Spectral Anticipations

Shlomo Dubnov

Music Department, University of California in San Diego
9500 Gilman Dr., La Jolla, CA 92037, email sdubnov@ucsd.edu

Introduction

This paper deals relations between randomness and structure in audio and musical sounds. Randomness, in layman sense, would be something that has an element of variation or surprise to it, while structure would be more predictable, rule-based or even deterministic. Dealing with noise, which is the “purest” type of randomness, one usually adopts the canonical physical or engineering definition of noise as signal with white spectrum, i.e. comprised of equal or almost equal energies in all frequencies. This seems to imply that noise is a complex phenomenon simply by the fact that it contains very many frequency components (mathematically, in order to qualify as random or stochastic process, the density of the frequency components has to be such that the signal would have a continuous spectrum, while periodic components would be spectral lines or delta functions). In contradiction to this reasoning comes 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 and organization. Musical notes or other repeating or periodic acoustic components in music are closer to being deterministic and could be considered as “structure”. But then, 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 situation, the ability to determine the structure of the signal can not be revealed by looking at signal spectrum alone. Therefore, the physical definition of noise as a
signal with smooth or approximately continuous spectrum seems to obscure other significant properties of signal versus noise, such as being a signal with, versus without temporal structure, respectively, or in other words signal that can or cannot be predicted.

The paper present a novel approach to (automatic) analysis of music based on an "anticipation profile". This approach considers dynamical properties of signals in terms of anticipation property, which is shown to be significant for discrimination of noise versus structure, 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 significant paradigm shift in approach to analysis of audio and musical signals. In our approach the signal is no more 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 this approach becomes characterization of a communication process over a time-channel between the music (acoustic signal’s present) and a listener (a system that has memory and prediction capabilities based on the signal’s past). The amount of structure is equated to the amount of information that is “transmitted” or “transferred” from signal 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 usage 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” type of
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 due to prediction is small, in the first case due to little variation in the signal to start with, while in the second case there is little reduction in uncertainty since prediction has little or no effect. Finally, it 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 paper are presented from a background of information theory and basis decomposition through to 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 presentation takes several approaches: a engineering or signal processing approach where notions of structure are related to basic questions about signal representation, an audio information retrieval approach where the ideas are applied for characterization of natural sounds, and a musical analysis approach, where time varying analysis of musical recordings is used to create an "anticipation profile" that describes their structure. This last aspect has interesting possibilities for exploring the relations between 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 a complex, multi-component musical signals. This measure uses concepts from Principle Components Analysis (PCA) and Independent Components Analysis (ICA) in order 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 anticipation of a complex signal from the sum of anticipations of its individual components. Accordingly, the term “complex signal” used throughout this paper 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 (human or machine) listener. The amount of information transmitted through this communication process depends on the nature of the music and the listener: The listener makes predictions, forms expectations; the music source generates new samples. Accordingly, we define structure to be the aspect of musical material that the listener can predict that, while noise or randomness is 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, image or even genetic codes are considered from information theoretic point of view, it is assumed that the specific data is a result of a production by an information source. The concept of information source allows description of many different sequences by a single statistical model which is their probability distribution. The idea of looking at a source, rather then a particular sequence, characterizes which types of sequences are probable and which are not, or in other words, what data is likely to be more frequently appearing then other. Information theory also teaches that “on 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 (the number of typical sequences divided by number of all possible sequences of 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 of only a small fraction of all possible sequences of “heads” and “tails”. In such a case, the proportion of the size of the typical set (i.e. a set comprising of mostly “heads”) relative to the number of all possible sequences (2 to the power of the number of coin flips we did) is low and the source will be considered to have little entropy or little uncertainty. More equally distributed sources (such as a flips of a fair coin) will have a larger typical set and high uncertainty.

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 relation between one type of data that is termed information source and another type of related but not identical data, called 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 hopefully very close to each other, 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.
Figure 1: Entropies $H(.)$ for separate variable $x$ and $y$ and the pair $(x,y)$, conditional entropies $H(x \mid y)$ and $H(y \mid x)$ and mutual information $I(x,y)$.

