An Investigation of the Second-and Higher-order Spectra of Music

David R. Brillinger and Rafael A. Irizarry
Statistics Department University of California Berkeley, CA 94720

Abstract

For a variety of musical pieces the following questions are addressed: Are the power spectra of 1/f form? Are the processes Gaussian? Are the higher-order spectra of 1/f form? Are the processes linear? Is long-range dependence present? Both score and acoustical signal representations of music are discussed and considered. Parametric forms are fit to sample spectra. Approximate distributions of the quantities computed are basic to drawing inferences. In summary, 1/f seems to be a reasonable approximation to the overall spectra of a number of pieces selected to be representative of a broad population. The checks for Gaussianity, really for bispectrum 0, in each case reject that hypothesis. The checks for linearity, really for constant bicoherence, reject that hypothesis in the case of the instantaneous power of the acoustical signal but not for the zero crossings of the signal or the score representation.

KEYWORDS: *Bicoherence, Bispectrum, Linear Process, Music, Parametric Models, Spectral Analysis*

1 Introduction

What is music? Probably nobody will ever give a final answer to this question, but something inside of us tells us when a sound we hear is music and when it is not. Most people hear the sound of cars passing by on a road and don't think it is music, but it only takes them a moment after a radio is turned on to identify the sound coming out as music. Certain sounds we classify as music others we do not. In this paper we examine some statistical properties of two different numerical representations of music to see if we can shine some light on the property that makes music, music.

We are able to process music in a data analytic fashion because the time is at hand when music can be treated directly as data to be analyzed by contemporary statistical procedures and packages.

The paper begins with a description of the two basic representations of music, moves on to some review of previous investigations, then presents the results of modeling the second-order spectra and finally employs higher-order spectra to assess Gaussianity and linearity.

The pieces investigated included: Baroque, Classical, Romantic, Atonal, Spanish Guitar, Jazz, Latin, Rock & Roll and Hip Hop.

2 Representations of Music

Certainly music is sound. Every sound we hear is the consequence of pressure fluctuations traveling through the air and hitting our ear drums. The signal representation takes this property of sound to represent music as a continuous function.

For years composers have transcribed the music they hear in their heads using what is known as common practice notation (CPN). We use such "numerical" representations of music for our analyses.

2.1 Signal Representation

The function describing the audible pressure fluctuations of air is called a "sound wave". The energy transmitted by this "sound wave" can be transformed into a voltage $Y(t)$, which will be a continuous function in time. Compact Disks are proof of how effective quantized samples of this function are. This time series $Y(t), 0 < t < T$, will be called the *signal representation* of music. Throughout this paper we will be using a discrete version of the function, $Y_t, t = 0, 1, ...$

When such fluctuations of air are approximately periodic we hear a sound with a definite musical pitch. Instruments play different pitches by changing the fundamental frequency of the "sound wave" they are creating, Pierce [15]. Some cultures, e.g. Western cultures, have quantized these pitches and created "notes". This has permitted composers to write with a notation that an instrumentalist can then turn into sounds. This notation provides the other representation of music, the *score representation.*

2.2 Score Representation

Most instruments known to us have the capability to play different "notes". In all "melodic" instruments, for example violins, pianos, trumpets, sitars, etc., as mentioned above different notes correspond to different fundamental frequencies or pitches. The pitch corresponding to 440 Hz has been called A (concert pitch A). Any frequency that holds a 2^n:1 relation with A is also called A, but in another octave. Western music uses the 12 tone equal-tempered scale in which the frequencies between, say 440Hz (concert pitch A) and 880Hz (an octave above concert pitch A) have been divided into 12 notes corresponding to frequencies with the same ratio between them. These 12 notes are $A, A\sharp$ (A sharp) , $B, C, C\sharp, D, D\sharp, E, F, F\sharp, G, G\sharp$ and that brings us back to A (an octave above). If you look at a piano the black keys correspond to the sharps and you will see a twelve white-black keys pattern repeating 7 times. Adjacent notes are said to be a half-step apart or a semi-tone away, see Pierce [15].

The human audible range can hear about 4 octaves below concert pitch A and about 6 octaves above (this is for the keenest of ears) . This means that there are about 100 notes that we can hear. Western composers have found a universal way of representing these notes, namely, what is known as common practice notation (CPN). Probably most if not all sheet music you have seen uses this notation. With this notation a composer tells a performer what pitch his instrument should play. Representing notes as numbers is now straightforward. The MIDI standard (see more detail below) assigns to concert pitch A the number 69 and for every adjacent note adds or subtracts one.

To transcribe a melody we also need the rhythm. CPN also provides symbols to denote how long each note is going to be played and also for how long nothing will be played (rests). In Western music the time domain is divided into measures and beats and into subbeats. For any given song one could find the smallest subdivision of the beat. We will call this a *tatum* (as defined by Bilmes in his master's thesis [1]), such that any distance between any two notes can be represented as k tatums, k an integer. This is usually easy to do by looking at the score. As an example consider a song that has 5 measures. Each measure is divided into 3 beats. Say that a tatum is equivalent to a quarter of a beat. Then each beat is 4 tatums long, each measure is 4×3 tatums and the song is $3 \times 4 \times 5$ tatums long.

Even with this representation we will not have a one-to-one correspondence with sounds. Each note can have millions of different sounds (timbre): loudness, tremolo, staccato, varying with different instruments, who's playing, (for some it is not hard to distinguish the timbre of two different players), etc.. The same occurs for the rhythm: decrescendos, accelerandos, rubato, swing, etc... Even though a one-to-one correspondence does not exist we can make a good approximation using the MIDI standard.

2.2.1 The MIDI Standard

The MIDI (Musical Instrument Digital Interface) standard is a hardware specification and communications protocol that allows computers, controllers, and synthesis gear to pass information amongst themselves, see Loy [13]. MIDI uses representations based on the concept of notes by defining a pitch and a velocity (volume) that go on and off. MIDI is mostly controlled by keyboard instruments which can be represented by a series of switches. Each separate key is treated as a switch. When a key is depressed, a *Note On* message is sent out, indicating the note associated with that key and with what velocity it was struck. When the key is released, a *Note Off* message is transmitted with the key number and velocity 0. In a similar way MIDI can be used to go from a score representation of sound to an acoustic signal. The way MIDI, together with sound synthesis techniques, converts scores to music is rather complicated. In the following section we present a method of converting a series of notes represented in a MIDI score to an acoustical signal representation of a sine tone instrument (i.e. an instrument with no harmonics).

2.2.2 Time Series Representation

The time series representation X_j, where j is the tatum number, is defined by $X_j =$ note at tatum j. This representation does not characterize the score exactly, since it makes no distinction between two contiguous identical notes with durations d_1 and d_2 and that note with duration $d_1 + d_2$.

As a numerical representation of a note we could use the MIDI-Note number. In this case an increase of a step would represent a jump to the note a semi-tone away. This presents a problem when dealing with rests. Rests do not have Midi-Note numbers. We couldn't just assign 0 to rests because then this would be representing a note corresponding to MIDI-number 0. Even though this note is below the audible range it does not correspond to 0 frequency thus its choice is quite arbitrary since notes with MIDI-note numbers smaller than 16 correspond to notes below the audible range and using equation (1) below we see that the MIDI-note number corresponding to 0 frequency is $-\infty$. One way to get around this is to prolong the duration of notes preceding rests. In a song with few rests of short duration this would not make much of a difference.

An alternative numerical representation, that is more in accordance with the signal representation, is using the fundamental frequency of the pitch determined by the notes. For example a note of MIDI-number X would be represented by frequency

$$440 \times 2^{\frac{X-69}{12}} = 8.175799 \times 2^{\frac{X}{12}} \text{ Hz.} \tag{1}$$

see Pierce [15]. In this case frequencies related to rests could be set to 0 since a sound wave with 0 frequency has no fluctuations and thus is silent. It would be interesting to note how robust our analysis is to this arbitrary assignment.

2.2.3 Marked Point Process Representation

Suppose we have a series of triplets, (Note,Duration,Volume), then we can construct an acoustical signal representation via the following definitions:

$$\begin{aligned} Y(t) &= \sum_j V_j h\left(\frac{t-\tau_j}{\sigma_j}\right) \cos \lambda_j (t-\tau_j) \qquad (2) \\ h(\cdot) &= \text{a taper function} \end{aligned}$$

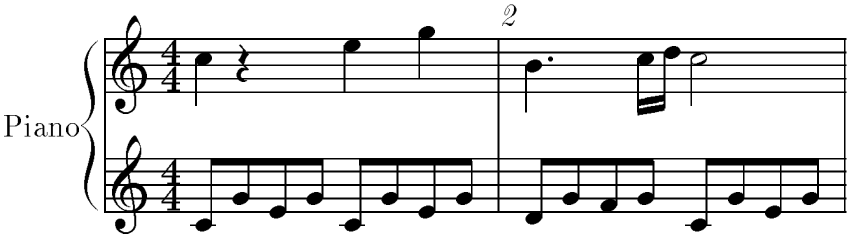
where τ_j is the time of commencement of the j-th note, λ_j is the frequency of the j-th note, V_j is the volume of the j-th note and σ_j is the duration of the j-th note. Here $\{\tau_j\}$ will be a point process corresponding to times of jumps between notes. For time t near τ_j the signal will look like a cosine wave of frequency λ_j and amplitude V_j. The units t here could be seconds as well as tatums in which case we could represent changes in tempo by using time maps that assign a duration in seconds to each tatum, see [21].

