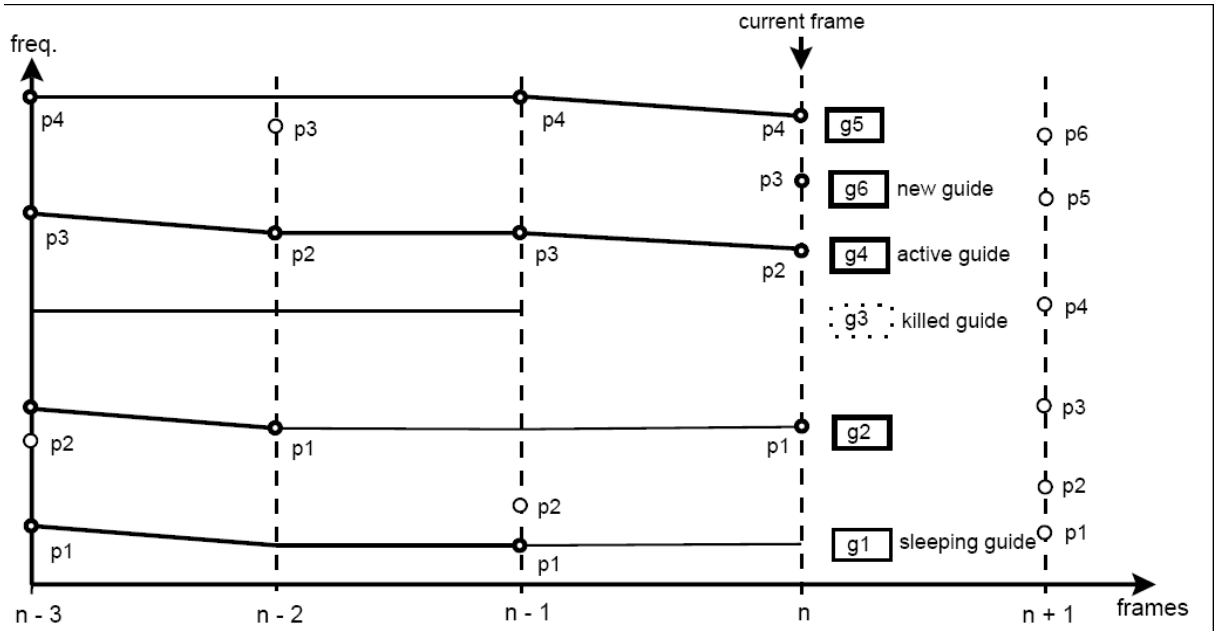

Figure 6: *Peak Continuation process. g represent the guides and p the spectral peaks.*

The algorithm described here is intended to track partials in a variety of sounds, although the behavior of a partial, and therefore the way to track it, varies depending on the signal. Whether we have speech, a harmonic instrumental tone, a gong sound, a sound of an animal, or any other, the time progression of the component partials will vary. Thus, the algorithm requires some knowledge about the characteristics of the sound that is being analyzed. In the current algorithm there is no attempt to make the process completely automatic and some of the characteristics of the sound are specified through a set of parameters. The specific parameters used in our implementation will be presented in the appendix.

The basic idea of the algorithm is that a set of "guides" advances in time through the spectral peaks, looking for the appropriate ones (according to the specified constraints) and forming trajectories out of them. Thus, a guide is an abstract entity which is used by the algorithm to create the trajectories and the trajectories are the actual result of the peak continuation process. The instantaneous state of the guides, their frequency and magnitude, are continuously updated as the guides are turned on, advanced, and finally turned off. For the case of harmonic sounds these guides are created at the beginning of the analysis, setting their frequencies according to the harmonic series of the first fundamental found, and for inharmonic sounds each guide is created when it finds the first available peak.

When a fundamental has been found in the current frame, the guides can use this information to update their values. Also the guides can be modified depending on the last peak incorporated. Therefore by using the current fundamental and the previous peak we control the adaptation of the guides to the instantaneous changes in the sound. For a very harmonic sound, since all the harmonics evolve together, the fundamental should be the main control, but when the sound is not very harmonic, or the harmonics are not locked to each other and we cannot rely on the fundamental as a strong reference for all the harmonics, the information of the previous peak should have a bigger weight.

Each peak is assigned to the guide that is closest to it and that is within a given frequency deviation. If a guide does not find a match it is assumed that the corresponding trajectory must "turn off". In

inharmonic sounds, if a guide has not found a continuation peak for a given amount of time the guide is killed. New guides, and therefore new trajectories, are created from the peaks of the current frame that are not incorporated into trajectories by the existing guides. If there are killed or unused guides, a new guide can be started. A guide is created by searching through the "unclaimed" peaks of the frame for the one with the highest magnitude. Once the trajectories have been continued for a few frames, the short ones can be deleted and trajectories with small gaps can be filled by interpolating the edges of the gaps.

The attack portion of most sounds is quite "noisy", and the search for partials is harder in such rich spectra. A useful modification to the analysis is to perform it backwards in time. The tracking process encounters the end of the sound first, and since this is a very stable part in most instrumental sounds, the algorithm finds a very clear definition of the partials. When the guides arrive at the attack, they are already tracking the main partials and can reject non-relevant peaks appropriately, or at least evaluate them with some acquired knowledge.

The peak continuation algorithm presented is only one approach to the peak continuation problem. The creation of trajectories from the spectral peaks is compatible with very different strategies and algorithms; for example, hidden Markov models have been applied (Garcia, 1992; Depalle, Garcia and Rodet, 1993). An N Markov model provides a probability distribution for a parameter in the current frame as a function of its value across the past N frames. With a hidden Markov model we are able to optimize groups of trajectories according to a defined criteria, such as frequency continuity. This type of approach might be very valuable for tracking partials in polyphonic sounds and complex inharmonic tones. In particular, the notion of "momentum" is introduced, helping to properly resolve crossing fundamental frequencies.

9. Stochastic Analysis

The deterministic component is subtracted from the original sound either in the time domain or in the frequency domain. This results in a residual sound on which the stochastic approximation is performed. It is useful to study this residual in order to check how well the deterministic component has been properly subtracted and therefore analyzed. If partials remain in the residual, the stochastic analysis models them as filtered noise and it will not sound good. In this case we should re-analyze the sound until we get a good enough residual, free of deterministic components. Ideally the resulting residual should be as close as possible to a stochastic signal. If the sound was not recorded in the ideal situation, the residual will also contain more than just the stochastic part of the sound, such as reverberation or background noise.

