Machine Audition: Principles, Algorithms and Systems

Wenwu Wang
University of Surrey, UK
Chapter 16
Musical Information Dynamics as Models of Auditory Anticipation

Shlomo Dubnov
University of California in San Diego, USA

ABSTRACT
This chapter investigates the modeling methods for musical cognition. The author explores possible relations between cognitive measures of musical structure and statistical signal properties that are revealed through information dynamics analysis. The addressed questions include: 1) description of music as an information source, 2) modeling of music–listener relations in terms of communication channel, 3) choice of musical features and dealing with their dependencies, 4) survey of different information measures for description of musical structure and measures of shared information between listener and the music, and 5) suggestion of new approach to characterization of listening experience in terms of different combinations of musical surface and structure expectancies.

GENERAL INTRODUCTION
The research on modeling musical cognition involves multiple perception modalities, with expectancies playing one of the central roles in shaping the experience of musical structure. An idea put forward many years ago by music theorists such as Meyer and Narmour states that listening to music consists of forming expectations and continual fulfillment or denial thereof (Narmour 1990, Meyer 1956). Recently several information theoretic measures of audio structure have been proposed in attempt to characterize musical contents according to its predictive structure (Dubnov, 2006; Dubnov et al. 2006; Potter et al., 2007; Abdallah and Plumbley 2009). These measures consider statistical relations between past, present and future in signals, such as predictive information that measures the mutual information between limited past and the complete future, information rate that measures information between unlimited past and the immediate present and predictive information rate that tries to combine past and future in view of a known present. Additional
models of musical expectancy structure build upon short-long term memory neural networks and predictive properties that are local to a single piece versus broader knowledge collected through corpus-based analysis.

The underlying assumption in investigation of Musical Information Dynamics is that the changes in information content, that could be measured in terms of statistical properties such as entropy and mutual information, correlate with musically significant events, which in parallel could be captured by cognitive processes related to music perception and acquired through exposure and learning of regularities present in a corpus of music in a certain style. These models may provide an explanation for the “inverted-U” relationship often found between simple measures of randomness (e.g. entropy rate) and judgments of aesthetic value (Rigau et al.).

In this chapter the authors will explore possible relations between cognitive measures of musical structure and statistical signal properties that are revealed through such information dynamics analysis. The questions they will try to address are: 1) description of music as an information source, 2) modeling of music–listener relations in terms of communication channel, 3) choice of musical features and dealing with dependencies, 4) survey of different information measures for description of musical structure and measures of shared information between listener and the music, and 5) suggestion of new approach to characterization of listening experience in terms of different combinations of musical surface and structure expectancies.

**STRUCTURE OF THIS CHAPTER**

After a brief introduction to the theories of expectancy in music and some historical as well as modern musicological and computational background, the authors will address the question of modeling music as an information source and listener as information receiver. From this approach they will develop a model of listening that is based on mutual information between the past and the present called Information Rate (IR) (Dubnov, 2006; Dubnov et al. 2006). This model will be extended to include Predictive Information (PI) (Bialek et al.) and Predictive Information Rate (PIR) (Abdallah and Plumbley, 2009).

The authors introduce a new notion of Information Gap as a measure of the salience of a present musical segment (instance in time) with respect to future and past of a musical signal. This measure combines notions of predictive information with a notion of momentary forgetfulness, determining saliency of the present instance in terms of how detrimental forgetfulness is on the ability to make predictions. They will show that the information gap unites the three notions of information dynamics through simple algebraic relations. Next they will consider application of IR for simple Markov chains (Cover and Thomas, 2006), and consider actual musical data from MIDI representation and cepstral audio features from recordings. Dealing with multiple features requires orthogonalization, which establishes the basis for vector IR.

