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, Window 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) and 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. Statistical techniques have been used in this field, 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 called sound waves or sound signals traveling through the air and hitting our ear drums. Physical theory and psychoacoustic concepts suggest that

Rafael A. Irizarry is Assistant Professor, Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205. The author thanks David Brillinger, Adrian Freed, and David Wessel for their mentoring; Rick Schoenberg, Scott Zeger, and the Associate Editor for helpful comments; and Ofer Licht and Matt Wright for help obtaining the data.

sound signals produced by *harmonic instruments* contain a component that is approximately periodic (Pierce 1992) and a non-periodic component which may be considered to be stochastic. Although sometimes referred to as noise, the stochastic component is believed to be an integral part of the sound (Chafe 1990). This belief has led sound researchers to use 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 a useful parameterization.

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. The estimation procedure provides a parametric representation with musical interpretations and many practical uses, as described in Section 6. There are accompanying audio versions of some of the examples mentioned in this paper on the author's web-page `http://biosun01.biostat.jhsph.edu/~ririzarr/Demo`.

2. SOUND ANALYSIS AND STATISTICS

For centuries, understanding sound has been of interest. In fact, the discovery of the relation between the lengths of strings on musical instruments and their pitch is commonly attributed to Pythagoras. 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, called *timbre*, permit humans to distinguish one instrument from another (Grey 1977). 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*. The reproduction of musical sounds without the use of an acoustical instrument, called *sound synthesis*, is also of interest. Mathews (1963) was one of the first to successfully make use of the information obtained from sound analysis to produce effective sound synthesis. Recently, interest has focused on using this information to facilitate the creation of new sounds based on an original. 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. 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 to represent them as data. The energy transmitted by a sound signal can be transformed into a fluctuating voltage $V(t)$, which is a continuous function in time. Tape recorders work by storing $V(t)$ on magnetic tape. One wants to have discrete data to facilitate statistical analysis. We take a discrete approximation of the continuous sound signal sampled at 44100 observations per second, as done by Compact Disc (CD) technology.

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 psychoacoustic 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.* We will call the frequency related to this periodicity the *fundamental frequency.* 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 concert pitch A or $A4$ (fourth usable A on a piano). Notes $(A, A\sharp, B, etc..)$ are defined so that the proportion of the frequencies of consecutive notes, said to be a semitone apart, is fixed. Apparently, the trained ear cannot distinguish two notes if they are less than 3 *cents* (1 cent = a hundredth of a tone) apart. See Pierce (1992, Chapter 2) for details.

More recent discoveries have been related to timbre. As early as the second half of the 19th century, physicists were interested in the *harmonic structure* of musical sound signals (Rayleigh 1894). Around this time von Helmholtz (1885) conducted an experiment that proved that signals produced by harmonic instruments had frequency components at multiples of the fundamental frequency, thus showing that the signals contained a strong periodic component with more 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 a wide variety of instruments.

Also in the 19th century, 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). Ohm was perhaps referring to the way the ear operates within very small time windows, and recent psychology and physiological experiments seem to confirm this; see Grey (1977) and Pierce (1992, chapter 7).

Von Helmholtz's discovery suggests that, within small time intervals, sound signals produced by harmonic instruments are periodic and hence 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 *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 ρ_ks and λ.

