Musical Sound Signal Analysis/Synthesis: Sinusoidal+Residual and Elementary Waveform Models

Xavier Rodet, IRCAM, 1 place Stravinsky, 75004, Paris, France. rod@ircam.fr

IRCAM, 1 place Stravinsky, 75004, Paris, France

Abstract

Several versions of Sinusoidal+Residual analysis/synthesis models have been developed for music applications. They have been very successful and are already found in commercial and experimental tools used by musicians as well as researchers. In this paper, we begin by presenting the principles of this now classical model. However, the standard version of the model suffers from limitations in various cases. Therefore, we discuss some improvements of the standard method designed in order to overcome these difficulties. We then present and compare other analysis techniques which use *Elementary Waveforms*, i.e. waveforms localized both on the frequency and the time axis. In particular, the High Resolution Matching Pursuit algorithm is proposed as a potentially successful new direction of research.

Keywords: Sound, analysis, synthesis sinusoidal, elementary-waveform.

1. Introduction

Some of the first attempts at sound synthesis were based on the method called *additive synthesis*, that is the summation of time-varying sinusoidal components [1]. Additive synthesis is accepted as perhaps the most powerful and flexible method. An important advantage of additive synthesis over digital sampling is that it allows the pitch and length of sounds to be varied independently [2]. The sample rate conversion technique [3] used in digital samplers achieves pitch alterations by changing the rate at which sounds are read from memory. But this results in changes in the duration of those sounds. Furthermore, because independent control of every component is available in additive synthesis, it is possible to implement models of perceptually significant features of sound such as inharmonicity and roughness. In commercial digital samplers, sound amplitude and pitch are readily controlled, but no fine control over the sound spectrum is possible for timbre manipulations, such as continuous changes in harmonicity.

Another important aspect of additive synthesis is the simplicity of the mapping of frequency and amplitude parameters into the human perceptual space. These parameters are meaningful and easily understood by musicians. This property is in contrast to, for instance, the parameter space of the frequency modulation synthesis method which maps awkwardly to the spectral domain through Bessel functions [4]. Recently, a new additive synthesis method based on spectral envelopes and Fast Fourier Transform has been developed [5]. Use of the inverse FFT reduces the computation cost by a factor in the order of 15 compared to oscillators. This technique renders possible the design of low cost real-time synthesizers allowing processing of recorded and live sounds, synthesis of instruments and synthesis of speech and the singing voice.

Consequently, it is not surprising that additive analysis and synthesis of musical signals have recently received a great deal of attention. Even though the, now classical, additive sinusoidal analysis is based on rather simple principles, it has been very successful. The first goal of this paper is to examine this method, the reasons of its success and its weaknesses. The second goal is to try to extrapolate from these conclusions some proposals for new research directions for musical signals analysis.

In section 2, we present additive sinusoidal analysis. The main drawbacks of the classical method are then explained in section 3 and some important improvements are briefly exposed. In section 4 we try to understand the advantages of additive analysis as well as its limitations. In section 5, under the term *Elementary Waveforms*, we present some new directions better suited for the analysis of musical sound signals.

2. The Additive Sinusoidal+Residual Model

Several similar sinusoidal models have been proposed for musical sound and speech signals [6, 7]. They often incorporate a non-sinusoidal residual part which can be waveform coded or modeled as a random signal.

2.1 Presentation of the standard sinusoidal model

In the standard model, the sinusoidal part s(t) is represented as the sum of I sine waves $c_i(t)$, called *sinusoidal partials*, with time-varying parameters:

$$s(t) = \sum_{i=1}^{I} c_i(t) \qquad (1)$$

With $a_i(t) \geq 0$ the amplitude and $\Phi_i(t)$ the phase of the sinusoidal partial:

$$c_i(t) = a_i(t) \cos(\Phi_i(t)) \qquad (2)$$

A first important assumption underlying (often implicitly) sinusoidal models is that $c_i(t)$ locally resembles a pure sinusoid. This means that $a_i(t)$ should be a slowly varying signal, i.e. a low pass signal with a bandwidth B_a and that $\Phi_i(t)$ is locally linear in t up to a small correction term $\varepsilon_i(t, t_0)$. If the *locality* around t_0 is defined as $t \in [t_0-\eta, t_0+\eta]$, then in this interval:

$$\Phi_i(t) = \Phi_i(t_0) + (t - t_0)\Omega_i(t_0) + \int_{t_0}^{t} \varepsilon_i(t; t_0)\, dt \qquad (3)$$

$$\frac{d\Phi_i(t)}{dt} = \Omega_i(t) = \Omega_i(t_0) + \varepsilon_i(t;t_0),$$
$$with \ \ \varepsilon_i(t_0;t_0) = 0 \qquad (4)$$

That $\varepsilon_i(t, t_0)$ is small can be stated more precisely by saying that $\Omega_i(t)$ is a slowly varying signal, i.e. a low pass signal or that $\cos(\Phi_i(t))$ is approximately a band limited signal for $t \in [t_0-\eta, t_0+\eta]$, with a bandwidth B_f around $\Omega_i(t_0)$. However, we will see in section 3 that the first assumption presented here above is only a rough approximation. In particular, the frequency behavior should not be formulated in terms of local variations of *relative value* but in terms of local variations of *relative slope*. Similarly, amplitude behavior should allow some fast variations such as are found in the attacks of percussive sounds.