To compute the frequency λ_j from midi-number X_j we use equation (1). (We used this conversion method to check for mistakes in the data entry. By converting the entered data and forming the signal produced by relation (2) we then played the signal through the speakers of a Sparc work-station using the Matlab command `sound`. See Appendix for some details.)

One reason the taper function is introduced in (2) is to avoid hearing clicks at instantaneous changes of pitch. It also expresses the restricted duration of a particular note.

2.2.4 An Example

The following is the common practice notation (CPN) for the first two bars of Mozart's Sonata in C-major, K545:



Note: The actual song starts with a half note C and no rest. We put in the rest for illustrative purposes.

The melody in these two bars is played by the right hand (shown in the upper clef). In this case the tatum would correspond to a *sixteenth* note, or a quarter of a beat. If the song were played at an Allegro tempo (about 144 quarter notes per minute) then a tatum would have a duration of 60(seconds)/144(quarter notes per beat) $\times$ 1/4(tatums per beat) $\approx$ 0.10 seconds. The note and duration in tatum pairs are the following: (C,4), (rest,4), (E,4), (G,4), (B,6), (C,1), (D,1), (C,8)

The time series representation using MIDI-numbers would be :

$$72, 72, 72, 72, \text{ NA, NA, NA, NA}, 76, 76, 76, 76, 79, 79, 79, 79,$$

$$71, 71, 71, 71, 71, 71, 72, 74, 72, 72, 72, 72, 72, 72, 72, 72$$

Here the NAs represent rests.

The time series representation using frequencies of the pitches would be:

$$523, 523, 523, 523, 0, 0, 0, 0, 659, 659, 659, 659, 783, 783, 783, 783,$$

$$493, 493, 493, 493, 493, 493, 523, 587, 523, 523, 523, 523, 523, 523, 523, 523$$

A marked point process representation with time measured in tatums characterizes the score. For the sonata we have $\{\tau_j, (V_j, \lambda_j, \sigma_j)\}$: {0,(1,523,4)}, {8,(1,659,4)}, {12,(1,783,4)}, {16,(1,493,6)}, {22,(1,523,1)}, {23,(1,587,1)}, {24,(1,523,8)}. (We have set the volume to 1, choice is completely arbitrary. This part of the score does not ask for certain notes to be played louder than others. In practice accents are always present.)

3 Some Previous Work

Electronic Musicians have used random processes to create melodies. Completely uncorrelated processes, with constant spectra, seem to create "melodies" with no structure. "Melodies" produced with random walks, i.e. spectrum $1/f^2$, seem to be too predictable. In between these two processes is so called $1/f$ Noise.

Voss studied the possibility of music having a $1/f$ spectrum [22, 23, 24]. He took the signal representations $Y(t)$ of a variety of songs and obtained the "instantaneous" audio power of music. In order to measure it, the audio signal $Y(t)$ was passed through a band-pass filter in the range 100 Hz to 10 kHz. The output voltage was squared, and filtered with a 20 Hz low-pass filter. Voss remarked that correlations of the resulting process represented

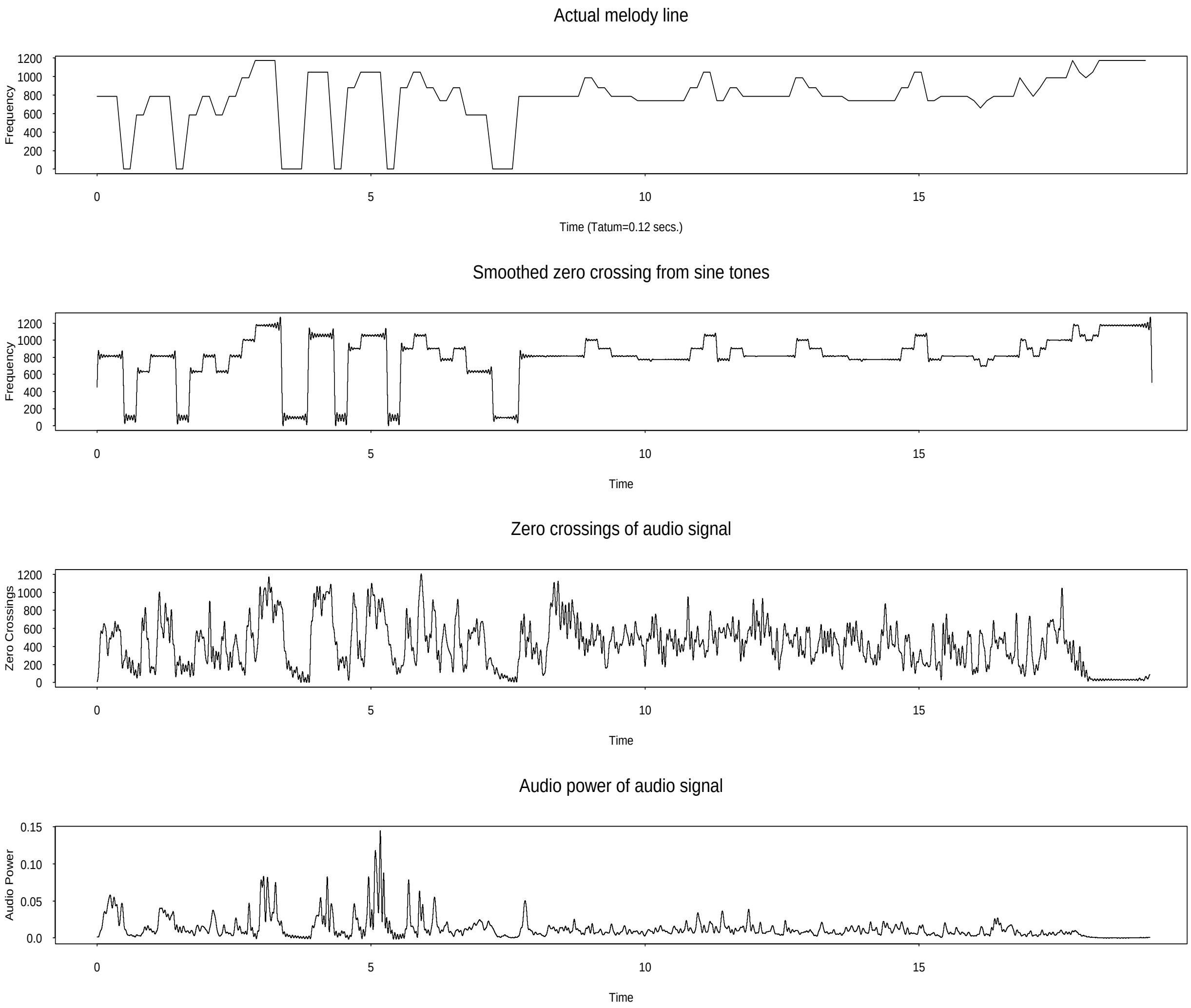


Figure 1: Score, smoothed zero crossings and instantaneous audio power of Eine Kleine Nachtmusik.

correlations of the audio power of successive notes. For a discussion of some properties of this filtering technique see the Appendix.

Another quantity Voss examined was the "instantaneous" frequency. He measured this by the rate, $Z(t)$, of zero crossings of the audio signal. He remarked that in the case of music, correlations of $Z(t)$ represented correlations in the frequencies of successive notes. This is reasonable because if say frequency λ dominates at time t, the signal will be approximately $\rho \cos(\lambda t + \phi)$ and the rate of zero crossings (or cycles) is $\lambda/2\pi$ per unit time. Of course problems may arise when more than one stream of notes is played at the same time, for example in Mozart's Sonata above you have the right hand playing a stream of notes corresponding to the melody and the left hand playing a stream of notes corresponding to the accompaniment.

These two methods, of seeking information of the melody from the audio signal, work well when the melody is being played by one instrument with no harmonics, as we see in

Figure 1. In a case where the audio signal contains more than one instrument and the sound produced by these instruments contains many harmonics, these methods do not work as well. In Figure 1 we see the 4 time series plots. The first is the score representation, using frequencies, of the first 10 measures of the melody line of Mozart's *Eine Kleine Nachtmusik*, the second is the smoothed zero crossings of the signal created using (2) on the score representation, the third is the smoothed zero crossing of the audio signal of an actual orchestra playing the song and finally the fourth plot is the "instantaneous" power of the audio signal. Notice how well the zero crossings method works when the sound signal contains only one instrument with no harmonics. In the third plot we see that the method doesn't work well when there is more than one instrument playing. Notice also that at the beginning of the song, when all the instruments are playing the same notes (first 4 measures), the method works better than when there is more than one stream present. (See the Appendix for the procedure used to obtain these figures).

In another formal study of music Hsu and Hsu [9] study the fractal nature of the intervals between successive notes. This corresponds to the intervals in the score representation using the MIDI-note numbers. If in (2) we used the MIDI-note numbers M_j instead of the frequencies λ_j, then these intervals would be defined by $I_j = M_{j+1} - M_j$ for $j = 1, ..., N$, where N is the number of notes in the whole piece.

4 Second-order Spectra

The second-order spectrum or power spectrum of a stationary process $Y(t)$, $-\infty < t < \infty$ is given by

$$cov\{Y(t+u), Y(t)\} = \int_{-\infty}^{\infty} \cos(\lambda u) f_2(\lambda) d\lambda \tag{3}$$

with u the lag. The physical meaning of the spectrum is that $f_2(\lambda)d\lambda$ represents the contribution to the variance or power of $Y(t)$ components with frequencies in the ranges $(\lambda, \lambda+d\lambda)$ and $(-\lambda, -\lambda+d\lambda)$.