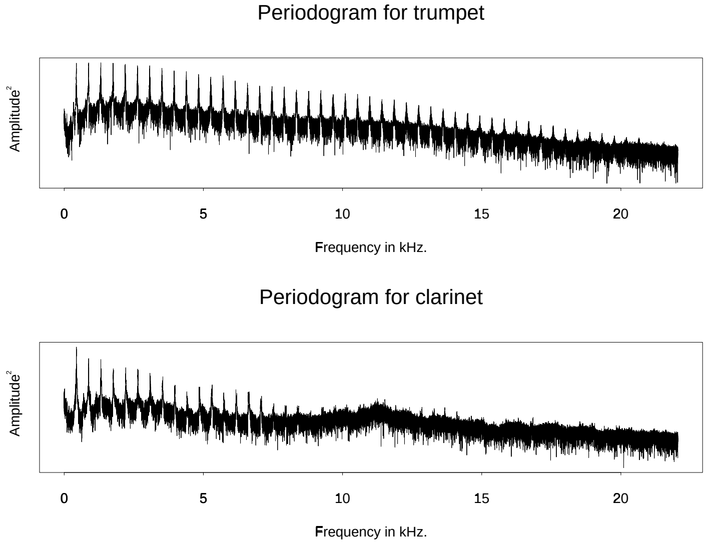


Figure 1. Periodogram for sound signals of a trumpet and a clarinet playing concert pitch A.

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:

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

If the data y_t is periodic as defined in equation (1), the periodogram will show peaks at frequencies $k\lambda$ with $k = 1, \dots, \lfloor \pi/\lambda \rfloor$ (Brillinger 1981). Computed periodograms of sound signals produced by harmonic instruments exhibit such peaks. Figure 1 presents the periodograms for the signals produced by a trumpet and a clarinet playing $A4$. Notice that peaks at the multiples of 440 Hz are observed in both cases and that the two periodograms look different. 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.

In the early 1960s, Risset and Mathews (1969) made a discovery that greatly advanced the understanding of timbre. By using 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 implied that the signals were not exactly periodic which is in agreement with Figure 1. We may verify this from the data by computing *spectrograms*, defined 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.

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 $C4$ (261.6 Hz), while the guitar is playing $D3$ (146.8 Hz). Dark shades of grey represent high power for the spectrogram. In the spectrograms we see that at all times the instruments show peaks at frequencies that are multiples of the fundamental frequency, thus 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. This verifies Risset and Mathews' discovery and suggests that the shape of the periodic component is changing. This is particularly clear in the case of the guitar where the higher

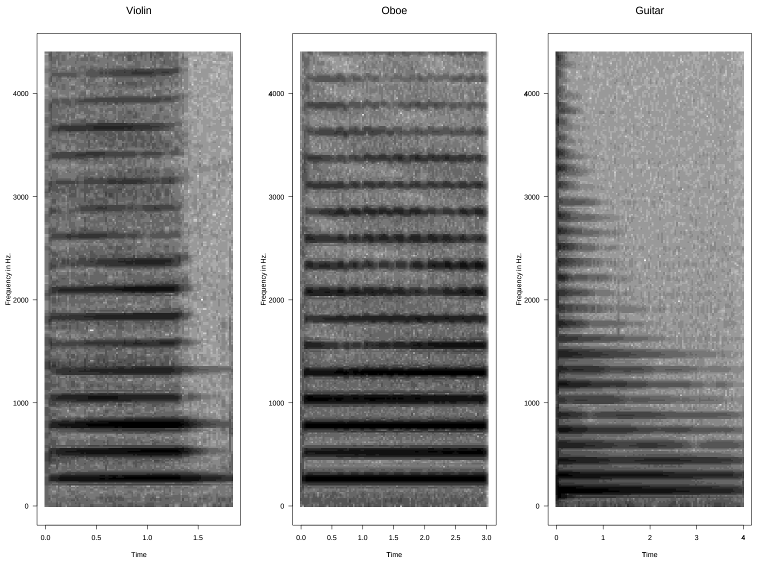


Figure 2. Spectrograms for harmonic instruments with $2M+1=20$ millisecond windows.

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/synthesis is to obtain a *musically meaningful parametric representation* of sound signals. Let $y_t, t=1,\dots,T$ be the discrete representation of a sound signal. We say $\beta_t, t=1,\dots,T$ is a parametric representation if it provides a source for reconstruction or synthesis of the original signal. For example, we could have $y_t \approx s\{t,\beta_t\}$ for some reconstruction function $s(\cdot)$. We say the transformation is musically meaningful if expressive sound transformations can be obtained through the parameters β_t.

