Spectral Anticipations

Shlomo Dubnov
Music Department
University of California at San Diego
9500 Gilman Drive
La Jolla, California 92037-0326 USA
sdubnov@ucsd.edu

This article deals with relations between randomness and structure in audio and musical sounds. Randomness, in the casual sense, refers to something that has an element of variation or surprise in it, whereas structure refers to something more predictable, rule-based, or even deterministic. When dealing with noise, which is the "purest" type of randomness, one usually adopts the canonical physical or engineering definition of noise as a signal with a white spectrum (i.e., composed of equal or almost-equal energies in all frequencies). This seems to imply that noise is a complex phenomenon simply because it contains many frequency components. (Mathematically, to qualify as random or stochastic process, the density of the frequency components must be such that the signal would have a continuous spectrum, whereas periodic components would be spectral lines or delta functions.)

In contradiction to this reasoning stands the fact that, to our perception, noise is a rather simple signal, and in terms of its musical use, it does not allow much structural manipulation or organization. Musical notes or other repeating or periodic acoustic components in music are closer to being deterministic and could be considered as "structure." However, complex musical signals, such as polyphonic or orchestral music that contain simultaneous contributions from multiple instrumental sources, often have a spectrum so dense that it seems to approach a noise-like spectrum. In such situations, the ability to determine the structure of the signal cannot be revealed by looking at signal spectrum alone. Therefore, the physical definition of noise as a signal with a smooth or approximately continuous spectrum seems to obscure other significant properties of signals versus noise, such as whether a given signal has temporal structure—in other words, whether the signal can be predicted.

The article presents a novel approach to (automatic) analysis of music based on an "anticipation profile." This approach considers dynamic properties of signals in terms of an anticipation property, which is shown to be significant for discrimination of noise versus structure and the characterization and analysis of complex signals. Mathematically, this is formalized in terms of measuring the reduction in the uncertainty about the signal that is achieved when anticipations are formed by the listener.

Considering anticipation as a characteristic of a signal involves a different approach to the analysis of audio and musical signals. In our approach, the signal is no longer considered to be characterized according to its features or descriptors alone, but its characterization takes into account also an observer that operates intelligently on the signal, so that both the information source (the signal) and a listener (information sink) are included as parts of one model. This approach fits very well into an information theoretic framework, where it becomes a characterization of a communication process over a time-channel between the music (present time of the acoustic signal) and a listener (a system that has memory and prediction capabilities based on a signal's past). The amount of structure is equated to the amount of information that is "transmitted" or "transferred" from a signal's past into the present, which depends both on the nature of the signal and the nature of the listening system.

This formulation introduces several important advantages. First, it resolves certain paradoxes related to the use of information theoretic concepts in music. Specifically, it corrects the naïve equation of entropy (or uncertainty) to the amount of "interest" present in the signal. Instead, we are considering the relative reduction in uncertainty caused by prediction/anticipation. Additionally, the measure has the desired "inverted-U function" behavior that characterizes both signals that are either nearly constant and signals that are highly random as signals that have little structure. In both cases, the reduction in uncertainty owing to prediction is small. In the first case, this is caused by the small amount of variation in the signal to start with, whereas in the second

Computer Music Journal, 30:2, pp. 63–83, Summer 2006

case, there is little reduction in uncertainty, because prediction has little or no effect. Finally, this formulation also clarifies the difference between determinism and predictability. Signals that have deterministic dynamics (such as certain chaotic signals) might have varying degrees of predictability, depending on the precision of the measurement or exact knowledge of their past. A listener, or any practical prediction system, must have some uncertainty about its past, and this uncertainty grows when trying to predict the future, even in case of a deterministic system.

The ideas in the article are presented starting from a background of information theory and basis decomposition, followed by the idea of a vector approach to the anticipation measurement, and concluding with some higher-level reflections on its meaning for complex musical signals. The presentation takes several approaches: an engineering or signal-processing approach, in which notions of structure are related to basic questions about signal representation; an audio information-retrieval approach, in which the ideas are applied for characterization of natural sounds; and a musical analysis approach, in which time-varying analysis of musical recordings is used to create an "anticipation profile" that describes their structures. This last aspect has interesting possibilities for exploring relationships among signal measurements and human cognitive judgments of music, such as emotional force.

The main contribution of our model is in the vector approach, which generalizes notions of anticipation for the case of complex, multi-component musical signals. This measure uses concepts from principle components analysis (PCA) and independent components analysis (ICA) to find a suitable geometric representation of a monaural audio recording in a higher-dimensional feature space. This pre-processing creates a representation that allows estimating the anticipation of a complex signal from the sum of anticipations of its individual components. Accordingly, the term *complex signal* used throughout this article describes the multi-component nature of a single-channel recording and should not be confused with multiple-channel recordings.

Our Model

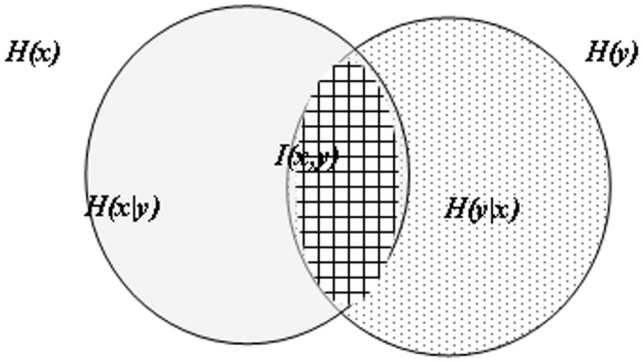
Our model of signal structure considers musical or audio material as an information source, which is communicated over time to a listener (either human or machine). The amount of information transmitted through this communication process depends on the nature of the signal and the listener. The listener makes predictions and forms expectations; the music source generates new samples. Accordingly, we define *structure* to be the aspect of musical material that the listener can predict, and *noise* or *randomness* as what the listener would consider as an unpredictable flow of data.

Information, Entropy, Mutual Information, and Information Rate

When a sequence of symbols, such as text, music, images, or even genetic codes are considered from an information theoretic point of view, it is assumed that the specific data is a result of production by an information source. The concept of the information source allows description of many different sequences by a single statistical model with a given probability distribution. The idea of looking at a source, rather than a particular sequence, characterizes which types of sequences are probable and which are not—in other words, what data is more likely to appear. Information theory also teaches that, "in the long run," some sequences become typical of the source, while others might turn very improbable or so rare that they would, in practice, never occur.

The logarithm of the relative size of the typical set (log of the number of typical sequences divided by log of number of all possible sequences of the same length) is called *entropy,* and it is considered as the characteristic amount of uncertainty inherent to the source. The larger the typical set, the higher the uncertainty, and vice versa. For instance, a biased coin that falls mostly on "heads" will produce sequences whose empirical average approaches the statistical mean, which comprises only a small fraction of all possible sequences of "heads" and

Figure 1. Entropies H*(·) for separate variables* x *and* y *and the pair (*x,y*), conditional entropies* H*(*x|y*) and* H*(*y|x*), and mutual information* I*(*x,y*).*



"tails." In such a case, the proportion of the size of the typical set (i.e., a set comprising mostly "heads") relative to the number of all possible sequences (the number two, raised to the power of the number of coin flips we performed) is low, and the source will be considered to have little entropy or little uncertainty. More equally distributed sources (such as multiple flips of a fair coin) will have a larger typical set and high uncertainly.

When information theory is used to describe a single information source, the concepts of entropy have direct relation to the coding size or compression lower bounds for that source. What is more interesting for our case is the aspect of information theory that deals with information channels, or the relationship between the *information source* and the *information sink.* In a physical communication channel, the source could be the voice of a person on one end of a phone line, the sink would be the voice emerging form the other end, and the channel noise would be the actual distortion caused to the source voice during the transmission process. Both signals are ideally similar, but not identical. The uncertainty about what was transmitted, remaining after we have received the signal, is characteristic of the channel. This notion is mathematically described using *mutual information,* which is defined as the difference between entropy of the source and conditional entropy between the source and the sink. If we consider source and sink as two random variables *x* and *y,* mutual information describes the cross-section between their entropies, as shown in Figure 1.

Denote by $H(x) = \sum P(x)\log P(x)$ the entropy of variable *x* (a source that has probability $P(x)$). This is depicted by the left (gray) circle in Figure 1; $H(y)$ is the right (dotted) circle of variable *y;* and $H(x,y)$ depicts the outer contour of both sources together. The conditional entropy $H(x|y)$ depicts the uncertainty about *x* if *y* is known (and vice versa for $H(y|x)$). The function $I(x,y)$ is the cross-section (grid) that indicates how much overlap in uncertainty occurs between *x* and *y,* that is, how much information one variable carries about the other. In extreme cases, if the two sources are independent, the two circles will have no overlap, and $I(x,y)=0$, accordingly. If *x* and *y* are exactly equivalent (i.e., knowing *x* is equivalent to knowing *y*), the two circles then will completely overlap each other, and $I(x,y) = H(x) = H(y) = H(x,y)$. The conditional entropy in such a case is zero, because knowledge of one variable completely describes the other, leaving no conditional uncertainty (i.e., $H(x|y) = H(y|x) = 0$). Mathematically, the above relations have simple algebraic expressions that are directly related to the marginal distributions $P(x)$, $P(y)$, and their common distribution $P(x,y)$.

$$I(X,Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$
$$= H(X) + H(Y) - H(X,Y) = \sum P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \quad (1)$$

Information Rate and Anticipation as "Capacity" of a "Time-Channel"

As we explained, mutual information is commonly used for theoretical characterization of the amount of information transmitted over a communication channel (see Figure 2). For our purposes, we will consider a particular type of channel that is different from a standard communication model in two important aspects. First, the channel is a time channel and not a physical transmission channel. The input to the channel is the history of the signal up to the current point in time, and the output is its next (present) sample. Second, the receiver must apply some algorithms to predict the current sample from its past samples. The information at the sink *y*

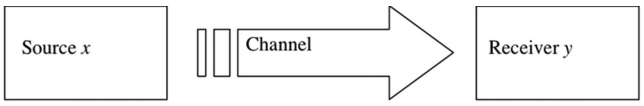


Figure 2. Information source, channel, and receiver.

consists of the signal history $x_1, x_2, \ldots, x_{n-1}$ available to the receiver prior to its receiving or hearing x_n. The transmission process over a noisy channel now has the interpretation of prediction/anticipation performance over time. The amount of mutual information depends on the "surprise" that the time-channel introduces to the next sample versus the ability of the listener to predict this surprise.