These definitions extend directly to the case of locally stationary process. Crudely the (overall) spectrum of the process (2) will be proportional to

$$\sum_j V_j^2 \sigma_j \delta(\lambda - \lambda_j) \tag{4}$$

and the process will have $1/f$ spectrum to the extent that $V_j^2\sigma_j$ falls of as $1/\lambda_j$. The process itself will be locally stationary with instantaneous frequency λ_j for t near τ_j.

4.1 Estimates

In the case that the stationary process Y_t has mean 0 a naive estimate of the spectrum is provided by the periodogram,

$$I_2^T(\lambda) = \frac{1}{2\pi T} |d^T(\lambda)|^2 \tag{5}$$

where

$$d^T(\lambda) = \sum_{t=1}^{T} \exp\{-i\lambda t\} Y_t \tag{6}$$

The periodogram is an asymptotically unbiased but inconsistent estimate (unless $f_2(\lambda) = 0$) since $Var[I_2^T(\lambda)] \approx f_2(\lambda)^2$ as $T \to \infty$.

If the series Y_t is mixing (see e.g. conditions in Brillinger [4]), the variates

$$I_2^T(\lambda_t)/f_2(\lambda_t),\ \lambda_t = 2\pi t/T \text{ for } t = 1, 2, \dots \tag{7}$$

are approximately independent exponentials with mean 1.

4.2 Parametric Modeling

Voss proposed that the spectrum of music has a $1/f$ (or $1/\lambda$) parametric form. Consider the problem of fitting parametric models to spectra. We can find estimates by maximizing the approximate log likelihood

$$L_T(\theta) = -\sum_{t=1}^{T-1} \left[log(f_2(\lambda_t;\theta)) + \frac{I_2^T(\lambda_t)}{f_2(\lambda_t;\theta)} \right],\ \lambda_t = \frac{2\pi t}{T} \tag{8}$$

see Dzhaparidze [6]. With θ estimated by $\hat{\theta} = \arg\max_\theta L_T(\theta)$, under certain conditions (including that the trispectrum is 0), $\hat{\theta}$ is consistent and asymptotically normal

$$\sqrt{T}(\hat{\theta}_T - \theta) \to N(0, \Gamma_\theta^{-1}) \tag{9}$$

as $T \to \infty$, where

$$\Gamma_\theta[k,l] = \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial\theta_k} \log f_2(\lambda;\theta) \frac{\partial}{\partial\theta_l} \log f_2(\lambda;\theta) d\lambda \tag{10}$$

The estimate is asymptotically efficient in the Gaussian case.

The goodness of fit of a particular parametric model may be assessed by graphing the estimate, $I_2^T(\lambda)$, as well as the parametric estimate $f_2(\lambda, \hat{\theta})$ surrounded by confidence bounds for the former. This will be done in the examples that follow.

4.2.1 Models for Spectra

We consider the following models for the overall power spectrum of music:

1. $f_2(\lambda; \alpha, \beta) = \frac{\alpha}{\lambda^\beta}$
2. $f_2(\lambda; \alpha, \beta) = \frac{\alpha}{1+\lambda^\beta}$
3. $f_2(\lambda; a, b, c, \alpha) = \frac{\lambda^a}{(1+\alpha\lambda^b)^c}$

4. $f_2(\lambda; a, b, .., \alpha, \beta, ..) = \frac{\lambda^{a}}{(1+\alpha\lambda^{b})^{c}} \frac{1}{(1+\beta\lambda^{d})^{e}}$

By choice of functional form and parameter values, these models are able to describe a fairly broad range of behavior.

In all the models $0 \leq \lambda \leq \pi$ with f_2 symmetric and of period 2π. Notice that the first model is the "1/f" Noise model.

4.3 Signal Representation

Songs from a variety of musical styles were chosen in our study of the power spectrum of the signal representation. These songs included:

1. Baroque: J.S. Bach, Cantata No. 211 (Coffee Cantata) BWV 211, *Recitativo: Wenn Du mir nicht den Coffee* and Cantata burlesque (Peasant Cantata) BWV 212, *Aria: Heute noch, lieber Vater, tut es doch.* Performed by baritone Kevin McMillan, soprano Dorothea Röschmann and *Le Violins du Roy* chamber orchestra.
2. Classical: J.F. Haydn, Sonata in D-mayor Hob. XVI/37, *Finale* and Sonata in F-mayor Hob. XVI/23, *Finale* . Both performed on Piano by Dominique Cornil.
3. Romantic: C. Debussy, Suite bergamasque L. 75, *Passepeid* and Images L. 87, *Lent.* Both performed on Piano by Zóltan Kocsis.
4. Atonal: A. Schoenberg, Orchesterstücke op. 16, *Vorgefühle* and Orchesterstücke op. 16, *Peripetie.* Both performed by Berlin Philarmonic.
5. Spanish Guitar: L. Milán, *Pavan No. 6* and *Pavan No. 5.* Both performed on Guitar by Andrés Segovia.
6. Jazz: Wayne Shorter, *Footprints* and Miles Davis, *Four. Footprints* performed by Miles Davis. *Four* performed by Sonny Rollins. In both cases we recorded just the head (In most Jazz tunes a song starts off with a fixed melody, called the head, and then improvisations are played).
7. Afro-Cuban: Juan Mesa, *Amalia* and Florencia Calle, *Baba Cuello Mao.* Both performed by *Los Muñequitos de Matanza.*
8. Rock and Roll: Chuck Berry, *Let it Rock* and Chuck Berry, *Bye Bye Johnny.* Both performed by Chuck Berry.
9. Hip-Hop/Rap: R. Stewart, E. Wilcox, R. Jackson, T. Hardson, R. Robinson and J. Martínez, *It's Jiggaboo Time* and *If I Were President.* Both Performed by *The Pharcyde.*

We sampled the audio signal of the mentioned songs at 8000 samples per second. The sampled signal was then filtered using the two methods of Voss described above. It is important to note that the units of the signal Y_t are arbitrary. (See the Appendix for the details of these computations.) First we determined Y_t to be the smoothed zero crossings of the signal. Then we calculated the periodogram of Y_t and minimized the negative of the approximate log likelihood given in equation (8) restricting λ to $(0, 20)$Hz since frequencies over 20Hz where filtered out. Using Powell's algorithm, see [16], the four models were fitted. The results of these fits in the case of Bach's *Coffee Cantata* can be seen in Figure 2. The goodness of fit may be assessed by the approximate 95% confidence intervals which are given as the dashed lines. Model 1 seems to fit well here. The same was done for the "instantaneous" audio power of the signals, the four models were fitted. The results of the fits for the *Coffee Cantata* can be seen in Figure 3. Again Model 1 seems to fit well.

We fitted the $1/f$ model for the smoothed zero crossings obtained from the signal representations of the 18 pieces listed above. A fit for each style can be seen in Figure 4. Approximate standard errors are calculated using equation (10). The $1/f$ model appears to be performing well.

The values obtained for $\hat{\beta}$ are given in Table 1. The fraction of points outside the (approximate) 95% intervals ranges from 4.6% to 6.3%.

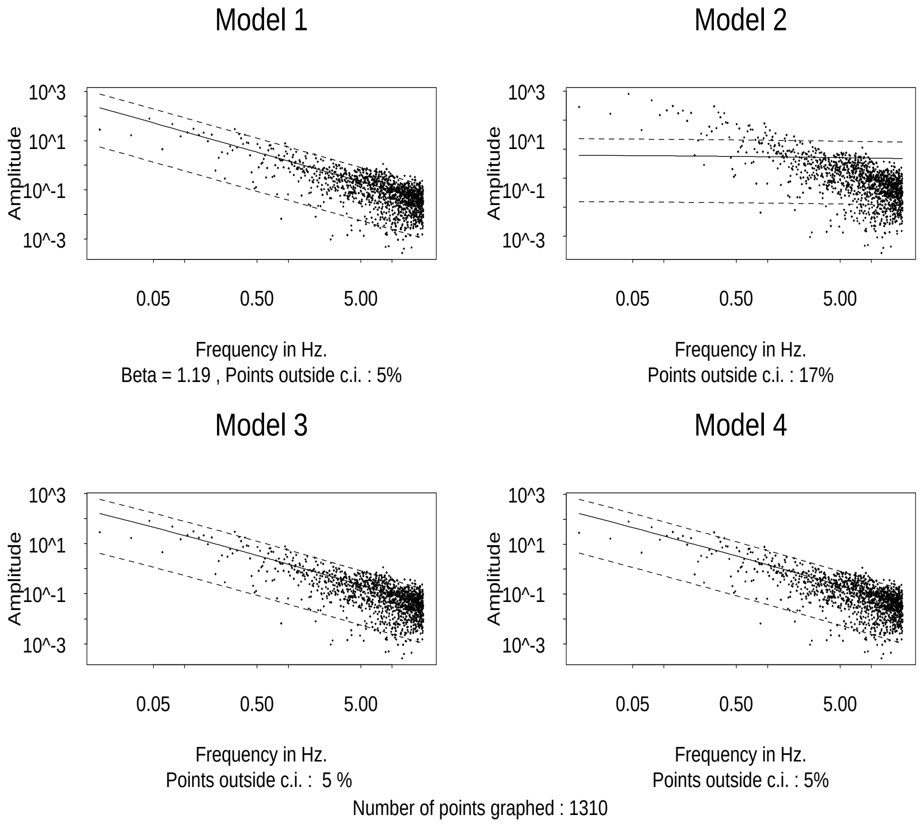


Figure 2: Fitted models for the smoothed zero crossings obtained from Bach's *Coffee Cantata* signal.