Denote by $H(x) = \sum P(x) \log P(x)$ the entropy of variable $x$ (a source that has probability $P(x)$), depicted by the left (gray) circle in the figure, $H(y)$ the right (dotted) circle of variable $y$, and $H(x,y)$ depicts the outer contour of both sources together. Conditional entropy $H(x \mid y)$ depicts the uncertainty about $x$ if $y$ is known or $H(y \mid x)$ vice versa. $I(x,y)$ is the cross-section (grid) that indicates how much overlap in uncertainty occurs between $x$ and $y$, i.e. 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 knowledge of $y$, 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 case is zero, since knowledge of one variable completely describes the other, leaving no conditional uncertainty, i.e. $H(x \mid y)=H(y \mid x)=0$.

Mathematically, the above relations have simple algebraic expression that are directly related to the marginal distributions $P(x)$, $P(y)$ and their common distribution $P(x,y)$.

$$I(X,Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

$$= H(X) + H(Y) - H(X,Y) = \sum P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$
Information Rate and Anticipation as Capacity of a “time-channel”

As explained above, mutual information is commonly used for theoretical characterization of the amount of information transmitted over a communication channel.

\[ I(x, y) = H(x) - H(x | y) \]

![Diagram of Information Source, Channel and Receiver](image)

Figure 2: Information Source, Channel and Receiver diagram. See text for more detail.

For our purposes we will consider a particular type of channel, which 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. Secondly, the receiver has to apply some algorithms to predict the current sample from its past. The information at the sink \( y \) 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 has now 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 Information Rate (IR, also to be called scalar-IR). IR is defined as the relative reduction of uncertainty about the present when considering its 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 IR equals to 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. amount of additional information that is added when one more sample of the process is observed

$$\rho(x_1, x_2, \ldots, x_n) = H(x_n) - H(x_n | x_{past}) = I(x_n, x_{past})$$

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

One can interpret IR as the amount of information that a signal carries into its future. This is quantitatively measured by the number 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 definition of structure versus noise for several example cases.

**Example cases:**

"Inverted U function" for the amount of structure:

A pure random signal cannot be predicted and thus has same uncertainty before and after prediction, resulting in zero IR. An almost constant signal on the other hand has small uncertainty $H(x)$, resulting in overall small IR as well. The types of signals that have high IR are the ones with a large difference between their uncertainty or variation without prediction $H(x)$ versus the remaining uncertainty after prediction $H(x | past of x)$. 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.

Relative “meaning” of noise:
Let us consider a situation where two systems are trying to form expectations about the same data. One system has a correct model, which allows making good predictions for the next symbol. In case when the uncertainty about the signal is large but the remaining uncertainty after prediction is small, the system manages to reveal the signal 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 case the quality of prediction, that can be measured by the amount 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 low IR.

**Signal Characterization:**

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

**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 loss of 1 bit in 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.

Equivalence of Scalar IR to Spectral Flatness Measure

Spectral Flatness Measure (SFM) [Jayant] is a well-known method for evaluation of the distance of a process from being a white noise. It is also widely used as a measure for “compressibility” of a process, as well as an important feature for sound classification [Allamanche]. Given a signal with power spectrum $S(\omega)$, SFM is defined as

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

and it is positive and less or equal to one, being equal to one 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 entropy rate is the limit of conditional entropy for large $n$, $H_r(x) = \lim_{n \to \infty} H(x_n | x_1, x_2, \ldots, x_{n-1})$. Using expressions for the entropy and entropy rate of Gaussian process, one arrives at the following relation

$$SFM(x) = \exp(-2\rho(x)).$$

Equivalently, one can express IR as a function of SFM

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

Figure 3 shows the close relation between spectral flatness and the IR measure for a jungle signal. 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, while segments that contain harmonic or narrow band noise have higher IR. It should be noted that this example contains alternating sounds of mostly pitched bird singing alternating with noisy bird cries, which is a “segmentation” type of processing that can be achieved 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.

