The Additive Sinusoidal Plus Residual Model: A Statistical Analysis

Rafael A. Irizarry
Department of Statistics and The Center for New Music and Audio Technologies (CNMAT)
University of California Berkeley, CA 94720
rafa@stat.berkeley.edu

Abstract

Many musical instruments' sounds can be represented by harmonic and additive noise signals. We are interested in separating these two elements of the sound. We fit a local harmonic model that tracks changes in pitch and of the amplitude of the harmonics. Deterministic changes in the signal suggest that different window sizes should be considered. Various ways to choose appropriate window sizes are studied.

1 Introduction

In this work we will set down a statistical model for sound signals produced by harmonic instruments based on previous work in sound synthesis. A summary of the previous work is given in section 2. The estimation procedure, based on weighted least square estimates within small time windows, is described in Section 3. Ways to choose amongst different window sizes to be used in the estimation and number of harmonics included in the model are presented in Section 4. Section 5 presents practical uses of the statistical methods.

2 Sound Analysis and Synthesis

Some of the first attempts at sound synthesis were based on *additive synthesis* (Risset and Mathews 1969). This has proven to be one of the most effective methods available until now (Rodet 1997). Sound signals are modeled as summations of time-varying sinusoidal components. Serra (1989) incorporated a non-sinusoidal residual part to the additive synthesis and modeled it as an additive random signal. Notice that under this assumption one is dealing with a signal plus noise statistical model.

$$y(t) = s\left[t;\beta(t)\right] + \epsilon(t), \text{ with } s\left[t;\beta(t)\right] = \sum_{k=1}^{K} a_k \cos(\phi_k(t)) \tag{1}$$

with $\beta(t) = (a_1(t), \ldots, a_K(t), \phi_1(t), \ldots, \phi_K(t))'$. An implicit assumption is that the signal $s[t;\beta]$ resembles a sum of pure sinusoids. Serra (1989) assumes the non-sinusoidal part, $\epsilon(t)$, to be a stationary autoregressive process.

One is interested in estimating $\beta(t)$. Estimation is done in two steps. In the first step the signal is divided into short, possibly overlapping segments, called *analysis frames*. For each segment, peaks of the periodogram of the tapered data are considered to be a possible indication of a sinusoidal partial. In the second step, referred to as *partial tracking*, peaks of successive analysis frames are grouped into *tracks*. For a particular track, say the k-th track, the frequencies at which the peaks occur are considered to be estimates of the sinusoidal partial associated with $\phi_k(t)$. See Rodet (1997) for details.

Notice that we assume the existence of deterministic sinusoidal components (the partials). The strong peaks seen in the periodogram of signals produced by harmonic instruments support this assumption. Partial tracking algorithms allow partials to exist at frequencies that are not multiples of a fundamental frequencies. In the case of harmonic instruments, estimates obtained for the deterministic sinusoidal signal when many non-harmonic partials are "tracked" are hard to interpret. Furthermore, the periodogram of a signal that is assumed to be stochastic will also be stochastic. A peak in the periodogram can be due to chance. Computing the statistical properties of periodogram peaks found by a tracking algorithm as the one presented in Rodet (1997) can be complicated. For this reason finding an algorithm for partial tracking that provides useful statistical estimates is not straightforward. We do not intend to search for such an algorithm. Instead, for the case of signals produced by harmonic instruments, we assume a harmonic version of the model defined by equation (1) and present an estimation procedure that provides estimates with "desirable" asymptotic properties (Irizarry 1998). These do appear useful in practice: not only is the risk of incorporating unwanted partial tracks reduced, but also the accuracy of our estimates may improve, as will be shown in Section 6.

3 Local Harmonic Model

For sound signals produced by harmonic instruments we propose using a model similar to the one in equation (1).

$$y(t) = s\left[t, \beta(t)\right] + \epsilon(t) \text{ with } s\left[t; \beta(t)\right] = \sum_{k=1}^{K} \left\{A_k(t)\cos(k\lambda(t)t) + B_k(t)\sin(k\lambda(t)t)\right\} \quad (2)$$

Here $\beta(t) = (A_1(t), B_1(t), \ldots, A_K(t), B_K(t), \lambda(t))'$. Two differences are that we will impose the constraint that the frequency of the partials are all multiples of a fundamental frequency and that the stochastic non-sinusoidal part will be allowed to be *locally stationary* (Dahlhaus 1997). Assuming stationary noise does not seem appropriate for sound signals. The local stationarity assumption provides a way to synthesize the non-sinusoidal part of the signal via simulations taking the non-stationarity into account. We will assume that the signal $s[t; \beta(t)]$ is *locally approximately sinusoidal*. Precise definitions and further assumptions needed for the asymptotic theory are given by Irizarry (1998).