Song	$\hat{\beta}$	SE	Tot. Pts.	% not in c.i.
Baroque 1	1.192739	0.0391	1310	5.0
Baroque 2	1.066802	0.0391	1310	5.8
Classical 1	1.190916	0.0276	2621	6.5
Classical 2	1.244281	0.0276	2621	5.8
Romantic 1	1.185097	0.0276	2621	4.7
Romantic 2	1.085337	0.0276	2621	5.7
Atonal 1	1.083865	0.0276	2621	4.8
Atonal 2	1.143627	0.0276	2621	5.2
Spanish Guitar	1.228339	0.0391	1310	4.6
Spanish Guitar	1.236005	0.0391	1310	6.0
Jazz 1	1.121480	0.0391	1310	5.4
Jazz 2	1.217540	0.0276	2621	5.6
Afro-Cuban 1	1.116375	0.0276	2621	5.8
Afro-Cuban 2	1.169583	0.0276	2621	6.3
Rock & Roll	1.000069	0.0276	2621	6.1
Rock & Roll	1.109331	0.0276	2621	5.6
Hip Hop	1.022285	0.0276	2621	5.8
Hip Hop	1.075118	0.0391	1310	6.0

Table 1 : Results of fitting the power spectrum.

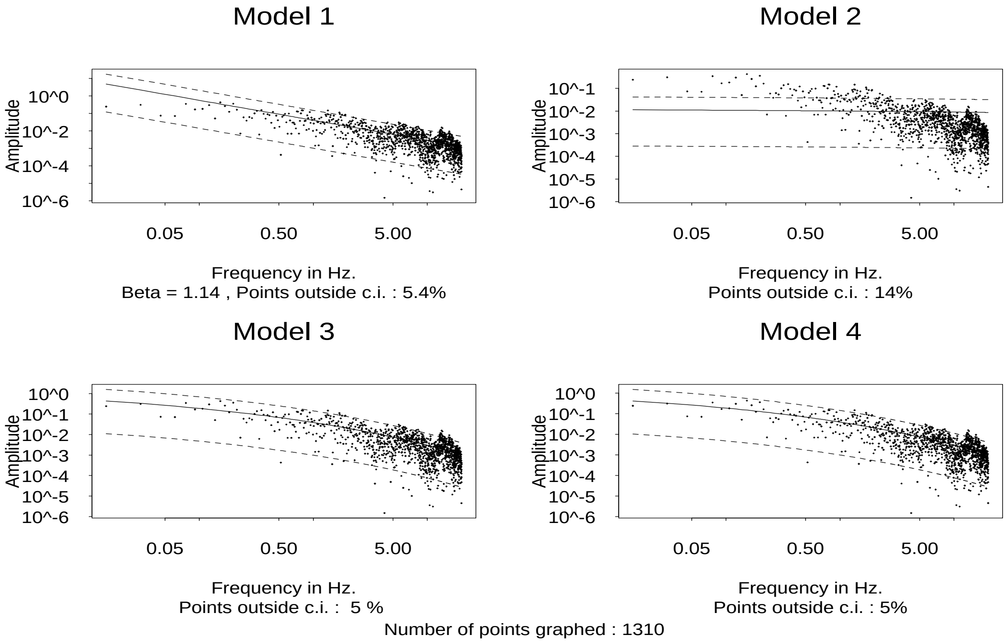


Figure 3: Fitted models for the "instantaneous" power obtained for Bach's *Coffee Cantata* signal

4.4 Score representation

We next looked at the time series expressions of various songs representative of several styles of music. In the following list we give the composer, title of composition, title of the specific part (when applicable), the key, time signature and tempo given in the score, and what type of note corresponds to a tatum.

1. Baroque
   (a) J.S. Bach Cantata No 211 (Coffee Cantata), Be Silent All, Recitativo : Wenn du mir nicht den Coffee, D-major, 4/4, Tempo = 70, Tatum = 1/16.
   (b) J.S. Bach, French Suites, Suite II, Courante, C-minor, 3/4, Tempo = 144, Tatum = 1/8.
2. Classical
   (a) F.J. Haydn, La Roxelane : Air and Variations, Theme, C-minor, 2/4, Tempo = 150 (Allegretto) , Tatum = 1/16.
   (b) F.J. Haydn, La Roxelane : Air and Variations, Var I, C-major, 2/4, Tempo = 150 (Allegretto), Tatum = 1/16.
   (c) F.J. Haydn, La Roxelane : Air and Variations, Var II, C-minor, 2/4, Tempo = 150(Allegretto), Tatum = 1/16.
3. Romantic
   (a) Claude Debussy, Suite Bergamasque, Passepeid, F-minor, 4/4, Tempo = 150 (Allegretto ma non troppo), Tatum = 1/8, (Note: this is an approximation to the melody. Triplets were ignored and replaced by the first note.)
4. Spanish Guitar
   (a) Luis de Milán, Pavan no. 5, in Tone VIII, "La bella Francesca"(Fol. [G vi^r]), G-minor, Complex meter varies between 2/4 3/4, Tempo = 120 (Allegro Moderato), Tatum = 1/16.
   (b) Luis de Milán, Pavan no. 6, in Tone VIII, (Fol. [G vi^r]), G-minor, Complex meter varies between 2/4 3/4, Tempo = 120 (Allegro Moderato), Tatum = 1/8.
5. Jazz
   (a) Miles Davis, Four, E♭-major, 4/4, Tempo = 178 (Medium Swing), Tatum = 1/8.
   (b) Wayne Shorter, Footprints, E♭-major, 6/4, Tempo = 178 (Medium Swing), Tatum = 1/8.

   In both cases we use the score of the head.
6. Latin
   (a) Pérez Prado, Mambo No. 5, E♭-major, 2/2, tempo = 240, tatum=1/8.
   (b) Pérez Prado, Mambo No. 8, F-major,2/2, Tempo=240, tatum=1/8.

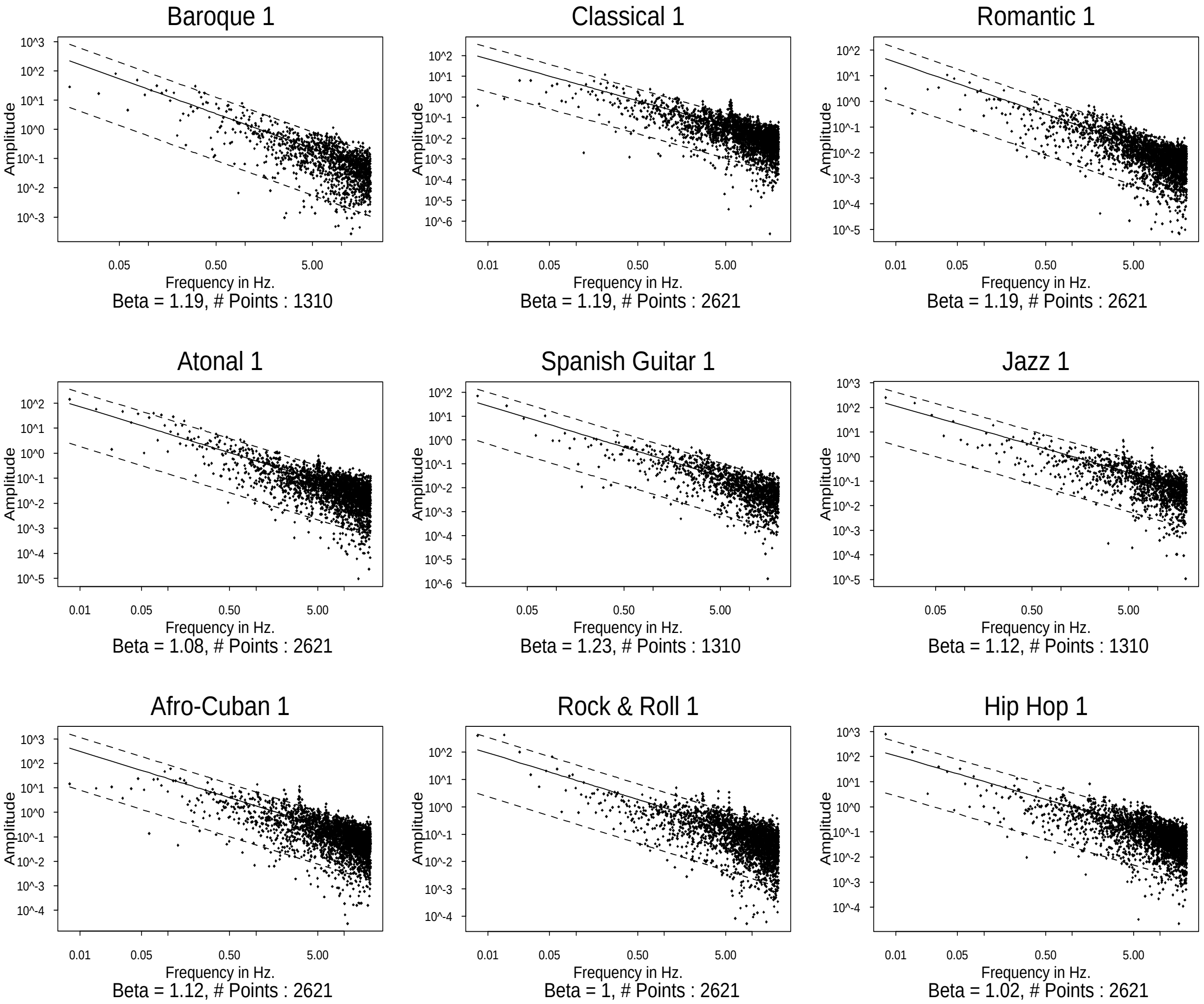


Figure 4: Fitted $1/f$ model for the smoothed zero crossings of 9 style types.