The ideas of von Helmholtz and Ohm motivate a method based on the periodogram. To characterize the sound, let $\rho_k = \sqrt{I(k\lambda)}$ and consider the parameterization $\beta = (\lambda, \rho_1, \dots, \rho_k)$. To obtain an approximation of the original sound from the parametric representation, we simply take $\hat{y}_t = \sum_k \rho_k \cos(k\lambda t)$. In this case the parameterization does not change with time, thus it is only 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.

Additive synthesis is a local version of the above approach. This method has proven to be one of the most effective methods available until now (Rodet 1997). In order to obtain a parametric representation for sound signals produced by an instrument playing one note an appropriate λ is chosen, then the signal is divided into $L = T/(2M+1)$ non-overlapping segments of suitable size $(2M+1)$ and the spectrogram is used to define $\rho_k(t) = \sqrt{I(t;k\lambda)}$ for all t in the lth segment with $k = 1,\dots,K$ (K chosen heuristically) and $I(t;k\lambda)$ defined as in (3). The parametric representation is then $\beta_t = \{\lambda, \rho_1(t),\dots,\rho_K(t)\}$. 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 von Helmholtz. Yet the "fitted" signal is never the same as the original. Possible reasons for this are added components that are not periodic in nature and which we will refer to as the stochastic component. Some examples are the sounds produced by fingers hitting keys, nails plucking strings, and surplus blown air. Some researchers have found 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 then, many have proposed and used similar models (Rodet 1997). In Serra and Smith (1991) the deterministic plus stochastic signal statistical model presented is

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

with $\{\epsilon_t\}$ a stationary autoregressive process. An implicit assumption is that the deterministic signal resembles a sum of sinusoids. In this case $\beta_t = \{\rho_1(t),\phi_1(t),\dots,\rho_K(t),\phi_K(t)\}'$ 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 which in this model need not be multiple of a fundamental frequency.

One is interested in estimating the $\phi_k(t)$s and $\rho_k(t)$s. Serra and Smith (1991) divide the signal into short (usually 256 data points), possibly overlapping segments called *analysis frames.* For each segment, the amplitude and frequency of peaks of the periodogram are recorded and considered as a possible indication of a *sinusoidal partial.* 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. 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. 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. 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, García, and Rodet 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 countertenor 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 suggest that for data $y_t, t = 1, \dots, T$, obtained from a short segment of a sound signal, a useful deterministic plus stochastic signal model is

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

$\beta = (A_1, B_1, \dots, A_K, B_K, \lambda)'$ and $\{\epsilon_t\}$ a stochastic component. In a musical context 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 1997). In Hannan (1973) 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 (2000) 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 paper 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.

In general, we allow the stationary noise $\{\epsilon_t\}$ to be correlated. However, we are using an estimation procedure commonly used for uncorrelated noise. Under general assumptions for $\{\epsilon_t\}$, Irizarry (2000) shows that the weighted least squares estimates have desirable asymptotic properties and provides the following approximations for the variances of the fundamental frequency and amplitude estimates:

$$\mathrm{Var}(\hat{\lambda}) \approx 4\pi T^{-3} \left\{ \frac{W_0^2 U_2 - 2W_0 W_1 U_1 + W_1^2 U_0}{(W_0 W_2 - W_1^2)^2} \right\} \left\{ \sum_{k=1}^{K} k^2 \rho_k^2 / f_{\epsilon\epsilon}(k\lambda) \right\}^{-1} \tag{8}$$

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

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 $f_{\epsilon\epsilon}(\omega) = \frac{1}{2\pi} \sum_u c_{\epsilon\epsilon} \exp\{-i\omega u\}, \ -\infty < \omega < \infty$ with $c_{\epsilon\epsilon}(u) = \mathrm{Cov}\{\epsilon_{t+u}, \epsilon_t\}$, the autocovariance function of $\{\epsilon_t\}$. If we use this asymptotic approximation to find standard errors and confidence intervals for our estimates, we need to estimate $f_{\epsilon\epsilon}(k\lambda)$ for $k = 1, \dots, K$. We may use the residuals to compute an estimate, for example, by using a *smoothed periodogram* (Brillinger 1981).