We want to estimate $\beta(t)$ for all $t \in [0, 1]$. We will describe how we will find an estimate $\hat{\beta}(t_0)$ for any $t_0 \in [0, 1]$. Consider a small enough segment, say h time units long, of the signal around t_0 so that one is able to assume that the signal is *approximately sinusoidal* within that segment. We will call the interval $(t_0 - h/2, t_0 + h/2)$ the *estimation window*.

Now assume that the parameters are constant in time within the estimation window, i.e. $\beta(t) = \beta(t_0)$ for $t \in [t_0 - h/2, t_0 + h/2]$. Estimate $\beta(t_0)$ by finding the value $\hat{\beta}$ that minimizes the weighted least squares equation:

$$S(\beta) = \sum_t w(t)\ \left(y(t) - s[t, \beta]\right)^2$$

Here the summation is over the times of each sample and the $w(t)$ is an appropriate weight function with support in $[t_0 - h/2, t_0 + h/2]$. By repeating this procedure for each t_0 we end up with an estimate $\hat{\beta}(t)$ of the function $\beta(t)$.

Many authors have studied the properties of estimates like these in the stationary equal weighted case (Hannan 1973). Estimates are shown to be consistent and asymptotic variance expressions are developed. The non-stationary and weighted case is presented by Irizarry (1998).

4 Dynamic Window Size Selection

For a particular sound signal a variety of deterministic factors may affect the smoothness of the parameter function $\beta(t)$. For example, a change in note creates a discontinuity in the fundamental frequency function $\lambda(t)$. Such phenomena suggest that h should not remain fixed for all $t \in [0, 1]$.

We want a criterion, based on the information provided by the data, that will permit us to choose amongst estimates obtained using different spans h_q and different number of parameters in the model $p = 2K + 1$.

Similar to criteria used to choose among competing statistical models, e.g., Akaike's Information Criteria (AIC) (Akaike 1973) and the Bayesian version (BIC) (Schwarz 1978), we develop the wBIC criterion (Irizarry 1998). The criteria is based on the weighted residual mean-square error but, because larger values of p and smaller values of h will tend to have smaller weighted residual mean-square, we add a penalty term.

$$\text{wBIC}(p, q) = N \log \hat{\sigma}^2_{p,q} + (V_{p,q}/W_0) \log N \quad (3)$$

where $\hat{\sigma}^2_{p,q}$ is the weighted mean-square error when estimating with p parameters and window span h_q, N is the total number of observations considered in the largest window, and

$$V_{p,q} \approx p\frac{U_0}{W_0} + \frac{2W_1[W_1U_0/W_0 - U_1] + W_0U_2 - W_2U_2}{W_2W_0 - W_1^2} \text{ with } W_n = \int t^n w(t)\, dt \quad U_n = \int t^n w(t)^2\, dt$$

Here $w(t)$ is the weight function used in the estimation. The second term on the right of equation (3) is a *penalty term* for large number of parameters and small window sizes. We pick the estimate that minimizes the wBIC.

5 Examples

Once we find an estimate $\hat{\beta}(t)$ we can construct estimates for the harmonic part of the signal with $s[t, \hat{\beta}(t)]$. The estimate of the non-sinusoidal signal is $\hat{\epsilon}(t) = s[t, \hat{\beta}(t)] - y(t)$. Sound examples 1,2, and 3 present the original, fitted, and residual signals for a violin. Examples 4, 5, and 6 are for a guitar.

For our estimation procedure to make sense we need to consider appropriate segments of sound signals. In the examples presented in this section, the wBIC of equation (3) is used to decide automatically how many partials to use in our model and how big a window size to consider for the estimation.