Information theoretic, statistical or corpus-based approaches rely on adaptive capabilities of the listening apparatus to extract rules from data rather than being pre-wired or explicitly taught through an expert. The authors will discuss the question of symbolic rules versus learning algorithms in the context of neural network model (Potter et al., 2007) that creates melodic pitch expectancies for minimalist music using Short and Long Term models. Short-term affective and long-term cognitive familiarity features will be discussed in relation to spectral feature expectancies and spectral repetitions (Dubnov et al., 2006) for a contemporary large orchestral piece. They will discuss model-based Bayesian approaches to surprise and salience detection in non-musical cases, such as Bayesian surprise in images (Itti and Baldi, 2005).
These experiments suggest that a principled approach is required for combination of expectancies from surface and structural predictions. Surface predictions are finding regularities in the data in a short-term stationary regime using information about the next sample from its immediate past. When used in a sliding window manner, an anticipation profile captures changes in surface expectancy over a complete musical piece. Clustering of musical features over larger periods leads to measure of expectancies for structural repetition. This creates an anticipation profile of expectancies for model change, capturing novelty and surprise in a musical form. The authors will show how both aspects of expectancy can be derived from Information Rate, depending on the underlying assumptions about the type of data distribution (type of musical source) and choice of the listening mechanism that is employed by the listener when creating these expectancies.

**EXPECTANCY IN MUSIC**

Modeling musical listening in terms of information processing has to deal with modeling of cognitive functions that try to “make sense” of incoming auditory data using various cognitive processes as well as relying on natural and learned schemata (Temperley, 2004; Huron, 2006; Narmour, 1990; Meyer, 1956). One of salient functions is forming anticipations on different time scales and structural levels of music information. The structure of music can be seen as a complex network of short and long-time relations between different music parameters, established through schemata related to repetitions and variations. In order to evoke a musical experience, a listening model needs to be able to recognize and predict such different structures in the audio signal. Moreover, response to music is unique in the sense that in addition to operating on specific perceptual categories such as sense of loudness, pitch, dissonance versus consonance, recognition of rhythm, tonality, and so on, an important aspect of musical “meaning” is related to the process of activating memories and forming expectations in an attempt to capture temporal structural information in the musical data. So “understanding” music is ultimately linked with a specification of the listening mechanism, such as a computer audition system, that encapsulates a set of listening operations. The process of allocating resources during the listening act considers music itself as an organization of sounds that creates an experience through a process of active listening and appraisal of its own abilities in discovering such organization. Following this line of thought, we distinguish here between two levels of listening: one of forming expectations about the immediate musical events, which, depending on the type of musical representation, could be either notes in a musical score or spectral features in a recording. We will term this type of information as Musical Surface, to distinguish it from Musical Structure or larger musical form that creates a different set of expectations related to likelihood of repetition of larger blocks of sonic materials during a musical piece.

Musical theorists suggest that meaning in music is related to emotions that arise through the process of implication-realization (Narmour 1990). In that sense, past musical material is framing our appraisal of what comes next, such as expecting a resolution after a melodic leap in a musical phrase. These expectations could be “natural”, i.e. described by simple mathematical relations, or be learned from cultural conventions, requiring complex syntactic models or be style specific. Going beyond melodic expectations, framing of larger scale expectancy can be established through conventions of style known to the listener through familiarity with typical patterns (schematic expectations), large scale repetition structure typical to certain musical forms, and even expectations framed as recognizable references to other musical works called veridical expectations (Huron, 2006).
Musical Information Dynamics as Models of Auditory Anticipation

Figure 1. Information Dynamics are represented in terms of different mutual information measures between Past, Present and Future

<table>
<thead>
<tr>
<th>X _\text{PRESENT}</th>
<th>Time “Channel” – Mutual Information</th>
<th>X _\text{PAST}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music</td>
<td></td>
<td>X _\text{FUTURE} | X _\text{PAST}</td>
</tr>
</tbody>
</table>

Box 1.

\[
\begin{align*}
\text{IR} & : \quad I(X_{\text{PRESENT}}, X_{\text{PAST}}) = H(X_{\text{PRESENT}}) - H(X_{\text{PRESENT}} | X_{\text{PAST}}) \\
\text{PI} & : \quad \text{I}^{\text{pred}}(T) = \lim_{T \to \infty} I(X_{\text{FUTURE}}(T^1), X_{\text{PAST}}(T)) \\
\text{PIR} & : \quad I(X_{\text{PRESENT}}, X_{\text{FUTURE}} | X_{\text{PAST}}) = H(X_{\text{FUTURE}} | X_{\text{PAST}}) - H(X_{\text{FUTURE}} | X_{\text{PRESENT}}, X_{\text{PAST}})
\end{align*}
\]