Speech and musical sounds always have random components, often heard as a noise, superposed for instance on the harmonic part. A second assumption often underlying a sinusoidal model is that the number I of sinusoidal partials is limited. Therefore, a purely sinusoidal model s(t) with slowly varying parameters can hardly represent all of a real signal x(t) and needs to be completed with a non-sinusoidal residual part r(t):

$$r(t) = x(t) - s(t) \qquad (5)$$

Another argument in favor of a non-sinusoidal residual part r(t) is that the residual should be considered as a random signal in case of transformations such as time compression or expansion. In consequence, classical representations of random signals are better suited for the residual. It is common to represent only the short-time magnitude frequency content of the residual by a spectral envelope $G(t,\omega)$ [7]. If n(t) is a white gaussian noise and $G(t,\omega)$ is the Fourier Transform of a time-varying impulse response $g(\theta,t)$, then the model of the residual is:

$$r(t) = \int_{-\infty}^{+\infty} n(\theta)\, g(t-\theta, t)\, d\theta \qquad (6)$$

This filtering can be implemented in the time domain or in the frequency domain. If $R(\omega,t)$ and $N(\omega,t)$ are the Short-Time Fourier Transforms of r(t) and n(t) respectively, then:

$$R(t,\omega) = N(t,\omega)G(t,\omega) \qquad (7)$$

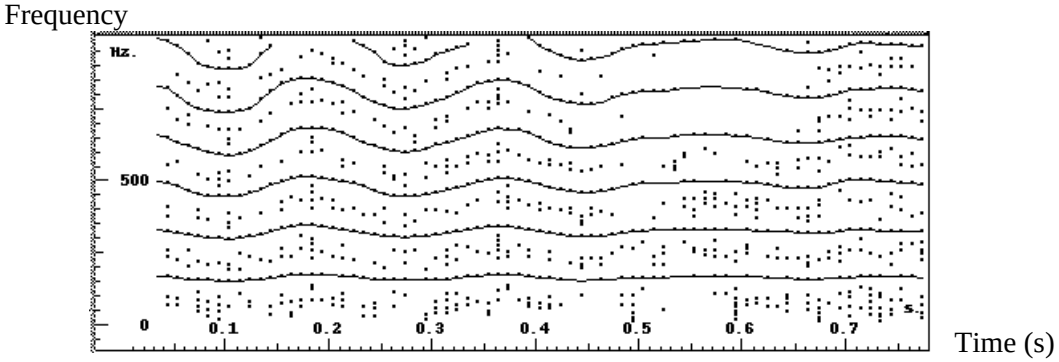

Fig. 1. Peaks found in successive analysis frames are grouped into tracks, i.e. sinusoidal partials.

2.2 Standard parameter estimation

Estimation of sinusoidal parameters is done in two steps. In the first step, a sliding window Short-Time Fourier Transform (STFT) is performed. Each peak of the magnitude spectrum is considered as the indication of a sinusoidal partial and its parameters are estimated. In the second step, peaks of successive analysis frames are grouped into tracks which are the sinusoidal partials under search (Figure 1).

Let x[n] be the analyzed discrete signal, h[n] the analysis window and $X(n,\omega_k)$ its STFT at time n and frequency ω_k:

$$X(n,\omega_k) = \sum_{m=-\infty}^{m=+\infty} x[m]h[n-m]e^{-jm\omega_k} \qquad (8)$$

Since we look at the STFT of a given frame around time n, let us drop index n and write $X(\omega_k)=X(n,\omega_k)$ A peak of the magnitude of the STFT obtained by Discrete Fourier Transform of size N, $|X(\omega_k)|$, is found at index λ when

$$|X(\omega_{\lambda-1})| < |X(\omega_\lambda)| > |X(\omega_{\lambda+1})| \qquad (9)$$

A peak at index λ supposedly indicates the presence of a sinusoidal partial at a near-by frequency Ω ($\Omega \approx \omega_\lambda$). The frequency $\Omega_i(t)$ and the amplitude $a_i(t)$ of this sinusoidal partial can be considered constant around time m, with values Ω and A respectively. Henceforth, the shape of $|X(\omega_k)|$ around index λ is the sampled shape of $|H(\omega-\Omega)|$, i.e. of the Fourier Transform of h[n] translated at the frequency Ω. To estimate Ω and A, since h[n] is usually a symmetric window, one can use the second order approximation of $H(\omega-\Omega)$ around Ω which is given by a quadratic function centered on Ω. This is why a quadratic approximation of $|X(\omega_k)|$, in the neighborhood of index λ, is performed [8]. Then the center and the maximum amplitude of this function are taken as an estimate of Ω and A respectively. Similarly, the local phase Φ of the sinusoidal partial is obtained as a weighted average of the phase of $X(\omega_k)$ in the neighborhood of index λ. The role of

the weighting factor is to increase the importance of high amplitude values $|X(\omega_k)|$ in order to diminish the effect of relatively low amplitude noise superposed on the sinusoidal partial.

The rather simple detection and estimation procedure described here above has weaknesses which are detailed in section 3. However, it should be underlined that it has known great success because it is fast and very robust. It can deal with the hundreds of sinusoidal partials encountered in musical signals and does not suffer from model-order determination and numerical or computational limitations which are common, for example, in parametric methods [9,10].

The second step of sinusoidal parameter estimation is the grouping of successive analysis frame peaks into tracks. This tracking is usually based on a heuristic approach [6, 7] which matches peaks in successive frames while allowing *deaths* and *births* of tracks. It simply grows trajectories, iteratively frame after frame, in the direction of increasing time: for each frame successively, the tracks which still appear in the current frame are possibly continued provided there is a convenient peak in the next frame according to an optimal frequency match. When needed, some tracks are terminated and new tracks arise. We will not detail this algorithm here since we prefer a well grounded statistical approach presented in section 3.4.

Estimation of the spectral envelope of the residual signal around time t, $G(\omega,t)$, can be done with any usual AR estimation technique [11]. Such a technique provides the P coefficients $\alpha_p(t)$ of an all-pole filter with magnitude transfer function $G(\omega,t)$. The coefficients $\alpha_p(t)$ are well suited for time-domain filtering of a noise n(t) at the synthesis stage. In practice, the $\alpha_p(t)$ are only estimated around successive times t_l, l=1, 2, 3,..., with a step t_{l+1}-t_l in the order of 5 to 20 milliseconds. Cepstral estimation can also be used on sliding window STFT $R(\omega,t)$ of r(t). Cepstral estimation provides cepstral coefficients which are well suited for frequency domain filtering of a noise n(t) at the synthesis stage. When estimating the spectral envelope $G(\omega,t)$, a nonlinear frequency scale, such as the *Mel* or the *Bark* scale, is appealing since is reflects some properties of human perception. Some authors [7, 12] have proposed to simply represent the magnitude short-time spectrum $|R(\omega,t)|$ by its mean value in channels distributed on such a nonlinear scale. This representation also is well suited for frequency domain filtering at the synthesis stage, since it requires only a product of the STFT of the noise n(t) by $|R(\omega,t)|$.