In Figure 1, we present a contour plot of the value of the wBIC obtained when fitting the model to a stretches of a violin playing C5 and C7, and stretches near the beginning and end of a guitar playing D3. Notice that a smaller number of harmonic is chosen for the higher pitch violin signal, and for the end of the guitar signal. Apparently higher harmonics die off faster in the guitar sound. The window size chosen for the violin is around 20 milliseconds. For the guitar it's around 60. This information may be useful in practice. Existing additive synthesis methods track many partials. Fitting only 15 will make estimation procedures faster and might possibly even provide more accurate estimates. Also, it provides an automatic way of choosing a window size.

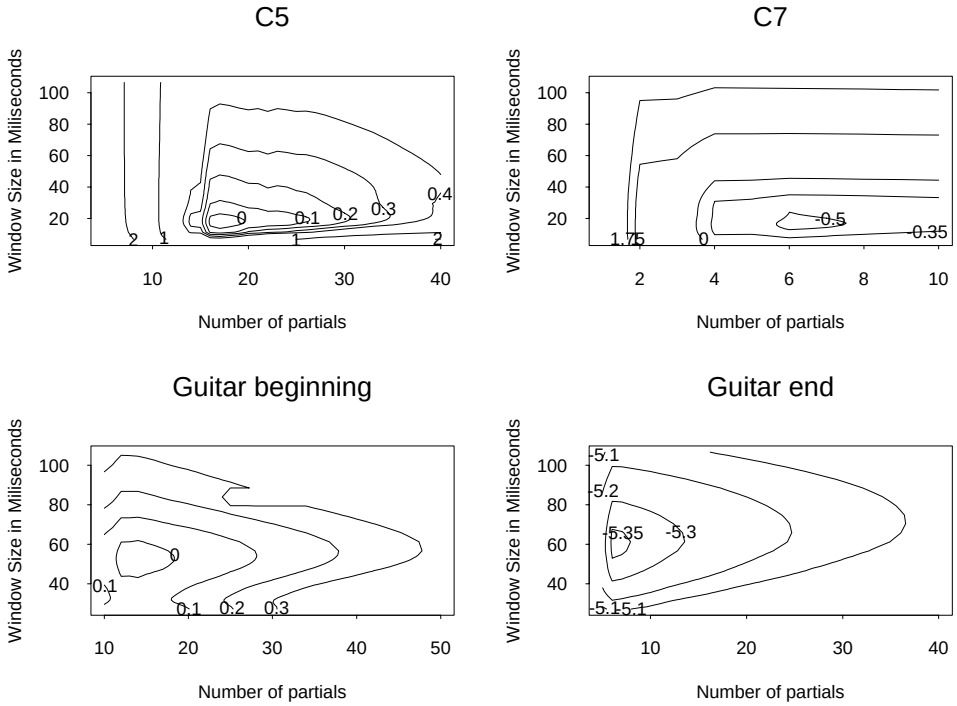

Figure 1: Contour plots of the wBIC for a violin playing c5 and C7 and for the beginning and end of guitar signal.

We next illustrate that dynamic window selection can improve our procedure via an example. The shakuhachi flute is a Japanese instrument characterized as being "noisy". The sound studied (sound example 7) contains a rapid change of pitch for the first 0.5 seconds, then the pitch is held steady for about 3.5 seconds, then a vibrato is played for about 0.5 seconds after which the pitch is held fixed again. The different behavior of the pitch function in different parts of the signal suggests that a fixed, large window size is inappropriate and thus that different window sizes should be used in different parts of the signal.

We fitted our model using a fixed window size of about 20 milliseconds, and the dynamic window procedure using the wBIC to choose between window sizes. In Figure 2 we see how the latter procedure, on average, chooses smaller window sizes during the parts of the signal where $\lambda(t)$ is not near constant. Finally, we notice the improvement of the dynamic window method by comparing the residual plots, also seen in Figure 2, of the fixed window (sound example 8) and dynamic window procedures (sound example 9).