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

2.3 Extension of IR for the case of multivariate (vector) process

In the case of complex signals, such as signal comprised of several components, signals described by sequences of spectral descriptors or sequences of feature vectors, we need to consider new type of IR that would be appropriate for dealing with 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 $X_1, X_2, \ldots, X_L$ and generalize the IR definition (to be called vector-IR) as
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, so that $S = TX$ and the components $S_1S_2...S_L$ after transformation are statistically independent. Using relations between entropies of a linear transformation of random vectors, it can be shown [Dubnov 2003] that IR may be calculated as a sum of IR’s of the individual components $s_i(n), i = 1..n$,

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

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 $s(t) = \sin(\omega t + \phi)$ 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

$$X_1 = [x_1, x_2, ..., x_n]^T, X_2 = [x_{n+1}, x_{n+2}, ..., x_{2n}]^T, ..., X_i = [x_{(i-1)n+1}, x_{(i-1)n+2}, ..., x_{in}]^T.$$

Concatenating $L$ signal segments into a matrix $X = [X_1X_2...X_L]$, 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 $\Lambda$ 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 4: Four basis vectors (shown as rows) obtained by SVD.

Figure 5 shows the corresponding expansion coefficients.

Figure 5: Expansion coefficients that correspond to the basis vectors of Fig. 7

We apply scalar-IR analysis to the first n-rows of the matrix $V^T$ using the SFM method of scalar-IR estimation. The sum of the scalar-IR’s makes the vector-IR. These estimates were compared to scalar IR of the individual sinusoidal and noise signals and their sum signal. In parenthesis 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-IR’s and transforming the resulting vector-IR back to SFM, i.e generalizedSFM = $\exp(-2(\text{vectorIR}))$ (this also equals to product of the individual column SFM’s). So, in principle, presence of a strong structural element (SFM close to zero) makes the generalized SFM small, indicating structure. This depends of course on the ability of SVD to separate out the structural components (since in practice SVD of sinusoidal+noise signal contained some errors in the structural components that it found, the resulting generalized-SFM is higher then the product of the SFM’s of the individual sinusoid or noise signals).

<table>
<thead>
<tr>
<th></th>
<th>Sinusoid</th>
<th>Noise</th>
<th>Sin+Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scalar IR (SFM)</strong></td>
<td>6.31 (3e-6)</td>
<td>8.7e-5 (0.99)</td>
<td>0.16 (0.72)</td>
</tr>
<tr>
<td><strong>Vector IR (SFM)</strong></td>
<td>8.18 (7e-8)</td>
<td>0.21 (0.657)</td>
<td>2.44 (0.007)</td>
</tr>
</tbody>
</table>

These results show that the amount of structure revealed by vector-IR (or generalized-SFM) for the sinusoidal+noise signal is significantly higher (SFM lower) then the structure estimated by scalar-IR on the sum signal. It should be noted also that vector IR estimation of noise components is imperfect, resulting in non-zero 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.

3. Audio Basis Representation
A common representation of audio is to transform the time signal into spectral representation using short-time Fourier transform (STFT). This processing is performed by applying 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.

3.1 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],[Lewicki],[Kim et al.] 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 feature extraction system is described in Figure 4. In order to assure independence between the transform coefficients for purpose of vector-IR analysis, additional statistical procedures have to be applied to the different transform coefficients (or in case of STFT different frequency channels).
Figure 6: Geometrical signal representation consisting of a transform operation followed by basis selection and data reduction.

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, to be described in the following section. Another common representation of spectral contents of audio signals is by means of cepstral coefficients [Oppenheim]. Cepstrum is defined as the inverse Fourier transform of a logarithm of the absolute value of Fourier transform of the signal \( C = \mathcal{F}^{-1}[\log(|\mathcal{F}\{x[n]\}|)] \). 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 due to pitch (the actual notes played) or other long-term signal correlations appear in the higher cepstral coefficients. Using part of the cepstrum allows an easy control over the type of spectral information that we would like to consider for 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 few strong components (basis functions). The rest of the signal (parts that contain noise, interference and etc.) are assumed to reside in a different subspace and hopefully are weak and approximately of equal energy. There are several methods for estimation of low rank models:

- Eigen-Spectral Analysis and Singular Value Decomposition (SVD).
• Principal Components Analysis (also known as Karhunen-Loeve Transform KLT)
• Independent Component Analysis methods (ICA)

Mathematical formulation of the model is the following: Assume that the measurements are collected into frames of n samples each, thus consisting a vector $x$. The model is written as:

$$X = AS + N$$

where $X$ are the actual measurements (resulting signal), $S$ are the expansion coefficients, sometimes also considered as the separate source components, $A$ is an $(m \times n)$ 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 $A$

\[
\begin{bmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_n 
\end{bmatrix} = \begin{bmatrix}
  a_1 & a_2 & \ldots & a_m 
\end{bmatrix} \begin{bmatrix}
  s_1 \\
  \vdots \\
  s_m 
\end{bmatrix} + \begin{bmatrix}
  n_1 \\
  n_2 \\
  \vdots \\
  n_n 
\end{bmatrix}
\]

