Local Harmonic Estimation in Musical Sound Signals

Rafael A. IRIZARRY

Statistical modeling and analysis have been applied to different music related fields. One of them is sound synthesis and analysis. Sound can be represented as a real-valued function of time. This function can be sampled at a small enough rate so that the resulting discrete version is a good approximation of the continuous one. This permits one to study musical sounds as a discrete time series, an entity for which many statistical techniques are available. Physical modeling suggests that many musical instruments' sounds may be characterized by a deterministic periodic and stochastic signal model. In this paper the interest is in separating these two elements of the sound and finding parametric representations with musical meaning. To do so a local harmonic model that tracks changes in pitch and in the amplitudes of the harmonics is fit. Deterministic changes in the signal, such as pitch change, suggest that different temporal window sizes should be considered. Ways to choose appropriate window sizes are studied. Amongst other things our analysis provides estimates of the harmonic signal and of the noise signal. Different musical composition applications may be based on the estimates.

KEY WORDS: Musical Sound Signals, Local Harmonic Model, Widow Size Selection

1. INTRODUCTION

Statistics has been applied in various ways to music. For example, various stochastic techniques have been applied in composition (Jones 1981). Stochastic techniques have also been used in forecasting unfinished works (Dirst and Weigend 1992). Voss and Clarke (1975) studied the spectral properties of different musical signals and speculated on the possibility of it being so called 1/f noise. In Brillinger and Irizarry (1998) this is studied in more detail, and in particular higher order statistics are examined. In this paper the particular application that will be examined in detail is the analysis of sound signals produced by musical instruments. In this field, statistical techniques have been used, for example, to separate the signals into what are believed to be deterministic and stochastic parts and to deconstruct the deterministic part into *harmonic components*.

Every sound we hear is the consequence of pressure fluctuations traveling through the air and hitting our ear drums. The function that describes the audible pressure fluctuations of air is called a sound wave or sound signal. Physical theory and psychoacoustic concepts suggest that sound signals produced by *harmonic instruments* contain a component that is approximately periodic (Pierce 1992). Physical theory also suggests that the sound signals produced by instruments contain a non-periodic component which may be considered to be stochastic, and is believed to be an integral part of the sound (Chafe 1990). This belief has led some sound researchers to model sound signals with deterministic plus stochastic signal models (Serra and Smith 1991). Parametric models are used to describe the deterministic part of the signal and statistical based procedures are used to estimate parameters. The estimates obtained provide a way to separate the deterministic and stochastic signals and furthermore provide a useful parameterization of the deterministic part.

In Section 2 we examine some of the procedures suggested in the sound analysis literature and briefly discuss some of their strengths and weaknesses from a statistical point of view. In sections 3, 4, and 5 we present a statistical model and an estimation procedure that we believe improves on the existing methods. This statistical procedure fits a harmonic model, within small segments of the signal, using weighted least squares. The estimation procedure provides a parametric representation with musical interpretations which provides many practical uses described in Section 6.

There are accompanying audio versions of some of the examples presented in this paper. They can be obtained from a CD available through the Graduate Assistant of the Department of Statistics, 367 Evans Hall # 3860, Berkeley CA 94720-3860.

2. SOUND ANALYSIS AND STATISTICS

For centuries, understanding sound has been of interest. The Greeks and others must have noticed from the earliest times that plucked strings vibrate. Various Greek philosophers associated fast motion with high pitch and slow motions with lower pitches. In fact, the discovery of the relation between the lengths of strings on musical instruments and their pitch is commonly attributed to Pythagoras (Pierce 1992).

Today the study of sound has become a popular research field and, with the advent of electronic music, a practical one too. Contemporary researchers are interested in, for example, the problem of determining what particular characteristics of the sound produced by musical instruments permit humans to distinguish one instrument from another (Grey 1977; Deutsch 1982; Hartman 1997). Musicians have called the characteristics that enable us to distinguish different instruments *tone quality*, *tone color* or *timbre*, (Pierce 1992, pp. 196–199). Sound signals have been recorded and analyzed, using different approaches, with the goal of understanding what defines timbre. We will call this type of study *sound analysis*. Recently scientists have also become interested in the creation or reproduction of musical sounds without the use of an acoustical instrument. This is called *sound synthesis*.

Mathews (1963) was one of the first to successfully make use of sound analysis to produce effective sound synthesis. Mathews used the computer to analyze the sound produced by musical instruments with perceivable pitch, and then used the information obtained from the analysis to reproduce the sound. We will call this procedure *sound analysis/synthesis*.

Recently, interest has focused on using this information to facilitate the creation of new sounds based on the original sound. In this paper we wish to analyze sound so as to be able to obtain some parametric representation that is musically meaningful and may be manipulated to either reproduce the original sound or a new version of it. This analysis of the sound may also provide insight to understanding timbre. With today's technology we are finally able to process sounds in a data analytic fashion because the time is at hand when music can be treated directly as a quantity to be analyzed by contemporary statistical procedures and packages.

In this section we discuss some of the statistical procedures that have been used in sound analysis/synthesis. We also discuss some of the physical and acoustical properties that motivate these methods.

2.1 Music as a Time Series

In order to speak about statistical analysis of sound signals, we need somehow to represent them as data. The energy transmitted by a sound signal traveling through air can be transformed into a fluctuating voltage $V(t)$, which is a continuous function in time. Tape recorders work by storing the voltage function $V(t)$ on magnetic tape. One wants to have discrete data to facilitate statistical analysis. The obvious procedure is to take a discrete approximation of the continuous sound signal. Compact Disc (CD) technology is proof of how well discrete approximations of sound signals can work. In the case of commercial CDs, the sampling rate is a standard at 44100 observation per second (44.1 kHz). CD technology works by storing the numbers associated with the discrete approximation as 0s and 1s on a 12-cm optical disc, called a CD (Roads 1996).

2.2 Psychoacoustics and The Physics of Musical Sounds

Although not all existing sound synthesis and analysis techniques have found it necessary to use models that are in agreement with physical theory and/or psychoacousitcal principals, most of them are essentially based on the physical properties of instruments and the way our brain perceives sounds.

The first important physical discovery related to music is that when fluctuations of air are approximately periodic, with period in the audible range, we perceive what musicians have defined as a *pitch* (Pierce 1992, Chapter 2). We will call the frequency related to this periodicity the *fundamental frequency* of the sound signal. Instruments play different pitches by changing the fundamental frequency of the sound signal they are creating. Some cultures, e.g. Western cultures, have quantized these pitches and created *notes.* The pitch corresponding to 440 Hz has been called an A *note*, also A 440 Hz concert pitch or A4 (fourth A on a piano). The frequencies between 440 Hz and 880 Hz have been divided into 12 notes ($A, A\sharp, B, C, C\sharp, etc.$.) such that the proportion f_1/f_2 is fixed for all consecutive notes $f_1 < f_2$. Consecutive notes are said to be a semitone apart. This implies that given a base frequency f_0 we can find the distance k in semitones of any frequency f via the equation $12 \log_2(f/f_0) = k$. Apparently the trained ear cannot distinguish two notes if they are less than 3 *cents* (1 cent = a hundredth of a tone) apart (Pierce 1992).

More recent discoveries have been related to timbre. As early as the later half of the 19th century physicists were interested in the *harmonic structure* of musical sound signal (Rayleigh 1894). Around this time a physicist named Hermann Ludwig Ferdinand von Helmholtz conducted an experiment that proved that sound signals produced by harmonic instruments had frequency components at multiples of the fundamental frequency (von Helmholtz 1885), showing that the signals being produced contained a strong periodic, or at least approximately periodic, component and that the periodic signal had more harmonic structure than a simple sinusoid. This discovery inspired physicists to seek an explanation for this phenomenon. In Fletcher and Ross (1991) mathematical models of the physical acoustics of instrument sound productions are presented for practically all orchestral instruments and various non-orchestral ones.