We used the frequency version of the time series representation of the score. Defining Y_t as follows:

$$Y_t = \text{frequency at tatum } t \quad (11)$$
$$= 0 \text{ if a rest occurred at time t} \quad (12)$$

With respect to the representation (2), Y_t would be the λ_j of the τ_j near t.

First we calculated the periodogram of Y_t and minimized the negative of the approximate log likelihood give in equation (8) above. The results for Bach's *Coffee Tocata* can be seen in Figure 5. Again the $1/f$ model is fitting well.

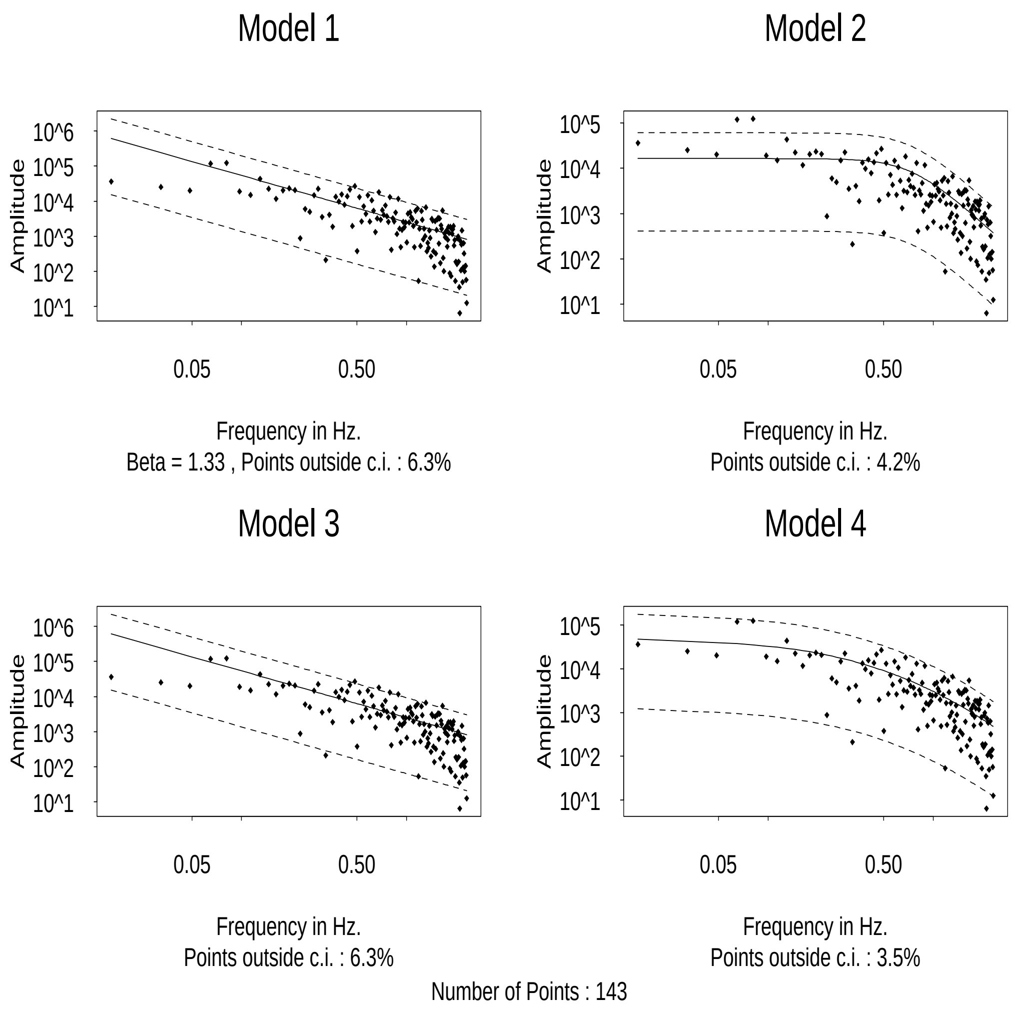


Figure 5: Fitted models for the time series representation using frequencies of the score representation of the *Coffee Cantata*.

Next we fitted the $1/f$ model to the 12 score representations listed above. The results can be seen in Figure 6. Again the $1/f$ model appears plausible. The values obtained for $\hat{\beta}$ are listed in Table 2. The lowest was for *Four* and the highest was for *Footprints* (the two Jazz tunes). Again the $1/f$ model appears to be performing reasonably.

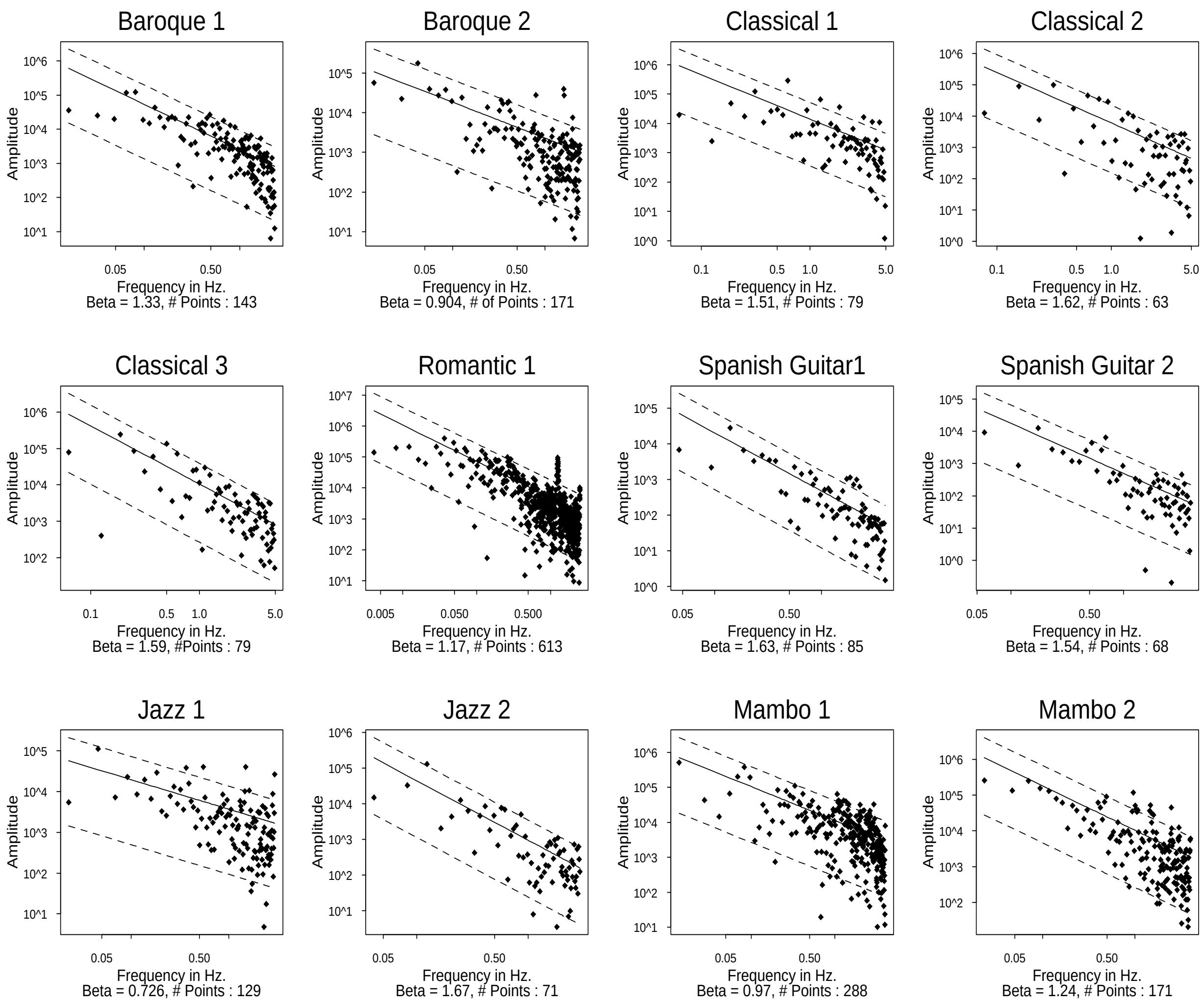


Figure 6: Fitted $1/f$ model for the scores of the 12 pieces listed.

Song	$\hat{\beta}$	SE	Tot. Pts.	% not in c.i.
Baroque 1	1.330	0.1180	143	6.3
Baroque 2	0.904	0.1080	171	8.2
Classical 1	1.510	0.1590	79	14
Classical 2	1.620	0.1780	63	18
Classical 3	1.590	0.1590	79	3.8
Romantic	1.170	0.0571	613	7.8
Spanish Guitar 1	1.630	0.1530	85	5.88
Spanish Guitar 2	1.540	0.1710	68	5.88
Jazz 1	0.726	0.1250	129	7.8
Jazz 2	1.670	0.1680	71	8.4
Mambo 1	0.970	0.0833	288	9.3
Mambo 2	1.240	0.1080	171	9.4

Table 2: Results of fitting the power spectrum.