Communicative Listening Act

The model of musical listening that we develop comprises of a pair music-listener that are connected through a virtual time-channel where present or next (future) musical observations enter the channel at the transmitter end and the past appears at the receiving end. This model is shown in Figure 1, and is described using the following variables: X \_\text{PAST}, X \_\text{PRESENT}, X \_\text{FUTURE}.

The idea is that the receiver (listener) holds some past information that represents both his ability to make predictions based on earlier heard materials belonging to the current musical piece, and possibly employing long term prior musical information acquired through training or exposure to other works in the same genre. Using the past, present and future we define three measures of anticipation: information rate (IR) (Dubnov 2006), predictive information (PI) (Bialek et al. 1999) and predictive information rate (PIR) (Abdallah and Plumbley, 2009) as shown in Box 1.

It should be noted that all three measures use slightly different definitions of what Past, Present and Future are. In case of IR, the present is a single observation and the past is of arbitrary length, which could be infinite. If we observed samples \( x_{[1:n]} = x_1, x_2, \ldots, x_n \) till the current moment \( n \), then the present is \( X_{\text{PRESENT}} = x_n \), and the past is \( X_{\text{PAST}} = x_{[1:n-1]} \). Since the beginning of time can be earlier then the start of the present musical piece, the past can extend to minus infinity. Another extension of IR that will be considered later is allowing a whole buffer of samples to exit in the present. In the case of PI, there is no clear notion of present, or one might say that present is included in the future as its first sample. What is special about PI is that it explicitly defines the duration of the past and the future, where \( T^1 \) is the extent of the future and \( T \) is the duration of the past, relative to some zero point where the two meet. Since PI was originally defined for continuous time, we will modify it for discrete time and define it as follows. If we assume that the present (zero time) point is \( n \), then the past...
Musical Information Dynamics as Models of Auditory Anticipation

and future become \( X_{\text{PAST}}(T) = x_{[n-T,n-1]} \) and \( X_{\text{FUTURE}}(T^+) = x_{[n,n+T^+-1]} \).

In the PIR approach, an explicit distinction is made between Past, Present and Future. In this case the definitions of the different random variables become \( X_{\text{PAST}} = x_{[n-1]} \), \( X_{\text{PRESENT}} = x_{n} \), and \( X_{\text{FUTURE}} = x_{[n+1,...]} \). It should be noted that the durations of the past or future are not defined, leaving it open to interpretation of the reader, or defined according to the specific problem at hand. Let us discuss these three measures in more detail.

In the first approach the past is used to predict the present. In this model there is no further guessing beyond the immediate present sample, and the overall listening act is measured in terms of its average ability to predict this sample given several past samples over some time period. Since the predictability of the present based on the past varies in time, the measure is only segment-wise stationary. Modeling of this situation requires two variables \( X_{\text{PRESENT}} \) and \( X_{\text{PAST}} \) that establish a sort of communication channel, where the mutual information between \( X_{\text{PRESENT}} \) and \( X_{\text{PAST}} \) represents the amount of information transferred from the past to the present. The listener tries to reduce the uncertainty of \( X_{\text{PRESENT}} \) by using information in \( X_{\text{PAST}} \), and his achievements in doing so are measured in terms of relative reduction in uncertainty, calculated as a difference between the entropy of \( X_{\text{PRESENT}} \) and conditional entropy of \( X_{\text{PRESENT}} \) given \( X_{\text{PAST}} \). We call this method Information Rate (IR) (Dubnov 2006). It should be noted that in order to account for mutual information between past and the present, the measure takes into account the overall of listener’s expectations about the relations between \( X_{\text{PRESENT}} \) and \( X_{\text{PAST}} \). In other words, IR measures the average reduction of uncertainty regarding the present using the past for all realizations of both, which in practice means that prediction errors are averaged over a larger macro-frame where multiple occurrences of past and present observations are available. The issue concerning the differences between predictions of the future given a particular past and measuring mutual information between all possible realizations of the two will be discussed in the section on PIR, and also in relation to the difference between Bayesian surprise and average Bayesian surprise.