In our case we are interested, in addition to the compaction or the noise removal property, also in independence of expansion coefficients $S$ in order to allow estimation of vector-IR. This can be achieved by means of Principle Components Analysis (PCA) for the case of a Gaussian multivariate process\(^1\). For more general type of multivariate processes (such as in case that the signals are non-stationary or non-Gaussian), more sophisticated methods such as independent components analysis (ICA) could be used [Cichocki and Amari]. For the purposes of the present

---

\(^1\) PCA procedure causes the components to be uncorrelated, which in general is not sufficient condition for statistical independence. In case of a Gaussian process, uncorrelated components are indeed independent.
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 singular value decomposition (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 $n$.

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

The problem of vector data decomposition into independent components is an active field of research. It should be noted that there are no closed form solutions for ICA for multi-component single-microphone signals. There are several works in the literature that attempt to perform one-microphone ICA [Roweis][Ozerov et al.] using adapted source-filter models, which are learned from prior examples of the single sources. These methods are not applicable to our purpose since training or prior adaptations are not possible.

4. 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. The first stage of analysis is deriving an appropriate geometrical representation, either as frames of audio samples or some
features extracted from these frames, such as STFT, Filter Banks, Cepstral, MFCC or some other method. The following step after initial transformation is basis decomposition, combined with data reduction by projection into a lower dimensional subspace.

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

4.1 Comparison of vector and scalar IR analysis for natural sounds

In this and the following sections we examine natural sounds that have a common characteristic of containing significant energy throughout all frequencies of the
spectrum. Figure 8 shows the magnitude spectrum on dB 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.

![Cheering Crowd Spectrogram](image)

**Figure 8: Cheering Crowd Spectrogram.**

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

The results of vector IR analysis were compared to scalar IR analysis of the signal. Additionally, a synthetic signal with power spectral density similar to that of the original cheering signal was constructed by means of passing a white noise through an appropriate filter. This synthetic signal, even though being non-white, has little overall structure since its different spectral bands (considered either as coefficients of STFT or outputs of a Filter Bank) 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 contain a single structured basis vector that captures the overall spectral shape, and remaining basis vectors having white (and thus non-structured) coefficients. Accordingly, 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, i.e. compare the real signal to a colored noise signal that does not have the detailed spectral envelope parameters for the colored noise signal were estimated using Linear Prediction (LP) with 8 filter coefficients.
temporal structure within its individual bands. The results of vector-IR and scalar-IR analysis of both signals are as follows,

<table>
<thead>
<tr>
<th></th>
<th>Cheering signal</th>
<th>Equivalent filtered noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector IR</td>
<td>10.3</td>
<td>1.9</td>
</tr>
<tr>
<td>Scalar IR</td>
<td>1.9</td>
<td>1.6</td>
</tr>
</tbody>
</table>

It can be seen that Vector IR efficiently detects the structure in the cheering signal, distinguishing it from the equivalent filtered noise. Additional improvement in terms of achieving a higher ratio between vector-IR of the original signal versus spectrally matching 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 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), while in our case we discard components according to amount of their IR structure.

Figure 10 shows the first four basis vectors of the spectral log-magnitude data. It can be seen, by visual inspection of the spectrogram in Figure 8, 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 1000, 3000 and 6000 Hz. Such visual inspection should be done by translating the colors in Figure 8 to values on the vertical axis of Figure 10 (amplitudes of the basis functions), with dark red corresponding to high amplitude and brighter red, yellow, green and blue corresponding to decreasing amplitudes, respectively.
The amount of structure (IR) of their respective expansion coefficients are 0.41, 0.78, 1.15, 1.04, 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.

4.2 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 descriptor provides characterization of a signal in terms of its overall “complexity”, i.e. considering the amount of structure that occurs in a sound when it is considered as a random process, i.e. as sound texture. In Figure 12

Figure 10: Four spectral magnitude basis functions.
we represent several sounds in a two dimensional plane consisting of vector-IR axis and 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 [Hayes] with 64 spectral bins.