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. For the sound signal of a violin, sampled at 44.1 kHz, playing the note $C4$, we consider a 50 millisecond stretch (2200 observations). In Figure 3 we see how the segment seems to be usefully periodic. We define weights with Tukey's triweight function and use the weighted least squares procedure, described in the previous section, to fit a harmonic model with K=15 to this data. The residual mean-square, which is 0.0003, is quite small compared to the total variance, which is 0.26, so it seems to be a reasonable fit. The residual plot, also seen in Figure 3, suggests that the noise could be considered stationary in the given stretch.

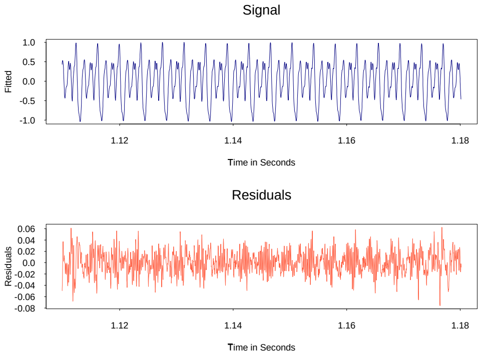


Figure 3. Local fit for the sound signal of a violin playing C4 and corresponding residuals. Smoothed periodogram estimates for the spectrum of the noise for the sound signal of a violin playing C4.

We 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 parameterizing 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 changes, vibrato, and tremolo, to mention a few. For this reason the stationary model described in Section 3 is not appropriate when considering long stretches of musical signals. 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), is to assume that for "small" segments $(t_0 - h, t_0 + h)$ the deterministic component is approximately periodic and the noise component is approximately stationary. We can write

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

and $\beta_0 = (A_{1,0}, B_{1,0}, \dots, A_{K,0}, B_{k,0}, \lambda_0)$ 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 $\hat{\mu}(t_0) = s(t_0, \hat{\beta}_0)$. 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 the estimates obtained for each t_0, namely $\hat{\beta}_t, t = 1, \ldots, T$. This provides a musically meaningful parametric representation of the signal since $\hat{\beta}_t$ contains an estimate of the local fundamental frequency $\hat{\lambda}_t$ and amplitudes $\hat{\rho}_{k,t} = \{\hat{A}^2_{k,t} + \hat{B}^2_{k,t}\}^{1/2}, k = 1, \ldots, K$.

The residuals may be defined via $\hat{\epsilon}_t = y_t - \hat{\mu}(t)$, $t = 1, \ldots, T$ and studied to assess the fit. For example, by converting them into a sound file, we may listen to the residuals and perform residual analysis by ear.

4.1 An Example

We run the analysis on the sound signal of an oboe playing $C4$ for a duration of 3 seconds. Listening to this sound we notice that the oboist is playing *vibrato* and *tremolo*, slight and rapid variations in volume and pitch respectively. We obtain estimates of $\mu(t_0)$ using $K = 15$ and 20 millisecond estimation windows for each $t_0 = 1, \ldots, T$. In Figure 4 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 known fact that the presence of noise components in a sound signal are stronger during the beginning of a note, or what musicians refer to as the *attack* (Masri and Bateman 1996). If we listen to the residuals we hear what sounds like the initial burst of air blown by the instrumentalist.