This notion of information transmission over time-channel is captured by the *information rate* (IR, also called the scalar-IR). This is defined as the relative reduction of uncertainty of the present when considering the past, which equals to the amount of mutual information carried between the past $x_{past} = \{x_1, x_2, \ldots, x_{n-1}\}$ and the present x_n. It can be shown using appropriate definitions of information for multiple variables, called *multi-information*, that the information rate equals the difference between the multi-information contained in the variables $x_1, x_2, \ldots, x_n$ and $x_1, x_2, \ldots, x_{n-1}$ (i.e., the amount of additional information that is added when one more sample of the process is observed):

$$\begin{aligned}\rho(x_1, x_2, \ldots, x_n) &= H(x_n) - H(x_n \mid x_{past}) = I(x_n, x_{past}) \\ &= I(x_1, x_2, \ldots, x_n) - I(x_1, x_2, \ldots, x_{n-1})\end{aligned} \quad (2)$$

One can interpret IR as the amount of information that a signal carries into its future. This is quantitatively measured by the number of bits that are needed to describe the next event once anticipation or prediction based on the past has occurred. Let us now discuss the significance of IR for the delineation of structure and noise for several example cases.

Example Case 1: "Inverted-U Function" for the Amount of Structure

A purely random signal cannot be predicted and thus has the same uncertainty before and after prediction, resulting in zero IR. An almost constant signal, on the other hand, has a small uncertainty $H(x)$, resulting in an overall small IR. High-IR signals have large differences between their uncertainty without prediction $H(x)$ versus the remaining uncertainty after prediction $H(x \mid \textit{past of } x)$. As mentioned in the Introduction, one of the advantages of IR is that in the "IR sense," both constant and completely random signals carry little information.

Example Case 2: The Relative "Meaning" of Noise

Let us consider a situation in which two systems attempt to form expectations about the same data. One system has a correct model, which allows good predictions for the next symbol. In the case when the uncertainty about the signal is large but the remaining uncertainty after prediction is small, the system manages to reveal the signal's structure and achieves its information-processing task, resulting in high IR.

Let us consider a second system that does not have the capability of making correct predictions. In such a case, the quality of prediction, which can be measured by the number of bits needed to code the next symbol, remains almost equal to the amount of bits required for coding the original signal without prediction. In such a case, the discrepancy between the two coding lengths is zero, and no information reduction was achieved by the system, resulting in a low IR.

Example Case 3: Signal Characterization

The previous discussion suggests that the IR is dependent on the nature of the information-processing system as well as the type of signal. Only in the case of an ideal system that has access to complete signal statistics does the IR become a characterization of the signal alone, independent of the information-processing system.

Example Case 4: Determinism Versus Predictability

An interesting application of IR is characterization of chaotic processes. Considering for instance a logistic map $x(n+1) = \alpha x(n)(1 - x(n))$, it is evident that knowledge of $x(n)$ provides complete information

for prediction of $x(n + 1)$. A closer look at the problem reveals that if the precision of measuring $x(n)$ is limited, the measurement error increases with time. For $\alpha = 4$ (chaos), the error approximately doubles every step, increasing the uncertainty by factor two, which amounts to a loss of one bit of information. This example shows that even complete knowledge of system dynamics might not suffice to make perfect predictions and that IR is dependent on the nature of measurement as well.

Example Case 5: Equivalence of Scalar-IR to Spectral Flatness Measure

The spectral flatness measure (SFM; Jayant and Noll 1984) is a well-known method for evaluating the "distance" of a process from white noise (see Appendix). It is also widely used as a measure for "compressibility" of a process, as well as an important feature for sound classification (Allamanche et al. 2001). Given a signal with power spectrum $S(\omega)$, the SFM is defined as

$$SFM = \frac{\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln S(\omega) d\omega\right)}{\frac{1}{2\pi}\int_{-\pi}^{\pi} S(\omega) d\omega} \tag{3}$$

and it is positive and less than or equal to one. It has a value of unity for a white-noise signal.

For large n, IR equals the difference between the marginal entropy and entropy rate of the sequence or signal $x(n)$, $\rho(x) = \lim_{n\to\infty} \rho(x_1, \ldots, x_n) = H(x) - H_r(x)$, where the entropy rate is the limit of conditional entropy for large n, $H_r(x) = \lim_{n\to\infty} H(x_n \mid x_1, x_2, \ldots, x_{n-1})$. Using expressions for the entropy and entropy rate of Gaussian process, one arrives at the following relation (Dubnov 2004):

$$SFM(x) = \exp(-2\rho(x)) \tag{4}$$

Equivalently, one can express IR as a function of SFM:

$$\rho(x) = -\frac{1}{2}\log(SFM(x)) \tag{5}$$

Figure 3 shows the close relationship between spectral flatness and the IR measure for a nature recording from a jungle. The figure shows the IR measure plotted on top of the signal spectrogram. (The values of IR were scaled so that it would display conveniently on top of the spectrogram.) It can be seen that signal segments that contain flat spectrum (either silences or bursts of high-bandwidth noise) correspond to low IR, and segments that contain harmonic or narrow-band noise have higher IR. It should be noted that this example contains sounds of mostly pitched bird singing alternating with noisy bird cries, which is a "segmentation" type of processing that can be performed using scalar-IR. In the next example, we will introduce the concept of vector-IR and consider a complex situation of simultaneously mixed periodic and noisy signals.

Extension of IR for Multivariate (Vector) Process

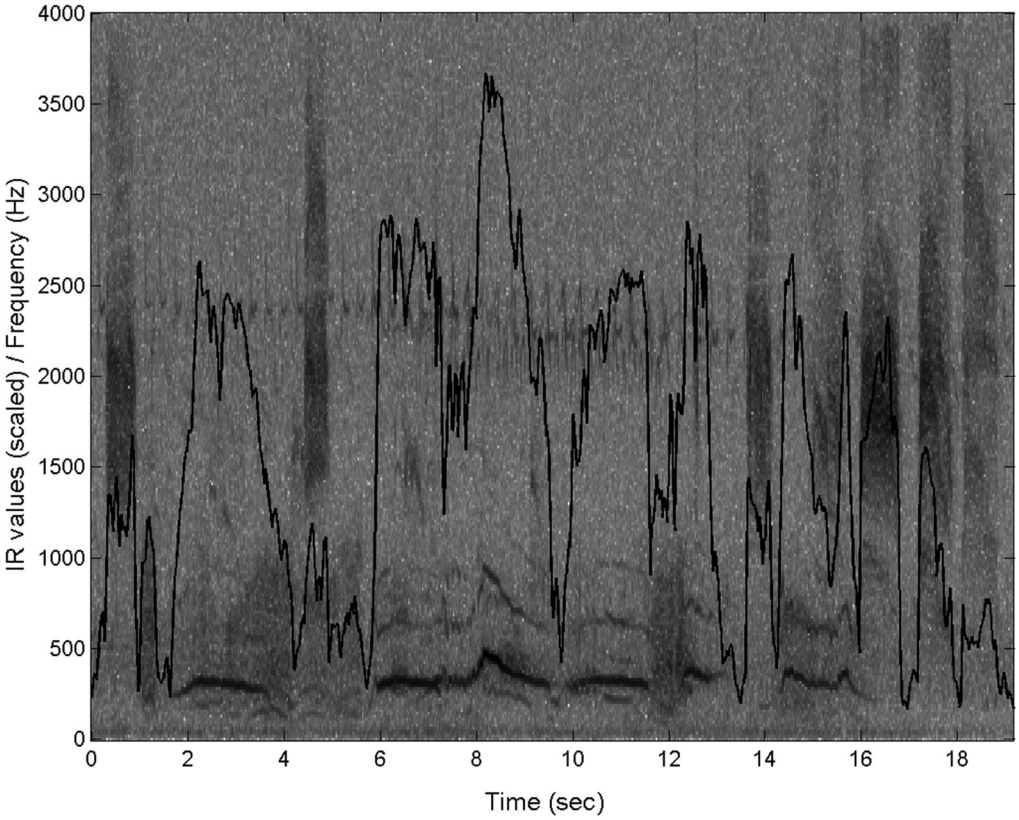
In the case of complex signals, such as those comprising several components, or described by sequences of spectral descriptors, or sequences of feature vectors, we must consider new type of IR that would be appropriate for dealing with a sequence of multiple variables described as vectors in a higher-dimensional space. We will discuss such representations in the context of audio basis and geometric signal representation in the next section.

Using capital-letter notation for vector variables, we denote a sequence of vectors by $X_1X_2 \ldots X_L$ and generalize the IR definition (to be called *vector-IR*) as

$$\rho(X_1, X_2, \ldots, X_L) = I(X_1, X_2, \ldots, X_L) - \{I(X_1, X_2, \ldots, X_{L-1}) + I(X_L)\} \tag{6}$$

The new definition of multivariate information rate represents the difference in information over L consecutive vectors minus the sum of information in the first L–1 vectors and the multi-information between the components within the last vector X_L. Let us assume that some transformation T exists, such that $S = TX$ and the components $S_1S_2 \ldots S_L$ after transformation are statistically independent. Using relations between entropies of a linear transformation of random vectors, it can be shown (Dub-

Figure 3. Signal-IR plotted on top of the signal spectrogram. See text for more detail.



nov 2003) that IR may be calculated as a sum of IRs of the individual components, $s_i(n)$, $i = 1 \ldots n$,

$$\rho_L(X_1, X_2, \ldots, X_L) = \sum_{i=1}^{n} \rho(s_i(i), \ldots, s_i(L)) \quad (7)$$

The vector-IR generalization allows identification of the structural elements of the signal as the components with high scalar-IR.