Also in the 19th century George Simon Ohm (of Ohm's law) conjectured that the human auditory system operates as a spectrum analyzer that displays the power spectrum of a complex tone and is insensitive to the relative phases of the components (Hartman 1997). By spectrum, Ohm was referring to the Fourier transform of the signal

$$f(\lambda) = \int_{-\infty}^{\infty} V(t) \exp\{it\lambda\} \, dt,$$

the power spectrum being the modulus $|f(\lambda)|^2$. At his time this conjecture was not accepted as plausible, but recent psychology experiments seem to confirm its validity; see Grey (1977) for an example, and Pierce (1992, chapter 7) for an overview. Furthermore, physiological studies of the ear have found evidence to support the validity of Ohm's conjecture, (Patterson et al. 1992). It is believed that a sort of *spectral analysis* is carried out in the cochlea to provide the brain with the necessary information to determine timbre see (Pierce 1992, chapter 7) .

We must notice that any musical signal $V(t)$ and its reverse $V(-t)$ will have the same power spectrum although in many cases the ear can definitely distinguish between a sound, say of a song, and its reverse. Ohm was perhaps referring to the way the ear operates within very small time windows. Green (1985) performed an experiment where subjects were played two different sounds separated in time by a number, Δt, of seconds, then the same sounds where played but in reverse. Subjects were asked to say if they could hear a difference. For Δt around 20 milliseconds the differences became hard to detect and after Δt was smaller than 1.5 milliseconds the difference was never noticed.

Helomotz's discovery suggests that within small time intervals sound signals produced by harmonic instruments are periodic and thus can be expressed as a sum of sinusoids of the form

$$\sum_{k=1}^{K} \rho_k \cos(k\lambda t + \phi_k) \tag{1}$$

with $K \leq \lfloor \pi/\lambda \rfloor$, t an index representing units of time, in this case $\{\text{sampling rate}\}^{-1}$ seconds, and $\lambda/2\pi$ the fundamental frequency in cycles per unit time. The K cosines included in the summation in (1) are referred to as *harmonic components*. The musical term for these components are the *harmonics* or *overtones*. Ohm's conjecture can then be summarized by saying that the timbre of short and "stable" segments of sound is determined only by the strength of the amplitudes, i.e. the ρ_ks, and the fundamental frequency λ which also determines the pitch.

Time series analysis provides a simple tool that allows us to check if the data obtained from sound signals are in agreement with this. For the data $y_t, t = 1, \dots, T$ obtained from the discrete representation of a sound signal the periodogram is defined by:

$$I(\lambda) = \frac{1}{2\pi T} \left| \sum_{t=1}^{T} \exp\{-i\lambda t\} y_t \right|^2, \; 0 \leq \lambda \leq \pi \tag{2}$$

If the data y_t is periodic as defined in equation (1) the periodogram will show peaks at frequencies $k\lambda$ with $k = 1, \ldots, \lfloor \pi/\lambda \rfloor$ (Bloomfield 1976). Computed periodograms of sound signals produced by harmonic instruments exhibit such peaks. Figure 1 presents the periodograms for the signal produced by a trumpet playing concert pitch A and for a clarinet playing that same note. Each signal is about 3 seconds long and the sampling rate is 44.1 kHz, so $T \approx 132300$. Notice that peaks at the multiples of 440 Hz are observed in both cases and that the two periodograms look different. The trumpet seems to have more noticeable harmonics. This is usually associated with a *brighter* timbre, which the trumpet has. However, relatively high values of the periodogram at non-harmonic frequencies suggest that a considerable amount of the variation of the signal is not explained by such harmonic components. This suggests that there is more to the sound signal than just a periodic component.

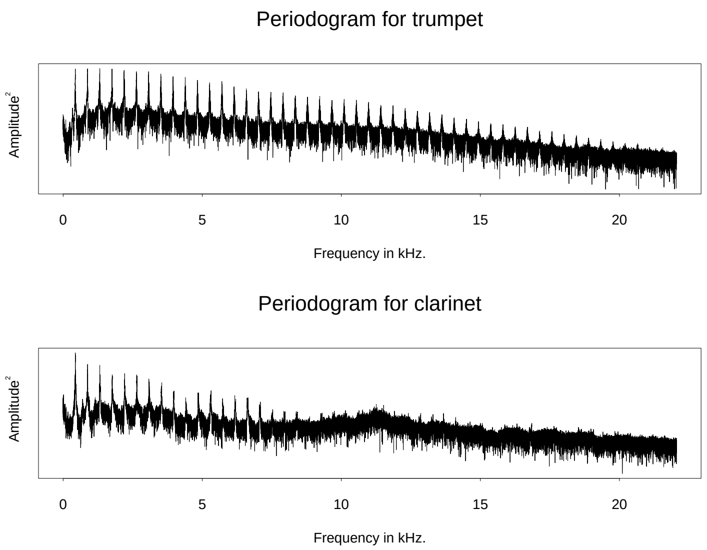

Figure 1. Periodogram for sound signals of a trumpet and a clarinet playing concert pitch A.

In the early 1960s, Risset and Mathews (1969) made a discovery that greatly advanced the understanding of timbre. Risset and Mathews used the computer to study the local behavior of the harmonic components of sound signals. They noticed that the intensity of the harmonic components varied substantially through time. This would imply that the signals are not exactly periodic which is in agreement with Figure 1.

We may verify this from the data by computing *spectrograms*. For data $y_t, t = 1, \ldots, T$, we define the spectrogram of a signal at time t_0 by:

$$I(t_0, \lambda) = \frac{1}{2\pi(2M+1)} \left| \sum_{t=t_0-M}^{t_0+M} \exp\{-i\lambda t\} y_t \right|^2 \tag{3}$$

Here $2M + 1$ is some suitable window size.

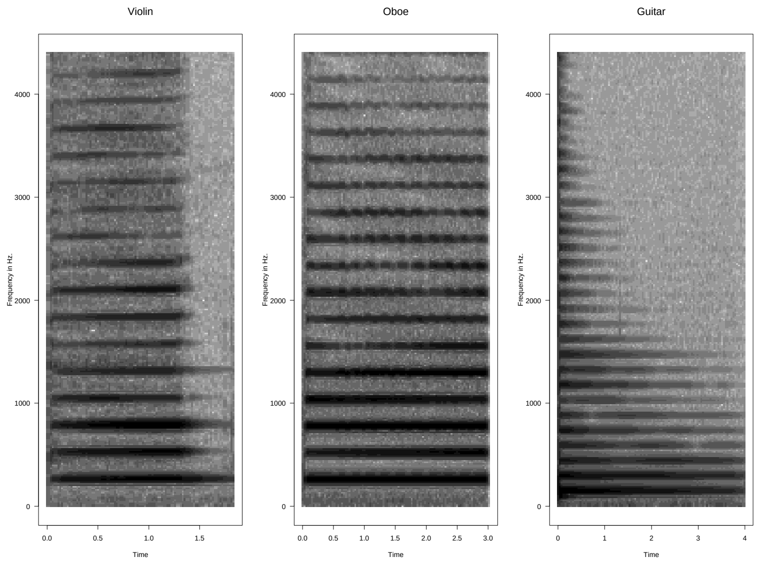

Figure 2. Spectrograms for harmonic instruments with $2M + 1 = 20$ millisecond windows.

In Figure 2 we see the spectrogram for the signals produced by three harmonic instruments; a violin, an oboe, and a guitar. The violin and oboe are both playing C (261.6256 Hz), while the guitar is playing D (146.8324 Hz). Dark shades of grey represents high power for the spectrogram. Notice that the spectrograms verify Mathews' discovery: At all times the instruments show peaks at frequencies that are multiples of the fundamental frequency, showing that the fundamental frequency is not changing much and that at least locally the signals are approximately periodic. The amplitudes of these harmonic components, however, are definitely varying through time in different ways, which suggests that the shape of the periodic component is changing. This is particularly clear in the case of the guitar where the higher harmonics "die off" more rapidly. Risset and Mathews conjectured that the time-varying character of the harmonic structure was important in determining timbre. These observations led Risset and Mathews to develop *additive synthesis*, a method for the analysis and synthesis of music sound signals described in the next section.

2.3 Sound Analysis and Synthesis

A main goal of sound analysis and synthesis is to obtain a *parametric representation* of sound signals. Consider $y_t, t = 1, \dots, T$ to be the discrete representation of the sound signal of interest, we say $\beta_t, t = 1, \dots, T$ is a parametric representation of the sound signal if it provides a source for reconstruction or synthesis of the original signal, for example that $y_t \approx s\{t, \beta_t\}$, and that meaningful sound transformations can be obtained through the parameters $\beta_t, t = 1, \dots, T$.