3. Improvements of the standard model

Many improvements have been proposed beyond the standard model presented here above. In the following, we will only present some of the most important ones. Others can be found, for example, in [13, 14, 15, 16, 17, 18, 19].

3.1 Peak detection and estimation

As mentioned in section 2.2, the presence in x(t) of a time-domain sinusoidal partial $c_i(t)$, around time t, is looked for by examining the STFT $X(m,\omega_k)$. Let us again drop index n and write $X(\omega_k)=X(n,\omega_k)$. If $H(\omega)$ is the Fourier Transform of the analysis window h[n], this problem can be viewed as the detection of the presence of a scaled and sampled version of the *signal* $H(\omega)$ in the *signal* $X(\omega_k)$. Therefore, it is natural to look for the maxima of the cross-correlation function Γ of H and X. If W is the bandwidth of the low-pass signal h[n], then $H(\omega)$ can be considered as negligible outside the interval [-W,W] and the computation of $\Gamma(\omega)$ is simplified:

$$\Gamma(\omega) = \sum_{k,\,|\omega-\omega_k|<W} H(\omega-\omega_k)X(\omega_k) \qquad (10)$$

Each maximum $|\Gamma(\Omega)|$ indicates a sinusoidal partial candidate at frequency Ω. An estimate of the amplitude a and of the phase Φ of the partial can then be derived. Defining at Ω a norm for H and X by:

$$|H|_\Omega^2 = \sum_{k,\,|\Omega-\omega_k|<W} |H(\Omega-\omega_k)|^2 \qquad (11)$$

$$|X|_\Omega^2 = \sum_{k,\,|\Omega-\omega_k|<W} |X(\omega_k)|^2 \qquad (12)$$

we obtain:

$$a = \frac{|\Gamma(\Omega)|}{|H|_\Omega^2} \quad , \quad \Phi = \mathrm{Arg}\{\Gamma(\Omega)\} \qquad (13)$$

Note that this computation also provides a measure v_Ω of the similarity between the observed peak and the peak which would result from a pure steady sinusoid (in which case v_Ω=1):

$$v_\Omega = \frac{|\Gamma(\Omega)|}{|H|_\Omega |X|_\Omega} \qquad (14)$$

A sinusoidal similarity measure (SLM) v_Ω <1 indicates the presence of noise or of other sinusoidal components in the neighborhood of Ω, or that the detected sinusoidal partial has fast varying parameters. The third case, fast variation, is examined in section 3.3 and 3.5. The second case, close-frequency partials, has been looked at by [20]. If one can disregard the two last cases, then the SLM v_Ω is similar to the so called *voicing* index of speech signals. But the SLM $v_\Omega(n)$ is here a function of two variables, the analysis time n and the frequency Ω. Errors on the usual speech voicing index at time n have serious consequences for speech coding or synthesis. On the other hand, at time n, errors on the SLM $v_{\Omega_0}(n)$ for some Ω_0 are of little consequence and are statistically compensated for by the v_Ω for all the other Ω values. This SLM

function $v_\Omega(n)$ has been very successfully used for speech coding or synthesis [21, 22] as well as for musical sound analysis and synthesis [23, 24, 25].

3.2 Parameter slope estimation

Several attempts have been made to extract information about the slope of sinusoidal parameters in a short time frame in order to overcome the limitation to mean values. Obtaining this information is not easy and usually suffers from uncertainty and errors. However, this information would be very useful when statistically combined with mean values, for example in the statistical approach presented in section 3.4.

In [26] a time-domain method is developed to estimate complex amplitudes (i. e. real amplitudes and phase deviations), when mean frequency values are known. Note that phase deviation is then equivalent to frequency variation. The model of the complex amplitude of the i^{th} sinusoidal partial is a low order (e.g. 3) polynomial of time n:

$$A_i[n] = \sum_{m=0}^{q} b_{i,m} n^m \qquad (15)$$

The model $\xi[n]$ of the signal is a sum of I sinusoidal partials. With $z_i=\exp(j\Omega_i)$:

$$\xi[n] = \sum_{i=1}^{I} \sum_{m=0}^{q} b_{i,m} n^m z_i^n \qquad (16)$$

Let $\underline{\xi}$ be the column vector $\underline{\xi}=(\xi[1],\xi[2],...,\xi[N])^t$ where N is the analysis frame length, $\underline{b}_i=(b_{i,0},b_{i,1},...b_{i,q})^t$, $B=(\underline{b}_1|\underline{b}_2|...|\underline{b}_I)^t$ and $\Psi=[\psi_1|\psi_2|...|\psi_I]$ where

$$\psi_i = \begin{bmatrix} 1 & 0 & ... & 0 \\ z_i & z_i & ... & z_i \\ z_i^2 & 2z_i^2 & ... & 2^q z_i^2 \\ ... & ... & ... & ... \\ z_i^N & Nz_i^N & ... & N^q z_i^N \end{bmatrix} \qquad (17)$$

Then we can write equation (16):

$$\underline{\xi} = \Psi B \qquad (18)$$

Using a least squares approach, minimization of

$$\sum_{n=0}^{N} (x[n] - \xi[n])^2 \qquad (19)$$

leads to:

$$B=[\Psi^h\Psi]^{-1}\Psi^h\underline{\xi} \qquad (20)$$

The method has been applied successfully by [26] to find the amplitude and phase variations of sinusoidal partials of percussive sound signals, after the mean frequencies were found by Prony's method. This simultaneous determination of amplitude and frequency variations would be a useful complement of the classical sinusoidal analysis where mean frequency values are estimated on spectral peaks.