To model the stochastic part of sounds, such as the attacks of most percussion instrument, the bow noise in string instruments, or the breath noise in wind instruments, we need a good time resolution and we can give up some frequency resolution. The deterministic component cannot maintain the sharpness of the attacks, because, even if a high frame-rate is used we are forced to use a long enough window, and this size determines most of the time resolution. When the deterministic subtraction is done in the time domain, the time resolution in the stochastic analysis can be improved by redefining the analysis window. The frequency domain approach implies that the subtraction is done in the spectra computed for the deterministic analysis, thus the STFT parameters cannot be changed (Serra, 1989).

In order to be able to perform a time domain subtraction, the phases of the original sound have to be preserved, this is the reason for calculating the phase of each spectral peak. But to generate a deterministic signal that preserves phases is computationally very expensive, as will be shown later. If we stay in the frequency domain, phases are not required and the subtraction of the spectral peaks from the original spectra, the ones that belong to partials, is simple. While the time domain subtraction is more expensive, the results are sufficiently better to to favor this method. This is done by first synthesizing one frame of the deterministic component which is then subtracted from the original

sound in the time domain. The magnitude spectrum of this residual is then computed and approximated with an envelope. The more coefficients we use, the better the modeling of the frequency characteristics will be.

Since it is the deterministic signal that is subtracted from the original sound, measured from long windows, the resulting residual signal might have the sharp attacks smeared. To improve the stochastic analysis, we can "fix" this residual so that the sharpness of the attacks of the original sound are preserved. The resulting residual is compared with the original waveform and its amplitude re-scaled whenever the residual has a greater energy than the original waveform. Then the stochastic analysis is performed on this scaled residual. Thus, the smaller the window the better time resolution we will get in the residual. We can also compare the synthesized deterministic signal with the original sound and whenever this signal has a greater energy than the original waveform it means that a smearing of the deterministic component has been produced. This can be fixed a bit by scaling the amplitudes of the deterministic analysis in the corresponding frame by the difference between original sound and deterministic signal.

Most of the problems with the residual, thus with the stochastic analysis, is in the low frequencies. In general there is more energy at low frequencies that there should be. Since most of the stochastic components of musical sounds mainly contain energy at high frequencies, a fix to this problems is to apply a high-pass filter to the residual before the stochastic approximation is done.

Once the analysis is finished we can still do some post-processing to improve the data. For example, if we had a "perfect" recording and a "perfect" analysis, in percussive or plucked sounds there should be no stochastic signal after the attack. Due to errors in the analysis or to background noise, the stochastic analysis might have detected some signal after this attack. We can delete or reduce this stochastic signal appropriately after the attack.

Next we describe the two main steps involved in the stochastic analysis; the synthesis and subtraction of the deterministic signal from the original sound, and the modeling of the residual signal.

10. Deterministic Subtraction

The output of the peak continuation algorithm is a set of peak trajectories updated for the current frame. From these trajectories a series of sinusoids can be synthesized which reproduce the instantaneous phase and amplitude of the partials of the original sound. Thus, it is possible to subtract the synthesized sinusoids from the original sound and obtain a residual which is substantially free of the deterministic part.

One frame of the deterministic part of the sound, $d(m)$, is generated by

$$d(m) = \sum_{r=1}^{R} \hat{A}_r \cos\left[m\hat{\omega}_r + \hat{\varphi}_r\right], \qquad m = 0,1,2,\ldots,S-1$$

where R is the number of trajectories present in the current frame and S is the length of the frame. To avoid "clicks" at the frame boundaries, the parameters $(\hat{A}_r, \hat{\omega}_r, \hat{\varphi}_r)$ are smoothly interpolated from frame to frame.

Let $(\hat{A}_r^{(l-1)}, \hat{\omega}_r^{(l-1)}, \hat{\varphi}_r^{(l-1)})$ and $(\hat{A}_r^{l}, \hat{\omega}_r^{l}, \hat{\varphi}_r^{l})$ denote the sets of parameters at frames $l-1$ and l for the r^{th} frequency trajectory (we will simplify the notation by omitting the subscript r). These parameters are taken to represent the state of the signal at time S (the left endpoint) of the frame.

The instantaneous amplitude $\hat{A}(m)$ is easily obtained by linear interpolation,

$$\hat{A}(m) = \hat{A}^{l-1} + \frac{(\hat{A}^{l} - \hat{A}^{l-1})}{S} m$$

where $m = 0,1,\ldots,S-1$ is the time sample into the l^{th} frame.

Frequency and phase values are tied together (frequency is the phase derivative), and both control the instantaneous phase $\hat{\theta}(m)$, defined as

$$\hat{\theta}(m) = m\hat{\omega} + \hat{\varphi}$$

Given that four variables affect the instantaneous phase: $\hat{\omega}^{(l-1)}$, $\hat{\varphi}^{(l-1)}$, $\hat{\omega}$, and $\hat{\varphi}$, we need three degrees of freedom for its control, but linear interpolation gives only one. Therefore, we need a cubic polynomial as an interpolation function,

$$\hat{\theta}(m) = \xi + \kappa m + \eta m^2 + \iota m^3$$

It is unnecessary to go into the details of solving this equation since they are described by McAulay and Quatieri (McAulay and Quatieri, 1986). The result is

$$\hat{\theta}(m) = \varphi^{(l-1)} + \hat{\omega}^{(l-1)} m + \eta m^2 + \iota m^3$$

where η and ι are calculated using the end conditions at the frame boundaries,

$$\eta = \frac{3}{S^2}\left(\hat{\varphi}^{l} - \hat{\varphi}^{l-1} - \hat{\omega}^{l-1} S + 2\pi M\right) - \frac{1}{S}\left(\hat{\omega}^{l} - \hat{\omega}^{l-1}\right)$$

$$\iota = -\frac{2}{S^3}\left(\hat{\varphi}^{l} - \hat{\varphi}^{l-1} - \hat{\omega}^{l-1} S + 2\pi M\right) - \frac{1}{S^2}\left(\hat{\omega}^{l} - \hat{\omega}^{l-1}\right)$$

This gives a set of interpolating functions depending on the value of M, among which we select the maximally smooth function. This is done by choosing M to be the integer closest to x, where x is

$$x = \frac{1}{2\pi}\left[\left(\hat{\varphi}^{l-1} + \hat{\omega}^{l-1} S - \hat{\varphi}^{l}\right) + \frac{S}{2}\left(\hat{\omega}^{l} - \hat{\omega}^{l-1}\right)\right]$$

Finally, the synthesis equation for frame l becomes

$$d^{l}(m) = \sum_{r=1}^{R^{l}} \hat{A}_r^{l}(m) \cos\left[\hat{\theta}_r^{l}(m)\right]$$

which goes smoothly from the previous to the current frame with each sinusoid accounting for both the rapid phase changes (frequency) and the slowly varying phase changes.