Example: Vector-IR Analysis of Mixed Sinusoidal and Noise Components

Given a sinusoidal signal $s(t) = \sin(\omega t + \varphi)$ and a white-noise signal $n(t)$, we consider the case of a mixed sinusoidal and noise signal $x(t) = s(t) + n(t)$. For this signal, we may find separate signal and noise spaces using singular value decomposition (SVD) using signal representation as a sequence of frames containing consecutive signal samples

$$\begin{aligned} X_1 &= [x_1, x_2, \ldots, x_n]^T, X_2 = [x_{n+1}, x_{n+2}, \ldots, x_{2n}]^T, \ldots, X_1 \\ &= [x_{(i-1)n+1}, x_{(i-1)n+2}, \ldots, x_{in}]^T \end{aligned} \quad (8)$$

Concatenating L signal segments into a matrix $X = [X_1 X_2 \ldots X_L]$, SVD provides a decomposition $X = U\Lambda V^T$, which gives the n basis vectors in columns of the U matrix (orthogonal n dimensional vectors), a diagonal matrix Λ that gives the n variances of the coefficients, and the first n rows of the matrix V^T that give the normalized expansion coefficients in this space.

Figure 4 shows the results of applying SVD to the sinusoidal and noise signal. The first two basis vectors are the two quadrature sinusoidal components. The remaining basis vectors are the noise components. Figure 5 shows the corresponding expansion coefficients.

We apply scalar-IR analysis to the first n rows of

Figure 4. Four basis vectors (shown as rows) obtained by SVD.

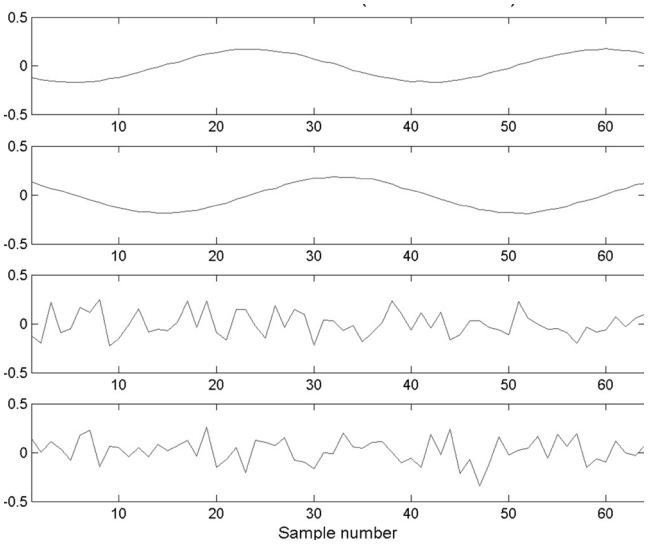
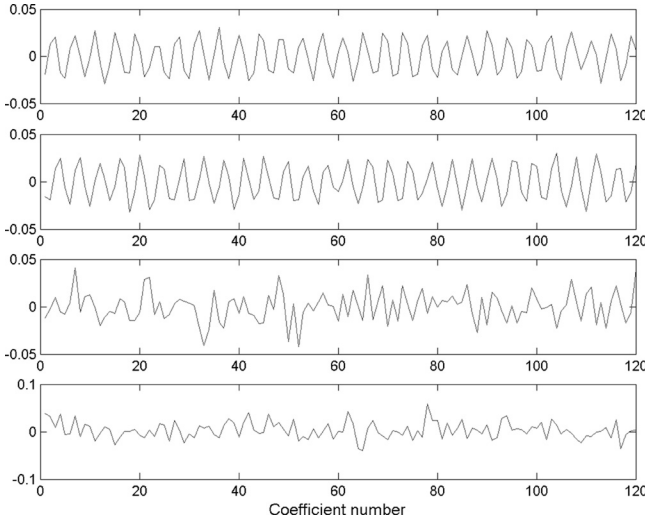


Figure 5. Expansion coefficients that correspond to the basis vectors of Figure 4.



the matrix V^T using the SFM method of scalar-IR estimation. The sum of the scalar-IRs yields the vector-IR. These estimates were compared to the scalar-IR of the individual sinusoidal and noise signals and their sum signal. In Table 1 in the paranthesis, we show the value of SFM for these signals, which is bounded between zero and one, making it somewhat easier to consider these values as a measure of the amount of structure. To read the SFM values, one should recall that SFM values close to one indicate noise and values near zero indicate structure. In the case of vector-IR, the SFM value is actually the *generalized-SFM,* which is a result of summing the individual scalar-IRs and transforming the resulting vector-IR back to SFM. We use the relationship

$$generalizedSFM = e^{-2 \cdot vectorIR} \qquad (9)$$

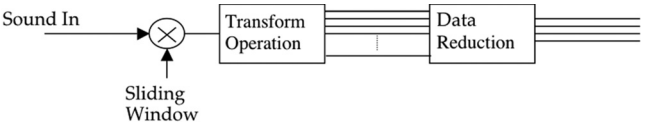
which also equals the product of the individual column SFMs. So, in principle, presence of a strong structural element (i.e., SFM close to zero) makes the generalized SFM small, indicating structure. This depends of course on the ability of SVD to separate the structural components (in practice, the SVD of the sines+noise signal contained some errors in the structural components that it found, resulting in a higher generalized-SFM than the theoretical product of the SFMs of the individual sinusoid or noise signals).

Table 1. Scalar-IR and Vector-IR Values for the Example Sinusoidal Component, Noise Component, and Combined Sines+Noise Signal

	Sinusoid	*Noise*	*Sines+Noise*
Scalar-IR (SFM)	6.31 (3×10^{-6})	8.7e-5 (0.99)	0.16 (0.72)
Vector-IR (SFM)	8.18 (7×10^{-8})	0.21 (0.657)	2.44 (0.007)

These results show that the amount of structure revealed by vector-IR (or generalized-SFM) for the sines+noise signal is significantly higher (and SFM lower) than the structure estimated by scalar-IR on the sum signal. It should also be noted that vector-IR estimation of noise components is imperfect, resulting in nonzero values. In general, there seems to be a tradeoff between the precision of estimation of structured and noise components using the SFM procedure. Using low-resolution spectral analysis in SFM estimation allows better averaging and improved estimation of the noise spectrum. This comes at the expense of poorer estimation of peaked or spectrally structured components. Using high-resolution spectral estimate causes better resolution of the spectral peaks, but creates a bias towards structured results.

Figure 6. Geometrical signal representation consisting of a transform operation followed by data reduction.



Audio Basis Representation

A common representation of audio results from transforming the time signal into the spectral domain using the Short-Time Fourier Transform (STFT). This processing is performed by applying the Fourier Transform to blocks of audio samples using a sliding-window that extracts short signal segments (also called *frames*) from the audio stream. It should be noted that each frame could be mathematically considered as a vector in a high-dimensional space. This approach can be generalized using additional types of transforms such as various types of filter banks, cepstral analysis, or auditory models, giving different types of features that might be better adapted to the specific processing task.

Spectral Audio Basis

When considering audio representations, a balanced tradeoff between reducing the dimensionality of data and retaining maximum information content must be achieved. For these reasons, various researchers and standards (Casey 2001; Lewicki 2002; Kim, Moreau, and Sikora 2004) have proposed feature extraction methods based on the projection of the transformed frames of the spectrum into a low-dimensional representation using carefully selected basis functions. A scheme of such a feature-extraction system is described in Figure 6. To assure independence between the transform coefficients for purpose of vector-IR analysis, additional statistical procedures must be applied to the different transform coefficients (or in the case of the STFT, different frequency channels).

Although Fourier channels are asymptotically independent, short-term statistics of different spectral channels may have significant cross-correlations. This dependence can be effectively removed using various methods that are described in the following section. Another common representation of spectral contents of audio signals is by means of cepstral coefficients (Oppenheim and Schafer 1989). The cepstrum is defined as the inverse Fourier transform of a logarithm of the absolute value of Fourier transform of the signal $C = \mathrm{F}^{-1}\{\log(|\mathrm{F}\{x(n)\}|)\}$. One of the great advantages of the cepstrum is its ability to capture different details of signal spectrum in one representation. For instance, the energy of the signal corresponds to first cepstral coefficient. Lower cepstral coefficients capture the shape of the spectral envelope or represent its smooth, gross spectral details. Detailed spectral variations such as spectral peaks corresponding to pitch (the actual notes played) or other long-term signal correlations appear in the higher cepstral coefficients. Selecting part of the cepstrum allows easy control over the type of spectral information that we submit to the IR analysis.

Clever choice of the transformation carries several advantages, such as energy compaction (compression by retaining the high-variance coefficients), noise reduction (projection onto separate signal and noise spaces), and improved recognition (finding salient features). The basic idea is that a signal of interest can be represented by linear combination of a few strong components (basis functions). The rest of the signal (parts that contain noise, interference, etc.) are assumed to reside in a different subspace and are ideally weak and approximately of equal energy.

There are several methods for estimation of low-rank models, such as Eigen-spectral analysis and singular value decomposition (SVD), PCA (also known as the Karhunen-Loeve Transform, or KLT), and ICA. For each model, we begin by assuming that the measurements are collected into frames of n samples each, thus consisting a vector x. The model is written as

$$X = AS + N \tag{10}$$

where X are the actual measurements, S are the expansion coefficients (sometimes also considered as the separate source components), A is an array of basis vectors, and N is an additive noise independent of S. Written explicitly, the former equation

represents X as a combination of basis vectors that are the columns of A

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [\bar{a}_1 \bar{a}_2 \ldots \bar{a}_m] \begin{bmatrix} s_1 \\ \vdots \\ s_m \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_n \end{bmatrix} \quad (11)$$

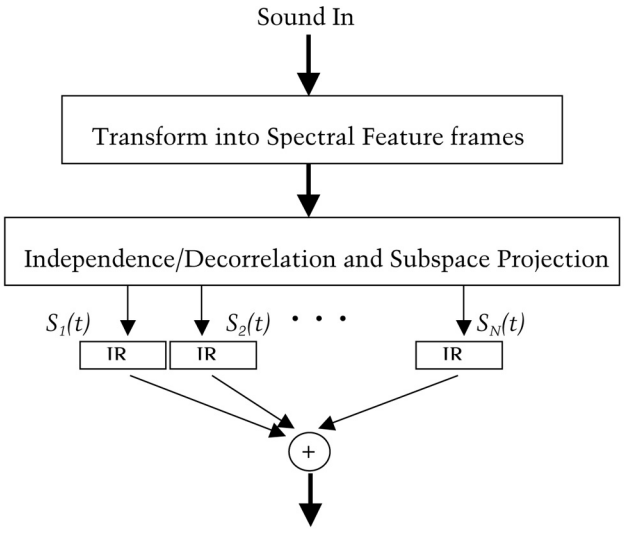
In our case, we are interested, in addition to the compaction or the noise removal property, in the independence of expansion coefficients S to allow estimation of vector-IR. This can be achieved by means of PCA for the case of a Gaussian multivariate process. It causes the components to be uncorrelated, which in general is not a sufficient condition for statistical independence. In the case of a Gaussian process, uncorrelated components are indeed independent.