Naturally, sinusoidal partial parameter evolution also appears in the STFT, leading to a distortion of peaks from the pure-sinusoid peak shape as mentioned in section 3.1. The distortion caused by linear frequency modulation (LFM) and exponential amplitude modulation (EAM), in a given STFT $X(\omega_k)=X(n,\omega_k)$, is examined in [27]. In the case of a pure sinusoid, the phase of the FFT bins, $\mathrm{Arg}\{X(\omega_k)\}$, in the vicinity of the corresponding peak, is constant. In the case of LFM or EAM modulation, this phase shows a variation $\sigma_f(\varpi_r)$ or $\sigma_a(\varpi_r)$ which is a function of the frequency ϖ_r relative to the peak center. Experimental measures show that LFM, up to 16 bins of modulation per frame, causes a variation $\sigma_f(\varpi_r)$ which is increasing with ϖ_r for small $|\varpi_r|$. Similarly, EAM, up to 6 dB of modulation per frame, causes a variation $\sigma_a(\varpi_r)$ which is proportional to ϖ_r for small $|\varpi_r|$. Therefore, when only one type of modulation appears and is not too large, it can be estimated from the phase spectrum. Furthermore, for small $|\varpi_r|$, the phase variations due to LFM and EAM are roughly additive. But, since $\sigma_f(\varpi_r)$ has even symmetry and $\sigma_a(\varpi_r)$ has odd symmetry, their cumulative effect produces a global variation $\sigma_f(\varpi_r)+\sigma_a(\varpi_r)$ with distinctive shapes according to the slope signs of $\sigma_f(\varpi_r)$ and $\sigma_a(\varpi_r)$. Therefore, simultaneous LFM and EAM can be estimated from the phase spectrum. The method has been successfully applied on simulated audio signals. However, for real musical signals, where additive interference from the more prominent peaks affects the phase profile across less prominent neighboring peaks, the described implementation is not resilient enough [27]. But it seems that the extracted slope information could be profitable at the partial tracking stage.

3.3 Reduced window length: parametric model of the STFT

In the standard sinusoidal analysis (section 2.2), sinusoidal partial candidates are found as peaks of the magnitude STFT $|X(m,\omega_k)|$ of the signal x[n], as given by equation (8). In this equation, h[n] is a classical window, such as the Hamming window. The computation is done with a Discrete Fourier Transform [28] and the ω_k, k=1, 2, ... K, are regularly spaced frequencies. Since we consider the STFT for a given n, let us simply write $X(\omega_k)=X(n,\omega_k)$. When two sinusoidal partials have nearby frequencies separated by Δf Hertz, in order that $|X(\omega_k)|$ exhibit two peaks, it is necessary that the window length L is large enough:

$$L > q/\Delta f \qquad (21)$$

where L is in seconds, q depends on the window main lobe width and is in the order of 3.5. Let us take, for example, a harmonic sound with a fundamental frequency of 110 Hz (which is heard as the note A2). Its sinusoidal partials are 110 Hz apart and L should be greater than 3.5/110, i.e. 32 ms. In polyphonic sound signals, partials can be

even much closer and the length L should be accordingly larger. Note that the minimum frequency distance between sinusoidal partials is often unknown. A large window is a great inconvenience when sinusoidal parameters vary substantially over a time segment L. In particular, fast transitions such as consonants or percussive attacks are smoothed. The problem is even worse for sinusoidal partials the frequency of which varies substantially. As an example, if the fundamental frequency varies by δ Hertz on the window length L, then the i^{th} partial varies by $i\delta$. When i is large, this important frequency modulation induces such a spreading of the spectrum of the i^{th} partial that the corresponding peak in $|X(\omega_k)|$ is smeared and its detection fails. Another weakness of the standard sinusoidal analysis is that it does not take into account the influence of sinusoidal partials close in frequency which slightly alter the estimation of frequency and amplitude of a given sinusoidal partial.

The method presented in [29, 30] remedies these difficulties by using a parametric model of the STFT of the signal and by allowing the window length to be as short as $2/\Delta f$. Let us still write $X(\omega_k)=X(n,\omega_k)$ for the STFT of the signal at time n. The model ξ of the signal is based on the assumption of a sum of sinusoidal partials with amplitude and frequency remaining constant over the window duration L, $a_i=a_i[n]$ and $\Omega_i=\Omega_i[n]$ and a local phase ϕ_i:

$$\xi[n] = \sum_{i=1}^{i=I} a_i \cos(\Omega_i n + \phi_i) \qquad (22)$$

Therefore, the model of the Fourier Transform is:

$$\Xi(\omega) = \sum_{i=1}^{i=I} \frac{a_i}{2} \left(\exp(j\phi_i) H(\omega - \Omega_i) + \exp(-j\phi_i) H(\omega + \Omega_i) \right) \qquad (23)$$

where $H(\omega)$ is the Fourier Transform of the analysis window h(n). The method consists of identifying the parameters for which the model best fits the observation $X(\omega_k)$ according to a least squares criterion. The identification is realized by an iterative algorithm which alternatively improves the estimates of amplitudes using the previous estimates of frequencies and improves the estimates of frequencies using the previous estimates of amplitudes [30]. Initial estimates are obtained from the standard sinusoidal analysis using a relatively long window with a small bandwidth (e. g. a rectangular window). At each iteration, the amplitude optimization is a simple linear problem. Since the frequency estimation problem is nonlinear, a simple linear optimization is performed at each iteration: the equation is linearized around the vector $\{\Omega_i-\omega_k,\ k=1, 2, ... K\}$ in order to lead to a linear problem. One difficulty is that the algorithm can then converge, not to the main-lobe maximum, but to a secondary maximum corresponding to a sidelobe. In order to avoid that, Depalle and Hélie [30] have designed and used a new family of analysis windows without sidelobes. Other improvements of the algorithm are given in the above two references. This algorithm is shown to converge rapidly to the correct parameter values even when it is initialized with rather poor approximations. It also remains efficient at low signal-to-noise ratios (e.g. 10 dB).