5 Third-Order Spectra

NonGaussian aspects of music do not appear to have been investigated. In this connection the bispectrum and bicoherence are pertinent parameters. They are useful in both discerning nonGaussianity and in examining for nonlinearity. Definitions and estimates are given in the Appendix.

Suppose the process Y_t is linear, that is

$$Y_t = \int a_{t-u} d\epsilon_u \tag{13}$$

where ϵ_t is a process with independent increments having mean 0, variance σ^2, and third moment κ. (In the Gaussian case $\kappa = 0$.) Then the power spectrum of Y_t is

$$f_2(\lambda) = \frac{\sigma^2}{2\pi}|A(\lambda)|^2 \tag{14}$$

and the bispectrum is

$$f_3(\lambda, \mu) = \frac{\kappa}{(2\pi)^2} A(\lambda)A(\mu)\overline{A(\lambda+\mu)} \tag{15}$$

where

$$A(\lambda) = \int e^{-i\lambda u} du \tag{16}$$

If, for example, $A(\lambda) = 1/\lambda^{\beta/2}$ and the process is linear, then the power spectrum is $1/2\pi\lambda^\beta$ and the bispectrum

$$\frac{\kappa}{(2\pi)^2} \cdot \frac{1}{\lambda^{\beta/2}} \cdot \frac{1}{\mu^{\beta/2}} \cdot \frac{1}{(\lambda+\mu)^{\beta/2}} \tag{17}$$

The spectrum may be estimated and examined to see if it is 0 (Gaussian process). Supposing that the denominator does not vanish, the bicoherence

$$|B(\lambda,\mu)|^2 = \frac{|f_3(\lambda,\mu)|^2}{f_2(\lambda)f_2(\mu)f_2(\lambda+\mu)} = \frac{\kappa^2}{(2\pi)^4} \cdot \frac{(2\pi)^6}{\sigma^6} = \gamma^2 \tag{18}$$

is defined and constant for this linear process case, see Brillinger [2]. In the case that the process Y_t is *reversible* (probabilistic properties of $\{Y_t\}$ and $\{Y_{-t}\}$ the same), the imaginary part of the bispectrum is identically 0. See Brillinger and Rosenblatt [3]. Reversibility is not the property of most music.

The process (2) will have nonzero bispectrum to the extent that the frequencies λ_j, present for t near τ_j, satisfy relations such as $\lambda_j + \lambda_{j'} = \lambda_{j''}$.

Under regularity conditions (including stationarity and mixing) estimates $f_2^T(\lambda)$, $f_3^T(\lambda,\mu)$ of the power and bispectra may be constructed that are asymptotically independent and normal. These may be used to form the bicoherence estimate, $|B^T(\lambda,\mu)|^2$, whose approximate statistical properties are indicated in the Appendix.

For a given sample value of the bicoherence, $|B^T(\lambda,\mu)|^2$, one may compute the approximate prob-value of achieving a value as large or larger in the null Gaussian case. The null

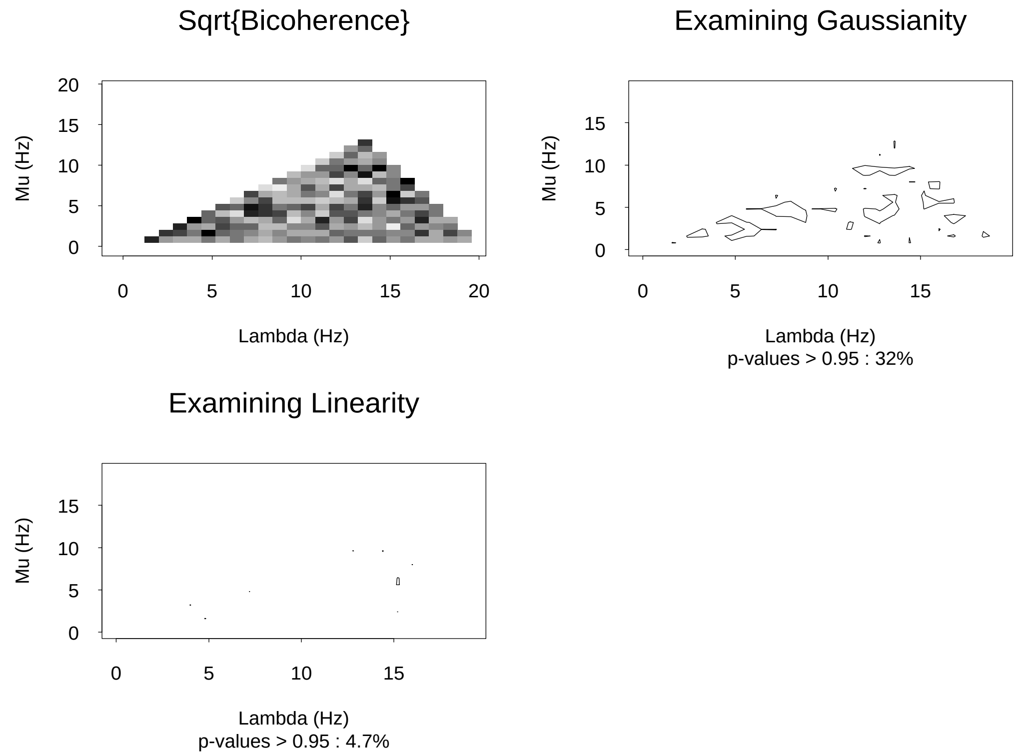


Figure 7: Square root of the bicoherence estimate and contour plots of prob-values for non-Gaussianity and non linearity respectively for the smoothed zero crossings of the *Coffee Cantata.*

distribution is an exponential, see Appendix. Such prob-values are contoured in Figures 7, 8 and 9, for the *Coffee Cantata*.

Likewise to assess the possibility that the basic process is linear, one may compute

$$\hat{\gamma} = \text{ ave } |B^T(\lambda, \mu)|^2 \tag{19}$$

with the average over all bicoherence estimates and then compute the approximate the prob-value corresponding to the deviate $||B^T(\lambda, \mu)|^2 - \hat{\gamma}|$. Again the prob-values are contoured for the *Coffee Cantata*. Details of the approximation are given in the Appendix.

These procedures, of using test statistics that are functions of (λ, μ), rather than some global statistic, have the advantage of indicating the character of departure if the null hypothesis appears rejected.

Nikias and Mendel [14] provide a review of higher order spectra and some of their uses.

5.1 Signal Representation

We checked for nonGaussianity and nonlinearity in the time series used in section 4. The series studied, *Coffee Cantata*, lasted 64.15 seconds and was sampled at 8000 Hz. After applying the Voss filter every 200th observation was retained, 2566 data points in all. The spec-

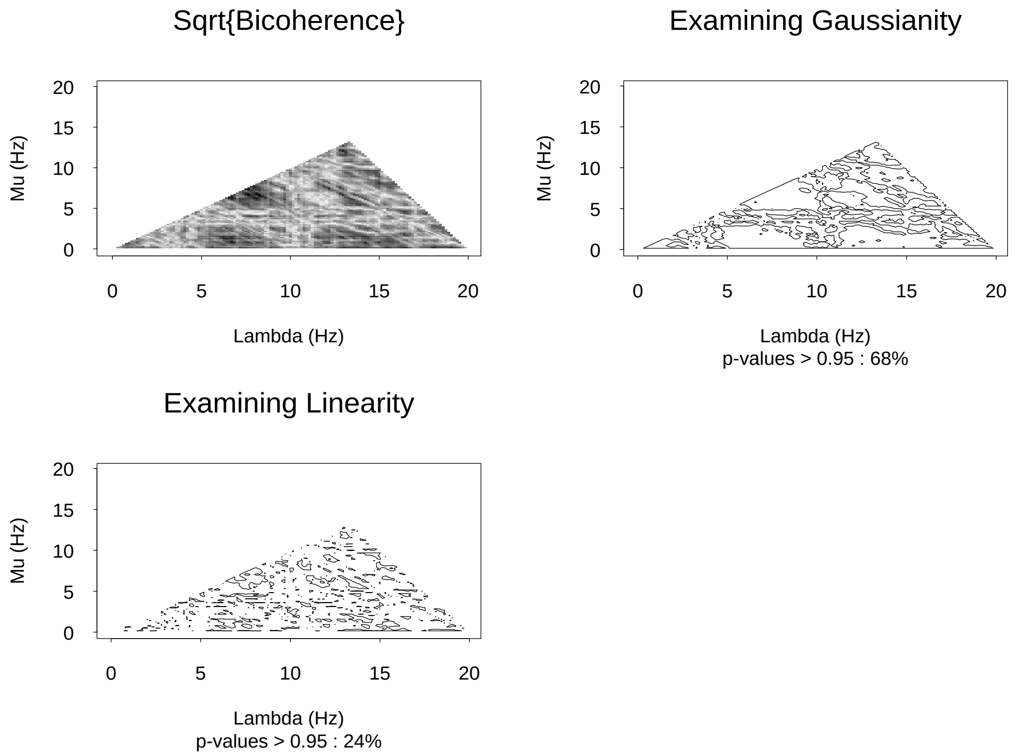


Figure 8: Square root of the bicoherence estimates and contour plots of prob-values for non-Gaussianity and non linearity respectively for the "instantaneous" power obtained from the *Coffee Cantata* signal.

tra were estimated from this data. The resulting estimates can be seen in Figure 7 and 8. For the zero crossings data, Gaussianity is being rejected, but not linearity. For the instantaneous power both Gaussianity and linearity are being rejected.