For more general types of multivariate processes (such as in case that the signals are non-stationary or non-Gaussian), more sophisticated methods such as ICA can be used (Cichocki and Amari 2002). For the purposes of the present discussion, we shall assume that for most practical purposes our signals could be assumed to be multivariate Gaussian. Accordingly, we consider PCA or its alternative formulation using SVD. Using matrix notation $X_1 X_2 \ldots$, the columns are arranged in the order of subsequent feature vectors in time. Applying audio basis modeling, these features are represented as a sequence of coefficients S, with basis functions given by basis vectors written as columns of matrix A and usually ignoring the remaining noise components n:

$$\left[X_1 X_2 \ldots\right] = A \begin{bmatrix} s_1(1) & s_1(2) & \ldots \\ s_2(1) & s_1(2) & \ldots \\ \vdots & \vdots & \ldots \\ s_m(1) & s_1(2) & \ldots \end{bmatrix} + n \quad (12)$$

The problem of vector data decomposition into independent components is an active field of research. It should be noted that there are no closed-from solutions for ICA for multi-component, single-microphone signals. There are several works in the literature that attempt to perform monaural ICA using adapted source-filter models that are learned from prior examples of the single sources (Roweis 2000; Ozerov et al. 2005). These methods are not applicable to our purpose, because training or prior adaptations are not possible. Another related approach is that of Independent Subspace Analysis (ISA), where factorization of the spectrogram is achieved through projection of the spectrogram matrix unto subsets of ICA vectors that are then clustered into sets that are believed to belong to individual sources (Casey and Westner 2000; Dubnov 2002; Fitzgerald, Coyle, and Lawlor 2002; Smaragdis and Brown 2003; Uhle, Dittmar, and Sporer 2003).

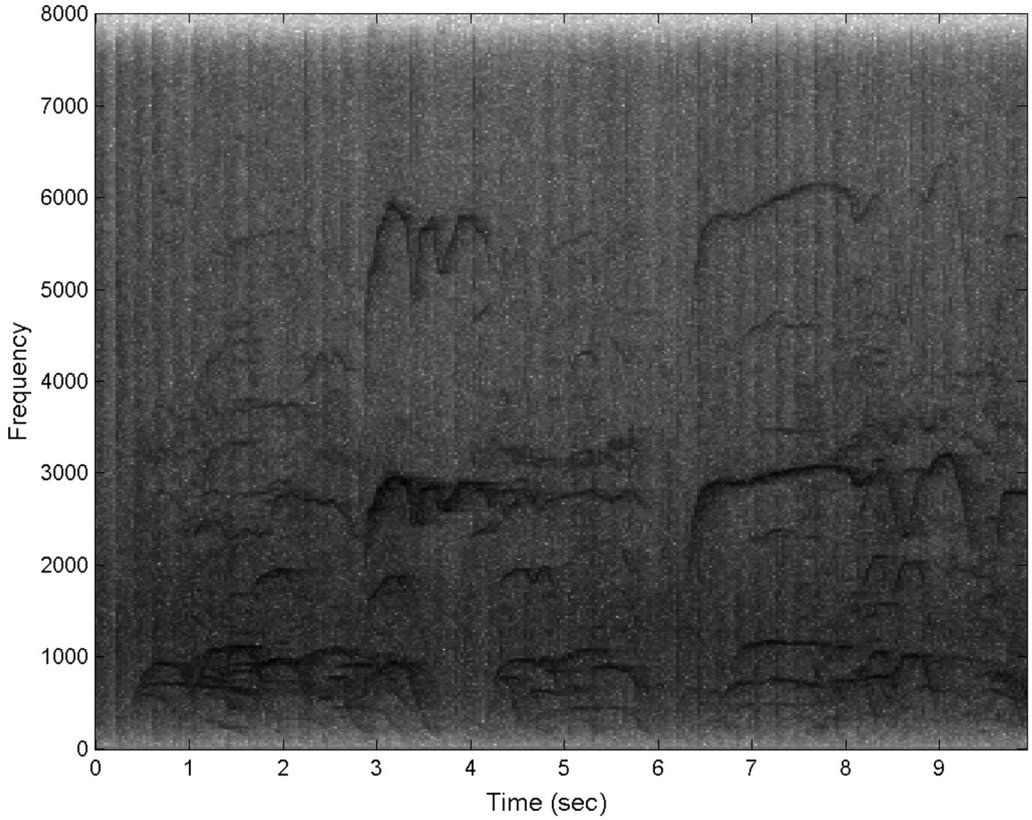
Figure 7. Anticipation algorithm flowchart.



Vector-IR Anticipation Algorithm

Having defined the various components that are necessary for IR analysis of complex signals, we present in Figure 7 the complete algorithm for analysis of IR using Audio Basis representations. In the first stage of analysis, we derive an appropriate geometrical representation, either as frames of audio samples or some features extracted from these frames, using the STFT, Filter Banks, Cepstral analysis, Mel-Frequency Cepstral Coefficients (MFCC), or some other method. Then, we perform basis decomposition, combined with data reduction by projection into a lower-dimensional subspace. The

Figure 8. Spectrogram of a cheering crowd.



final step consists of separate estimation of the IR of the individual components, according to the principles of IR estimation for multivariate/vector processes described earlier. Having described the different stages in the anticipation algorithm, we now describe the application of this method to audio characterization and musical analysis.

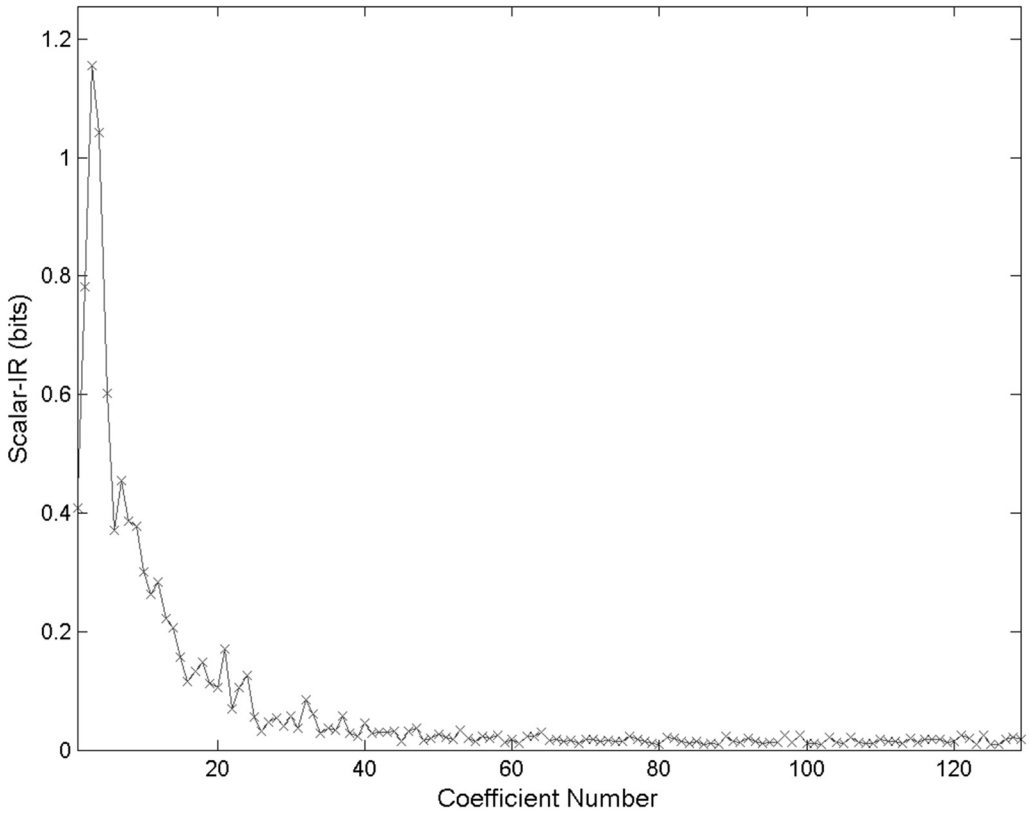
Comparison of Vector and Scalar-IR Analysis for Natural Sounds

In this and the following sections, we examine natural sounds that contain significant energy throughout all frequencies of the spectrum. Figure 8 shows the magnitude spectrum on a decibel scale (log-magnitude) of a cheering crowd sound. This sound contains a dense mixture of different types of sounds. Hand clapping can be seen as vertical lines, while the vocal exclamations appear as high-amplitude spectral lines varying in time.

Performing vector-IR analysis shows that different coefficients contain different IR structure. Figure 9 shows the result of IR analysis of a cheering-crowd signal sampled at 8 kHz using an FFT of size 256 with 50 percent overlap. A total of 129 spectral bins are retained (half the total bins due to symmetry of the STFT, plus the first bin). The IR analysis consists of decorrelation of log-magnitude spectral matrix using SVD and evaluation of scalar-IR separately for each of the expansion coefficients. Scalar estimation of IR of each of the coefficients consists of estimating SFM from the power spectral density of the coefficient time series using the Welch method (Hayes 1996) with 64 spectral bins.

The results of vector-IR analysis were compared to scalar-IR analysis of the signal. Additionally, a synthetic signal with a power spectral density simi-

Figure 9. Vector-IR analysis of cheering crowd signal using spectral log-magnitude features and SVD-derived bases.