The ideas of Helmholtz and Ohm motivate a method based on the periodogram. To characterize the sound let $\rho_k = \sqrt{I(k\lambda)}$, the intensity of the kth harmonic as measured by the periodogram with λ the fundamental frequency, and consider the parameterization, $\beta = (\lambda, \rho_1, \dots, \rho_k)$. To obtain the approximation of the original sound from the parametric representation we simply take $\hat{y}_t = \sum_k \rho_k \cos(k\lambda t)$. Notice that in this case the parameterization does not change with time, thus this parameterization has proven to be useful when considering short segments. Instruments' sounds have been *synthesized* using this representations. When the sounds are of relatively long duration (say, more than 100 milliseconds) the results are not satisfactory, in the sense that the synthetic sounds obtained sounded quite different from the original. This of course is due to the fact that in general sound signals are not exactly periodic if long stretches are considered.

As mentioned above, Risset and Mathews (1969) suggested a local approach to improve on this method that became an important step towards accurate digital sound synthesis. Risset and Mathews named their approach *additive synthesis*, which has proven to be one of the most effective methods available until now (Rodet 1997). In fact, similar models have been proposed for speech signals (McAualay 1986). Sound signals are modeled as summations of sinusoidal components with time-varying amplitudes. In order to obtain a parametric representation for sound signals produced by a trumpet playing one note Risset and Mathews choose an appropriate λ as the fundamental frequency, then divided the signal into $L = T/(2M+1)$ non-overlapping segments of suitable size $(2M+1)$ and used the spectrogram to define

$$\rho_k(t) = \sqrt{I(t; k\lambda)} \text{ for all } t \text{ in the } l\text{th segment}$$

with $k = 1, \dots, K$ (K was chosen heuristically) and $I(t; k\lambda)$ defined as in (3). The parametric representation is then $\beta_t = \{\lambda, \rho_1(t), \dots, \rho_K(t)\}$ (in general, Risset and Mathews chose L much smaller than T in which case the parameterization provided a reduction in dimension). The signal reconstruction is $\hat{y}_t = \sum_k \rho_k(t) \cos(k\lambda t)$.

Sounds have been synthesized using this parametric representation and have been found to greatly improve on the method suggested by Helmholtz. However, the "fitted" signal is never the same as the original. Possible sources for this are added components that are not periodic in nature and we will refer to these as the stochastic or noise component. Some examples are the sound produced by fingers hitting keys, nails plucking strings, and surplus blown air. The presence of noise components in a sound signal seem to be stronger during the beginning of a note, or what musicians refer to as the *attack* (Masri and Bateman 1996). Some researchers have founded it important to model this in order to obtain a more accurate reconstruction of the sound. We now turn our attention to a statistical model based on additive synthesis that takes the stochastic component into account.

2.4 Deterministic Plus Stochastic Signal Models

Serra and Smith (1991) incorporated a non-periodic component to additive synthesis and modeled it as an additive random signal to account for the variation not described by the additive synthesis model. Since, many have proposed and used similar models (DePalle and Poirot 1991; Depalle et al. 1993; Rodet 1997). Such models are characterized by a deterministic plus stochastic signal statistical model. In Serra and Smith the model presented is

$$Y_t = \sum_{k=1}^{K} \rho_k(t) \cos\{\phi_k(t)\} + \epsilon_t, t = 1, \dots, T. \tag{4}$$

Serra and Smith (1991) assumes $\{\epsilon_t\}$ to be a stationary autoregressive process. An implicit assumption is that the deterministic signal resembles a sum of sinusoids. Notice that in this case the $\beta_t = \{\rho_1(t), \dots, \rho_K(t), \phi_1(t), \dots, \phi_K(t)\}', t = 1, \dots, T$ can be thought of as the parametric representation of the deterministic part of the sound.

We can think of $\omega_k(t) = d\phi_k(t)/dt$ as the instantaneous frequency of the kth harmonic. Notice that in this model the frequency components $\omega_k(t), k = 1, \dots, K$, need not be at frequencies that are multiples of a fundamental frequency.

One is interested in estimating $\phi_k(t)$ and $\rho_k(t)$ for $k = 1, \dots, K$. In Serra and Smith (1991) estimation is done in two steps. In the first step the signal is divided into short (usually 256 data points), possibly overlapping segments, called *analysis frames*. For each segment the peaks of the periodogram of the data, which are tapered, are considered as possible indication of a *sinusoidal partial*. The amplitude and frequency of each peak is recorded. In the second step, peaks of successive analysis frames are grouped into tracks. For a particular track, say the kth track, the frequencies at which the peaks occur are considered to be estimates of the instantaneous frequency $\omega_k(t)$ of the kth sinusoidal component. Notice that we may estimate the $\phi_k(t)$'s from the estimates of $\omega_k(t)$. This tracking is usually based on a heuristic approach that matches peaks of consecutive frames by the proximity of the frequencies associated with them. A procedure described in Depalle, García and Rodet (1993) , takes into account that under model (4) periodogram peaks are random quantities and perform the tracking by globally optimizing over the set of all tracks via a hidden Markov model. A stretch of a trumpet sound (playing concert pitch A) of length 0.4 seconds, was analyzed using this method, the results are shown in Figure 3.

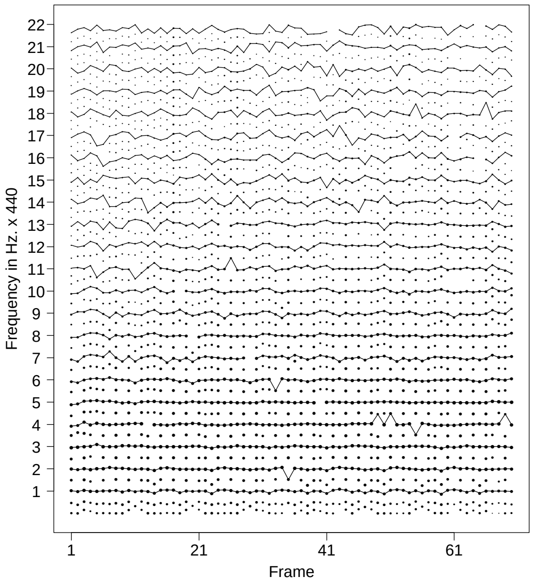

Figure 3. Periodogram maxima of analysis frames with partial tracks of a 0.4 second stretch of the sound signal of a trumpet playing A4.

The validity of this method under the model defined in (4) and the statistical properties of the estimates obtained are not discussed by Depalle, García and Rodet (1993). We now discuss some of the statistical considerations.

In model (4) the existence of a deterministic component in the signal Y_t is assumed. As mentioned above, in many examples this deterministic component appears to be periodic within small segments of the signal. If this is the case, strong peaks will appear in the periodogram at frequencies that are multiples of a fundamental frequency as seen in Figures 1 and 2. However, the above mentioned partial tracking algorithms allow partials to exist at frequencies that are not multiples of a fundamental frequencies. Furthermore, the assumed model implies that the periodogram quantities are random, i.e. a peak can be due to random variability. In particular, when only 256 observations are used when computing the periodogram the variation can be relatively large. Estimates obtained for the deterministic component when many non-harmonic partials are "tracked", as done by Depalle, García and Rodet (1993) are hard to interpret from a statistical point of view. Computing the statistical properties of periodogram peaks found by such tracking algorithms can be complicated, even when using a simple statistical model. For this reason finding an algorithm for partial tracking that provides useful estimates is not straightforward and will not be discussed in this paper. Instead, in Sections 3, 4, and 5, we present a statistical model that assumes the deterministic components of the signal are locally periodic as opposed to simply a sum of sinusoidal components that are not necessarily related in a harmonic fashion.