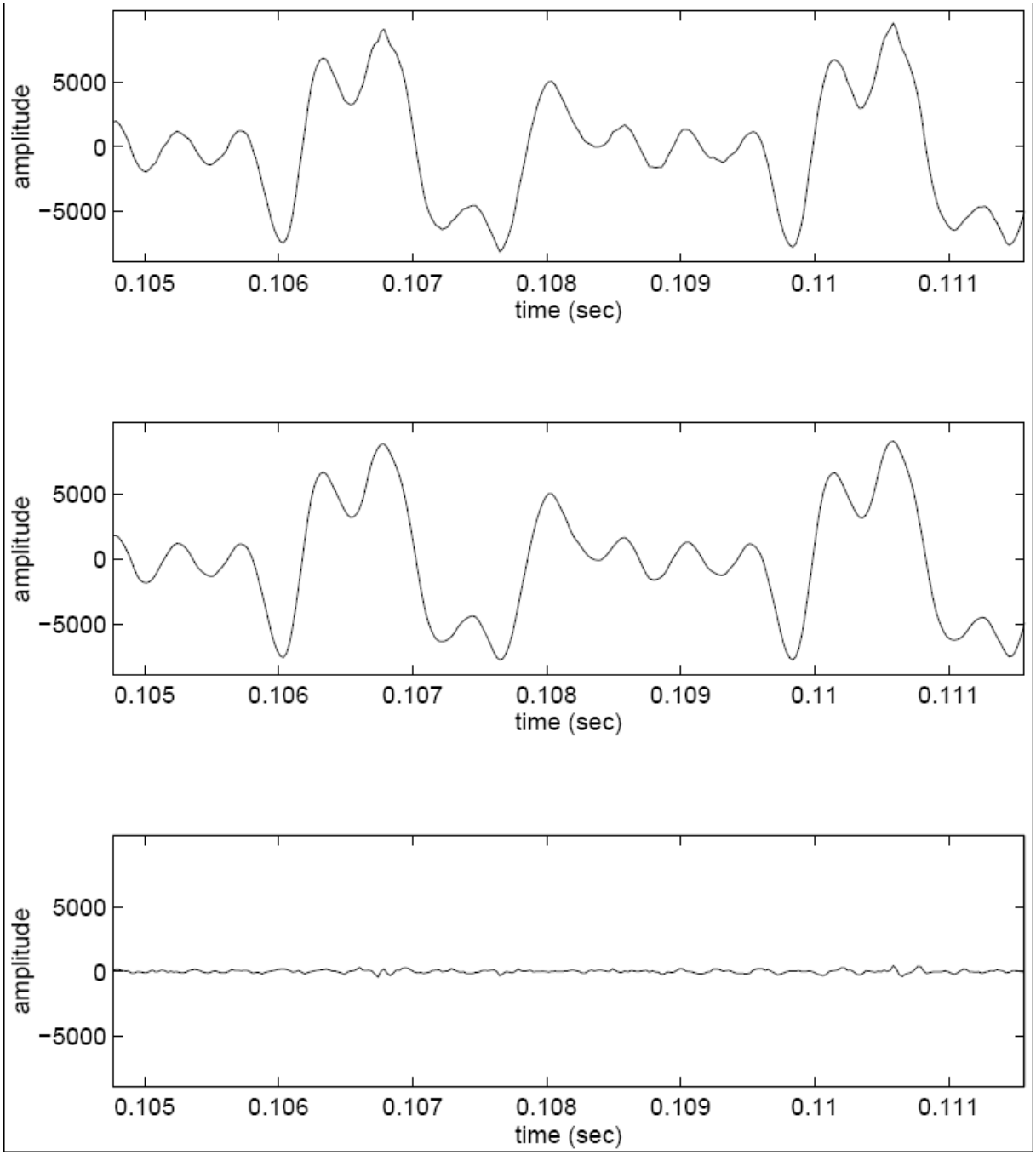

Figure 7: *Deterministic subtraction. a. original sound. b. deterministic synthesis. c. residual sound.*

The synthesized deterministic component can be subtracted from the original sound in the time domain by

$$e(n) = w(n) \times (s(n) - d(n)) \qquad n = 0,1,\ldots, N-1$$

where $e(n)$ is the residual, $w(n)$ a smoothing window, $s(n)$ the original sound, $d(n)$ the deterministic component, and N the size of the window. We already have mentioned that it is desirable to set N smaller than the window-size used in the deterministic analysis in order to improve the time resolution of the residual signal. While in the deterministic analysis the window-size was chosen large enough to obtain a good partial separation in the frequency domain, in the deterministic subtraction we

are especially looking for good time resolution. This is particularly important in the attacks of percussion instruments.

Tests on this residual can be performed to check whether the deterministic plus stochastic decomposition has been successful (Serra, 1994a). Ideally the resulting residual should be as close as possible to a stochastic signal. Since the autocorrelation function of white noise is an impulse, a measure of correlation relative to total power could be a good measure of how close we are to white noise,

$$c = \frac{\sum_{l=0}^{L-1} |r(l)|}{(L-1)r(0)}$$

where $r(l)$ is the autocorrelation estimate for L lags of the residual, and c will be close to 0 when the signal is stochastic. A problem with this measure is that it does not behave well when partials are still left in the signal; for example, it does not always decrease as we progressively subtract partials from a sound. A simpler and sometimes better indication of the quality of the residual is to measure the energy of the residual as a percentage of the total sound energy. Although a problem with this measure is that it cannot distinguish subtracting partials from subtracting noise, and its value will always decrease as long as we subtract some energy, it is still a practical measure for choosing the best analysis parameters.

This sound decomposition is useful in itself for a number of applications. The deterministic component is a set of partials, and the residual includes noise and very unstable components of the sound. This technique has been used (Chafe, 1990; Schumacher and Chafe, 1990) to study bow noise in string instruments and breath noise in wind instruments. In general, this decomposition strategy can give a lot of insight into the makeup of sounds.

The residual component is the part of the instrumental sounds that the existing synthesis techniques have a harder time reproducing, and it is especially important during the attack. A practical application would be to add these residuals to synthesized sounds in order to make them more realistic. Since these residuals remain largely invariant throughout most of the instrumental range, only a few residuals would be necessary to cover all the sounds of a single instrument.

11. Stochastic Approximation

One of the underlying assumptions of the current model is that the residual is a stochastic signal. Such an assumption implies that the residual is fully described by its amplitude and its general frequency characteristics. It is unnecessary to keep either the instantaneous phase or the exact spectral shape information. Based on this, a frame of the stochastic residual can be completely characterized by a filter, i.e., this filter encodes the amplitude and general frequency characteristics of the residual. The representation of the residual for the overall sound will be a sequence of these filters, i.e., a time-varying filter.