In a second scenario, the listener tries to predict the entire future based on the past. It is a more demanding task since the prediction quickly deteriorates as we try to guess more then just a few next observations. One of the reasons for considering this scenario is that it alleviates the need to define what the present is. In some sense, the future is as long as the correlation between the past and the future allows it to be. Since in principle the uncertainty about a big segment of sound grows proportionally with the size of the segment, it can be shown that this view, called Predictive Information (PI) (Bialek, 1999), actually reveals the so called “intensive” part of the entropy of the available past, where the difference between intensive and extensive quantities is that extensive quantity grows with the size of the system (like mass for instance) while intensive quantity remains constant with size (like heat). So if we write the entropy as a function of time, then \( H(T) \) can be expressed in terms of a combination of extensive and intensive parts as follows

\[
H(T) = H_0(T) + H_1(T).
\]

Using this expression, PI can be written in terms of relation between entropies of the past, present and the combined time interval.
Since the intensive part does not grow in time, 
\[ \lim_{T \to \infty} \frac{H_i(T)}{T} = 0 , \] for sufficiently long future \( T^i \) the intensive parts of \( H(T^i) \) and \( H(T + T^i) \) cancel out:

\[ I_{\text{pred}}(T) = \lim_{T \to \infty} I(X_{\text{FUTURE}}(T^i), X_{\text{PAST}}(T)) = H_i(T) \]

So by looking into infinite future not only the extensive parts cancel out, but also the intensive parts of the future disappear, leaving only the intensive part related to the past at hand. Looking into larger blocks of future is important in the case of long correlations. In the following we will show that in the case of low order Markov models, predictive information in fact equals information rate, where the present is defined by the size of the memory of the Markov process. PI measure is described in Figure 2.

A third scenario assumes that the present can be distinguished from both the past and the future. In such a case we might think of a listener who holds past instance of music information in his memory. The question now becomes one of measuring information contained in the present segment regarding its ability to predict the entire future, considering that the past is already known. To explain this notion better we would need to use the following relation between relative entropy, also called Kullback-Liebler (KL) distance, and mutual information. Mutual information between two random variables \( X \) and \( Y \) can be written in terms of KL distance as:

\[ I(X, Y) = D[P(X, Y) || P(X)P(Y)] . \]

This can be shown to arise directly from the definitions of KL distance and definitions of entropy and conditional entropy

\[ D(X, Y) = \sum P(X) \log \frac{P(X)}{P(Y)} , \]
\[ H(X) = -\sum P(X) \log P(X) , \]
\[ H(X \mid Y) = -\sum P(X, Y) \log P(X \mid Y) \]

respectively.

It should be noted that conditional entropy averages log-conditional probability of \( X \) given
Musical Information Dynamics as Models of Auditory Anticipation

Box 3.

\[
I(X, Y) = D[P(X, Y) \| P(X)P(Y)] = \sum P(X, Y) \log \frac{P(X, Y)}{P(X)P(Y)} = \\
= \sum_y P(Y)\sum_x P(X \mid Y = y) \log \frac{P(X \mid Y = y)}{P(X)} = E\{D[P(X \mid Y = y) \| P(X)]\}_{P(Y)}
\]

Box 4.

\[
I(X = x, Y \mid Z = z) = D[P(Y \mid X = x, Z = z) \| P(Y \mid Z = z)]
\]

Box 5.

\[
I(X, Y \mid Z) = E\{D[P(Y \mid X = x, Z = z) \| P(Y \mid Z = z)]\}_{P(X, Z)}
\]

\(Y\) over both \(X\) and \(Y\). If instead of considering every possible event \(Y\) we assume a specific outcome \(Y = y\), conditional entropy can be considered as averaging of the entropy of \(X\) over every outcome of \(Y\) (see Box 2). This also allows writing mutual information in terms of averaging of KL between the probability of \(X\) for specific outcome of \(Y\) and probability of \(X\) without seeing that outcome, as shown in Box 3.