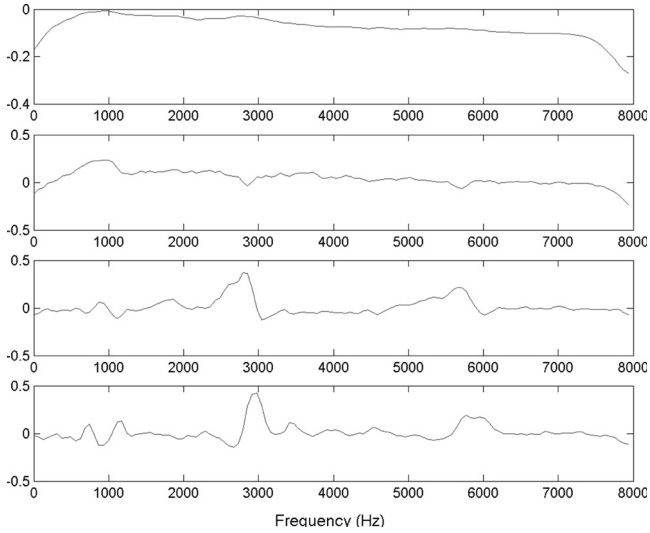
lar to that of the original cheering-crowd signal was constructed by means of passing white noise through an appropriate filter, whose parameters were estimated using linear prediction (LP) with eight filter coefficients. This synthetic signal, even though non-white, has little overall structure, because its different spectral bands (considered either as coefficients of STFT or outputs of a filterbank) lack a time structure. The log-magnitude spectral matrix of such a signal can be described as a rank-one constant matrix corresponding to single basis vector that captures the overall spectral shape, and a noise matrix that represents the variations between the different bands at different times. The SVD of such matrix contains a single structured basis vector that captures the overall spectral shape, and remaining basis vectors having white (and thus non-structured) coefficients. The purpose of testing IR for this synthetic signal was to examine the robustness of vector-IR analysis to the influence of an overall spectral shape, that is, to compare the real signal to a colored noise signal that does not have the detailed temporal structure within its individual bands. The results of vector-IR and scalar-IR analysis of both signals are given in Table 2.

Table 2. Results of Vector-IR and Scalar-IR Analysis of the Original Cheering-Crowd Signal and a Corresponding Synthetic Signal

	Original Signal	*Equivalent Filtered Noise*
Vector-IR	10.3	1.9
Scalar-IR	1.9	1.6

It can be seen that vector-IR efficiently detects the structure in the original cheering-crowd signal, distinguishing it from the corresponding filtered

Figure 10. First four spectral magnitude basis functions.



noise. Additional improvement in terms of achieving a bigger difference between vector-IR of the original signal versus spectrally matched noise signal can be achieved by discarding coefficients whose scalar-IR values fall below a threshold value. Using threshold values of 0.05 and 0.1 results in zero vector-IR for the matching noise signal, and vector-IR values of 8.5 and 8.0, respectively, for the original cheering-crowd signal. It should be noted that this method of dimension reduction is different from the usual methods that discard components according to their variance (energy); here, we discard components according to amount of their IR structure.

Figure 10 shows the first four basis vectors of the spectral log-magnitude data. Visual inspection of the spectrogram in Figure 8 shows that the first component roughly corresponds to an average overall spectral envelope, and the next basis vectors capture some of the spectral patterns that correspond to high energy peaks around 1, 3, and 6 kHz. Such visual inspection should be done by translating the gray levels in Figure 8 to values on the vertical axis of Figure 10 (amplitudes of the basis functions), with dark lines corresponding to high amplitude and brighter shades corresponding to weaker amplitudes.

The amount of structure (IR) of their respective expansion coefficients are 0.41, 0.78, 1.15, 1.04, as shown in Figure 9. It should be noted that the first basis function that has been ranked first in terms of energy (variance) by SVD actually has little structure in terms of IR.

Characterization of Natural Sounds Using Vector-IR

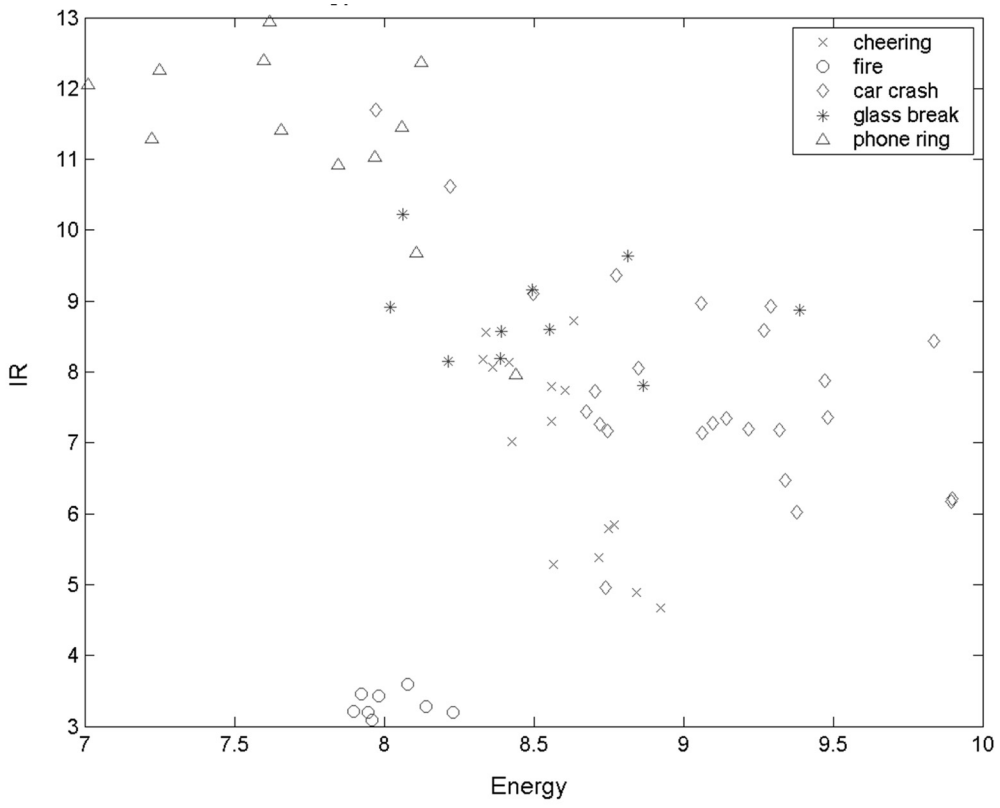
One possible application of IR analysis is as a descriptor for complex sounds such as natural sounds and sound effects. Such a descriptor provides characterization of a signal in terms of its overall "complexity," that is, considering the amount of structure that occurs in a sound when it is considered as a random process (as sound texture). In Figure 11, we represent several sounds in a two-dimensional plane consisting of a vector-IR axis and a signal energy axis. In this analysis, the IR estimation was preformed using cepstral features, excluding the first energy related coefficient, which was used for energy estimation. IR estimation was done using 30 cepstral coefficients, with frames 512 samples long with 50 percent overlap, and scalar-IR estimation using SFM with power spectral estimation using Welsh method with 64 spectral bins.

One may note that the energy and IR characteristics in Figure 11 correspond to our intuitive notions about the character of these sounds. For instance, fire sounds are the most noise-like, with little variation in energy. This can be compared to a phone-ringing sound that is highly "anticipated," and car crash, glass break, and cheering crowd that have intermediate levels of IR, with the car crash having the largest energy variation.

Anticipatory Listening and Music Applications

In this section, we reach what may be the most interesting and speculative application of the IR method, exploring the relationship between the IR measure and concepts of auditory and musical anticipation (Meyer 1956). In previous sections, we have applied a single IR analysis to a complete signal, resulting in a single number that describes the overall anticipation property of the sound. When listening to longer audio signals or music, the properties of time-evolving signals cannot be summarized into a single "anticipation number." Accordingly,

Figure 11. Energy and IR distribution of different natural sounds.



we extend the method of IR analysis to the non-stationary case by applying IR in a time-varying fashion.

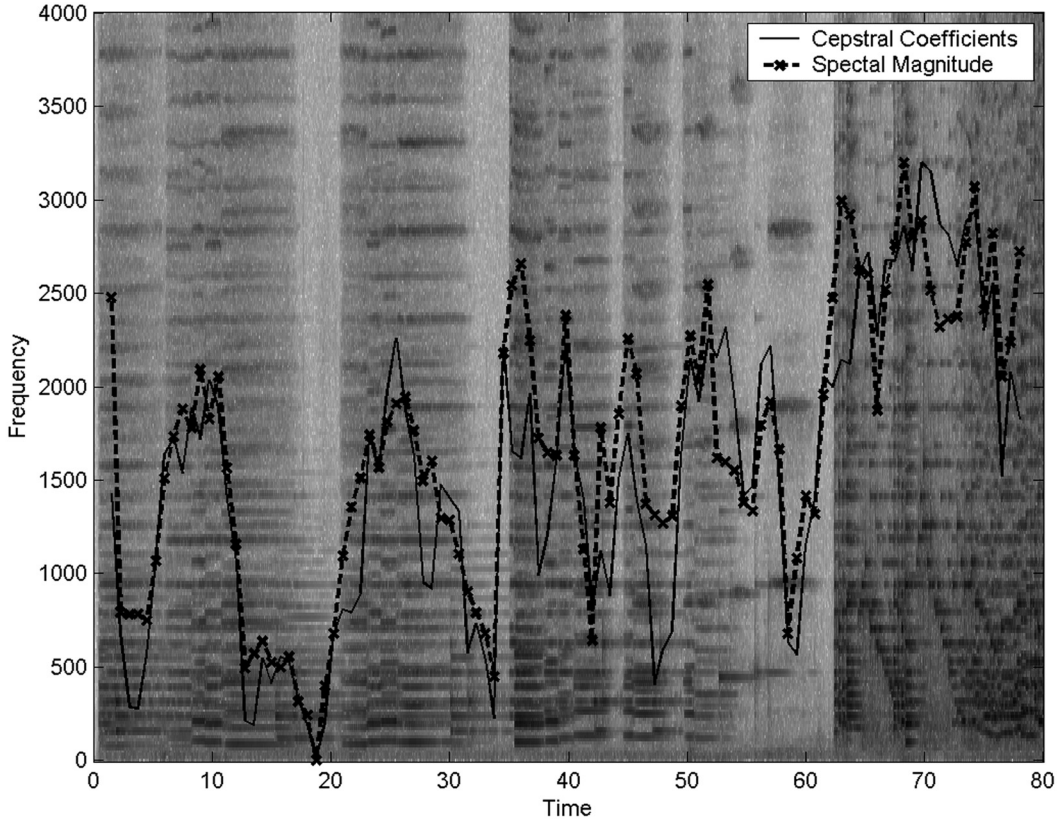
In the following example, we represented a sound in terms of a sequence of spectral envelopes, represented by spectral or cepstral coefficients. These vectors are grouped into macro-frames and submitted to separate IR analyses, resulting in a single IR value for each macro-frame. When the IR graph is plotted against time, one obtains a graph of IR evolution over the course of the musical signal; we call this the *anticipation profile.* Using different analysis parameters, anticipation profiles could be used to analyze affective vocalizations (Slaney and McRoberts 1998) or musical sounds (Scheirer 2000). For the purpose of vocalization analysis, the IR macro-frames are 1 sec long with 75 percent overlap. For analysis of musical recordings, longer frames are required, varying from 3 sec for solo or chamber instrumental music to 30 sec for complex orchestral textures. (The reason for this will be explained in the next section.)