3.4 Statistical approach of partial tracking

During the second step of the analysis (see section 2.2), peaks found in successive analysis frames have to be grouped into partial tracks (Fig. 1). Some of the peaks do belong to partial tracks while others are spurious peaks (due to non-sinusoidal components for instance). The standard approach (section 2.2) works well enough for some categories of sounds (harmonic, voiced, and slow time-varying sounds), but fails in presence of multiple harmonic structures, inharmonic partials, crossing partials, voiced/unvoiced transitions, and large frequency variations. Furthermore, this procedure takes into account frequency proximity only, neglecting other sinusoidal parameters, i.e. amplitude, phase and sinusoidal similarity measure SLM (section 3.1).

However, the procedure described in [31, 32] copes with these problems by globally optimizing the set of tracks. The peak tracking problem is formulated in terms of a Hidden Markov Model (HMM) [33]. The optimization is performed in a given time interval T according to a statistical criterion of slope continuity for all the sinusoidal parameters. Therefore, the optimal set of trajectories is found as the highest probability state sequence, by means of the Viterbi algorithm [33]. Note that the use of parameter slopes rather than parameter values, while being consistent with the first assumption of a sinusoidal model (see section 2.1), enables one to track time-varying partials as easily as constant ones, and solves the problem of detecting crossing trajectories.

We shall only indicate here a few features of the algorithm. Since the number of tracks can be in the hundreds, the biggest problem is to reduce computational complexity. Therefore, the Viterbi algorithm is applied on a window length of T frames, which slides frame by frame, and some constraints on index combinations, maximum number of tracks, etc., are added. Furthermore, the algorithm considers only the possible combinations of peaks between successive frames. Sinusoidal parameters are used to compute state transition probabilities [31,32] which favor slope continuity and disfavor spurious peaks. At time m, there are h_m peaks $P_m[i]$, $1 < i \le h_m$. Each track is labeled by an index greater than zero. The problem is to associate an index $D_m[i]$, $1 < i \le h_m$, to each peak $P_m[i]$. When a peak $P_m[i]$ is considered as a spurious one, it is associated with a null index $D_m[i] = 0$. A state S_m is defined by an ordered pair of vectors (D_{m-1}, D_m) and the

observation is defined by an ordered pair of integers (h_{m-1}, h_m). The optimal sequence of states $S_m=(D_{m-1}, D_m)$ is found by means of the Viterbi algorithm, which maximizes the joint probability of state and observation sequences leading to a globally optimal solution. Then the tracks are defined by the sequence of vectors D_m from the state sequence.

This algorithm has been implemented at IRCAM by G. Garcia. Other computational cost reductions have been applied. In particular, the Viterbi algorithm has been replaced by a more efficient one taking advantage of the factorised structure of transition probabilities and eliminating computational redundancy. IRCAM's HMM tracking algorithm has been successfully used for sound analysis, processing and synthesis for research and for musical creation. As an example, it is possible to analyze polyphonic music comprising simultaneously several instruments, chords and percussion sounds.

3.5 Fast transients

As mentioned in section 2.1, a sinusoidal model is based on an assumption of bounded local variations of *relative slope* of sinusoidal parameters. However, it should allow some fast amplitude variations such as found in the attacks of percussive sounds. Standard sinusoidal analysis based on STFT requires a rather long signal window (typically 30 ms) which smears such fast transients. To overcome this difficulty, Masri [27] detects fast transient instants and takes them into account when positioning analysis windows. The aim is to guarantee that spectra on either side of the fast transient (which are essentially different) are never captured in the same window. Furthermore, the method disallows any peak linking or spectral interpolation (for the residual part) across the event boundary. During the synthesis stage, a fast crossfade is performed at the event boundary to retain the abruptness of the original sound. In particular, at the synthesis stage, the crossfade length is kept constant even though the sound is time-stretched. The method has been successfully applied to mixtures of continuous and percussive sounds and preserves perceptual properties of both types of sounds.

4. Discussion of the sinusoidal model

The sinusoidal+residual model has been very successful for musical signal analysis, processing and synthesis. Several commercial and experimental systems are currently used by musicians [7, 23, 34, 35]. Let us present some of the reasons for this success. A first one is the nature of musical sound signals. They often are composed of damped sinusoids of quasi-steady frequency (percussive sounds) or have relatively long and steady harmonic sustained parts. It is clear that a sinusoidal model is well adapted to represent a steady harmonic sound. It is probable that the nature of human perception of musical sound signals constitutes another reason. Human perception is extremely precise in steady sustained parts where sinusoidal analysis is the most efficient, and apparently less precise in fast transients where sinusoidal analysis is less efficient. It seems also that localisation, or better, *redundancy* in time and localisation in frequency, of sinusoidal analysis largely contributes to its quality. In particular, each peak of the STFT is modeled independently and hence precisely when the peak is due to a quasi-sinusoid since corresponding spectral peaks are easy to measure accurately. Not only estimation errors are small but they tend to be statistically distributed in frequency and time, amounting only to a nearly-inaudible level of distortion.

However, there are sound signals for which sinusoidal analysis does not seem so well adapted, typically signals where excitation departs from periodicity. Curiously enough, sinusoidal analysis is also used for speech signals even though they often fall in the last category. The classical model of glottal speech production [36] consists of short pulses filtered through the vocal tract. Firstly, variations of vocal tract transfer function can be appreciable at the time scale of three glottal periods. Secondly, time locations of pulses can be far from periodic. We already noted in section 3.3 that high rank partials cause difficulties even for small fundamental variations. Figure 2 shows another case, i.e. a speech waveform resulting from irregular pulses, as often occurs at the end of a sentence (it is sometimes called *vocal fry*). Sinusoids make sense when, in a given frequency band, a waveform repeats periodically at least three times. However, signals like the one in figure 2 suggest the use of other methods based on waveforms better localized in time when needed, sometimes called *Elementary Waveforms* (see section 5).

Finally, the standard noise source and filter model represents non-sinusoidal and random components in a very unsatisfactory way [12]. Moreover, the fact that two totally different analysis techniques are needed is a weakness which leads to difficulties since the separation of an unknown number of time-varying sinusoidal components and random components is not based on any solid grounding.