This can be extended also for the case of the variables \(X\) and \(Y\) being conditional on a third variable \(Z\). Assigning times labels to three variables as \(X = X_{\text{present}}, Y = X_{\text{future}}, Z = X_{\text{past}}\), this measure becomes the instantaneous predictive information rate that measures the distortion (in KL sense) between future distribution of observations with or without knowing specific observations of the present \(X\), but with the past \(Z\) being known in both cases (see Box 4). PIR, defined as mutual information between present and future given the past, can be obtained by averaging over instantaneous predictive information over all possible outcomes of past and present (see Box 5).

It can be shown that in this scenario PIR actually represents a difference between two predictive information situations – one where the past and future are adjacent to each other, and the other where there is a gap between past and future that we call present. By doing so, we realize that PIR is a measure of difference between two types of prediction, one that uses the present and one that skips it to predict the future. We will use a notion of information gap to show how detrimental forgetfulness or absentmindedness of a listener might be on information transmission from past to present. It can be shown that in the case of simple Markov processes, there is a “sweet spot” in terms of correspondence between predictive information rate and the entropy rate of the process, where entropy rate is actually the lower bound on the entropy that exploits all possible knowledge about the process that is available through knowledge of the Markov transition matrix. So in cases where
Musical Information Dynamics as Models of Auditory Anticipation

the structure is very loose or very strong, i.e. when entropy rate is close to zero or close to maximum, PIR is low, and a peak occurs at some intermediate values where many but not all continuations are possible for the different steps. The intuition behind this measure is that forgetting one step in such scenario is detrimental in the mid-range, while for totally random or totally predictive processes being absent minded for a moment does not have much effect on the ability to predict the future (zero in one and perfect prediction in the second extreme case).

Informational Aesthetics

Eighty years ago, Birkhoff formalized the notion of beauty by introducing the aesthetic measure (M), defined as the ratio between order and complexity. According to this measure, aesthetic feelings originate from discovery of some harmonious interrelations inside an object, dependent on its initial complexity. He identified three successive phases in the aesthetic experience related to a preliminary effort of attention, which increases proportionally to the object’s complexity, and eventually the feeling of value or aesthetic measure coming from this effort through discovery of structures such as harmony, symmetry, or order factors which seems to be necessary for evoking of an aesthetic effect. This measure was later formulated in terms of information theory (Bense 1969) using the notion of compressed versus uncompressed representation. Complexity is measured in terms of uncompressed data size, while order becomes the difference between this and size of data when compression is applied to it. We will modify this idea and replace compression with conditional entropy using past knowledge. In such a case the past is used to “compress” the present, or render it into a more compact representation. This can be written with $X$ representing the present and $Z$ the past or prior knowledge as follows

$$M = \frac{O}{C}$$

$$C : Complexity = H(X)$$

$$O : Order = H(X) - H(X \mid Z)$$

$$M = \frac{H(X) - H(X \mid Z)}{H(X)} = I(X, Z) / H(X)$$

It is interesting to note that in this information theoretical setting, the order component in aesthetic measure has close resemblance to IR measure of the listening act. In this case the ability to compress the signal is considered as a function of time, where the past is used to discover order in the present. Complexity becomes a measure of entropy without using prediction, and the difference between the two, which is the mutual information between past and present, becomes the measure of order. The complete expression of $M$ includes also a normalization factor that bounds the values of IR to be between zero and one. Simple algebraic manipulation shows that $M$ is one minus the ratio of conditional entropy divided by entropy. So in all respects, the ability to reduce uncertainty about an object using some efficient encoding, such as prediction or employing other rules to discover redundancies can be considered as a measure of the efficiency of the viewing or listening system when it comes do “deal” with a new visual or musical work. How this efficiency is translated into pleasure, aesthetics or some other experiential effect will be a question to be dealt with later.