5.2 Score Representation

Some similar computations were done for the score representation. In estimating the bicoherence we took 10 stretches. As a consequence the stretches were short, ranging from 12 to 120 points. The resulting estimates for the score of the *Coffee Cantata* can be seen in Figure 9, now graphing the 50% and 90% contours. In this case Gaussianity appears rejected but not linearity.

6 Discussion and Conclusions

We began with the question of what makes music, music. To address it we considered whether certain parametric forms fitted well, whether associated time series were Gaussian and whether they were linear. A broadly ranging selections of pieces were analyzed. The

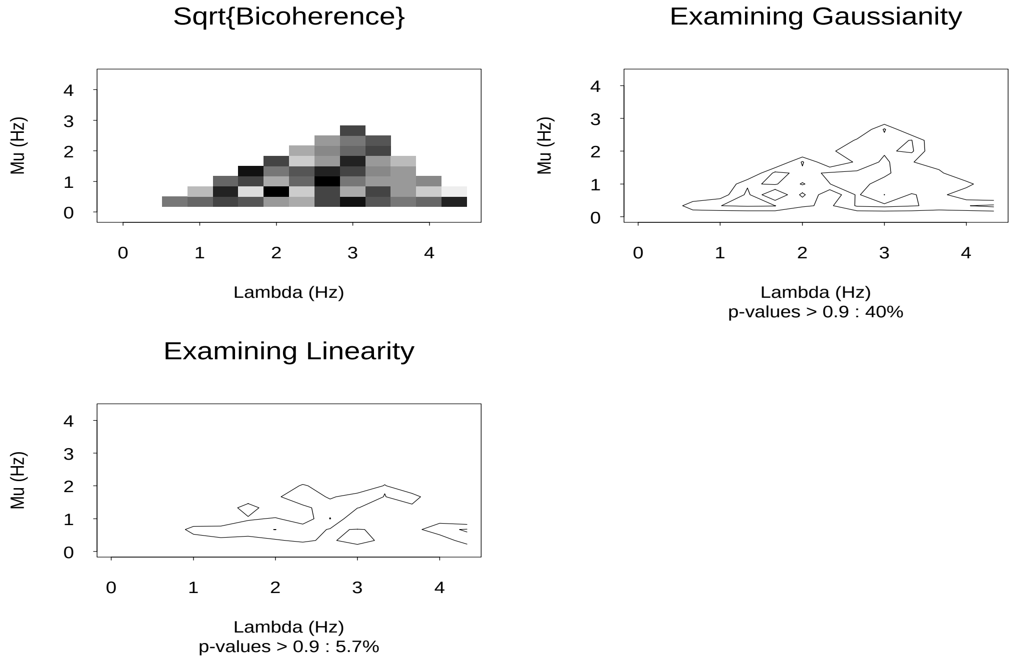


Figure 9: Square root of the bicoherence estimates and contour plots of prob-values for non-Gaussianity and non linearity for the score of the *Coffee Cantata*

model $1/f^{\beta}$ with β near 1 appeared to fit the scores well, as opposed to alternatives allowing more curvature or flatness at low frequencies. The same can be said for the derived processes of zero crossings and instantaneous power. In each, the hypothesis of Gaussianity (really 0 bispectrum) was rejected. The conclusions regarding linearity were not so clear.

We have acted as if the processes involved were stationary. To the extent that they are not, the parameters and estimates may be treated as if they are focussed on an average of instantaneous spectra obtaining for the processes involved. The statistical packages of Matlab and S-plus were employed.

In future work the trispectrum will be considered. It will allow further assesment of linearity. Other values for the passband of the lowpass filter, here 20 Hz, will also be considered. The signal computations were carried through only for *Coffee Cantata*. The other pieces will be studied as well.

7 Acknowledgments

We thank David Wessel and Steven Clark for a variety of helpful remarks. We thank Ofer Licht for help in entering the CD music. The work of David R. Brillinger was supported by NSF grants DMS-9300002 and DMS-9625774. The work of Rafael A. Irizarry was supported by a National Science Foundation Graduate Research Fellowship.

A Appendix

A.1 Third Order Spectra

We provide the basic definitions and properties in order that others can directly reproduce such study.

A.1.1 Definitions and Estimates

Bispectral Analysis is of use in discerning nonGaussianity of a time series and also in examining the series for nonlinearity.

Let $Y_t, t = 0, \pm 1, \pm 2, ...$ denote a stationary time series. Let it have mean c_1, covariance function $c_2(u)$ and third moment function

$$c_3(u,v) = E\left([Y_{t+u} - c_1][Y_{t+v} - c_1][Y_{t+u+v} - c_1]\right) \tag{20}$$

The *bispectrum* at *bifrequency* (λ, μ) is defined by

$$f_3(\lambda, \mu) = \frac{1}{(2\pi)^2} \sum \sum c_3(u,v) e^{-i(u\lambda + v\mu)} \tag{21}$$

and the *bicoherence* by

$$|B(\lambda, \mu)|^2 = \frac{|f_3(\lambda, \mu)|^2}{f_2(\lambda) f_2(\mu) f_2(\lambda + \mu)} \tag{22}$$

The fundamental domain of these parameters is $0 \leq \mu \leq \lambda$, $\lambda + \mu/2 \leq \pi$.

There are a variety of fashions by which the bispectrum may be estimated. A convenient one is: let the data be broken into L stretches of length V, so that $T = LV$. Next compute the tapered Fourier Transform of the l^{th} stretch,

$$d^V(\lambda; l) = \sum_{v=0}^{V-1} h\left(\frac{v+1}{V+1}\right) Y_{lV+v} e^{-iv\lambda} \tag{23}$$

for $l = 0, ..., L-1$. Then form the third order periodogram of the l^{th} stretch

$$I_3^V(\lambda, \mu; l) = \frac{1}{(2\pi)^2 V h_3} d_Y^V(\lambda; l) d_Y^V(\mu; l) \overline{d_Y^V(\lambda + \mu; l)} \tag{24}$$

where $h_3 = \int h(u)^3 du$. The estimate of the bispectrum is now

$$f_3^T(\lambda, \mu) = \frac{1}{L} \sum_{l=0}^{L-1} I_3^V(\lambda, \mu; l) \tag{25}$$

In forming the estimate of the bicoherence, the power spectrum is estimated by similarly averaging the second order periodograms of the L stretches.

In our empirical work no taper was employed, rather the series were prefiltered by fitting an autoregressive, prior to computing the spectral quantities. Such a linear filtering retains the 0 bispectral property of a Gaussian process and the linearity property of a linear process.

A.1.2 Statistical properties of the estimates

Suppose that the bispectrum is estimated, as above, by averaging the third-order periodograms of L contiguous segments of length V of a series of length T=LV. Then, for (λ, μ) not on the boundary of the fundamental domain, $f_3^T(\lambda, \mu)$ is asymptotically complex normal with mean $f_3(\lambda, \mu)$ and variance

$$\frac{h_6}{h_3^2} \cdot \frac{1}{2\pi} f_2(\lambda) f_2(\mu) f_2(\lambda + \mu) \frac{V}{L} \tag{26}$$

provided V, $L/V \to \infty$ as $T \to \infty$. It is noteworthy that for consistency a large number, L, of individual stretches will be required. Further estimates at distinct frequencies are asymptotically independent.

It follows that when $f_3(\lambda, \mu) = 0$, $|f_3^T(\lambda, \mu)|^2$ is asymptotically

$$\frac{h_6}{h_3^2} \cdot \frac{V}{L} \cdot \frac{1}{2\pi} f_2(\lambda) f_2(\mu) f_2(\lambda + \mu) \chi_2^2/2 \tag{27}$$

which result may be used to examine the hypothesis $f_3(\lambda, \mu) = 0$. In the examples, prob-values based on this distribution are graphed.

In the case that $f_3(\lambda, \mu) \neq 0$ the variate $|f_3^T(\lambda, \mu)|^2$ will be approximately normal with mean $|f_3(\lambda, \mu)|^2$ and variance

$$2 \frac{h_6}{h_3^2} \cdot \frac{V}{L} \cdot \frac{1}{2\pi} f_2(\lambda) f_2(\mu) f_2(\lambda + \mu) |f_3(\lambda, \mu)|^2 \tag{28}$$

In other words the large sample distribution of the bicoherence estimate

$$|B^T(\lambda, \mu)|^2 = \frac{|f_3^T(\lambda, \mu)|^2}{f_2^T(\lambda) f_2^T(\mu) f_2^T(\lambda + \mu)} \tag{29}$$

will be approximately exponential with mean

$$\frac{h_6}{h_3^2} \cdot \frac{V}{L} \cdot \frac{1}{2\pi} \tag{30}$$

when $f_3(\lambda, \mu) = 0$. It will be approximately normal with mean, the bicoherence,

$$|B(\lambda, \mu)|^2 = \frac{|f_3(\lambda, \mu)|^2}{f_2(\lambda) f_2(\mu) f_2(\lambda + \mu)} \tag{31}$$

and variance

$$2 \frac{h_6}{h_3^2} \cdot \frac{V}{L} \cdot \frac{1}{2\pi} \frac{f_3(\lambda, \mu)|^2}{f_2(\lambda) f_2(\mu) f_2(\lambda + \mu)} \tag{32}$$

when $f_3(\lambda, \mu) \neq 0$. The quantity $|B^T(\lambda, \mu)|$ will then be approximately normal with mean $|B(\lambda, \mu)|$ and variance

$$\frac{1}{2} \cdot \frac{h_6}{h_3^2} \cdot \frac{1}{2\pi} \cdot \frac{V}{L} \tag{33}$$

This approximation follows via the delta-method.