The filter design problem is generally solved by performing some sort of curve fitting in the magnitude spectrum of the current frame (Strawn, 1980; Sedgewick, 1988). Standard techniques are: spline interpolation (Cox, 1971), the method of least squares (Sedgewick, 1988), or straight line approximations (Phillips, 1968). For our purpose a simple line-segment approximation to the log-magnitude spectrum is accurate enough and gives the desired flexibility.

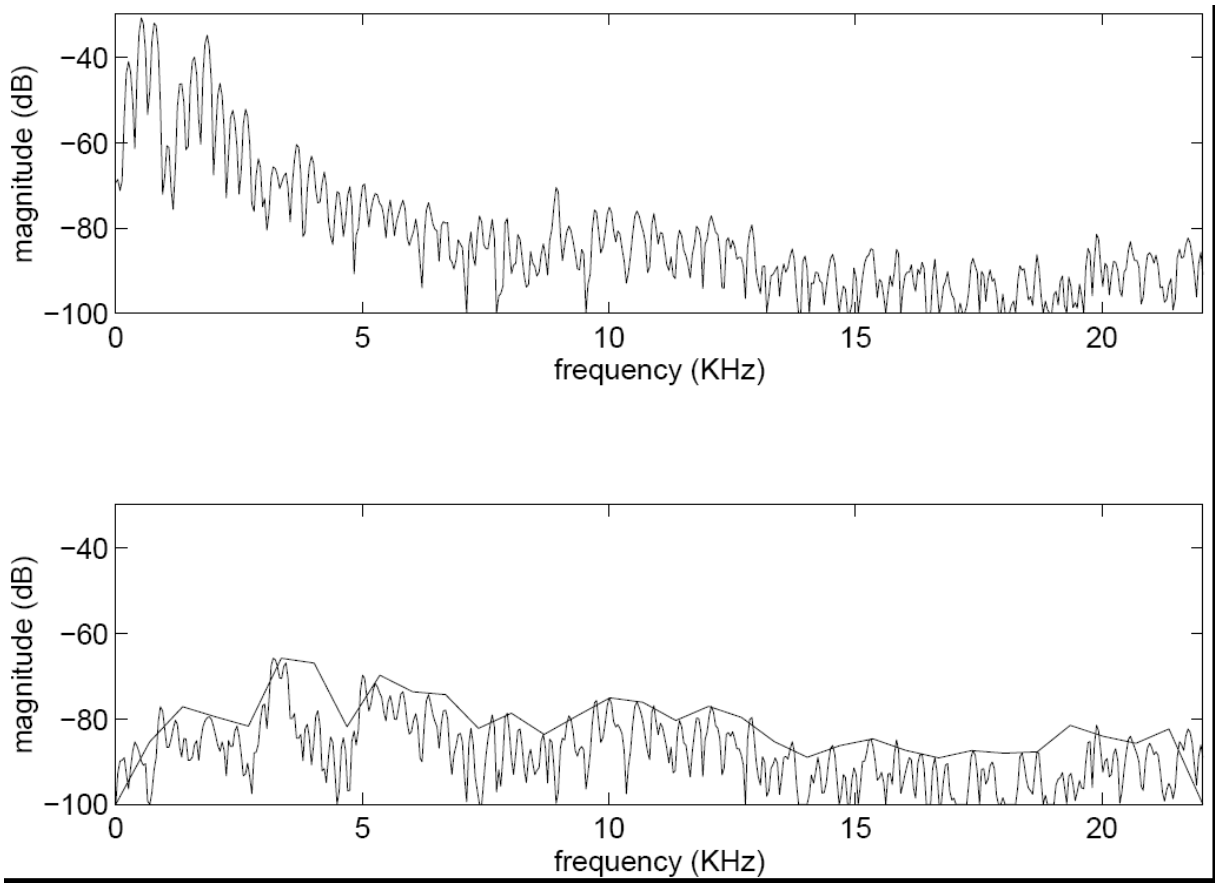

Figure 8: *Stochastic approximation from the sound in Fig. 7. a. Original spectrum. b. Residual spectrum and its line-segment approximation.*

One way to carry out the line-segment approximation is to step through the magnitude spectrum and find local maxima in each of several defined sections, thus giving equally spaced points in the spectrum that are connected by straight lines to create the spectral envelope. The accuracy of the fit is given by the number of points, and that can be set depending on the sound complexity. Other options are to have unequally spaced points, for example, logarithmically spaced, or spaced according to perceptual criteria.

Another practical alternative is to use a type of least squares approximation called linear predictive coding, LPC (Makhoul, 1975; Markel and Gray, 1976). LPC is a popular technique used in speech research for fitting an nth-order polynomial to a magnitude spectrum. For our purposes, the line-segment approach is more flexible than LPC, and although LPC results in less analysis points, the flexibility is considered more important.

12. Representation of the Analysis Data

The output of the deterministic analysis is a set of amplitude and frequency functions with one breakpoint for every frame. From these functions a series of sinusoids can be synthesized which reproduce the deterministic part of the sound. The phase trajectories are not kept because they are unnecessary in the final synthesis, they are perceptually irrelevant in most cases, and they make it harder to perform modifications. However, we have found some situations in which the preservation of phases has made a difference in the quality of the resynthesis. These are: badly analyzed sounds, very low instrumental tones, and some vocal sounds. In the case of badly analyzed sounds, some of the trajectories may actually be tracking non-deterministic parts of the signal, in which case the phase of

the corresponding peaks is important to recover the noisy characteristics of the signal. In the case when the analyzed sound has a very low fundamental, maybe lower than 30 Hz, and the partials are phase-locked to the fundamental, the period is perceived as a pulse and the phase of the partials is required to maintain this perceptual effect. Also in the case of some vocal sounds, the higher partials have a high degree of modulation that cannot be completely recovered from the frequency and magnitude information of the partials, but that seems to be maintained when we add the phases of the peaks.