In current sound analysis research it is common to give estimates of sinusoidal parameters without indications of their uncertainties. The asymptotic variances of the estimates resulting from the models provides a way to obtain standard errors for our estimates. For some applications and the details see Irizarry (1998). A particular application is to asses the possible advantage of the harmonic model over the model defined by (1) when harmonic constraint $\omega_k = k\lambda$ is not used. If we assume the harmonic model is true, then for each k, both estimators $\hat{\omega}_k$ and $k\hat{\lambda}$ are consistent estimates of the frequency related to the k-th partial $k\lambda$. The advantage of the latter is that it has smaller asymptotic variance. In Irizarry (1998) expressions are given for the variance of the estimates. We can use this to estimate the relative efficiency $\text{EFR}_k^2 = \text{var}(\hat{\omega}_k)/\text{var}(k\hat{\lambda})$. For one example we obtain$\text{EFR}_1 \approx 6.4$.

An example of analyzing a pipe organ sound with reverberation by fitting a model like (2) but with two fundamental frequency, can be found in Irizarry (1998). Sound examples 10, and 11 are the signals for a pipe organ with

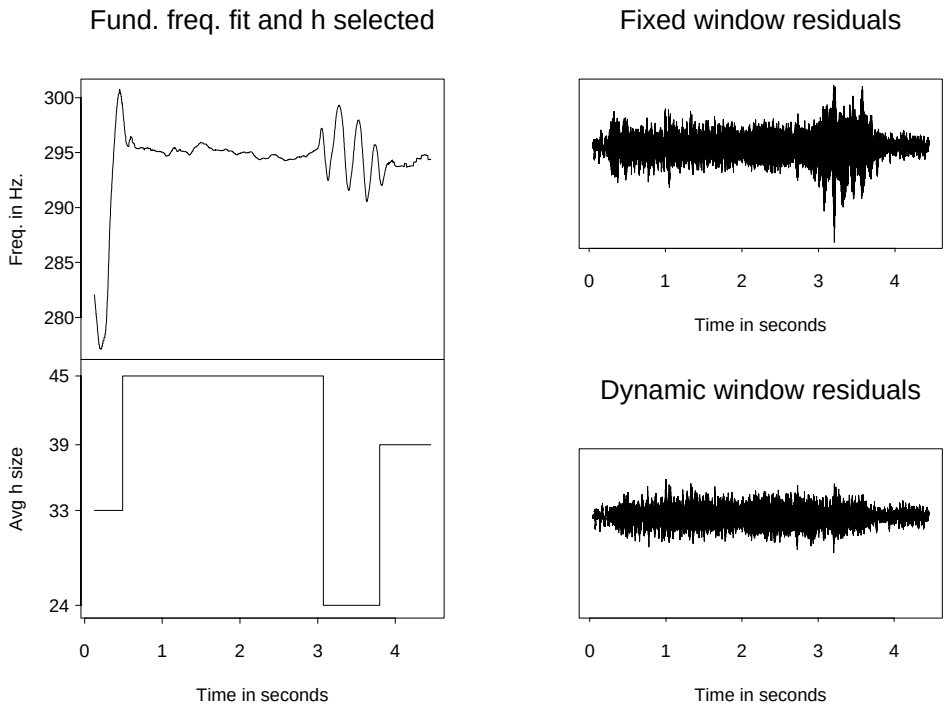

Figure 2: Comparison of fixed and dynamic window procedures.

reverberation and the residuals when fitting a one-fundamental model. Sound example 12 are the residuals using a two-fundamentals model on the part of the sound corresponding to the second note.

6 Conclusion

We have seen how musical sound signals can be be modeled with a statistical local harmonic model. The estimates obtained when fitting the model provide a useful parametric representation. The asymptotic variances of the estimates resulting from the models provide a way to obtain standard errors for our estimates. The wBIC provides a way to choose appropriate window sizes as well as the number of parameters in the model. Estimates found using the dynamic window procedure seem to provide some improvement over the fixed window procedure estimate.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *in* B. Petrov and B. Csaki (eds), *Second International Symposium on Information Theory*, Academiai Kiado, Budapest, pp. 267–281.

Dahlhaus, R. (1997). *Maximum Likelihood Estimation and Model Selection for Locally Stationary Processes*, Institute of Statistical Science Academia Sinica, Heidelberg, Germany.

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

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

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

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

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics* **6**(2): 461–464.

Serra, X. (1989). *A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition*, PhD thesis, Stanford University.