MUSIC AS AN INFORMATION SOURCE

Dealing with uncertainty and redundancy as musical features requires some sort of statistical approach to describing the music itself. Over the years many attempts were done to capture the structure in music and audio in terms of statisti-
Musical Information Dynamics as Models of Auditory Anticipation

...cal models. 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 we observe is only one of many possible realizations that can be produces by that information source. This approach allows description of many variations or families of different sequences through a single statistical model. The idea of looking at a source, rather than a particular sequence characterizes which types of sequences are probable and which are not, or in other words, we care more about what types of data are more or less likely to appear rather then any particular realization. For instance, constructing histograms of note appearances in a tonal music can be used to address problems such as key-finding (Temperley, 2004), i.e. the process of inferring the key from a pattern of notes.

To do so, a musical piece, represented in terms of the musical score, is divided into segments, such as individual measures, and the number of pitch-classes (notes “wrapped” inside one octave range) are recorded. These patterns of probability are then compared to some earlier trained distributions of pitch-classes where the tonality is known. The Bayesian approach to key-finding then asks the question of how likely it is for the current empirical distribution of pitches to be generated by one of the known keys. In this respect, a key is considered as a random number generator, absent the notion of time, producing notes so that some are more likely (like the tonic, dominant or subdominant notes) then others, with alterations (non tonal notes) having especially low probability. In order to consider the possibility that the key may change from one segment to the next, an additional probability is given to the likelihood of a key-change, information obtained again through some prior training, sometimes also using musical knowledge about possible chord progressions and modulation represented as a musical grammar. An overall most likely pass through a trellis of possible key probabilities and key-changes can be derived using dynamic programming approaches, resulting in a profile of most likely key changes over the entire piece. Such Bayesian perspective leads to a simple, elegant, and highly effective model of key-finding process; the same approach can also be extended to other aspects of music perception, such as metrical structure and melodic structure. Bayesian modeling also relates in interesting ways to a number of other musical issues, including musical tension, ambiguity, expectation, and the quantitative description of styles and stylistic differences (Temperley, 2004).

In this respect, Leonard B. Meyer’s remarked on the fundamental link between musical style, perception, and probability: “Once a musical style has become part of the habit responses of composers, performers, and practiced listeners it may be regarded as a complex system of probabilities.... Out of such internalized probability systems arise the expectations—the tendencies—upon which musical meaning is built.... The probability relationships embodied in a particular musical style together with the various modes of mental behavior involved in the perception and understanding of the materials of the style constitute the norms of the style” (Meyer, 1957).

**Texture Modeling**

A more difficult question arises when dealing with natural sounds, such as audio textures, or considering music that has no clear definition of structural categories, such as tonality. Moreover, one might also consider aspects of listening to tonal music in terms of statistics of spectral features without relating it to musical properties such as tonality or rhythm. One such application is Audio Texture modeling (Zhang et al., 2004) that attempts to synthesize long audio stream from an example audio clip. The example clip is first analyzed to extract its basic constituent patterns. An audio stream of arbitrary length is then synthesized by recombination of these patterns as building blocks.
To discover structure one needs to define a distance measure between instances of audio at different times. Recurrence matrix, also known as similarity matrix is a method for showing structural relations in terms of a matrix that shows pair-wise similarity between audio features at all time instances.

Two common methods for constructing similarity matrix are (1) normalized dot product of the feature vectors (Foote 2001), or (2) $e^{-\beta d}$, where $d$ is a distance function and $\beta$ is a scaling factor that controls how “fast” increasing distance translates to decreasing similarity.

Figure 3 shows a similarity matrix of an example sound using dot product of cepstral feature vectors of a recording of J.S. Bach Prelude in G major from Book I of the Well Tempered Clavier. The bright red areas in the matrix correspond to high similarity and dark or blue areas are different. We will use this similarity matrix for partitioning the sound into perceptually similar groups, and will relate it to aspects of information dynamics and perception of familiarity.

Converting similarity matrix $S_{ij}$ into Markov probability matrix is done by assuming that instead of continuing from a current location to the next segment, a jump to another location in the piece is possible if the contents of those target segments are similar (Lu et al. 2002). So instead of a linear progression through the piece, the musical piece is viewed in terms of a generative Markov process where every segment can have multiple next step targets, with probability of jump being proportional to the distance between the next step and all other segments in that piece.