2.5 Applications

Obtaining parametric representations of sounds leads to many musical applications, (Rowe 1994). One example is timbre morphing, which is the process of combining two or more sounds to create a new sound with intermediate timbre and duration. Morphing can be used to create interesting sounds that are not found in nature, but that have the characteristics of naturally occurring sounds. An interesting example is the recreation of a castrato voice (Depalle et al. 1995). This was done to produce a sound-track for the film Farinelli, the famous 18th century castrato. To simulate Farinelli's voice, the voice of a counter tenor and a soprano were analyzed and parametric representations were obtained. These two representations were combined in a way that produced a timbre similar to that of a castrato.

3. HARMONIC REGRESSION MODELS

Both physics and data analysis seem to suggest that musical sound signals produced by harmonic instruments are locally periodic. Psychoacoustical experiments suggest that the harmonic components of such signals are important in determining timbre. Further analyses seem to suggest that there is also a stochastic component included in these signals. This suggests that for data $y_t, t = 1, \dots, T$ in a short segment of a sound signal a useful deterministic plus stochastic signal model would be

$$y_t = s(t;\beta) + \epsilon_t \quad t = 1,\dots,T \text{ with } s(t;\beta) = \sum_{k=1}^{K} \{A_k \cos(k\lambda t) + B_k \sin(k\lambda t)\}, \tag{5}$$

$\beta = (A_1,\dots,A_K,B_1,\dots,B_K,\lambda)'$ and $\{\epsilon_t\}$ a stochastic component. In a musical context, as indicated in the previous section, we call λ the pitch and the terms being added in (5) the harmonics or overtones with amplitudes defined by $\rho_k = (A_k^2 + B_k^2)^{1/2}$.

Many signals in nature have been statistically analyzed via sinusoidal regression models like the one defined by (5) (Brillinger 1977). In Walker (1971), Hannan (1971, 1973, 1974), and Brown (1990) similar models are studied under the assumption that the noise $\{\epsilon_t\}$ is stationary. Assuming certain regularity conditions these authors find estimates that are asymptotically equivalent to least squares estimates. Consistency is shown and asymptotic variance expressions are developed. See Irizarry (1998) for a review.

Since we are fitting this model in order to obtain estimates of parameters that may vary with time, it is only natural to consider window based estimates. The weighted least squares method consists of choosing $\hat\beta$ to minimize the criterion

$$S(\beta) = \sum_{t=1}^{T} w(t/T)\,\{y_t - s(t,\beta)\}^2\,. \tag{6}$$

Here $w(s)$ is a non-negative weight function with support $[0,1]$. In the remainder of this sections we use the following constants

$$W_n = \int_0^1 t^n w(t)\,dt \quad \text{and} \quad U_n = \int_0^1 t^n w(t)^2\,dt \tag{7}$$

for n=0,1, and 2.

Under some regularity condition Irizarry (1998) develops expression for the asymptotic variance of $\hat\beta$. This provides approximations for the variance of the fundamental frequency and amplitude estimates. Namely that

$$\mathrm{Var}(\hat\lambda) \approx 4\pi T^{-3}\left\{\frac{W_0^2U_2 - 2W_0W_1U_1 + W_1^2U_0}{(W_0W_2 - W_1^2)^2}\right\}\left\{\sum_{k=1}^{K} k^2\rho_k^2/f_{\epsilon\epsilon}(k\lambda)\right\}^{-1} \tag{8}$$

$$\mathrm{Var}(\hat\rho_k) \approx 4\pi f_{\epsilon\epsilon}(k\lambda)T^{-1}\left(\frac{U_0}{W_0^2}\right) \tag{9}$$

where W_0, W_1, W_2, U_0, U_1, and U_2 are defined by (7) and $f_{\epsilon\epsilon}$ is the power spectrum of the stationary noise ϵ_t defined by

$$f_{\epsilon\epsilon}(\lambda) = \frac{1}{2\pi} \sum_u c_{\epsilon\epsilon} \exp\{-i\lambda u\}, \quad -\infty < \lambda < \infty$$

with $c_{\epsilon\epsilon}(u) = \mathrm{Cov}\{\epsilon_{t+u}, \epsilon_t\}$, the autocovariance function of the stationary process $\{\epsilon_t\}$.

3.1 An Example

We illustrate the appropriateness of fitting a harmonic model as that defined by (5) to short segments of musical sound signals with an example.

One may convert digital sound signals into ASCII files readable to S-Plus (see Appendix for details). For the sound signal of a violin, sampled at 44.1 kHz, playing the note C4 we consider a 50 millisecond stretch (2200 observations). Notice in Figure 4 how the signal, within the considered segment, does seem to be usefully periodic.

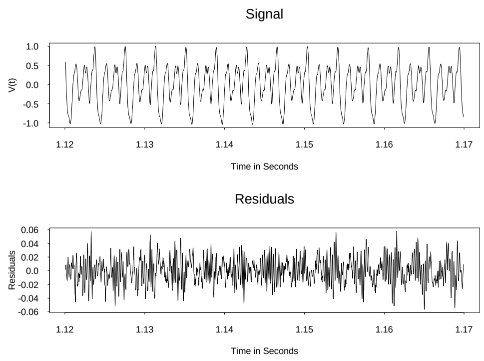

Figure 4. Local fit for the sound signal of a violin playing C4 and corresponding residuals. Smoothed periodogram estimates for the spectrum of the noise for the sound signal of a violin playing C4

We fit a harmonic model, using K=15, to the stretch of data presented in Figure 3 and obtain a *reasonable* fit. The residual mean-square is $\hat{\sigma}^2 = 0.0003$. Comparing this to the total variance of the original signal 0.26 shows that the fitted model explains a large amount of the variation. The residual plot, also seen in Figure 4, suggests that the noise could be considered stationary in the given stretch.

The estimate of the fundamental frequency is 262.795 Hz. The note being played is C4 which if the instrument is tuned to A 440 Hz is equivalent to 261.6256 Hz Using (8) and (9) our estimate is not included in the approximate 99% confidence interval around the frequency corresponding to C4, $[261.5921, 261.6590]$. Is the violin out of tune? In section 6 we discuss possible musical interpretations of standard errors. Notice also that the estimates of the amplitudes could be used as a possible description of the timbre of this sound. The approximate standard errors of the estimates could be used when comparing these amplitude estimates for two different sounds.

We have seen an example of how useful a periodic deterministic plus stochastic signal model can be for describing sound signal data within short time segments. This motivates the use of a local regression type estimation procedure for parametrizing sound signals of long duration.

4. LOCAL HARMONIC ESTIMATION

In the case of sound signals of long duration, the harmonic structure appears to change in time as the performer changes the sound being produced by the instrument. Examples of such changes are note change, vibrato, and tremolo, to mention a few. For this reason the stationary model mentioned in Section 3 is not appropriate when considering long stretches of musical signals. As mentioned, models have been suggested in the signal processing literature (Risset and Mathews 1969; Serra and Smith 1991) that take this into account. We now present a statistical model and an estimation procedure for these types of signals.

We model the sound signal produced by a harmonic instrument with a deterministic plus stochastic signal model

$$Y_t = \mu(t) + \epsilon_t, t = 1, \dots, T$$

with $\mu(t)$ the deterministic component and ϵ_t the stochastic noise.

Say we are interested in estimating only $\mu(t_0)$. The *local harmonic estimation* approach, similar to local regression (Cleveland and Devlin 1988) and local likelihood estimation (Tibshirani and Hastie 1987), is to assume that there is some segment $(t_0 - h, t_0 + h)$ such that the deterministic component within this segment of the signal is approximately periodic and the noise component is approximately stationary. In this case we have

$$Y_t \approx s(t;\beta_0) + \epsilon_t \text{ for } t \in (t_0 - h, t_0 + h) \text{ with } s(t;\beta_0) = \sum_{k=1}^{K} \{A_{k,0}\cos(k\lambda_0 t) + B_{k,0}\sin(k\lambda_0 t)\} \tag{10}$$

and $\beta_0 = (A_{1,0}, \dots, A_{K,0}, B_{1,0}, \dots, B_{k,0}, \lambda_0)$ and $\{\epsilon_t\}$ a stationary process. We obtain an estimate $\hat{\beta}_0$ of β_0 using weighted least squares by giving positive weight only to points within the estimation segment $(t_0 - h, t_0 + h)$. To obtain an estimate of $\mu(t_0)$ we simply use equation (10) and let $\hat{\mu}(t_0) = s(t_0, \hat{\beta}_0)$. With some adjustments to the window function $w(u)$ when t_0 is near the end points of the signal, we may repeat this procedure for each $t_0 = 1, \dots, T$ and obtain estimates $\hat{\mu}(t), t = 1, \dots T$. Furthermore, we can define a parametric representation using our estimates obtained for each t_0 and form $\hat{\beta}_t, t = 1, \dots, T$. Notice that this provides a musically meaningful parametric representation of the signal since each $\hat{\beta}_t$ contains an estimate of the local fundamental frequency $\hat{\lambda}_t$, and the coefficients needed to form estimates of the local amplitudes of the harmonic components $\hat{\rho}_{k,t} = \{\hat{A}_{k,t}^2 + \hat{B}_{k,t}^2\}^{1/2}, t = 1, \dots, T$ for $k = 1, \dots, K$.