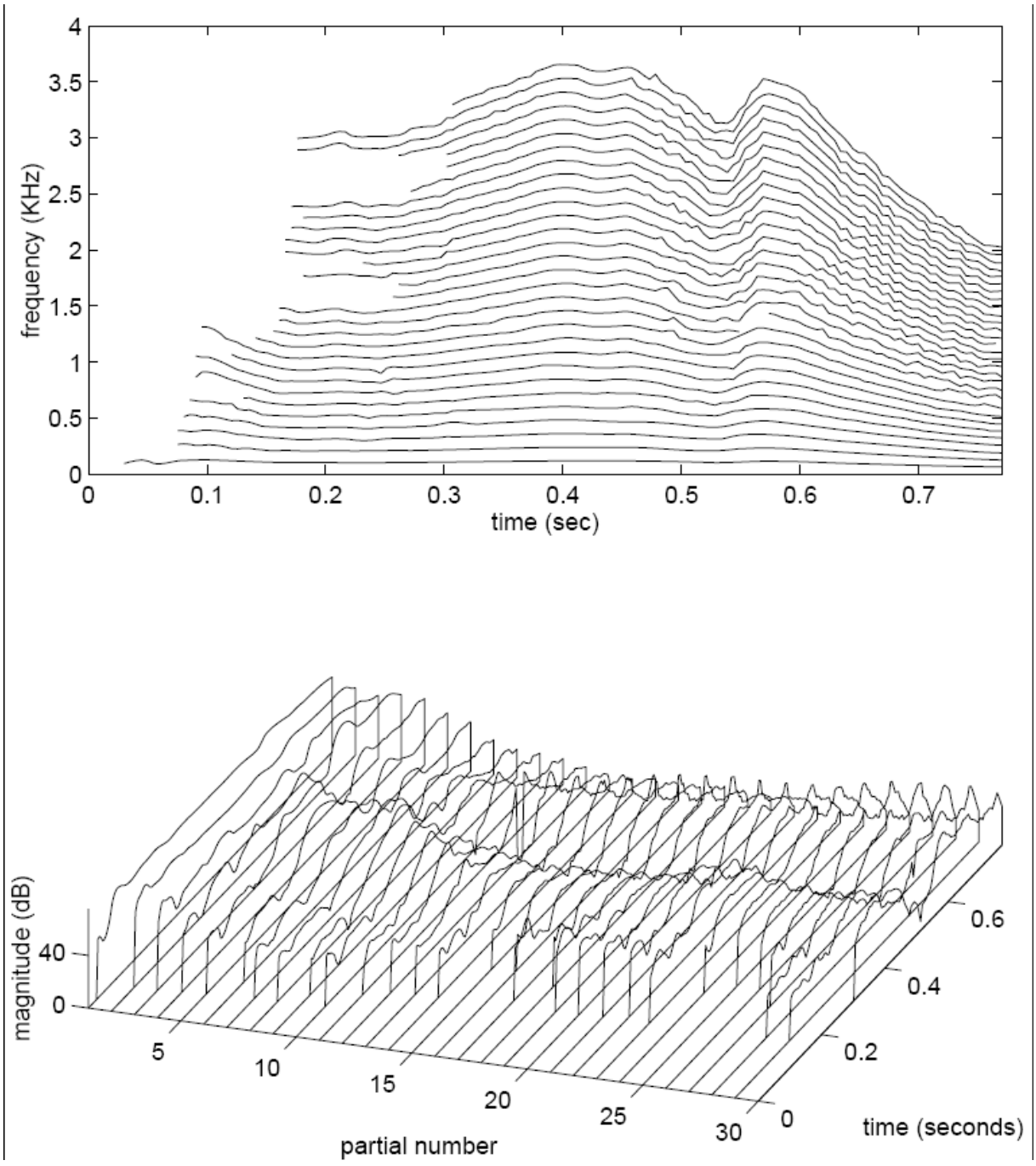

Figure 9: *Analysis representation of a vocal sound. a. Deterministic frequencies. b. Deterministic magnitudes.*

The resulting amplitude and frequency functions can be further processed to achieve a data reduction of the representation or to smooth the functions. A data reduction strategy is to perform a line-segment approximation on each function, thus reducing the number of breakpoints (Grey, 1975; Strawn, 1980).

For the purpose of easy manipulation of the representation it is useful to have equally spaced points along each function, and thus it may be better to keep one breakpoint per frame as returned by the analysis, unless data reduction is a priority. Another alternative for data reduction is to combine groups of similar functions into a single one, thus reducing the number of functions (Laughlin et alt., 1990).

The stochastic analysis returns an envelope for every frame. These envelopes can either be interpreted as a series of envelopes or frequency-shaping filters, one per frame, or as time-varying, equally spaced band-pass filters, each one centered at each breakpoint. It is convenient to normalize these envelopes by dividing them by their average magnitude so that we can control the spectral shape of the noise independently of its time-varying magnitude.