This matrix represents statistics of the data in terms of probability of transition from frames $i$ to $j$. Denoting by $X_i$ the feature vector that summarizes the observations in macro-frame $i$, we derive probability for transition from frame $i$ to $j$ from similarity matrix as follows:

$$P(j \mid i) = \frac{S(X_{i+1}, X_j)}{\sum_j S(X_{i+1}, X_j)}$$

A stationary vector can be derived through eigenvector analysis of the transition matrix $P = P_{ij}$, finding a vector $\vec{p}^*$ so that $\vec{p}^* P = \vec{p}^*$ (Cover and Thomas, 2006). Note that due to the indexing convention chosen for the transition matrix, Markov process operates by left side matrix multiplication. The stationary vector then is a left (row) eigenvector with an eigenvalue that equals to one. The meaning of stationary distribution is that we are looking at a situation where the transitions between the states settle into a “stable” set of probabilities. In terms of information dynamic analysis, we will assume that $\vec{p}^*$ represents the knowledge of the listening system about the musical form. This knowledge will then be used to form a prior for estimation of structure related expectations. In the following we will use
this approach to describe musical from based on similarity between instances of sounds (marcoframes) whose duration is of the order or magnitude of “perceptual present”, i.e. between 3 or 7 seconds for short chamber works (solo or small ensemble) and up to 10 or 12 seconds in the case of large scale orchestral works.

**Dictionary Based Approaches**

Dictionary based models try to address the problem of first-order Markov chains that assume dependence on the last symbol only. Higher-order Markov models assume a longer context, so that generating the next symbol depends on several symbols back into the past. It was shown (Brooks et al, 1993) that at very low orders—such as the second order or so-called bigram—Markov models generate strings that do not recognizably resemble strings in the corpus, while at very high orders, the model simply replicates strings from the corpus. Dictionary-based methods can be used to model the musical (information) source in terms of a lexicon of motifs and their associated prediction probabilities.

To generate new instances (messages), these models “stochastically browse” the prediction tree in the following manner: Given a current context, check if it appears as a motif in the tree. If found, choose the next symbol according to prediction probabilities. If the context is not found, shorten it by removing the oldest (leftmost) symbol and go back to the previous step. By iterating indefinitely, the model is capable of producing a sequence of symbols that presumably corresponds to a new message originating from the same source. In some cases, this procedure might fail to find an appropriate continuation and end up with an empty context, or it might tend to repeat the same sequence over and over again in an infinite loop. The methods for finding such dictionaries include lossless compression methods based on Lempel-Ziv parsing method or lossy methods such as Probabilistic Suffix Trees (Dubnov et al, 2003)

**Audio Oracle**

Other methods for generalizing an instance of audio into a generative model include the Audio Oracle (AO) (Dubnov et al., 2007) that outputs an automaton that contains pointers to different locations in audio recording that satisfy certain smoothness of continuation and similarity criteria. For synthesis the resulting automaton is loaded into an audio generation module that randomly traverses the automaton outputting a new audio stream by concatenation of audio frames that appear on transitions between AO states. The unique property of the AO is in the construction of a graph that points from every moment in the sound to another moment in that same sound that has the longest common past (common suffix). This assures a sense of continuity and smooth concatenation during synthesis, while creating interesting variations in the output audio stream. AO uses string-matching algorithm known as Factor Oracle (Allauzen et al., 1999; Assayag and Dubnov, 2004), generalizing it to the case of imprecise matching for dealing with sequences of feature vectors rather then exact matching of discrete symbols. In comparison to the Audio Texture method of previous paragraph, AO enjoys from longer overlapping segments between concatenated instances due to its longest common suffix structure. This process was extended to query by example over audio databases by defining a distance measure between AO and new sound. This method has applications to synthesis by using a sound query as a “guideline” for recreating the best matching concatenative synthesis from segments found in a database (Cont et al., 2007). This may lead to new characterization of information dynamics in terms of the ability to predict portions of a new sound by finding best matching segments in a database of prior models.
SURFACE VERSUS STRUCTURE: LONG AND SHORT TERM INFORMATION DYNAMICS