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

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 phone ringing that is highly “anticipated”, and car crash, glass break and cheering that have intermediate levels of IR, with car crash having the most variation in terms of energy.\(^3\)

\(^3\) Variability here is considered in terms of variability across different samples or different instances sounds from the same class. It is not related to the variability of spectral contents or spectral anticipation that is analyzed by IR.
5. 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 relation between IR measure and the concepts of auditory and musical anticipation [Meyer]. In previous sections we have applied a single IR analysis to a complete signal, resulting in a single number that described 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, we extend the method of IR analysis to the non-stationary case by applying IR in a time varying fashion.

In the following we represented the 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, that we call “anticipation profile”. Using different analysis parameters, anticipation profile could be used to analyze affective vocalizations [Slaney],[SpeechComm] or musical sounds [Scheirer]. For purpose of vocalization analysis the IR macro-frames are 1 second long with 75 percent overlap. For analysis of musical recordings longer frames are required, varying from 3 seconds for solo or chamber instrumental music and up to 30 seconds for complex orchestral textures (the reason for such big frames will be explained in the next section).

5.1 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 Schumann Piano Quartet E-flat Major I. The features were estimated over signal frames
of 20 milliseconds duration. The macro-frame for IR analysis was 3 seconds 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.

![Vector IR Analysis](image)

**Figure 12: Music analysis: graph of anticipation profile (estimated by Vector IR) using cepstral (solid) and spectral magnitude (dot) features, displayed over a spectrogram of musical excerpt (Schumann Piano Quartet E-flat Major).**

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 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. In order to show situation where vector-IR analysis give a unique glimpse into musical structure that other features do not allow, we present in Figure 13 results of vector-IR analysis of
an acoustic recording of a MIDI rendering of a Bach Prelude in G from the first book of the Well Tempered Clavier.

Figure 13: Music analysis: graph of anticipation profile (estimated by Vector IR) using 30 cepstral features, displayed over a spectrogram of musical excerpt (Bach Prelude in G from book I of Well Tempered Clavier). The acoustic signal was created by computer rendering of a MIDI file.

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 variations. As can be observed in Figure 13, vector-IR still detects significant changes in the music, creating an anticipation profile that seems to capture significant aspects of its structure.

5.2 Anticipation Profile and Emotional Force
In order to evaluate the significance of the IR method for music analysis, a comparison between anticipation profile derived from automatic signal analysis and human perception of musical contents is required. In recent experiment large
amounts of data concerning human emotional response when listening to a performance of a contemporary orchestral musical work “Angel of Death” by R. Reynolds was collected during live concerts [McAdams et al, 2002]. During these concerts listeners were assigned a response box with a sliding cursor that allowed continuous analog ratings to be made on a scale of emotional force [Smith]. 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 since the cursor provided a slight resistance to moving into or out of this zone. Continuous data from response boxes were converted to MIDI format (integer scale from 0 to 127) and recorded simultaneously with the musical performance. Figure 13 presents a comparative graph of anticipation profile resulting from IR 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.
Figure 14: Graphs representing human judgments of Emotional Force (Red) and IR analysis (blues) of audio recordings of two musical performances.

The analysis was done using analysis frame size of 200 milliseconds with macro-frames 3 seconds long, with no overlap between the macro-frames segments. These IR values were additionally smoothed using a 10-segment-long moving-average filter, resulting in an effective analysis frame of 30 seconds with a 3 seconds interval between analysis values. The reason for the extra averaging was to remove fast variations in vector-IR analysis and assuming that a 30 second smoothing better matches the rate of change in human judgments for such complex orchestral piece.

As can be seen from Figure 5, certain portions of the IR curve fit closely to the EF data, while other portions differ significantly. It was found that correlation of the IR data and the EF were 63% and 47% 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 et al., 2006].