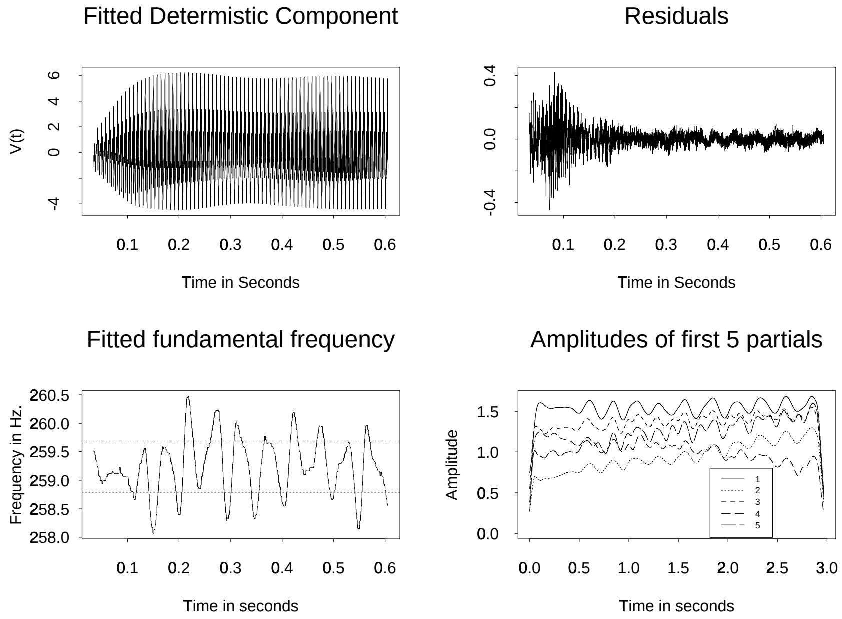


Figure 4. Estimated fundamental frequency and amplitude of first five partials for the sound signal of an oboe playing $C4$.

In Figure 4 we also see $\hat{\lambda}(t)$ and $\hat{\rho}_k(t)$ for $k = 1, \ldots, 5$. The dotted lines in Figure 4 are 3 cents away from the average fundamental frequency (259.25 Hz), which would be considered a $C4$. The Figure seems to suggest that there are variations in pitch perceivable to the ear and that there are variations in the amplitudes of the first five partials. This is in agreement with what we hear: vibrato and tremolo. Here, we are using the fact that the approximate standard errors (not shown) are small compared to the range of the estimates. In Section 6.1 we further discuss standard errors and their musical interpretation.

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 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.

5. DYNAMIC WINDOW SIZE AND NUMBER OF HARMONICS SELECTION

For a particular sound signal a variety of 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 is gradual changes in volume. Such phenomena suggest that the estimation window sizes should not remain fixed and that we need to choose an appropriate window size h for all estimation times t_0.

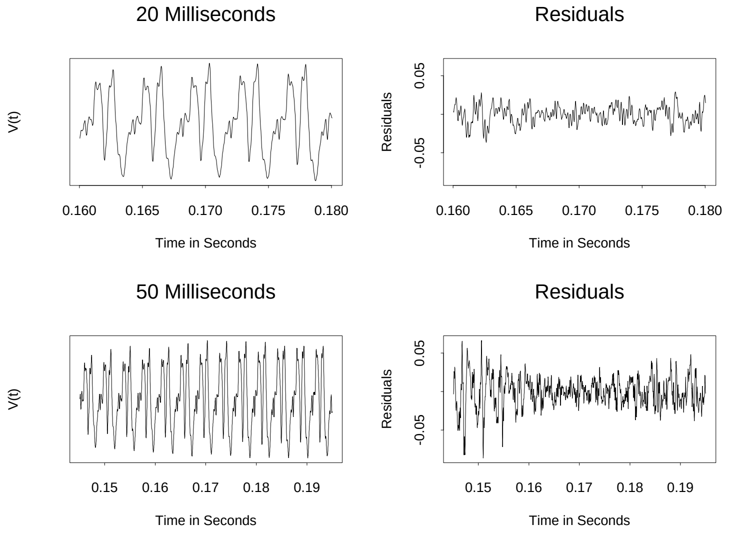


Figure 5. Comparison of two stretches of different duration of the sound signal of a violin playing $C4$.

In Figure 5 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 5, produced from fitting a harmonic model with $K = 15$, 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.