Anticipation Profile of Musical Signals

A comparative vector-IR analysis of musical signal using 30 cepstral and 30 magnitude spectral basis components is presented in Figure 12 for an excerpt from the first movement of Schumann's *Piano Quartet in E-flat Major.* The features were estimated over signal frames of 20 msec duration. The macro-frame for IR analysis was 3 sec long with 75 percent overlap between successive analysis frames. Figure 12 shows vector-IR results for the two representations, overlaid on top of a signal spectrogram.

As can be seen from the figure, both methods result in similar anticipation profiles. Varying the order of the model (changing the data reduction or so called cepstral "liftering" number) between 20 and

Figure 12. Graph of the anticipation profile (estimated by vector-IR) using cepstral (solid) and spectral magnitude (dotted) features, displayed over a spectrogram of a musical excerpt (the first movement of Schumann's Piano Quartet in E-flat Major*).*



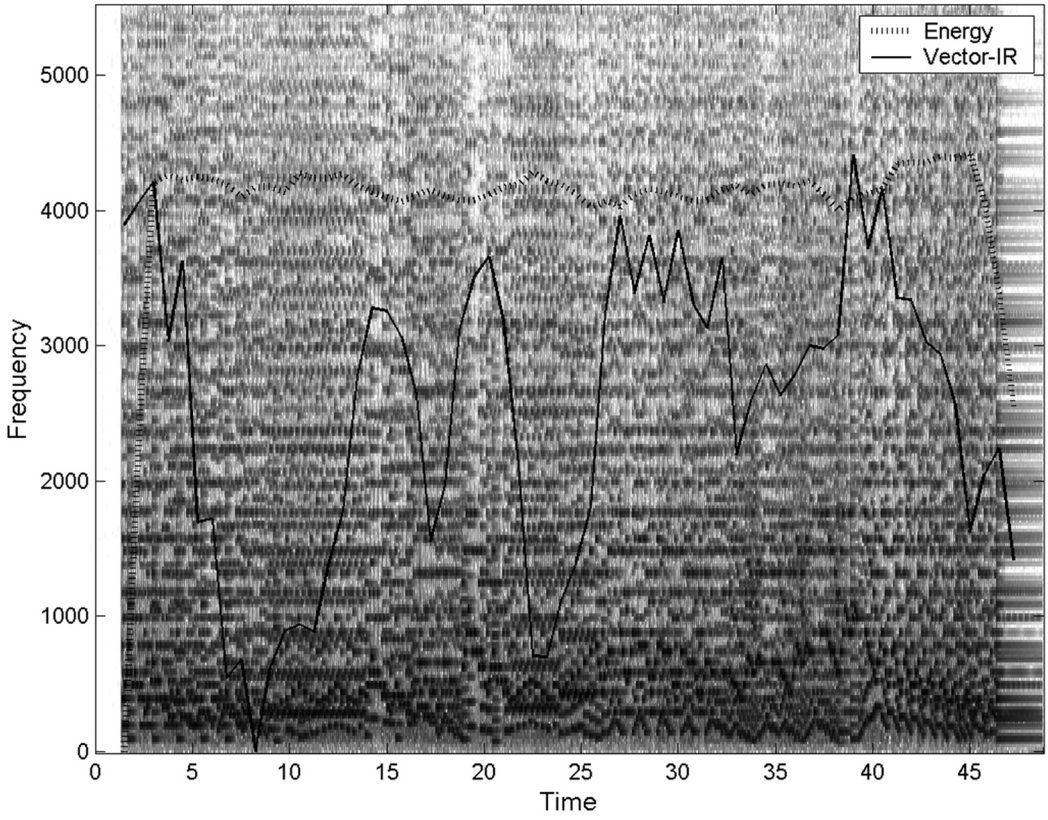
60 components has little effect on the results, indicating that the system is quite robust to changes in the representation detail.

When considering the results of this analysis, one might argue that the nature of music of the Schumann piano quartet is such that its structure might be detected using simpler methods, such as energy or other existing voice-activity detectors. To show a situation where vector-IR analysis gives a unique glimpse into musical structure that other features do not allow, we present in Figure 13 the results of vector-IR analysis of an acoustic recording of a MIDI rendering of Bach's *Prelude in G Major* from Book I of the *Well-Tempered Clavier.*

The synthetic nature of the computer playback of a MIDI file is such that the resulting acoustic signal lacks almost any expressive inflections, including dynamics. Moreover, the use of a synthetic piano creates a signal with little spectral variation. As can be observed in Figure 13, vector-IR still detects significant changes in the music. Analysis of similarities between the IR curves and the major sections of the score from a music-theoretical point of view seem to indicate that the anticipation profile captures relevant aspects of the musical structure. For instance, the first drop of the IR graph corresponds to the opening of the prelude, ending at the first cadence and modulation to the dominant. The following lower part of IR graph corresponds to two repetitions of an ostinato pattern. Then the two peaks in IR seem to be related to reappearance of the opening theme with harmonic modulations, ending in the next IR valley at the repeating melodic pattern on the parallel minor. The next increase in IR corresponds to development section on the dominant, followed by a final transition to cadence, with climax and final descent along the cadential passage.

To have a better impression of the correspondence between the variation in music structure and texture to our analysis, we present in Figure 14

Figure 13. Graph of anticipation profile (estimated by vector-IR) using 30 cepstral features, displayed over a spectrogram of a musical excerpt (Bach's Prelude in G Major *from Book I of the* Well-Tempered Clavier*). The acoustic signal was created by computer rendering of a MIDI file.*



again the same graph of anticipation profile, this time overlaid on top of MIDI notes. (The crosses indicate note onsets, with actual note numbers not represented in the graph.)

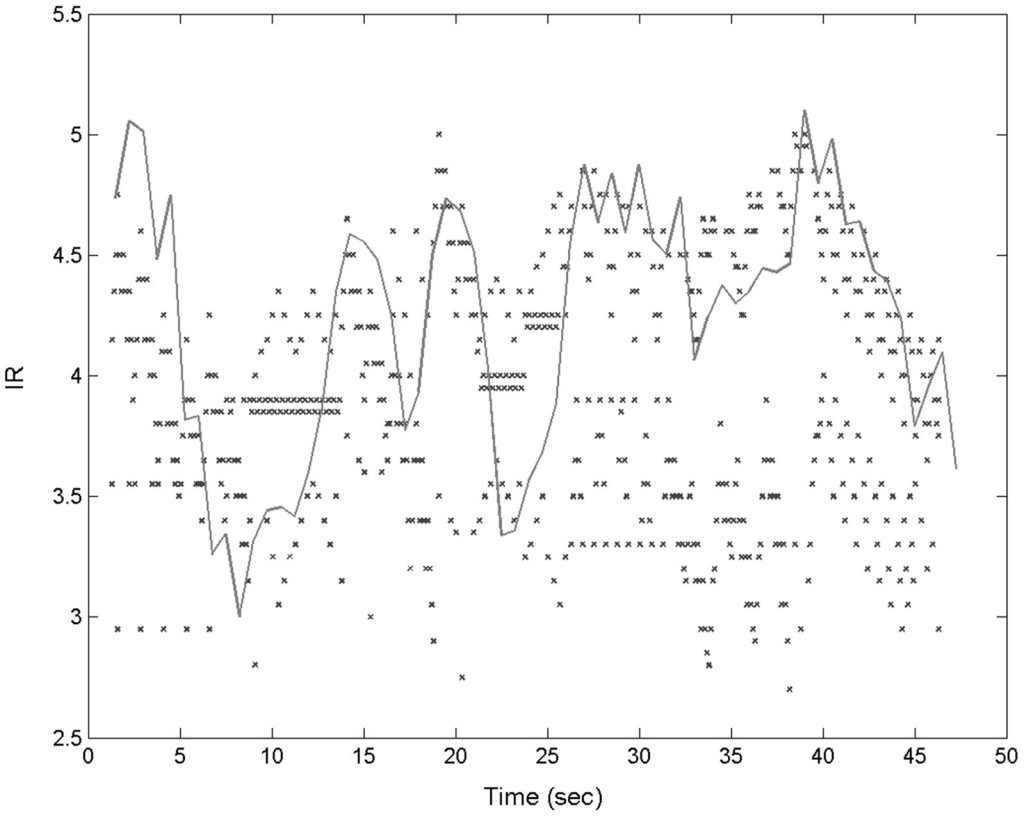
It appears that changes and repetitions of music materials are detected by IR analysis of the acoustic signal. (It should be noted that our anticipation analysis does not involve note or pitch detection or any use of the score of MIDI information.)