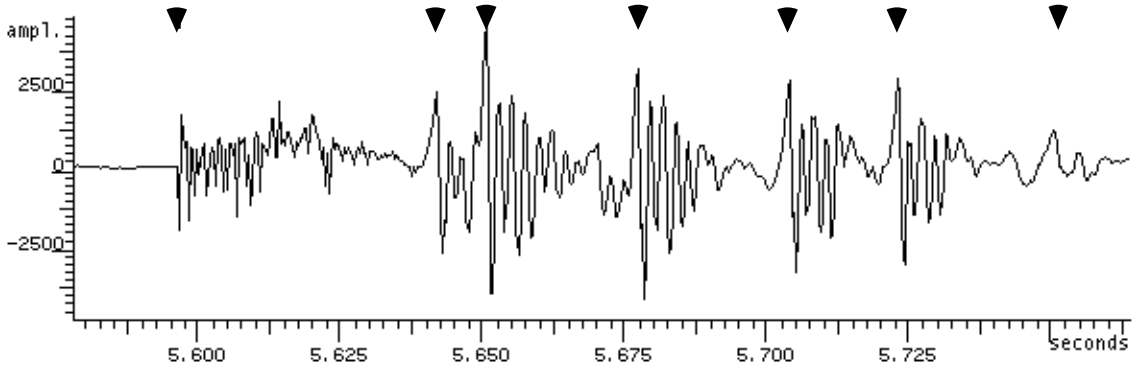

Figure 2. A speech waveform resulting from irregular pulses, as often occurs at the end of a sentence.

5. Elementary Waveform analysis

5.1 Presentation

Under the name Elementary Waveforms (EW) we group a certain number of methods using waveforms well localized in frequency and time which are overlapped and added to construct a signal. As a first example, Formant Waveforms (called FOF from the French *Formes d'Onde Formantiques*) [37] have been used for speech and musical signal synthesis. A FOF analysis method has been proposed in [38]. Localisation in frequency is adapted according to formant regions of the signal under analysis. Pitch Synchronous Overlap Add (PSOLA) is one of the most successful method for speech synthesis [39]. In PSOLA analysis, segments extend over two pitch periods exactly. However, these segmented waveforms are not localized in frequency and, usually, no further analysis is done.

In [40], a narrow band-pass filter bank is used to ensure localisation in frequency. The signal at the output of each filter is segmented at successive minima of its amplitude envelope. Each segment is considered as an EW. The method has been used for speech analysis and synthesis.

Note that sinusoidal analysis starts with a STFT at arbitrary regularly spaced times, then looks for specific peak patterns in the STFT. On the other hand, some EW analyses start with arbitrary regularly spaced band-pass filtering, then look for specific patterns in filter outputs. Matching Pursuit, presented here below, does not favor time or frequency but, at each step, looks for the position of an EW in time and in frequency, as well as for a scale, which are optimal according to the properties of the signal under analysis.

5.2 Matching Pursuit (MP)

Usual time-frequency and time-scale analysis methods, such as STFT [28] or Wavelets [41] perform a decomposition of signals on a given fixed basis. Therefore, the analysis spreads some important structures of musical signals on many basis vectors. Regrouping the results of the decomposition of these structures, for recognition or processing, becomes difficult. For instance, musical signals include fast transients which are well represented by short waveforms and sustained parts which are more efficiently represented by long waveforms with short frequency support. We have seen in sections 2 and 3 that the usual analysis methods lead to difficulties with transients. New adaptive approaches have been developed in order to choose the decomposition vectors depending upon signal properties (e.g. [42]), but they still use an orthogonal basis. Therefore, some important structures still tend to be spread on many vectors.

Pursuit algorithms, such as Matching Pursuit (MP) [43] or Basis Pursuit [44] have been designed to overcome these difficulties. The decomposition vectors are selected among a *redundant* family, called a *dictionary*, of EWs well localized in frequency and time. In MP, the EWs which constitute the dictionary have three parameters, a scale factor s, a time position u and a modulation frequency ω (note that, unlike in Wavelets, scale and frequency *are* independent). With $\gamma=(s,u,\omega)$:

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-u}{s}\right) e^{j\omega t} \quad , \text{ where} \qquad (24)$$

$$g(t) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{t^2}{2\sigma^2}\right) \qquad (25)$$

is a Gaussian function with unit norm. A MP is an iterative algorithm which decomposes a signal x over dictionary vectors as follows. Let us write $R^n x$ a *residue* at step n, starting from $R^0 x = x$. At each step n, the vector selected in the dictionary is the one which matches $R^n x$ best, i.e. such that:

$$\left|C(R^n x, g_{\gamma_n})\right| = \sup_{\gamma \in Y} \left|C(R^n x, g_\gamma)\right| \qquad (26)$$

where Y is the dictionary of all possible values for γ and $C(x,g_\gamma)$ is a correlation function which measures the similarity between x and g_γ. The residue for the next step is then:

$$R^{n+1} x = R^n x - C(R^n x, g_{\gamma_n}) g_{\gamma_n} \qquad (27)$$

Finally, the signal is represented as:

$$x = \sum_{n=0}^{+\infty} C(R^n x, g_{\gamma_n}) g_{\gamma_n} \qquad (28)$$

In [43], the correlation function C is the inner product $C(x,g_\gamma)=\langle x,g_\gamma\rangle$. This decomposition is relatively fast to compute, gives a good resynthesis with a limited number of vectors and exhibits the different structures of the signal at different scales [45, 46]. However, these references show that the chosen correlation function C leads to inadequate representations of some structures, such as a sinusoid the envelope of which varies rapidly. Therefore, a High Resolution MP (HRMP) algorithm is introduced. It uses a different correlation function which allows the pursuit to emphasize local fit over global fit at each step. HRMP performs a better time-resolution than MP so that, in audio applications, attack-patterns recognition or processing is improved.