Also, for different segments the number of harmonic components that seem meaningful varies, as seen for the guitar in Figure 2. Previous estimation procedures (Depalle, García, and Rodet 1993) usually fit many sinusoidal components. However, fitting too many parameters may result in a saturated model. A decision that we need to make is how many harmonic components K to consider when estimating. The number of "significant peaks" in the periodogram plot may be used to obtain a general idea of how many harmonics to consider. However, this procedure is quite arbitrary. We may use z-tests to reduce the value of K in situations where many $\hat{\rho}_k$ are not "statistically significant" for the larger value of k. Yet this strategy is also a bit arbitrary so we do not intend to use hypothesis testing as a tool for choosing how many harmonics to include in our model. In any case, we have found it to be useful as a descriptive illustration of why considering different number of harmonics for different sound signals may be appropriate.

5.1 Information Criteria

Due to the large amount of data and points of estimation, using procedures like cross-validation to choose h and K is unrealistic (at least in 2000). We want a criteria that will permit us to automatically choose from amongst different possible estimates. Irizarry (1999) presents useful criteria for model selection for this situation. 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 a useful criterion is

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

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 the WBIC criterion satisfies the conditions for consistent model selection criteria for harmonic models presented in Wang (1991). Furthermore, Irizarry (1999) presents simulations showing that this criteria performs relatively well at choosing window sizes with small MSE.

5.2 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 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.

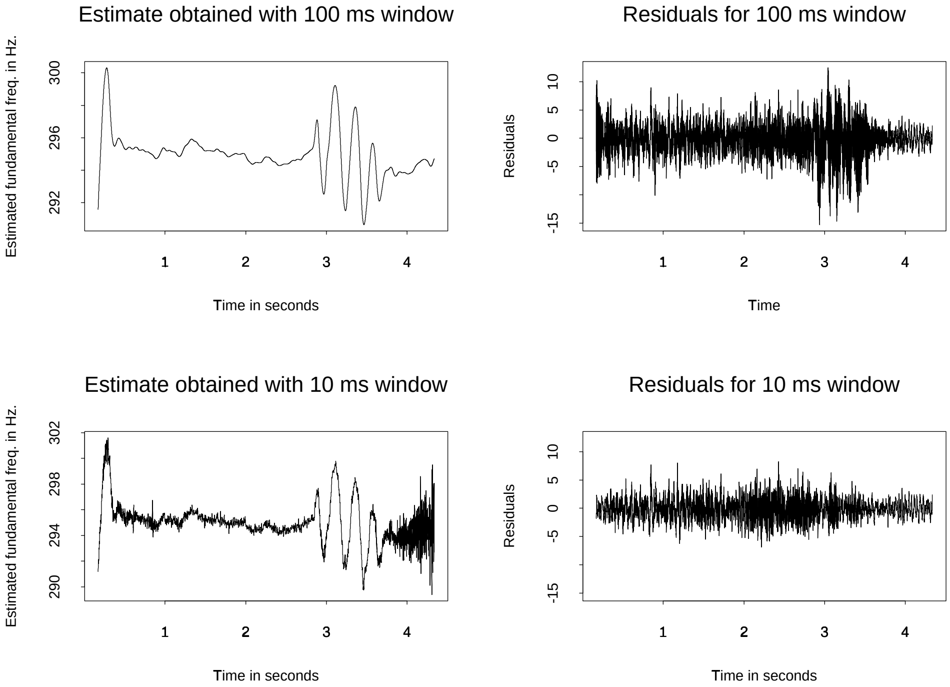
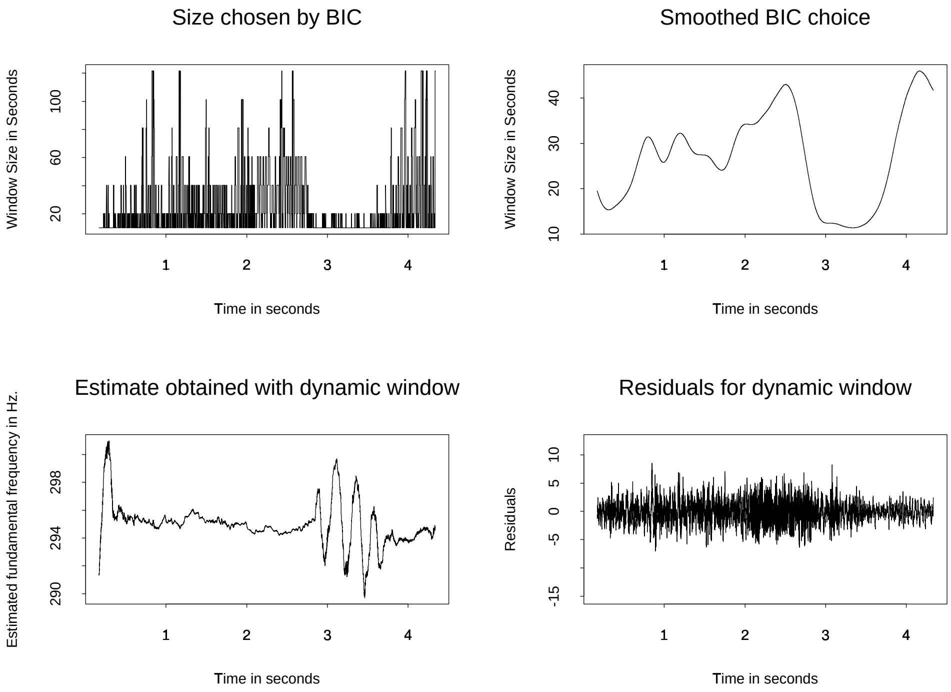