A.1.3 Related Work

Rosenblatt and Van Ness [18] developed various asymptotic properties of bispectral estimates, as did Brillinger [2] for higher-order spectral estimates. Huber et al. [10] considered the estimation of the bicoherence and in particular suggested approximating its distribution, when the population value was 0, by a χ_2^2. Elgar and Guza [7] investigated the accuracy of this approximation. Rao and Gabr [17] and Hinich [8] proposed global bispectrum-based tests for the nonGaussianity and nonlinearity of a stationary process. Rao and Gabr structured the problem as assessing whether all components of a multivariate normal have the same mean. Hinich (see also Brocket et. al. [5]) based tests on the interquantile range of sample bicoherence values. Terdik and Math [20] note that the bispectrum of a process, such that the linear predictor is the quadratic, satisfies a particular algebraic identity and use this to assess possible linearity.

A.2 Obtaining the Numerical Representations

A.2.1 Signal Representation

A song was chosen from a Compact Disc. It was down-loaded into a Mono .au file sampled at 8000 Hz using a CD-ROM and software for the Sparc machines. To go from stereo to a mono signal the two channels were averaged. Our statistical analysis was done mostly by S-Plus which can't read .au files. We altered Thau's program `xplay`, a sound player for Sun Sparc machines, which can handle AIFF, .au, and some WAVE files, so that it would save a file with the numbers corresponding to the sampled signal in a file readable to S-Plus. Due to technical details of the way Compact Discs are recorded and the way `xplay` works the units of the sampled audio signal are completely arbitrary.

To obtain the smoothed zero crossings or "intantaneous" pitch and power from the sampled signals we wrote C programs that performed the zero crossing calculation, the band-pass filtering, the squaring and the lowpass filtering relatively quickly. The filtering was done by calculating the FFT of the signal, setting the coefficients of the pertinent frequencies to zero and then performing the inverse FFT.

A.2.2 Score Representation

A piece of music was selected. Then using an EMU Proteus Keyboard, a Mac Power Book 520 and a program we wrote in Max (see Rowe [19] for some information on Max) we saved the midi-numbers and duration into text files. We used MIDI-note number 36 (lower than the lowest note in any of the score) to denote rests. These files where made readable to S-Plus and Matlab using Pearl. Using S-Plus we created various functions that converted the raw data into objects of the time series and marked point process representations respectively.

A.2.3 Time Series Representation

For each score we had two time series; one for the MIDI-numbers and the other for the frequencies. To do the analysis on the MIDI-number time series we took care of the rests by extending the previous notes over the duration of the rests. In the case of a song starting on a rest we simply ignored that part of the song.

For the case of the respective frequencies we assigned frequency 0 to the rests. This representation is probably more representative of the music, so we focused our attention on it.

A.2.4 Playing the data

To play the data we used Matlab. The command *sound* takes as an argument a vector. This vector is taken to be a signal sampled at 8192 Hz (at least on Sparcs, it varies for other computers). Each k-th element of the vector is taken to be the sample at time $\frac{k}{8192}$ seconds. The following Matlab code performs the work.

```
%note is a vector containing the MIDI-note numbers(rests are 36)
%dur is a vector conatining the durations of these notes
%rate is the sample rate
for i = 1:length(note);
        t = [0:rate*dur(i)];
        window = 1-cos(2*pi/(rate*dur(i))*t); %TUKEY's WINDOW
        t = t/rate*2*pi;
        if note(i) == 36  %in our case MIDI-number 36 in the raw data
                          %represented rests
           signal=[signal,t*0];
        else
           y = window .* sin(note(i)*t);
           signal = [signal,y];
        end;
end;
sound(signal)
```

A.3 Notes on the Voss Technique

Suppose the signal may be written

$$Y(t) = R(t)\cos\left(Z(t)t + \phi(t)\right) \tag{34}$$

with $Z(t)$ the instantaneous frequency, $R(t)$ a slowly changing amplitude and $\phi(t)$ a slowly changing phase. For $Z(t)$ in the passband of .1 to 10 kHz, after filtering the signal will remain essentially (34). With squaring it becomes

$$R(t)^2\left[1 + cos\left(2Z(t)t + 2\phi(t)\right)\right]/2 \tag{35}$$

After the lowpass filtering to $[0, .01]$ kHz one has approximately

$$R(t)^2/2 \tag{36}$$

in other words essentially the squared envelope of the signal (34).

In terms of the representation (2) one has approximately

$$V_j^2 h\left(\frac{t-\tau_j}{\sigma_j}\right)^2/2 \tag{37}$$

for t near τ_j, provided λ_j is in the band $[.1, 10]$ kHz. A spectrum analysis of this will bring out the periodicity properties of the point process $\{\tau_j\}$, which in music can be regarded as the rhythmic structure. If the probability distribution of the $(\tau_{j+1} - \tau_j)$ is long-tailed then the fitted process (2) can have $1/f$ spectra, see Lowen and Teich [12]. Also notice that V_j^2 is related to the accents or "dynamics" of the music.

References

[1] J. Bilmes, Timing is of the Essence: Perceptual and Computational Techniques for Representing, Learning, and Reproducing Expressive Timing in Percussive Rhythm, Master's Thesis Massachusetts Institute of Technology, 1993.

[2] D.R. Brillinger, "An introduction to polyspectra.", Ann. Math. Statist., Vol. 36, 1965, pp. 1351-1374.

[3] D.R Brillinger and M. Rosenblatt, "Computation and interpretation of the k-th order spectra", in *Spectral Analysis of Time Series*, edited by B. Harris, John Wiley, 1967, pp. 189-232.

[4] D.R. Brillinger, *Time Series: Data Analysis and Theory*, New York, Holt, Reinhart and Winston 1975.

[5] P.L. Brockett, M.J. Hinich and D. Patterson, "Bispectral-based tests for the detection of Gaussianity and linearity in time series", J. Amer. Statist. Assoc., Vol. 83, 1988, pp. 657-664.

[6] K. Dzhaparidze, *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series*, Springer-Verlag, New York, 1986.

[7] S. Elgar and R.T. Guza, "Statistics of bicoherence", IEEE Trans. Acoustics, Speech and Signal Proc., Vol. 36, 1988, pp. 1667-1668.

[8] M.J. Hinich, "Testing for Gaussianity and linearity of a stationary time series", J. Time Series Analysis, Vol. 3, 1982, pp. 169-176.

[9] K.J. Hsu and A.J. Hsu, " Fractal geometry of music", Proc. Natl. Acad. Sci. USA, Vol. 87, 1990, pp. 938-941.

[10] P.J. Huber, B. Kleiner, T. Gasser and G. Dumermuth "Statistical methods for investigating phase relations in stationary stochastic processes", IEEE Trans. Audio Electro AU-19, 1971, pp. 78-86.

[11] K.S. Lii, M. Rosenblatt and C. Van Atta, "Bispectral measurements in turbulence", J. Fluid Mech., Vol. 77, 1976, pp. 45-62.

[12] S.B. Lowen and M.C. Teich, "Fractal renewal processes generate 1/f noise", Physical Rev., Vol. 47, 1993, pp. 992-1001.

[13] G. Loy, "Musicians make a standard: the MIDI phenomenon", Computer Music Journal, Vol. 9(4), reprinted in *The Music Machine*, ed. C. Roads, Cambridge, The MIT Press, 1989.

[14] C.L. Nikias and J.M. Mendel, "Signal processing with higher-order spectra", IEEE Signal Processing Magazine, 1993, pp. 10-37.

[15] J.R. Pierce, *The Science of Musical Sound*, New York, Freeman 1992.

[16] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipes in C : the Art of Scientific Computing*, Cambridge University Press, 1992.

[17] T. Subba Rao and M. Gabr, "A test for linearity of stationary time series", J. Time Series, Vol 1, 1980, pp. 145-148.

[18] M. Rosenblatt and J.W. Van Ness, "Estimation of the bispectrum", Ann. Math. Statist, Vol. 36, 1976, pp. 1120-1136.

[19] R. Rowe, *Interactive Music Systems*, Cambridge, MIT Press, 1994.

[20] Gy. Terdick and J. Màth, "A new test of linearity for the based on its bispectrum", University of Debrecen , Technical Report No. 95/152, 1996.

[21] D. Wessel, D. Bristow and Z. Settel, "Control of phrasing and articulation in synthesis", ICMC Proceedings, 1987, pp. 108-116.

[22] R.F. Voss and J. Clarke, " '1/f Noise' in music and speech", Nature, Vol. 258, 1975, pp. 317-318.

[23] R.F. Voss and J. Clarke, " 1/f Noise in music: music from 1/f Noise", J. Accous. Soc. Am., Vol. 63, 1978, pp. 258-263.

[24] R.F. Voss, "Fractals in nature: from characterizations to simulation" in The Science of Fractal Images, eds. by Peitgen, H.-O. and D. Saupe, Berlin, Springer-Verlag,1988, pp. 29-69.