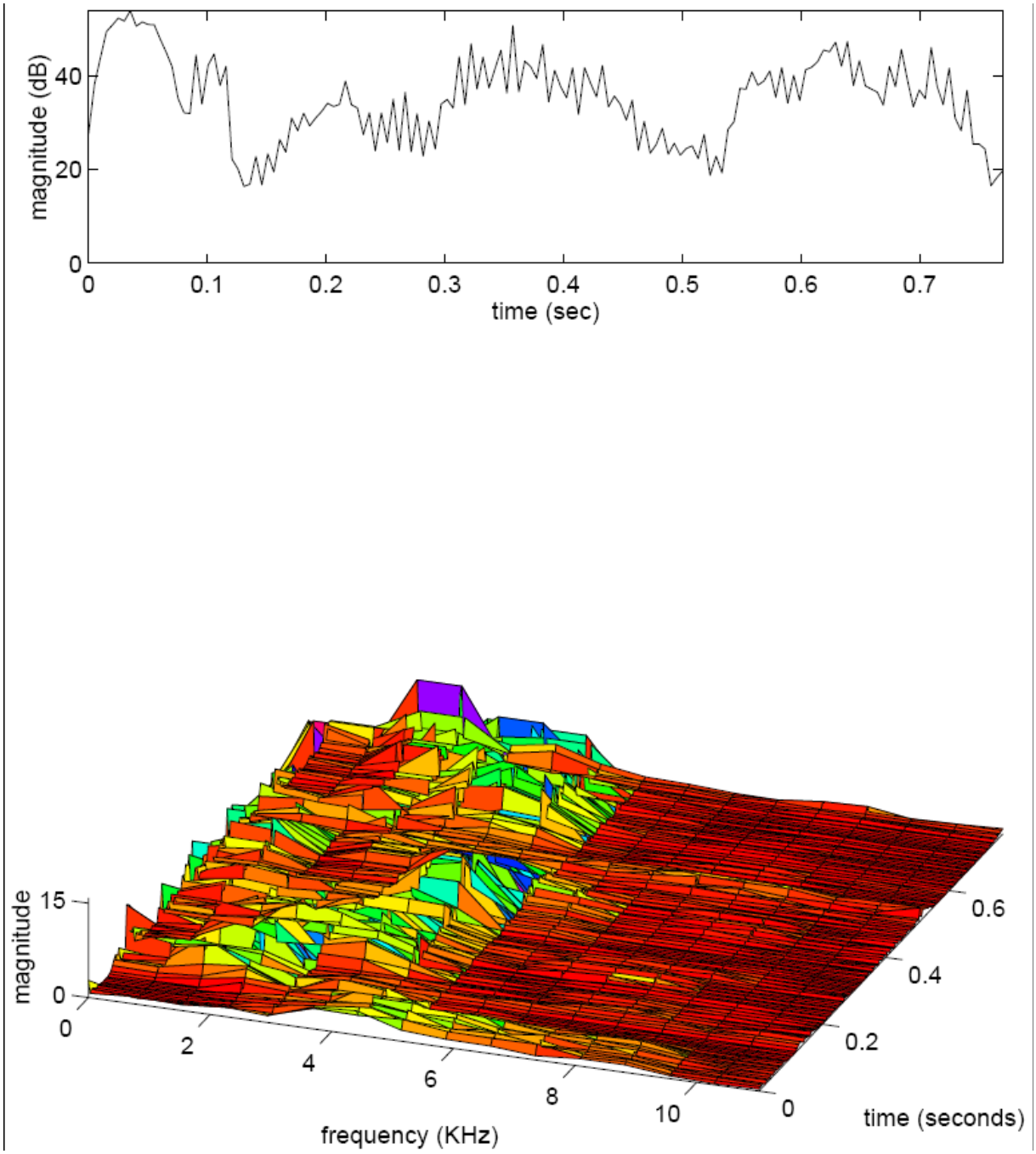

Figure 10: *Analysis representation of a vocal sound. a. Stochastic magnitude. b. Stochastic coefficients.*

12. Modifications of the Analysis Data

One of the main considerations in setting the analysis parameters is the potential for manipulating the resulting representation. For this goal, we would like to have a representation with a small number of partials and stochastic coefficients, and each of the functions (amplitudes and frequencies for the partials, gain and coefficients for the noise) should be as smooth as possible. In most cases there will be a compromise between perceptual identity from the original sound versus flexibility of the representation. Depending on the transformation desired this will be more or less critical. If we only want to stretch the sound a small percentage or transpose it a few Hz this will not be an issue. But when drastic changes are applied, details that were not heard in the straight resynthesis will become prominent and many of them will be perceived as distortion. For example, whenever the amplitude of a very soft partial is increased or its frequency transposed, since its measurements where not very accurate, the measurement errors that were not heard in the straight resynthesis, will probably come out.

The representation resulting from the analysis is very suitable for modification purposes, permitting a great number of sound transformations. For example, time-scale modifications are accomplished by resampling the analysis points in time and results in slowing down or speeding up the sound while maintaining pitch and formant structure. Due to the stochastic and deterministic separation, this representation is more successful in time-scale modifications than other spectral representations. With it, the noise part of the sound remains “noise” no matter how much the sound is stretched, which is not true with a sinusoidal representation.

In the deterministic representation each function pair, amplitude and frequency, accounts for a partial of the original sound. The manipulation of these functions is easy and musically intuitive. All kinds of frequency and magnitude transformations are possible. For example, the partials can be transposed in frequency, with different values for every partial and varying during the sound. It is also possible to decouple the sinusoidal frequencies from their amplitude, obtaining effects such as changing pitch while maintaining formant structure.

The stochastic representation is modified by changing the shape of each of the envelopes and the time-varying magnitude, or gain. Changing the envelope shape corresponds to a filtering of the stochastic signal. Their manipulation is much simpler and more intuitive than the manipulation of a set of all-pole filters, such as those resulting from an LPC analysis.

Interesting effects are accomplished by changing the relative amplitude of the two components, thus emphasizing one or the other at different moments in time. However we have to realize that the characterization of a single sound by two different representations, which are not completely independent, might cause problems. When different transformations are applied to each representation it is easy to create a sound in which the two components, deterministic and stochastic, do not fuse into a single entity. This may be desirable for some musical applications, but in general it is avoided, and requires some practical experimentation with the actual representations.

One of the most impressive transformations that can be done is by interpolating the data from two or more analysis files, creating the effect of “sound morphs” or “hybrids” (Serra, 1994b). This is most successful when the analysis of the different sounds to be hybridized where done as harmonic and all the functions are very smooth. By controlling how the interpolation process is done on the different parts of the representation and in time, a large number of new sounds will result. This type of sound processing has been traditionally called cross-synthesis, nevertheless a more appropriate term would be sound hybridization. With this spectral modeling method we can actually explore the timbre space created by a set of sounds and define paths to go from one sound to another.

The best analysis/synthesis computation is generally considered the one that results in the best perceptual identity with respect to the original sound. Once this is accomplished, transformations are performed on the corresponding representation. For musical applications, however, this may not be

always desirable. Very interesting effects result from purposely setting the analysis parameters "wrong". We may, for example, set the parameters such that the deterministic analysis only captures partials in a specific frequency range, leaving the rest to be considered stochastic. The result is a sound with a much stronger noise component.

Although this representation is powerful and many musically useful transformations are possible, we can still go further in the direction of a musically powerful representation based on analysis. The goal is to be able to control the perceptually relevant musical parameters of a sound, and the current representation is still far from that. Steps in this direction consist on extracting parameters, such as spectral shape, vibrato, overall amplitude and frequency evolution, from the current representation. These parameters can be extracted, modified and added back into the analysis data before the synthesis is done, without any degradation of the resulting sound. This process is easily implemented when the input sound is a single note, in which case the musical parametrization can be quite complete. Figure 11 shows a block diagram of the steps that should be done, but a discussion of the details involved in each of the steps is beyond this presentation.

13. Deterministic Synthesis

The deterministic component is generated with additive synthesis, similar to the sinusoidal synthesis that was part of the analysis, with the difference that now the phase trajectories are discarded. By not considering phase, this synthesis can either be done in the time domain or in the frequency domain. We will first present the more traditional time domain implementation.

The instantaneous amplitude $\hat{A}(m)$ of a particular partial is obtained by linear interpolation,

$$\hat{A}(m) = \hat{A}^{l-1} + \frac{(\hat{A}^{l} - \hat{A}^{l-1})}{S} m$$

where $m = 0,1,\ldots,S-1$ is the time sample in the l^{th} frame.

The instantaneous phase is taken to be the integral of the instantaneous frequency, where the instantaneous radian frequency $\hat{\omega}(m)$ is obtained by linear interpolation,

$$\hat{\omega}(m) = \hat{\omega}^{l-1} + \frac{(\hat{\omega}^{l} - \hat{\omega}^{l-1})}{S} m$$

and the instantaneous phase for the rth partial is

$$\hat{\theta}_r(m) = \hat{\theta}_r(l-1) + \hat{\omega}_r(m)$$

Finally, the synthesis equation becomes

$$d^{l}(m) = \sum_{r=1}^{R^{l}} \hat{A}_r^{l}(m) \cos\left[\hat{\theta}_r^{l}(m)\right]$$

where $\hat{A}(m)$ and $\hat{\theta}(m)$ are the calculated instantaneous amplitude and phase.