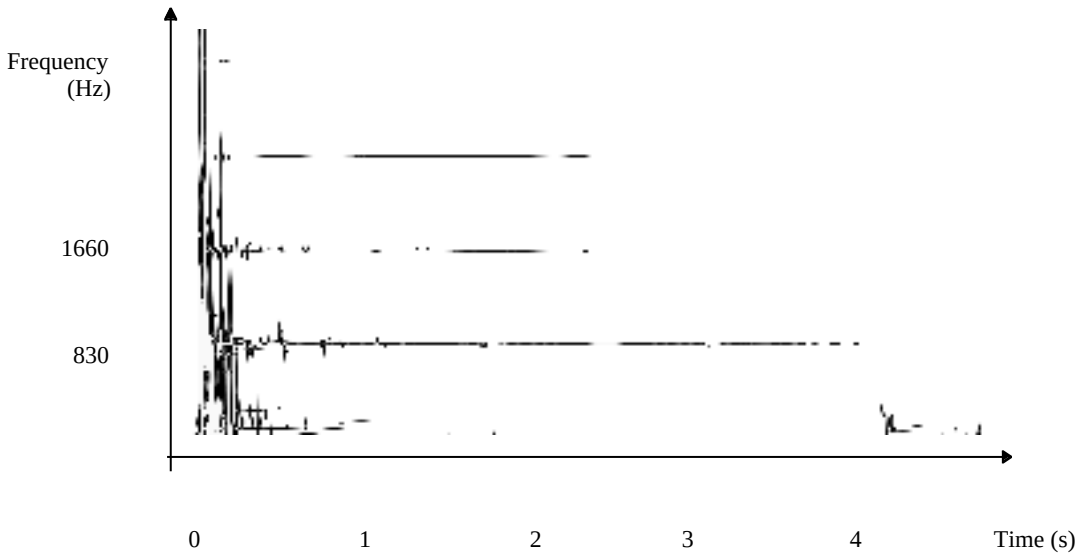

Fig. 3. Time-frequency distribution of a G5 sharp piano note, obtained with HRMP

The time-frequency distribution of a G5 sharp piano note, obtained with HRMP, is displayed in figure 3. One can easily distinguish long horizontal lines due to large-scale vectors well defined in frequency around 830 Hz, 1660 Hz, etc.. They correspond to the damped sinusoidal quasi-harmonic modes of the string. However, vertical lines corresponding to small-scale transient structures are visible at the attack and at the release of the damper of the piano. This example shows how HMRP provides a time-frequency representation adapted to the specificities of sound signals. The elements of this representation are easily related to perceptually important structures such as fast transients, or sustained sinusoidal partials.

6. Conclusion

The principles of sinusoidal+residual analysis have been exposed in order to better explain the reasons for its success and of its weaknesses. Some important improvements to the standard analysis technique have also been presented. Then, we have critically examined the overall method. Whereas sinusoidal representation is quite adequate for some musical signals, it seems inadequate for others. Localisation in time and frequency and adaptation to signal properties is found to be an advantage of the method while the grouping of local features in larger scale sinusoidal partials, on a statistical basis, is shown to be very efficient. However, when the sinusoidal model is inadequate, other methods may give better results. In particular, more emphasis on local fit to the signal is advantageous for adaptive algorithms such as HRMP as well as for the sinusoidal model. But many aspects still need continued development. For instance, repeated pulses with a changing period cannot be easily modeled in a sinusoidal method. A good solution in HRMP analysis for this problem has also not yet been found. Similarly, random components require a totally different analysis technique in the sinusoidal+residual case. In HRMP, these components lead to a large number of short duration Elementary Waveforms which cannot be distinguished a priori from sinusoidal components and still need to be grouped in order that processing or recognition can be performed.

References

1. Risset JC, Mathews MV. Analysis of musical-instrument tones. Physics Today, 22(2):23-30, Feb. 1969.

2. Quatieri ThF, McAulay RJ. Shape Invariant Time-Scale and Pitch Modification of Speech. IEEE Trans. on Signal Processing, Vol. 40 No. 3, March 1992.