The residuals may be defined via

$$\hat{\epsilon}_t = y_t - \hat{\mu}(t), t = 1, \dots, T. \tag{11}$$

These can be studied to assess the fit. For example, this data may be converted into a sound file. We may then listen to the residuals and assess the fit by what we here, i.e. residual analysis by ear.

4.1 An Example

As an example we run the analysis on the sound signal of an oboe playing C4 (261.6256 Hz) for a duration of 3 seconds. Listening to this sound we notice that the oboist is playing both a *tremolo* and a *vibrato*, slight and rapid variations in volume and pitch respectively. We obtain estimates of $\mu(t)$ by keeping the number of partials in the model and the estimation window sizes constant for each $t_0 = 1, \dots, T$, at 15 partials and 20 milliseconds respectively. In Figure 5 we see the estimated deterministic signal and the residuals. The residual plot suggests a reasonable fit. The larger variation of the residuals during the beginning of the sound is in agreement with the statement made in Section 2 regarding the attack (if we listen to the sound generated by the residuals we hear what sounds like the initial burst of air blown by the instrumentalist).

In Figure 5 we also see estimated fundamental frequencies $\hat{\lambda}(t)$ and the estimates for the amplitudes $\hat{\rho}_k(t)$ for $k = 1, \dots, 5$. The estimated fundamental frequency is between 258 Hz and 260.5 Hz, close to the frequency related to C4, (261.6256 Hz). The oboist might be playing a bit out of tune. The dotted lines in Figure 5 are 3 cents away from the average fundamental frequency (259.25 Hz). The Figure seems to suggest that there are variations in pitch perceivable to the ear and that there are variation in the amplitudes of the first five partials. This is in agreement with what we hear, a tremolo and a vibrato.

In this and many other cases studied, the sounds of the original signals y_t and the estimated deterministic component $\hat{\mu}(t)$ were almost indistinguishable. When amplified, the sound of residuals sounded much as

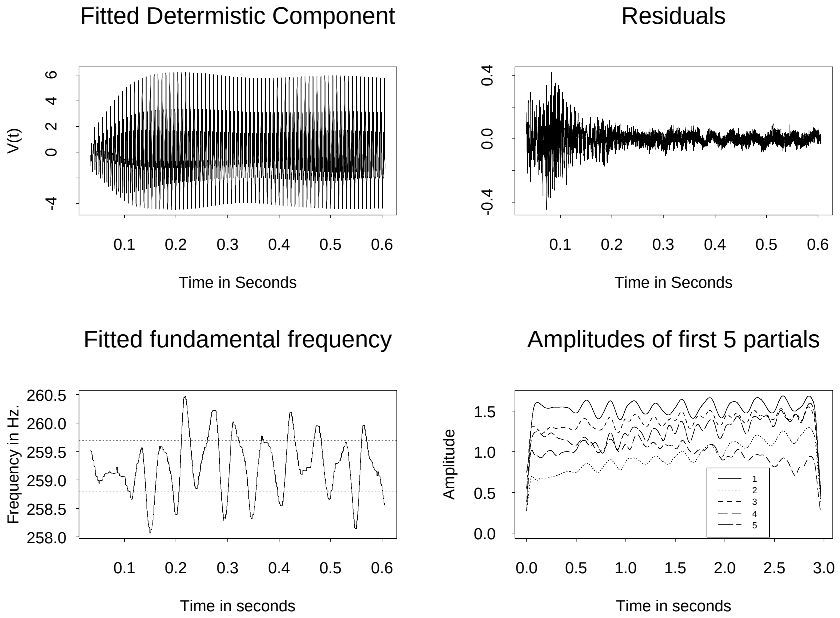

Figure 5. Estimated fundamental frequency and amplitude of first five partials for the sound signal of a violin playing C4.

we expected: specifically a sound like that of air and spit going through a tube for the saxophone, clarinet and trumpet, a screechy metallic sound for a violin, a pluck with no clear tone for the guitar, etc.. We can further assess the fit by studying second-order properties of the residuals. See Irizarry (1998) for details.

5. DYNAMIC WINDOW SIZE SELECTION

For a particular sound signal a variety of deterministic factors may affect how good of an approximation it is to assume the deterministic component is periodic within a given segment. For example, a change in note creates a discontinuity in the fundamental frequency and thus any segment containing the time change will not be periodic. Another example includes, gradual or sudden changes in volume. Such phenomena suggest that the estimation windows sizes h should not remain fixed for all estimation times t_0. For two values t_0 and t_1, the approximately periodic assumption might be appropriate within the interval $(t_0 - h, t_0 + h)$, but not within $(t_1 - h, t_1 + h)$. For each t_0 one needs to make a decision on the size of the span h of the estimation segment or *smoothing window*. Also, for different segments of sound signals the number of harmonic components that seem meaningful varies, as seen for the guitar in Figure 2.

5.1 Heuristic Window Size and Number of Harmonics Selection

For the estimation procedure presented in the previous section to make sense we need to consider appropriate stretches of sound signals. For each local fit we need to find a segment of signal that is appropriate for the approximately periodic assumption.

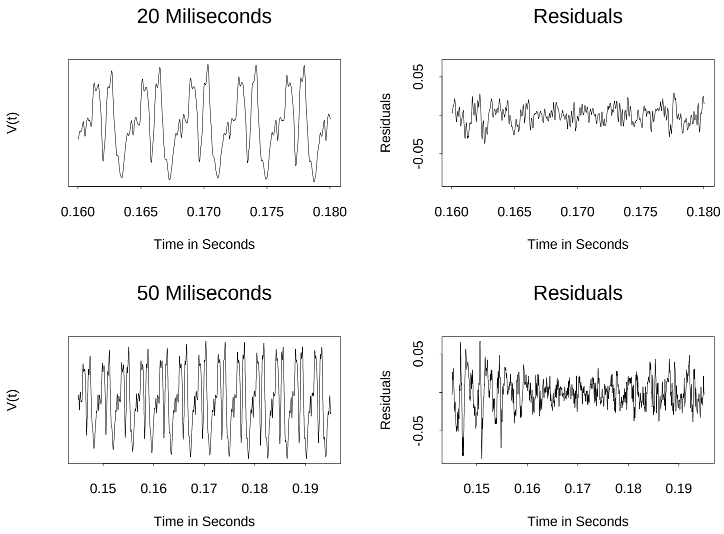

Figure 6. Comparison of two stretches of different duration of the sound signal of a violin playing C4.

In Figure 6 we see two stretches, taken from an early part of the signal ($t_0 = 0.17$ seconds) of a violin playing C4, one with a duration of 20 milliseconds and the other with a duration of 50 milliseconds. Notice that in the second plot the deterministic component does not appear to be approximately periodic, but rather that the total amplitude is growing with time. Looking at the residuals, also seen in Figure 6, produced from fitting a harmonic model with 15 harmonics, we see that they do not appear to be stationary for the 50 millisecond stretch. In this case we would pick the 20 millisecond stretch over the 50 millisecond one.

Another decision that we need to make is how many harmonic components K to consider in our model. Previous estimation procedures, (Depalle et al. 1993), usually fit many sinusoidal components. If the sampling rate is N Hz and the fundamental frequency is λ Hz at least $N/2\lambda$ and as many as $N/2$ components are considered (frequencies above $N/2$ are aliased). Fitting too many parameters may result in a saturated model. We need to decide how many harmonics to include in our model.