Anticipation Profile and Emotional Force

To evaluate the significance of the IR method for music analysis, a comparison between anticipation profiles derived from automatic signal analysis and human perception of musical contents is required. In a recent experiment, large amounts of data concerning human emotional responses when listening to a performance of a contemporary orchestral musical work (*Angel of Death* by Roger Reynolds) was collected during live concerts (McAdams et al. 2002). During these concerts, listeners were assigned a response box with a sliding fader that allowed continuous analog ratings to be made on a scale of emotional force (Smith 2001). Listeners were instructed that positive or negative emotional reactions of similar magnitude were to be judged at the same level of the emotional force scale. The ends of the emotional force scale were labeled "weak" and "strong." In addition, a small "I don't know" region was provided at the far left end of this scale that could be sensed tactilely, as the cursor provided a slight resistance to moving into or out of this zone.

Continuous data from response boxes were converted to MIDI format (on an integer scale from 0 to 127) and recorded simultaneously with the musical performance. Figure 15 presents a comparative graph of the anticipation profile resulting from IR

Figure 14. Graph of anticipation profile (estimated by vector-IR from an audio recording) displayed on top of MIDI note onsets of the Bach Prelude.



analysis of the audio signal (the concert recording) and a graph of the average human responses termed "Emotional Force" (EF) for two versions of the orchestral piece.

Analyses were performed using analysis frame sizes of 200 msec with macro-frames 3 sec long, with no overlap between the macro-frame segments. These IR values were additionally smoothed using a 10-segment-long moving-average filter, resulting in an effective analysis frame of 30 sec with a 3-sec interval between analysis values. The extra averaging removed fast variations in vector-IR analysis, assuming that a 30-sec smoothing better matches the rate of change in human judgments for such a complex orchestral piece.

As can be seen from Figure 15, certain portions of the IR curve fit closely to the EF data, whereas other portions differ significantly. It was found that correlation of the IR data and the EF were 63 percent and 47 percent for top and bottom graphs, respectively. Combined with additional signal features such as signal energy, higher correspondence between signal information analysis and Emotional Force judgments was achieved. The analysis shows strong evidence that signal properties and human reactions are related, suggesting applications of these techniques to music understanding and music information-processing systems. Full details of the experiments and the additional information analysis methods will appear elsewhere (Dubnov, McAdams, and Reynolds in press).

Thoughts About Self-Supervised Brain Processing Architecture

We would like to consider briefly in this section the possible relationships between anticipation analysis and self-supervised architectures of brain processing, particularly in relation to higher cognitive

Figure 15. Two graphs representing human judgments of Emotional Force (solid) and IR analysis (dot) of audio recordings of two musical performances.

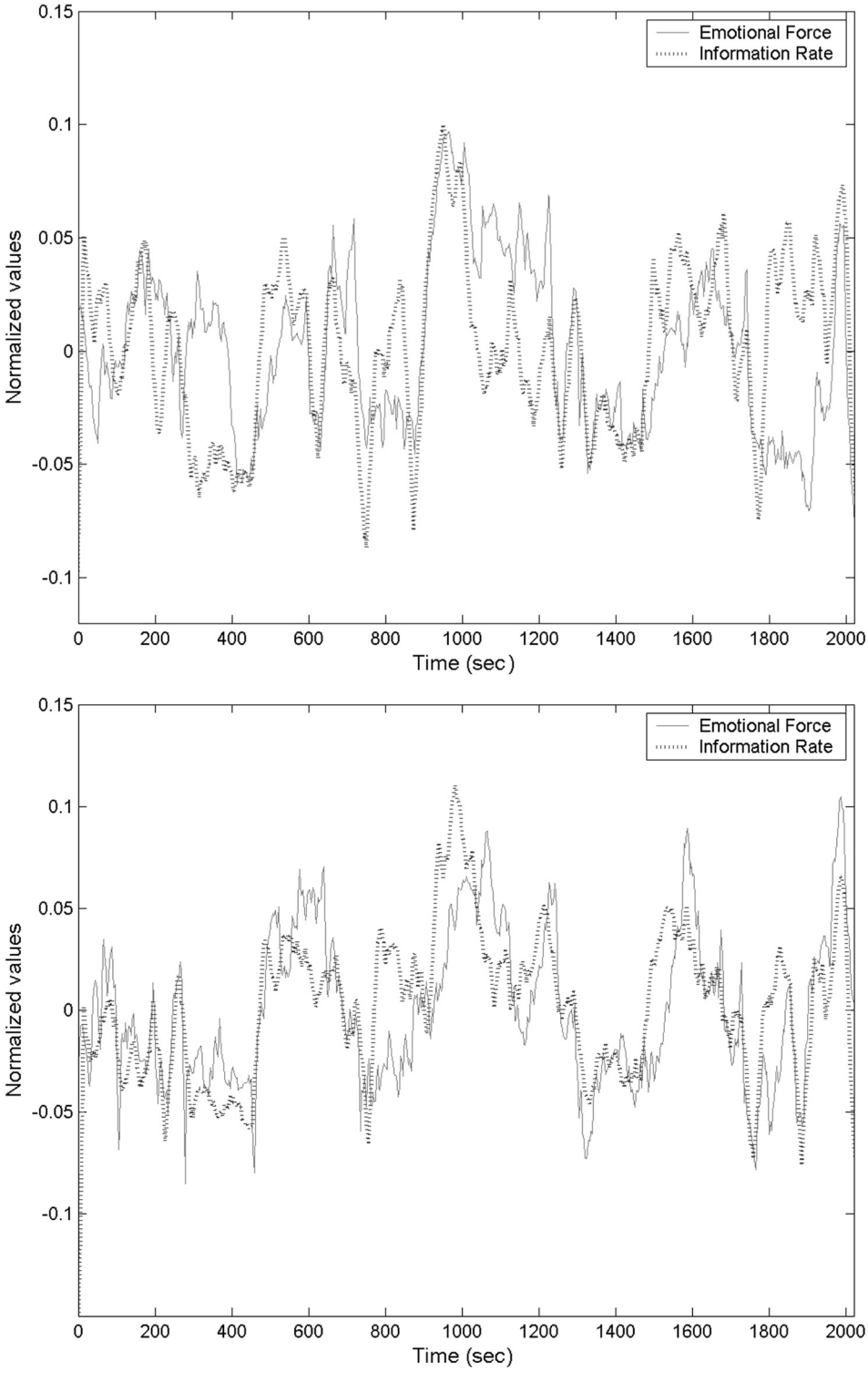
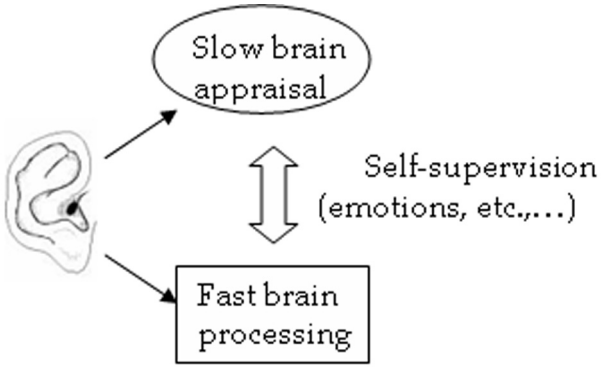


Figure 16. Information-processing architecture that considers emotion as a self-supervision communication within the brain.



aspects of musical information processing. One such architecture that was suggested for emotional processing assumes an existence of two separate brain mechanisms that interact in analysis of complex information (Huron 2005): a "fast brain" that deals primarily with pre-processing of sensory information, forming perceptions from the stream of constantly impinging sensory excitations, and a second component, the "slow brain," that interacts with the fast brain by performing appraisal of the fast-brain performance. This architecture is described in Figure 16.

In the context of our computational model, we consider the functions of the fast brain to be related to basic pattern-recognition actions, such as feature extraction and data reduction. Anticipation could be considered as an appraisal that is done by the slow brain. That is, anticipation involves evaluating the utility of the past perception for explaining the present in terms of assigning a score to the fast-brain functions according to the "relative reduction of uncertainty about the present when considering its past" (a quote from the definition of IR).

It should be noted that selection of informative features is commonly done in other signal-processing applications, such as speech understanding or computer vision, mostly in terms of classification performance. That is, features are considered to be informative if they have significant mutual information with the class labels for a particular recognition task. In music, the task of recognition is secondary, and it is only natural to assume that anticipation, rather then recognition, is a more appropriate task for describing music information-processing operations. In other words, information contents for music signals should be measured not by mutual information between signal features and a set of signal labels (recognition task), but as mutual information between past and present features (anticipation task).

Conclusion

This article presents a novel approach to (automatic) analysis of audio and music based on an "anticipation profile." The ideas are developed from a background of information theory and basis decomposition, through the idea of a vector approach to the anticipation measurement, and ending with some higher-level reflections on its meaning for complex musical signals. The anticipation profile is estimated by evaluating the reduction in the uncertainty (entropy) of a variable resulting form prediction based on the past. Algorithms for estimation of the anticipation profile were presented, with applications for audio signal characterization and music analysis. Discussions show that this measure also resolves several ill-defined aspects of the structure-versus-noise problem in music and includes the listener as an integral part of structural analysis of music. Moreover, the proposed anticipation measure might be related to more general aspects of appraisal or self-supervision of information-processing systems, including aspects of emotional brain architecture.

References

Allamanche, E., et al. 2001. "Content-Based Identification of Audio Material Using MPEG- 7 Low Level Description." *Proceedings of the 2001 International Symposium on Music Information Retrieval.* Bloomington, Indiana: Indiana University, pp. 197–204.

Casey, M. A. 2001. "MPEG-7 Sound Recognition Tools." *IEEE Transactions on Circuits and Systems for Video Technology* 11(6):737–747.

Casey, M. A., and A. Westner. 2000. "Separation of Mixed Audio Sources by Independent Subspace Analysis." *Proceedings of the 2000 International Computer Music Conference.* San Francisco, California: International Computer Music Association, pp. 154–161.