Figure 6. Estimated fundamental frequency when using fixed window sizes of 100 and 10 milliseconds and the respective residuals.

This particular sound is interesting to study for two reasons. First, the noisy character of the sound makes the partial tracking techniques, described in Section 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, the 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 6 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 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 6, especially towards the end.


Size chosen by BIC

Smoothed BIC choice

Estimate obtained with dynamic window

Residuals for dynamic window

Figure 7. Selected window size (in milliseconds) and smoothed version of the selected windows; estimated fundamental frequency when using a dynamic window size and respective residuals.

We fitted a local harmonic model with $K = 15$ using the WBIC to choose between window sizes of 10, 20, 40, 60, 80, 100, and 120 milliseconds. In Figure 7 we see how the window size that minimizes the WBIC vary as the signal progresses. Notice in particular how, on average, smaller window sizes are chosen during the parts of the signal where the fundamental frequency is not near constant. Because the WBIC choices are so variable, we compute a smoothed version of window sizes by using a kernel smoother, also seen in Figure 7. 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 7. We see the improvement this provides over the fit obtained with the 100 millisecond window procedure by comparing the residuals, seen in Figures 6 and 7. We see the improvement it provides over the fit obtained with a 10 millisecond window by comparing the estimated fundamental frequencies, also seen in Figures 6 and 7.

6. APPLICATIONS

6.1 Variability of the Estimates

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 presented in Section 3 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.

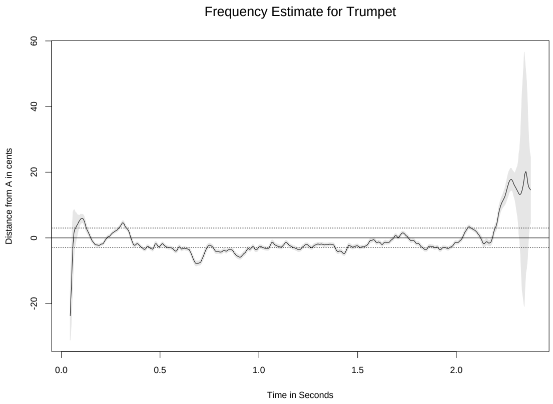


Figure 8. Pitch estimate for trumpet sound and confidence intervals.