A very efficient implementation of additive synthesis, when the instantaneous phase is not preserved, is based on the inverse-FFT (Rodet and Depalle, 1992; Goodwin and Rodet, 1994). While this approach looses some of the flexibility of the traditional oscillator bank implementation, specially the instantaneous control of frequency and magnitude, the gain in speed is significant. This gain is based on the fact that a sinusoid in the frequency domain is a sinc-type function, the transform of the window used, and on these functions not all the samples carry the same weight. To generate a sinusoid

in the spectral domain it is sufficient to calculate the samples of the main lobe of the window transform, with the appropriate magnitude, frequency and phase values. We can then synthesize as many sinusoids as we want by adding these main lobes in the FFT buffer and performing an IFFT to obtain the resulting time-domain signal. By an overlap-add process we then get the time-varying characteristics of the sound.

The synthesis frame rate is completely independent of the analysis one. In the implementation using the IFFT we want to have a high frame rate, so that there is no need to interpolate the frequencies and magnitudes inside a frame. As in all short-time based processes we have the problem of having to make a compromise between time and frequency resolution. The window transform should have the fewest possible significant bins since this will be the number of points to generate per sinusoid. A good window choice is the Blackman-Harris 92dB because its main lobe includes most of the energy. However the problem is that such a window does not overlap perfectly to a constant in the time domain. A solution to this problem (Rodet and Depalle, 1992) is to undo the effect of the window by dividing by it in the time domain and applying a triangular window before performing the overlap-add process. This will give a good time-frequency compromise.

14. Stochastic Synthesis

The synthesis of the stochastic component can be understood as the generation of a noise signal that has the frequency and amplitude characteristics described by the spectral envelopes of the stochastic representation. The intuitive operation is to filter white noise with these frequency envelopes, that is, performing a time-varying filtering of white noise, which is generally implemented by the time-domain convolution of white noise with the impulse response of the filter. But in practice, the easiest and most flexible implementation is to generate the stochastic signal by an inverse-FFT of the spectral envelopes. As in the deterministic synthesis, we can then get the time-varying characteristics of the stochastic signal by an overlap-add process.

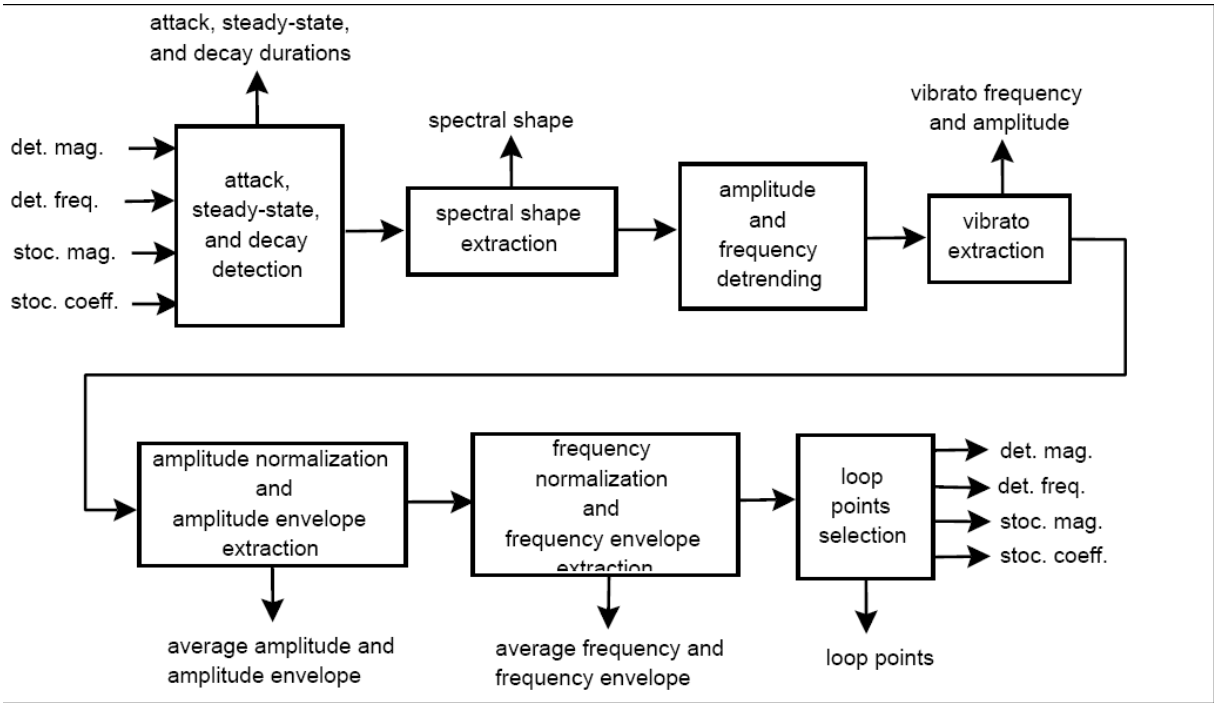

Figure 11: *Extraction of musical parameters from the analysis representation of a single note of an instrument.*

Before the inverse-FFT is performed, a complex spectrum (i.e., magnitude and phase spectra), has to be obtained from each frequency envelope. The magnitude spectrum is generated by linear

interpolating the spectral envelope to a curve of length $N/2$, where N is the FFT-size, and multiplying it by the average magnitude, gain, that was extracted in the analysis. There is no phase information in the stochastic representation, but since the phase spectrum of noise is a random signal, the phase spectrum can be created with a random number generator. To avoid a periodicity at the frame rate, new values are generated at every frame.

By using the IFFT method for both the deterministic and the stochastic synthesis it could be possible to use a single IFFT to generate both components. That is, adding the two spectra in the frequency domain and computing the IFFT once per frame. The problem to be solved is that in the noise spectrum there has not been any window applied and in the deterministic synthesis we have used a Blackman-Harris 92dB. Therefore we should apply this window in the noise spectrum before adding it to the deterministic spectrum. This would imply to convolve the transform of the Blackman-Harris 92dB by the noise spectrum, but with this operation there is no speed gain compared with performing the two IFFT separate and adding the deterministic and stochastic components in the time domain. This could be simplified by only convolving the most significant bins of the window transform.

15. Conclusions

Modeling sounds by their time-varying spectral characteristics is a well known powerful tool. But only a few of the possible approaches are musically useful. Our discussion has been focused by the musical goal of getting a general and intuitive sound representation based on analysis, from which we can manipulate musical parameters while maintaining the perceptual identity with the original sound when no transformations are made. The sinusoids plus noise, or deterministic plus stochastic, model gives us a powerful starting point in this direction with many musical applications and possibilities for further development. In this article we have presented the basic concepts involved in obtaining this representation and we have discussed ways to transform and synthesize sounds from the analyzed data.