Cichocki, A., and S. Amari. 2002. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications.* New York: Wiley.
Cover, T. M., and J. A. Thomas. 1991. *Elements of Information Theory.* New York: Wiley.
Dubnov, S. 2002. "Extracting Sound Objects by Independent Subspace Analysis." Paper presented at the 2002 Audio Engineering Society International Conference, Espoo, Finland, 17 June.
Dubnov, S. 2003. "Non-Gaussian Source-Filter and Independent Components Generalizations of Spectral Flatness Measure." *Proceedings of the International Conference on Independent Components Analysis* (*ICA2003*), Nara, Japan, pp. 143–148.
Dubnov, S. 2004. "Generalization of Spectral Flatness Measure for Non-Gaussian Linear Processes." *IEEE Signal Processing Letters* 11(8):698–701
Dubnov, S., S. McAdams, and R. Reynolds. In press. "Structural and Affective Aspects of Music from Statistical Audio Signal Analysis." *Journal of the American Society for Information Science and Technology.*
Fitzgerald, D., E. Coyle, and B. Lawlor. 2002. "Sub-Band Independent Subspace Analysis for Drum Transcription." Paper presented at the 5th International Conference on Digital Audio Effects, Hamburg, Germany, 26–28 September.
Hayes, M. 1996. *Statistical Signal Processing and Modeling.* New York: Wiley.
Huron, D. 2005. "Six Models of Emotion." Available online at http://csml.som.ohio-state.edu/Music829D/Notes/Models.html.
Jayant, N. S., and P. Noll. 1984. *Digital Coding of Waveforms.* Upper Saddle River, New Jersey: Prentice-Hall.
Kim, H., N. Moreau, and T. Sikora. 2004. "Audio Classification Based on MPEG-7 Spectral Basis Representations." *IEEE Transactions On Circuits and Systems for Video Technology* 14(5):716–725.
Lewicki, M. S. 2002. "Efficient Coding of Natural Sounds." *Nature Neuroscience* 5(4):356–363.
McAdams, S., et al. 2002. "Real-Time Perception of a Contemporary Musical Work in a Live Concert Setting." Paper presented at the 7th International Conference on Music Perception and Cognition, Sydney, Australia, 17–21 July.
Meyer, L. B. 1956. *Emotion and Meaning in Music.* Chicago: Chicago University Press.
Oppenheim, A. V., and R. W. Schafer. 1989. *Discrete-Time Signal Processing.* Upper Saddle River, New Jersey: Prentice Hall.
Ozerov, A., et al. 2005. "One Microphone Singing Voice Separation Using Source-Adapted Models." Paper presented at the 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, 17 October.
Roweis, S. 2000. "One Microphone Source Separation." *Neural Information Processing Systems* 13:793–799.
Scheirer, E. D. 2000. *Music Listening Systems.* Ph.D. Thesis, Massachusetts Institute of Technology.
Slaney, M., and G. McRoberts. 1998. "Baby Ears: A Recognition System for Affective Vocalizations." Paper presented at the 1998 International Conference on Acoustics, Speech, and Signal Processing, Seattle, 12–15 May.
Smaragdis, P., and J. C. Brown. 2003. "Non-Negative Matrix Factorization for Polyphonic Music Transcription." Paper presented at the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, 21 October.
Smith, B. K. 2001. "Experimental Setup for *The Angel of Death.*" Available online at http://www.ircam.fr/pcm/bks/death.
Uhle, C., C. Dittmar, and T. Sporer. 2003. "Extraction of Drum Tracks from Polyphonic Music Using Independent Subspace Analysis." Paper presented at the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation, Nara, Japan, 1–4 April.

Appendix

Given a signal with power spectrum $S(\omega)$, the SFM is defined as

$$SFM = \frac{\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln S(\omega) d\omega\right)}{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln S(\omega) d\omega}$$

Rewriting it as a discrete sum gives

$$SFM(x) = \frac{\exp\left(\frac{1}{N}\sum_i \ln S(\omega_i)\right)}{\frac{1}{N}\sum_i S(\omega_i)} = \frac{\left(\prod_{i=1}^{N} S(\omega_i)\right)^{\frac{1}{N}}}{\frac{1}{N}\sum_i S(\omega_i)} \qquad (13)$$

which shows that SFM can be viewed as the ratio between the geometric and arithmetic means of signal spectra, thus being positive and less than or equal to unity. The SFM equals unity only if all

spectrum values are equal, thus meaning a flat spectrum or a white-noise signal.

Information Redundancy

Given a random variable x, with probability distribution $f(x)$, the entropy of the distribution is defined (Cover and Thomas 1991) as

$$H(x) = -\int f(x)\log f(x)dx \tag{14}$$

For the joint distribution of two variables x_1, x_2, the joint entropy is defined as

$$H(x_1, x_2) = -\int f(x_1, x_2)\log f(x_1, x_2)dx_1dx_2 \tag{15}$$

The average amount of information that the variable x_1 carries about x_2 is quantified by the mutual information

$$I(x_1, x_2) = H(x_1) + H(x_2) - H(x_1, x_2) \tag{16}$$

Generalization of the mutual information for the case of n variables yields

$$I(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} H(x_i) - H(x_1, x_2, \ldots, x_n) \tag{17}$$

This function measures the average amount of common information contained in variables $x_1, x_2, \ldots, x_n$. Using the mutual information, we originally define the information rate (IR), denoted ρ, to be the difference between the common information contained in the variables $x_1, x_2, \ldots, x_n$ and the set $x_1, x_2, \ldots, x_{n-1}$, namely, the additional amount of information that is added when one more variable is observed:

$$\rho(x_1, x_2, \ldots, x_n) = I(x_1, x_2, \ldots, x_n) - I(x_1, x_2, \ldots, x_{n-1}) \tag{18}$$

Because in our application we are considering time-ordered samples, this redundancy measure corresponds to the rate of growth of the common information as a function of time. It can be shown that the following relationship exists between redundancy and entropy:

$$\rho(x_1, x_2, \ldots, x_n) = H(x_n) - H(x_n \mid x_1, x_2, \ldots, x_{n-1}) \tag{19}$$

This shows that redundancy is the difference between the entropy (or uncertainly) about isolated x_n and the reduced uncertainty of x_n if we know its past. In information theoretic terms, and assuming a stationary process, this measure equals the difference between the entropy of the marginal distribution of the process x_n and the entropy rate of the process, equally for all n.

Relationship Between SFM and IR

To assess the amount of structure present in a signal in terms of its information content, we observe the following relationships between signal spectrum and entropy. Entropy of a "white" Gaussian random variable is given by

$$H(x) = \ln\sqrt{2\pi e\sigma_x^2} = \frac{1}{2}\ln\left(\frac{1}{2\pi}\int S(\omega)d\omega\right) + \ln\sqrt{2\pi e} \tag{20}$$

whereas the entropy rate of a Gaussian process (the so called Kolmogorov-Sinai Entropy) is given by

$$\begin{aligned} H_r(x) &= \lim_{N\to\infty}\frac{1}{N}H(x_1, \ldots, x_N) = \lim_{N\to\infty}\frac{1}{N}H(x_N \mid x_1, \ldots, x_{N-1}) \\ &= \frac{1}{4\pi}\int \ln S(\omega)d\omega + \ln\sqrt{2\pi e} \end{aligned} \tag{21}$$

According to the previous section, IR is defined as a difference between the marginal entropy and entropy rate of the signal $x(t)$, $\rho = H(x) - H_r(x)$. Inserting the expressions for entropy and entropy rate, one arrives at the following relation:

$$SFM(x) = \exp(-2\rho(x)) = \frac{\exp\left(\frac{1}{2\pi}\int \ln S(\omega)d\omega\right)}{\frac{1}{2\pi}\int S(\omega)d\omega} \tag{22}$$

One can see here that IR is equal to one-half of the logarithm of SFM.

Vector-IR as Sum of Independent Component Scalar-IR

Given a linear transformation $X = AS$ between blocks of the original data (signal frame of feature vector X)

and its expansion coefficients S, the entropy relationship between the data and coefficients is $H(X) = H(S) + \log|\det(A)|$. For a sequence of data vectors, we evaluate the conditional IR as the difference between the entropy of the last block and its entropy given the past vectors. (This is a conditional entropy, which becomes the entropy rate in the limit of an infinite past.) Using the standard definition of multi-information for signal samples $x_1 \ldots x_{nL}$,

$$I(X_1, X_2, \ldots, X_L) = \sum_{i=1}^{Ln} H(x_i) - H(x_1, \ldots, x_{Ln}) \tag{23}$$

we originally define and develop the expression for vector-IR as

$$\begin{aligned}
&\rho_L(X_1, \ldots, X_L) = I(X_1, \ldots, X_L) - I(X_1, \ldots, X_{L-1}) - I(X_L) = \\
&= \sum_{i=(L-1)n+1}^{Ln} H(x_i) - H(X_1, \ldots, X_L) + H(X_1, \ldots, X_{L-1}) - I(X_L) = \\
&= \sum_{i=(L-1)n+1}^{Ln} H(x_i) - H(X_L \mid X_1, \ldots, X_{L-1}) - I(X_L) = \\
&= H(X_L) - H(X_L \mid X_1, \ldots, X_{L-1})
\end{aligned} \tag{24}$$

This shows that the vector-IR can be evaluated from the difference of the entropy of the last block and the conditional entropy of that block given its past.

Using the transform relationship, one can equivalently express vector-IR as a difference in entropy and conditional entropy of the transform coefficients $\rho(X_1, \ldots, X_L) = H(S_L) - H(S_L \mid S_1, \ldots, S_{L-1})$. (Note that the dependence upon determinant of A is cancelled by subtraction.) If there are no dependencies across different coefficients and the only dependencies are within each of the coefficients sequences as a function of time (i.e., the trajectory of each coefficient is time-dependent, but the coefficients between themselves are independent), we arrive at the relationship

$$\begin{aligned}
H(S_L) &= \sum_{i=1}^{n} H(s_i(L)) \\
H(S_L \mid S_1 \ldots S_{L-1}) &= \sum_{i=1}^{n} H(s_i(L) \mid s_i(1) \ldots s_i(L-1))
\end{aligned} \tag{25}$$

Combining these equations gives the desired result:

$$\rho_L^n(X_1, X_2, \ldots, X_L) = \sum_{i=1}^{n} \rho(s_i(i), \ldots, s_i(L)) \tag{26}$$

MATLAB Code

MATLAB code of the IR analysis, along with specific examples of sound analysis used in this article, can be found online at http://music.ucsd.edu/~sdubnov/InformationRate.