The number of "peaks" in the periodogram plot may be used to obtain a general idea of how many harmonics to consider. As we saw in Figure 1, the clarinet seems to have fewer "significant" harmonics than the trumpet. We may also use the estimation procedure described in Section 4 in the following way.

We examine stretches of about 50 millisecond durations of a clarinet and a trumpet playing concert pitch A (440 Hz). The plots in Figure 7 suggest that the approximately sinusoidal assumption is appropriate in

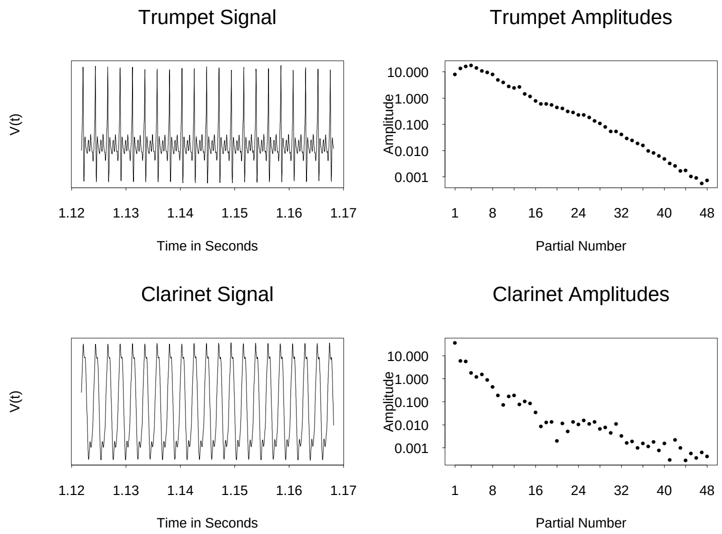

Figure 7. Stretches of 50 millisecond duration of the sound signals of a trumpet and a clarinet playing concert pitch A and estimated amplitudes

these stretches. We fit a harmonic model with 48 harmonics to each segment. In Figure 7, we also see the estimated amplitudes for each partial of the harmonic model. Notice that the estimated amplitudes of the higher harmonics are close to 0, relative to the amplitude estimated for the lower harmonics.

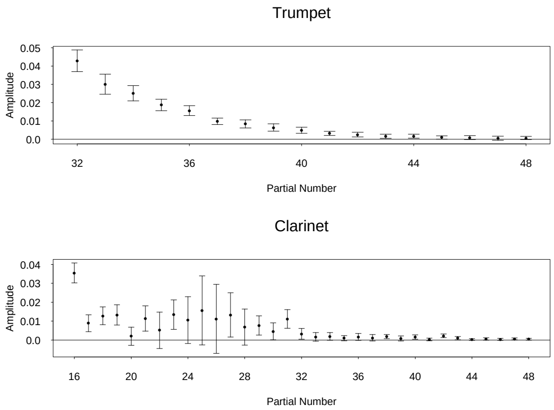

Figure 8. Estimated amplitudes of higher harmonics with estimated 99% confidence intervals.

In Figure 8 we see the estimates of the higher amplitudes surrounded by 99% confidence intervals. The confidence intervals are constructed using the asymptotic approximation for the variance of the estimates given by equation (9). Notice that in some cases 0 is included in such intervals. As we expected the clarinet has fewer amplitudes that appear "statistically significant". One problem with this z-test-type approach is that the amplitudes ρ_k are by definition non-negative. We do not intend to use hypothesis testing as a tool to choose how many harmonics to include in our model, but rather as a descriptive illustration of why consider different number of harmonics for different sound signals may be appropriate. In the next section we present criteria that will be used to decide among the different choices.

5.2 Information Criteria

Due to the large amount of data and points of estimation, using procedures like cross-validation to choose window sizes and number of harmonics in practice is unrealistic (at least in 1999). We want a criteria that will permit us to automatically choose amongst different possible estimates. Irizarry (1998) presents intuitive arguments that lead to a useful criteria for model selection. If we assume that the errors in the approximate model in Section 4 follow a Gaussian-type distribution for the data of a segment $(t_0 - h, t_0 + h)$ then two useful criteria are:

$$\text{WAIC(K,h)} = \log \hat{\sigma}^2 + p/N_h$$

$$\text{WBIC(K,h)} = \log \hat{\sigma}^2 + \{2K \log(N_h/2) - p\}/N_h$$

where p is the *equivalent number of parameters*, in this case defined by $(U_0/W_0)2(K+1)$ with U_0 and W_0 defined by (7), and N_h the *equivalent number of observations*, in this case defined by $N_h = \sum_{t=0}^{2h} w(t/2h) \approx 2hW_0$. We define the estimated variance for a given segment with h, as

$$\hat{\sigma}^2 = (N_h - p)^{-1} \sum_{t=0}^{2h} w(t/2h) \left\{ y_{t_0-h+t} - s(t_0 - h + t, \hat{\beta}_0) \right\}^2,$$

with $\hat{\beta}_0$ the weighted least squares estimate described in Section 4. Notice that both these criteria satisfy the conditions for consistent model selection criteria presented in Wang (1991) for harmonic models with a more general assumption for the noise process $\{\epsilon_t\}$.

5.3 An Example

The shakuhachi flute is a Japanese instrument characterized as being “noisy”. The sound of the performer blowing is one of its distinguishing characteristics. By listening to the sound signal studied in this example we notice that it is characterized by a rapid change of pitch for the first half second, then the pitch is held steady for about 3.5 seconds, then a vibrato is played for about half a second after which the pitch is held fixed again.

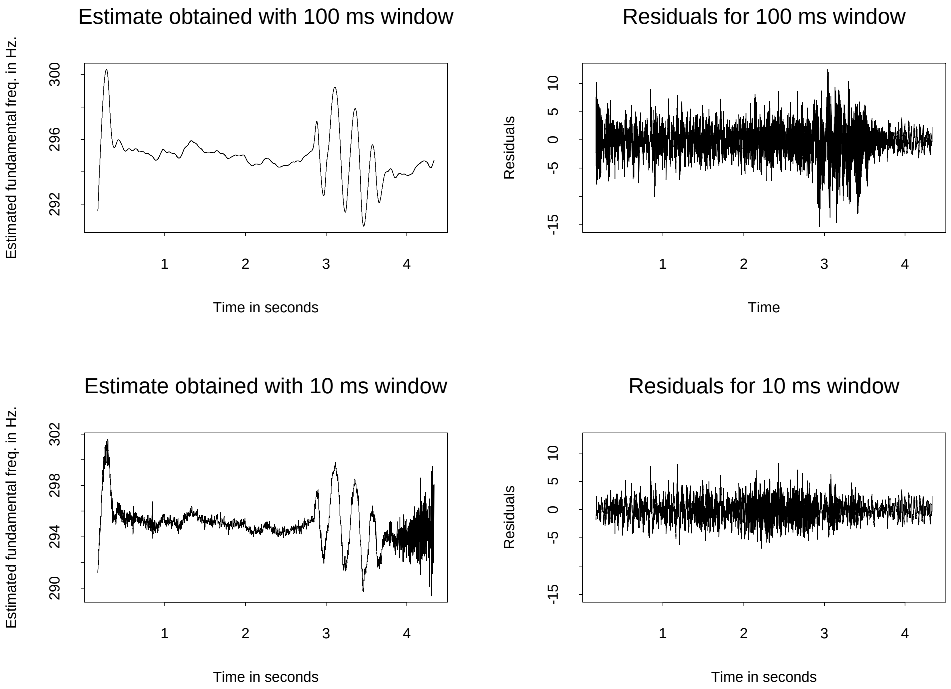

Figure 9. Estimated fundamental frequency when using fixed window sizes of 100 and 10 milliseconds and the respective residuals.

This particular sound is interesting to study for two reasons. First, the noisy character of the sound makes the partial tracking techniques, described in Chapter 2, difficult to implement because it is hard to discern significant peaks in the periodogram with small amounts of data. Second, the different behavior of the pitch function in different parts of the signal suggests that a fixed window size may be inappropriate and thus that different window sizes should be used in different parts of the signal.