A recording was made by a professional trumpet player playing (or trying to play) concert pitch A (440 Hz). In Figure 8 we see approximate point-wise 99% confidence intervals around our estimated pitch. The figure suggests that for most of the signal the trumpet player is “statistically significantly out of tune”. However, if we hear the signal it seems to be in tune. Is the statistical variation of our estimates unreasonably “small”? In Figure 8 we also see that for most of the signal the distance between the estimate and 440 Hz is not “statistically significantly” bigger than 3 cents, so our estimates are actually in agreement with what we hear. The trumpet player may not be playing an exact 440 A, but it is close enough for our ears not to perceive the difference. Notice that if we consider the estimated pitch ignoring its statistical variability, one could conclude that at the beginning and at the end the trumpet player is more out of tune than during the middle. However, during these sections the standard error is also bigger, reflecting the possibility that the larger deviation is due to chance.

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.

This interpretation of variability should be considered with caution since confidence intervals estimates are constructed using asymptotic approximations. In our case, T is usually between 800 and 2000 observations; hence there is a possibility that the variances of our estimates are larger than the approximations used. Also, the approximations made by our model for the deterministic part of the signal may not be as good as we would want them to be. 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, since 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. A possible approach is to follow the ideas presented in the literature on time series models with time-varying parameters, for example the approach followed by West, Prado, and Krystal (1999).

6.2 Applications to Musical Composition

We may use the parametric representation $\hat{\beta}_t$ provided by our method to create new sounds. In general, we can create a new signal based on the estimates of the original via $z_t = \sum_{k=1}^{K} r_k(t)\hat{\rho}_{k,t} \cos\{k\, l(t)\hat{\lambda}_t\, \tau(t)\}$. We can now: change pitch through the function $l(t)$ (pitch modification), change duration of certain parts of the signal using a time substitution function (Wessel 1987) $\tau(t)$ (time scale modification), and 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 which are all available on the aforementioned web page. For example, our analysis provides a way of bringing the hidden soprano out of an oboe sound, mainly by adding jitter to the even harmonics as well as a way to turn the sound produced by a professional violin player into a beginner's by "amplifying" the estimated stochastic component of the sound signal.

7. CONCLUSIONS AND EXTENSIONS

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 has many practical uses, such as in musical composition applications. A possible extension to the procedure is to represent the time-varying coefficients parametrically. For example, we could model amplitudes of the harmonics of instruments like the guitar with a decaying exponential. This would permit us to consider larger window sizes. Furthermore, we could fit flexible models through the use of splines and fit them through a single optimization.

The procedure presented in this paper need not be restricted to sound signals. The procedure may be useful for analyzing other types of data with approximately periodic behavior, such as biological data following Circadian patterns and EEG data. The procedure would not only provide a smooth version of the data, but also a parametric representation that may have useful interpretation. 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 data or residuals in situations where large amounts of data are available and visual plots are not quite useful.

REFERENCES

Brillinger, D. R. (1977), Comments on "Consistent nonparametric regression," *Annals of Statistics,* 5, 622–623.

——— (1981), *Time Series Data Analysis and Theory*, San Francisco: Holden–Day.

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, 653–656.

Chafe, C. (1990), "Pulsed noise in self-sustained oscillations of musical instruments," in *IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 1157–1160.

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*, pp. 94–97.

——— (1995), "The recreation of a castrato voice, Farinelli's voice," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 242–245.

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

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

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

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

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

Irizarry, R. A. (1999), "Information and posterior probability criteria for model selection in local likelihood estimation," Technical Report MS99-05, Johns Hopkins University, Department of Biostatistics.

——— (2000), "Asymptotic distribution of estimates for a time-varying parameter in a harmonic model with multiple fundamentals," *Statistica Sinica,* (to appear)

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

Maganza, C. and Caussé, R. (1986), "Bifurcations, period doubling and chaos in clarinet-like systems," *Europhysics Letters*, 1, 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*, pp. 100–104.

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

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

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

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

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.*

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, 12–24.

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.

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

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

West, M., Prado, R., and Krystal, A. (1999), "Evaluation and comparison of EEG traces: Latent structure in non-stationary time series," *Journal of the American Statistical Association*, 94, 1083–1095.