5.3 Some thoughts about self-supervised brain processing architecture

We would like to consider briefly in this section the possible relations between anticipation analysis and self-supervised architectures of brain processing, in relation to higher cognitive 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]; 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 so-called slow brain, that interacts with the fast brain by performing appraisal of the fast brain performance. This architecture is described in Figure 14. 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, i.e. 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).
Figure 14: Information processing architecture that considers emotion as a self-supervision communication between the brain and the self

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, i.e. 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 than 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 paper presents a novel approach to (automatic) analysis of audio and music based on an "anticipation profile". The ideas in the paper are developed from a background of information theory and basis decomposition through to 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 from prediction based on the past. Algorithms for estimation of anticipation profile are presented, with applications for audio signal characterization and music analysis. Discussions show that this measure also resolves several ill-defined aspects of the structure vs. 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:**


Huron, D., *Six Models of Emotion*,


Smith, B. K. *Experimental setup for The Angel of Death.*

Appendix:

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

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

Rewriting it as a discrete sum gives

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

which shows that SFM can be viewed as the ratio between the geometric and arithmetic means of signal spectra, thus being positive and less or equal to one. SFM equals one only if all spectrum values are equal, thus meaning a flat spectrum or a white noise signal.

### A.1 Information Redundancy

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

$$H(x) = -\int f(x) \log f(x) dx$$

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

$$H(x_1, x_2) = -\int f(x_1, x_2) \log f(x_1, x_2) dx_1 dx_2$$

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

$$I(x_1, x_2) = H(x_1) + H(x_2) - H(x_1, x_2)$$

Generalization of the mutual information for the case of $n$ variables is

$$I(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} H(x_i) - H(x_1, x_2, \ldots, x_n)$$
This function measures the average amount of common information contained in variables $x_1, x_2, \ldots, x_n$. Using the mutual information we define marginal information redundancy (sometimes called just Information Redundancy or IR) to be the difference between the common information contained in the variables $x_1, x_2, \ldots, x_n$ and the set $x_1, x_2, \ldots, x_{n-1}$, i.e. the additional amount of information that is added when one more variable is observed.

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

Since 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 relation exists between redundancy and entropy

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

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 to the difference between the entropy of the marginal distribution of the process $x_n$ and the entropy rate of the process, equally for all $n$.

### A.2 The relation between SFM and IR

In order to assess the amount of structure present in a signal in terms of its information content, we observe the following relations between signal spectrum and its entropy. Entropy of a “white” Gaussian random variable is given by

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

while the entropy rate of a Gaussian process (the so called Kolmogorov-Sinai Entropy) is given by
\[
H_r(x) = \lim_{N \to \infty} \frac{1}{N} H(x_1, \ldots, x_N) = \lim_{N \to \infty} \frac{1}{N} H(x_N \mid x_1, \ldots, x_{N-1}) = \frac{1}{4\pi} \int \ln S(\omega) d\omega + \ln \sqrt{2\pi e}
\]

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

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

One can see equally that IR is equal one half of a logarithm of SFM.

**A.3 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 relations 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 entropy rate in the limit of an infinite past). Using the standard definition of multi-information for signal samples \(x_1 \ldots x_{nL}\),

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

we write the vector IR as

\[
\rho^n_L(X_1, \ldots, X_L) \triangleq I(X_1, \ldots, X_L) - I(X_1, \ldots, X_{L-1}) - I(X_L) =
\]

\[
= \sum_{i=(L-1)n+1}^{Ln} H(x_i) - H(X_1, \ldots, X_L) + H(X_1, \ldots, X_{L-1}) - I(X_L) =
\]

\[
= \sum_{i=(L-1)n+1}^{Ln} H(x_i) - H(X_L \mid X_1, \ldots, X_{L-1}) - I(X_L) =
\]

\[
= H(X_L) - H(X_L \mid X_1, \ldots, X_{L-1})
\]
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 relation one can equivalently express vector IR as a difference in entropy and conditional entropy of the transform coefficients \( \rho^n_L(X_1, \ldots, X_L) = H(S_L) - H(S_L \mid S_1, \ldots, S_{L-1}) \) (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 the each coefficient sequence 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 relation

\[
H(S_L) = \sum_{i=1}^{n} H(s_i(L))
\]

\[
H(S_L \mid S_1 \ldots S_{L-1}) = \sum_{i=1}^{n} H(s_i(L) \mid s_i(1) \ldots s_i(L-1))
\]

Combining these equations give the desired result

\[
\rho^n_L(X_1, X_2, \ldots, X_L) = \sum_{i=1}^{n} \rho(s_i(i), \ldots, s_i(L))
\]