For various segments of the shakuhachi flute signal both WAIC and WBIC criteria suggested that we fit a local harmonic model with 15 harmonics. We fit such a model using fixed window sizes of 100 milliseconds and then 10 milliseconds. In Figure 9, we see the estimated fundamental frequencies and residual plots for these two window sizes. For the estimate obtained with the larger window size we notice that during the vibrato part, i.e. the part where the pitch is not near constant, the fit is not as good (the residuals are bigger). If we try to fix this problem by considering a smaller window size, then the estimated fundamental frequency seems to be too variable, as seen in Figure 9, especially towards the end.

We fitted a local harmonic model with 15 harmonics to the shakuhachi flute using the WBIC to choose between window sizes of 10, 20, 40, 60, 80, 100, and 120 milliseconds. In Figure 10 we see how the window size that minimize the WBIC vary as the signal progresses. Notice in particular how, on average, smaller

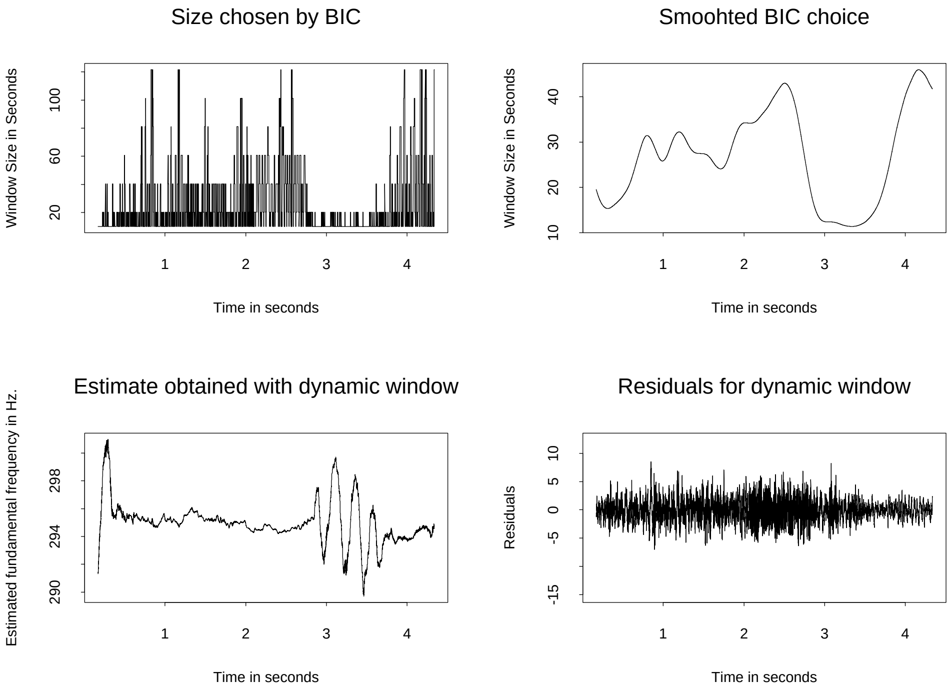

Figure 10. Selected window size (in milliseconds) and smoothed version of the selected windows; estimated fundamental frequency when using a dynamic window size and respective residuals.

window sizes are chosen during the parts of the signal where the fundamental frequency is not near constant, as we would expect. Because the WBIC choices are so variable we compute a smoothed version of window sizes by using a kernel smoother, also seen in Figure 10. The dynamic window local harmonic estimate is obtained by following the procedure defined in Section 4 using window sizes h defined by the smoothed WBIC choices. We present the estimate of the fundamental frequency and the residuals obtained using this procedure in Figure 10. We notice the improvement this provides over the fit obtained with the 100 millisecond window procedure by comparing the residuals as seen in Figures 9 and 10 and the improvement it provides over the fit obtained with a 10 millisecond window by comparing the estimated fundamental frequencies, also seen in Figures 9 and 10.

6. APPLICATIONS

6.1 Variability of the Estimates

In the music technology literature a variety of pitch estimators are presented (Junhar 1997). When such procedures are used to estimate the fundamental frequency of sound signals produced by one instrument playing one note, the movement of these estimates are sometimes explained with deterministic factors. If we assume the presence of such a stochastic element, the possibility exists that the variation over time of the pitch estimates can be explained by chance as opposed to deterministic reasons.

In current sound analysis research it is common to give estimates of harmonic parameters without indications of their uncertainties. The asymptotic variances of the estimates resulting from the models provides a way to obtain approximate standard errors and to construct approximate confidence intervals for our estimates. It is interesting to speculate on the meaning of these quantities in a musical context. In the following section we discuss a possible interpretation of these statistical variation measures.

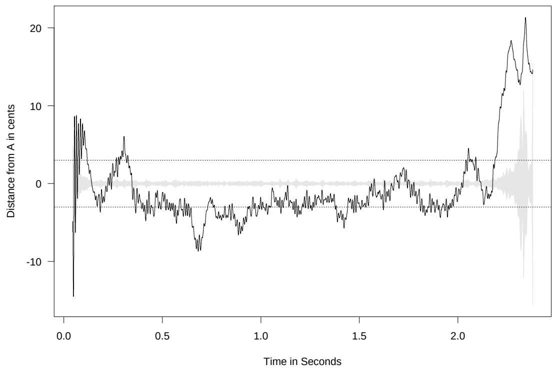

Figure 11. Pitch estimate for trumpet sound and confidence interval around 440 Hz

A recording was made by a professional trumpet player playing (or trying to play) concert pitch A (440 Hz). In Figure 11 we see that the estimated pitch found by our procedure is closer than 3 cents from concert pitch A for most of the signal. Figure 11 also shows approximate 99% confidence intervals around 440 Hz. The figure suggest that for most of the signal the trumpet player is statistically significantly out of tune. Is the statistical variation of our estimates unreasonably "small"? One possible interpretation is the following: In general, the human ear/brain is quite accurate at determining pitch. Suppose our brain "estimates" pitch in a similar manner to our procedure and that the stochastic part of the signal made the variation in this "pitch estimate" large. Then changes in pitch might be detected even when hearing a sound with deterministic constant pitch.

If we consider the estimated pitch found by our procedure ignoring its statistical variability one could conclude that at the beginning and end of the trumpet signal, the trumpet player is more out of tune than during the middle. However, we must also notice that during these sections the standard error is also bigger reflecting the possibility that the larger deviation is due to chance.

This interpretation of variability should be considered with caution. Confidence intervals estimates are constructed using the asymptotic approximation. In our case T is usually between 800 and 2000 observations. Thus there is a possibility that the variance of our estimates might be quite different from the approximation used. Simulation and bootstrap methods can be used to check this. Furthermore, we obtain variance estimates under the assumption of additive noise. For many instruments this assumption appears inappropriate. The noise seems to be signal related and possibly not additive (Maganza and Caussé 1986). Finding variance structures under assumptions like these is a subject of future work.

6.2 Applications to Musical Composition

Our analysis method provides a parametric representation $\hat{\beta}_t$ with estimates of what can be thought of as the fundamental frequency $\hat{\lambda}_t$ and amplitude $\hat{\rho}_{k,t}, k = 1, \dots, K$ for each $t = 1, \dots, T$. We can create new sounds by altering the parameter functions in different ways. In general, we can create a new signal based on the estimates of the original via

$$z_t = \sum_{k=1}^{K} r_k(t) \hat{\rho}_{k,t} \cos\{k\, l(t) \hat{\lambda}_t t\}, t = 1, \dots, T \tag{12}$$

where $r_k(t)$ and $l(t)$ are functions that will permit us to control the change we wish to perform on the original sound. We can now

- Change pitch through the function $l(t)$ (pitch modification).
- Change duration of certain parts of the signal by using $z\{\tau(t)\}$where τ is a time substitution function (Wessel 1987). For example, if $\tau(t) = 2t$, the signal lasts half as long as the original. (time scale modification)
- Change the energy of a specific harmonic through the functions $r_k(t)$ in order to change the timbre of the sound (timbre modification).