3. Smith JO, Gossett P. A Flexible Sampling-Rate Conversion Method. In: Proc. IEEE ICASSP, vol. 2 , San Diego, March 1984. pp. 19.4.1-19.4.2
4. Corrington MS. Variation of Bandwidth with Modulation Index in Frequency Modulation. In: Selected Papers on Frequency Modulation. Edited by Klapper. Dover, 1970
5. Rodet X, Depalle Ph. A new additive synthesis method using inverse Fourier transform and spectral envelopes. In: Proc. of ICMC, San Jose, California. Oct. 1992. pp. 410-411
6. McAulay RJ, Quatieri ThF. Speech analysis/synthesis based on a sinusoidal representation. In: IEEE Trans. on Acoust., Speech and Signal Proc., vol ASSP-34. 1986. pp. 744-754
7. Serra X. A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. Philosophy Dissertation, Stanford University, Oct. 1989
8. McIntyre CM, Dermott DA. A New Fine-Frequency Estimation Algorithm Based on Parabolic Regression. In: IEEE-ICASSP 1992. pp. 541-544
9. Laroche J, Rodet X. A new Analysis/Synthesis system of musical signals using Prony's method. In: Proc. ICMC, Ohio, Nov. 89.
10. Laroche J. The use of the Matrix-Pencil method for the spectrum analysis of musical signals, J. Acoust. Soc. America, Vol. 94 No. 4., Oct. 1993.
11. Kay SM. Modern Spectral Estimation: Theory and Application. Prentice Hall, 1988.
12. Goodwin M. Residual modeling in music analysis-synthesis. In: Proc IEEE-ICASSP, Atlanta, GA, May 1996. pp. 1005-1008
13. George EB, Smith JT. Analysis-by-Synthesis/Overlap-Add Sinusoidal Modeling Applied to the Analysis and Synthesis of Musical Tones. In: J. Audio. Eng. Soc., Vol. 40, No. 6, June 1992
14. Laroche J, Stylianou Y, Moulines E. HNS: Speech modification based on a harmonic+noise model. In: Proc. IEEE-ICASSP-93, Apr. 1993. Vol. II. pp. 550-553
15. Ravera B, d'Alessandro C. Double Frequency and Time-Frequency Analyses of Modulated Speech Noises. In: Signal Processing VII: Theories et Applications. Edited by M. Holt, C. Cowan, P. Grant. W. Sandham. 1994
16. McAulay RJ, Quatieri ThF. Sinusoidal Coding. In: Speech Coding and Synthesis. Edited by W. B. Kleijn and K.K. Paliwal. Elsevier Science B.V. 1995
17. Pielemeier WJ, Wakefield GH. A high-resolution time-frequency representation for musical instrument signals. J. Acoust. Soc. Amer. 99(4). 1996
18. Ding Y, Qian X. Sinusoidal and Residual Decomposition and Residual Modeling of Musical Tones Using the QUASAR Signal Model. In: Proc. Intern. Comp. Music. Conf. ICMC'97, Thessaloniki, Greece, Sep. 1997. pp. 35-42
19. Verma TS, Levine SN. Meng TH. Transient Modeling Synthesis: a flexible analysis/synthesis tool for transient signals. In: Proc. Intern. Comp. Music. Conf. ICMC'97, Thessaloniki, Greece. Sep. 1997. pp. 164-167
20. Maher RC, Beauchamp JW. Fundamental frequency estimation of musical signals using a Two-Way Mismatch procedure. J. Acoust. Soc. Am. Vol. 95 No.4. pp. 2254-2263
21. Griffin DW, Lim JS. A New Model-Based Speech Analysis/Synthesis System. In: Proc. IEEE-ICASPP 1985. pp. 513-516
22. Rodet X, Depalle Ph, Poirot G. Speech Analysis and Synthesis Methods Based on Spectral Envelopes and Voiced/Unvoiced Functions. In: Proc. European Conference on Speech Tech. Edinburgh, U.K., Sept. 87. pp. 155-158
23. Rodet X, Depalle Ph, Poirot G. Diphone Sound Synthesis based on Spectral Envelopes and Harmonic/Noise Excitation Functions. In: Proc. ICMC-88. Kohln, Germany. Sept. 1988. pp. 313-321
24. Doval B, Rodet X. Fundamental Frequency Estimation and Tracking using Maximum Likelihood Harmonic Matching and HMMs. In: Proc. IEEE-ICASSP 93. pp. 221-224
25. Doval B. Estimation de la Fréquence Fondamentale des signaux sonores. PhD. Thesis. Université Paris-6. Paris, 1994
26. Laroche J. Etude d'un système d'analyse et de synthèse utilisant la méthode de Prony. PhD thesis. Télécom Paris. Paris, Oct. 89
27. Masri P. Computer Modeling of Sound for Transformation and Synthesis of Musical Signal. PhD thesis. University of Bristol. Dec. 1996
28. Rabiner LR, Schafer RW. Digital Processing of Speech Signals. Englewood Cliffs, NJ. Prentice Hall. 1978
29. Depalle Ph, Tromp L. An improved additive analysis method using parametric modeling of the short-time Fourier transform. Proceedings of International Computer Music Conference (ICMC'96), Clear Water Bay, Hong-Kong. August 1996. pp. 297-300
30. Depalle Ph, Hélie T. Extraction of spectral peak parameters using a short time Fourier transform modeling and no sidelobe windows. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, Oct. 1997
31. Depalle Ph, García G, Rodet X. Tracking of partials for additive sound synthesis using hidden Markov models. In: Proc. IEEE ICASSP-93. Minneapolis, Minnesota, Apr. 1993. pp. 225-228
32. Depalle Ph, García G, Rodet X. Analysis of Sound for Additive Synthesis: Tracking of Partials Using Hidden Markov Models. In: Proceedings of International Computer Music Conference (ICMC'93). Oct. 1993. pp. 94-97
33. Rabiner LR, Juang B-H. An introduction to Hidden Markov Models. IEEE ASSP Magazine, Jan. 1986

34. Fitz K, Haken L, Holloway B. Lemur - A Tool for Timbre Manipulation. In: Proc. Int. Comp. Music Conf. 1995. Banff, Sept. 1995. pp. 158-161
35. Rodet X, Lefèvre A. Macintosh graphical interface and improvements to generalized Diphone control and synthesis. In: Proc. ICMC'96. Hong Kong. Aug. 1996. pp. 336-338
36. Fant G. Acoustic Theory of Speech Production. Mouton, 1970
37. Rodet X. Time-Domain formant-wave-function synthesis. In: Simon JC ed. Spoken Language Generation and Processing. 1980. D. Reidel Publishing Company. Dordrecht, Holland. pp. 429-441
38. d'Alessandro C, Rodet X. Synthèse et Analyse-Synthèse par Fonctions d'Onde Formantiques. J. Acoustique 2 (1989) pp. 163-168
39. Moulines E, Charpentier F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In: Speech Commun. 16 1990. pp 453-467
40. Liénard JS. Speech Analysis and Reconstruction Using Short-Time Elementary Waveforms. In: Proc. IEEE-ICASSP 1987. Dallas.
41. Kronland-Martinet R. The Wavelet Transform for Analysis, Synthesis, and Processing of Speech and Music Sound. Computer Music Journal, vol. 12:4 1988. pp. 11-20
42. Coifman R, Wickerhauser MV. Entropy based algorithms for best basis selection. IEEE Trans. Inform. Theory, 38 (2):713-718, March 1992
43. Mallat S, Zhang Z. Matching Pursuit with time-frequency dictionaries. IEEE Trans. Signal Process. 41(12):3397-3415, Dec. 1993
44. Chen S, Donoho DL. Atomic decomposition bt basis pursuit, Technical report, Statistics Department, Stanford University, 1995.
45. Gribonval R, Bacry E, Mallat S, Depalle Ph, Rodet X. Analysis of sound signal with high resolution matching pursuit. In: Proceedings of the IEEE Conference on Time-Frequency and Time-Scale Analysis (TFTS'96). Paris, France. June 1996. pp. 125-128
46. Gribonval R, Depalle Ph, Rodet X, Bacry E, Mallat S. Sound signal decomposition using a high resolution matching pursuit. In: Proceedings of International Computer Music Conference (ICMC'96). Clear Water Bay, Hong-Kong. August 1996. pp. 293-296