An interesting direction in which to continue the work on spectral modeling using the sinusoids plus noise model is to go beyond the representation of single sounds and towards the modeling of entire timbre families. Such as all the sounds generated with an acoustic instrument. Thus, being able to represent their common characteristics by a common set of data and keeping separate only the perceptual differences. As part of this process it is also important to model the articulation between notes, so that we can generate expressive phrasing based also on analysis. The result is a powerful synthesis technique that has both the sound identity properties of sampling and the flexibility of FM.

References

Allen, J.B. 1977. “Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform.” *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(3):235--238.

Chafe, C. 1990. “Pulsed Noise in Self-sustained Oscillations of Musical Instruments.” *Proceedings of the IEEE Int. Conf on Acoustics, Speech, and Signal Processing*, Alburquerque, NM, 1990.

Cox, M. G. 1971. “An algorithm for approximating convex functions by means of first-degree splines.” *Computer Journal*, vol. 14, pp. 272--275.

Depalle, Ph., G. Garcia and X. Rodet. 1993. “Analysis of Sound for Additive Synthesis: Tracking of Partials Using Hidden Markov Models.” *Proceedings of the 1993 International Computer Music Conference*. San Francisco: Computer Music Association.

Doval, B., and X. Rodet. 1993. “Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs.” *Proceedings of the ICASSP ‘93*, 221--224.

Garcia G. 1992. “Analyse des Signaux Sonores en Termes de Partiels et de Bruit. Extraction Automatique des Trajets Frèquentiels par des Modèles de Markov Cachès.” *Mèmoire de DEA en Automatique et Traitement du Signal*, Orsay, 1992.

General Electric Co. 1977. “ADEC Subroutine Description.” Heavy Military Electronics Dept., Syracuse NY, 13201, June 1977.

Goodwin, M. and X. Rodet. 1994. “Efficient Fourier Synthesis of Nonstationary Sinusoids.” *Proceedings of the 1994 International Computer Music Conference*. San Francisco: Computer Music Association.

Grey, J.M. 1975. *An Exploration of Musical Timbre*. Ph.D. Dissertation, Stanford University.

Harris, F. J. 1978. “On the use of windows for harmonic analysis with the discrete Fourier transform,” Proceedings IEEE, vol. 66, pp. 51-83.

Hess, W. 1983. *Pitch Determination of Speech Signals*. New York: Springer-Verlag.

Laughlin, R., Truax, B. and B. Funt. 1990. “Synthesis of Acoustic Timbres using Principal Component Analysis.” *Proceedings of the 1990 International Computer Music Conference*. San Francisco: Computer Music Association.

Maher, R. C. and J. W. Beauchamp. 1994. “Fundamental Frequency Estimation of Musical Signals using a two-way Mismatch Procedure.” *Journal of the Acoustical Society of America* 95(4):2254--2263.

Makhoul, J. 1975. “Linear Prediction: A Tutorial Review.” *Proceedings of the IEEE* 63:561--580.

Markel, J.D. and A.H. Gray. 1976. *Linear Prediction of Speech*. New York: Springer-Verlag.

McAulay, R.J. and T.F. Quatieri. 1984. “Magnitude-only Reconstruction using a Sinusoidal Speech Model.” *Proceedings of the 1984 IEEE International Conference on Acoustics, Speech and Signal Processing*. New-York: IEEE Press.

McAulay, R.J. and T.F. Quatieri. 1986. “Speech Analysis/Synthesis based on a Sinusoidal Representation.” *IEEE Transactions on Acoustics, Speech and Signal Processing* 34(4):744--754.

Moorer, J. A. 1973. “The Hetrodyne Filter as a Tool for Analysis of Transient Waveforms.” Memo AIM-208, Stanford Artificial Intelligence Laboratory, Computer Science Dept., Stanford University.

Moorer, J. A. 1977. “Signal Processing Aspects of Computer Music—A Survey.” *Computer Music Journal* 1(1):4--37.

Moorer, J. A. 1978. “The Use of the Phase Vocoder in Computer Music Applications.” *Journal of the Audio Engineering Society* 26(1/2):42--45.

Phillips, G. M. 1968. “Algorithms for Piecewise Straight Line Approximation.” *Computer Journal* vol. 11, pp. 211--212.

Piszczalski, M. and B. A. Galler. 1979. “Predicting Musical Pitch from Component Frequency Ratios.” *Journal of the Acoustical Society of America* 66(3):710--720.

Portnoff, M.R. 1976. “Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform.” *IEEE Transactions on Acoustics, Speech and Signal Processing* 24(3):243--248.

Rodet, X. and P. Depalle. 1992. “Spectral Envelopes and Inverse FFT Synthesis.” *93rd Convention of the Audio Engineering Society*. San Francisco, October 1992.

Schumacher, R. T., and C. Chafe. 1990. “Detection of Aperiodicity in Nearly Periodic Signals.” *Proceedings of the IEEE Int. Conf on Acoustics, Speech, and Signal Processing*, Alburquerque, NM, 1990.

Sedgewick, R. 1988. *Algorithms*. Reading, Massachusetts: Addison-Wesley.

Serra, X. 1989. *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*. Ph.D. Dissertation, Stanford University.

Serra, X. and J. Smith. 1990. “Spectral Modeling Synthesis: A Sound Analysis/Synthesis System based on a Deterministic plus Stochastic Decomposition.’’ *Computer Music Journal* 14(4):12--24.

Serra, X. 1994a. “Residual Minimization in a Musical Signal Model based on a Deterministic plus Stochastic Decomposition.” *Journal of the Acoustical Society of America* 95(5-2):2958--2959.

Serra, X. 1994b. “Sound Hybridization Techniques based on a Deterministic plus Stochastic Decomposition Model.” *Proceedings of the 1994 International Computer Music Conference*. San Francisco: Computer Music Association.

Smith, J.O. and B. Friedlander. 1984. “High Resolution Spectrum Analysis Programs.” TM no. 5466-05, Systems Control Technology, Palo Alto CA, April 1984.

Smith, J.O. and X. Serra. 1987. “PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds based on a Sinusoidal Representation.” *Proceedings of the 1987 International Computer Music Conference*. San Francisco: Computer Music Association.

Strawn, J. 1980. “Approximation and Syntactic Analysis of Amplitude and Frequency Functions for Digital Sound Synthesis.” *Computer Music Journal* 4(3):3--24.

Terhardt, E., G. Stoll, and M. Seewann. 1982. “Algorithm for Extraction of Pitch and Pitch Salience from Complex Tonal Signals.” *Journal of the Acoustical Society of America* 71(3):679--688.

Wolcin J.J. 1980. “Maximum A Posteriori Line Extraction: A Computer Program.” USC TM no. 801042, March 20, 1980.

Wolcin J.J. 1980. “Maximum A Posteriori Estimation of Narrowband Signal Parameters.” NUSC TM no. 791115, June 21, 1979, *Journal of the Acoustical Society of America* 68(1):174--178.