Using this technique we have constructed various interesting sound examples. For example, it is said that if one listens to an oboe very closely one can hear the voice of a soprano singing at an octave above the fundamental frequency of the oboe signal. Our analysis provides a way of bringing the hidden soprano out, mainly by adding jitter to the even harmonics. The sound produced by a beginner on a violin may have a "screechy" quality. This is the sound of the bowing that is not converted into a harmonic signal. In this case the noise characterizes the sound as that of a beginner playing the violin as opposed to a professional. This leads us to believe that by "amplifying" the estimated stochastic component of the sound signal played by a professional violin player it will sound like a beginner's violin sound. Again, our analysis permits us to do this. For details on how the reader may acquire and listen to the signals analyzed here, see the Appendix below.

7. CONCLUDING REMARKS

We have presented a statistical procedure that permits us to decompose musical sound signals into locally harmonic and stochastic signals. The procedure provides a parametric representation with musical interpretations which have many practical uses, for example musical composition applications. However this procedure needs not to be restricted to sound signals. The procedure may be useful for analyzing other type of data with approximately periodic behavior, such as biological data following Circadian patterns, functional MRI data, evoke potential data, and others. The procedure would not only provide a smooth version of the data, but a parametric representation that may have useful interpretation. For example a fundamental frequency estimate in heart beat data may be used to assess the condition of the heart. Listening to the residuals in order to detect lack of fit, residual analysis by ear, may also be useful in fields other than musical sound analysis. Hidden periodicities, non-stationarity, and other phenomena may be detected by hearing the data or residuals in situations where large amounts of data are available and visual plots are not quite useful.

APPENDIX: CONVERTING SOUND SIGNALS INTO ASCII FILES

A variety of sound signals were recorded in a recording studio at the Center for New Music and Audio Technology (CNMAT). These recordings were converted into digital signals using an Analog to Digital Converter (ADC). Software synthesis programs generated a sound file as their output (a sound file is simply a data file stored on a disk). After all the samples for a composition are calculated the sound file can be played and heard through a Digital to Analog Converter (DAC). For the details on how sound can be recorded into computer sound files see Road (1996) .

Many different sound file formats exist. The differences between such formats are mainly the rate at which the sound signal was sampled and the number of bits per stored sample. This information is usually contained in the header of such files. The sound files used in this work were sampled at 44.1 kHz and used 16-bit words. The particular format used by the SGI computers, at CNMAT, is the Audio Interface File Format (aiff). The CNMAT aiff files are in stereo (two channels are recorded). For this work only the left channel was used and for these files the left and right channels were practically equal. The aiff sound files used in this work are available via ftp from the CNMAT server `http://www.cnmat.berkeley.edu`.

The sound files can be converted into text files readable to Splus using the sound exchange program `sox` (free software available for UNIX and MS-DOS on the Internet). Similarly, any data set contained in a text file can be converted into a sound signal.

References

Bloomfield, P. (1976). *Fourier Analysis of Time Series: An Introduction.* John Wiley and Son, New York.

Brillinger, D. R. (1977). Comments on “consistent nonparametric regression”. *Annals of Statistics*, 5:622–623.

Brillinger, D. R. and Irizarry, R. A. (1998). An investigation of the second- and higher- order spectra of music. *Signal Processing*, 65:161–179.

Brown, E. N. (1990). A note on the asymptotic distribution of the parameter estimates for the harmonic regression model. *Biometrika*, 77(3):653–656.

Chafe, C. (1990). Pulsed noise in self-sustained oscillations of musical instruments. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 1157–1160, Albuquerque, NM.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610.

Depalle, P., García, G., and Rodet, X. (1993). Analysis of sound for additive synthesis: Tracking of partials using hidden markov models. In *Proceedings of the International Computer Music Conference*, pages 94–97, Waseda University Center for Scholarly Information. International Computer Music Association.

Depalle, P., García, G., and Rodet, X. (1995). The recreation of a castrato voice, Farinelli's voice. In *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 242–245, New Paltz, NY. IEEE.

DePalle, P. and Poirot, G. (1991). SVP: A modular system for analysis, processing and synthesis of sound signals. In *Proceeding of the International Computer Music Conference*, pages 161–164, Montreal, Canada. International Computer Music Association.

Deutsch, D. (1982). *Psychology of Music*. Academic Press, New York.

Dirst, M. and Weigend, A. (1992). Baroque forecasting: On completing J.S. Bach's last fugue. In Weigend, A. S. and Gershenfeld, N. A., editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*, Santa Fe Institution Studies in the Sciences of Complexity, pages 151–172. Addison-Weseley, Reading, MA.

Fletcher, N. H. and Rossing, T. D. (1991). *The Physics of Musical Instruments*. Springer-Verlag, New York.

Green, D. (1985). Temporal factors in psychoacoustics. In Miohelsen, A., editor, *Time Resolution in Auditory Systems*. Springer–Verlay, New York.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbre. *Journal of the Acoustical Society of America*, 62:1270–1277.

Hannan, E. J. (1971). Non-linear time series regression. *Journal of Applied Probability*, 8:767–780.

Hannan, E. J. (1973). The estimation of frequency. *Journal of Applied Probability*, 10:510–519.

Hannan, E. J. (1974). Time series analysis. *IEEE Transactions on Auto Control*, AC-19:706–715.

Hartman, W. M. (1997). *Signals, Sound, and Sensation*. AIP Press, Woodburg, New York.

Irizarry, R. A. (1998). *Statistics and Music: Fitting a Local Harmonic Model to Musical Sound Signals*. PhD thesis, University of California, Berkeley.

Jones, K. (1981). Compositional applications of stochastic processes. *Computer Music Journal*, 5(2):381–396.

Junhar, J. (1997). Advanced pitch detection algorithms. In Kocur, D., Levicky, D., and S., M., editors, *3rd International Conference on Digital Signal Processing*, Kosice, Slovakia. DSP.

Maganza, C. and Caussé, R. (1986). Bifurcations, period doubling and chaos in clarinet-like systems. *Europhysics Letters*, 1(6):295–302.

Masri, P. and Bateman, A. (1996). Improved modelling of attack transient in music analysis-resynthesis. In *Proceedings of the International Computer Music Conference*, pages 100–104, Hong Kong. International Computer Music Association.

Mathews, M. V. (1963). The digital computer as a musical instrument. *Science*, 142(11):553–557.

McAualay, R. J. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions of Acoustics, Speech and Signal Processing*, ASSP-34(4):744–754.

Patterson, R., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (1992). Complex sounds and auditory images. In Cazals, Y., Demany, L., and Horner, K., editors, *Auditory Psychology and Perception*, pages 426–446. Pergamon, Oxford.

Pierce, J. (1992). *The Science of Musical Sound.* Freeman, New York.

Rayleigh, J. (1894). *The Theory of Sound.* Dover, New York. Reprinted 1945.

Risset, J.-C. and Mathews, M. V. (1969). Analysis of musical-instrument tones. *Physics Today*, 22(2):23–30.

Roads, C. (1996). *The Computer Music Tutorial.* The MIT Press, Cambridge, MA.

Rodet, X. (1997). Musical sound signals analysis/synthesis: Sinusoidal+residual and elementary waveform models. In *Proceedings of the IEEE Time-Frequency and Time-Scale Workshop (TFTS'97)*, Coventry, UK. IEEE.

Rowe, R. (1994). *Interactive Music Systems.* The MIT Press, Cambridge, Massachusetts.

Serra, X. J. and Smith, J. O. (1991). Spectral modeling synthesis: A sound analysis/synthesis system based on deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82:559–567.

von Helmholtz, H. (1885). *On the Sensation of Tone.* (trans. A.J. Ellis), London.

Voss, R. F. and Clarke, J. (1975). '1/f noise' in music and speech. *Nature*, 258:317–318.

Walker, A. M. (1971). On the estimation of a harmonic component in a time series with stationary independent residuals. *Biometrika*, 58:21–36.

Wang, X. (1991). An AIC type estimator for the number of cosinusoids. *Journal of Time Series Analysis*, 14(4):433–440.

Wessel, D. (1987). Control of phrasing and articulation in synthesis. In *Proceedings of the International Computer Music Conference*, pages 108–116. International Computer Music Association.