Pearce and Wiggins (2006) evaluate a statistical model of musical pitch perception that predicts the expectation generated by monodic tonal melodies. Their computational system is based on n-gram models commonly used in statistical language modeling (Manning & Schutze, 1999). An n-gram is a sequence of n symbols and an n-gram model is simply a collection of such sequences each of which is associated with a frequency count. During the training of the statistical model, these counts are acquired through an analysis of a corpus of sequences (the training set) in the relevant domain. When the trained model is exposed to a new example, the frequency counts associated with n-grams are used to estimate a probability of the next symbol given the n -1 preceding symbols. The experiments of Pearce and Wiggins were conducted on a set of synthetic examples, comparing statistical prediction to that of listeners who were asked to rate continuation tones following a two-tone context.

A more elaborate application of this approach was done by (Potter et al 2007). Their expectancy model was built from two memory models, one short-term memory (STM) and one long-term memory (LTM). Each model takes as its musical surface sequences of musical notes as written in the score, defined in terms of properties such as onset time, pitch, duration and key. The representation scheme can also express derived features (such as pitch interval) and interactions between features, using the “viewpoint” practice of Conklin and Witten (1995). The LTM was trained on a database of about 900 tonal melodies, while the STM, conversely, had no prior knowledge, and learned dynamically, only from the current piece of music. In this way, the author claim to model “typical human Western musical experience” (LTM) and “on-going listening to unknown music” (STM). The model is applied to study of Philip Glass’ Gradus and Two Pages (both written around 1968), which are monodic and isochronous pieces, the first for solo saxophone and the second for a keyboard (synthesizer). In their paper the authors report a detailed analysis of how the model predictions correspond to musically meaningful events as analyzed by a human expert.

The same Philip Glass works were also analyzed using Markov models and evaluating their performance in terms of predictive-information measures (Abdallah and Plumbley, 2009). The Markov model was learned from the current piece (“on-going listening”) with one elaboration that allowed the transition matrix to vary slowly with time in order to track changes in the musical structure. This could be considered as a model where the listener’s belief state is represented in terms of a probability distribution over all possible Markov transition matrices, and this belief is slowly updated during the course of listening to the piece. In order to learn the space of Markov models the probability to observe a transition matrix is represented as a product of Dirichlet distributions, one for each column. At each time step, the distribution over the space of transition matrices undergoes two changes – the probabilities slightly broaden to represent a “forgetting” or diffusion process on one hand, and they are also updated to represent more truthfully the next observation. This update for the Dirichlet distribution is done by incrementing by one the Dirichlet parameter corresponding to matrix element (i,j) upon observing a symbol i following symbol j in the data.

In the case of Two Pages, close correspondence is shown between the information measures and the structure of the piece, and in particular, between the six ‘most surprising moments’ as evaluated by a human expert and so called “model information rate” that measures the relative entropy or Kullback-Liebler (KL) divergence between prior and posterior model distributions $D(P(\theta | X = x, Z = z) \parallel P(\theta | Z = z))$, i.e. the
change in the distribution over model parameter space when the present observation \( X=x \) has been incorporated into the model. Gradus, which is much less systematically structured than Two Pages did not give such a clear picture. Regarding surface information dynamics, and specifically the predictive information rate, the authors’ findings were inconclusive.

## INFORMATION DYNAMICS IN AUDIO SIGNALS

In the previous works the methods of information dynamics were applied to symbolic sequences of musical notes, mostly monophonic (one voice). In order to analyze information contents in audio signals several steps need to be done. First features need to be extracted from the audio signal that are both audibly and structurally significant. Second, the statistical model needs to be determined, and finally the amount of fit between the predictions of the model and the actual data need to be determined in order to be able to estimate information dynamics.

### Spectral Anticipation

Spectral Anticipation (SA) is a method of Information Rate (IR) analysis of audio recording based on prediction of spectral features. Justification for this choice of features is that spectral information is an important descriptor of the audio contents and that a listening apparatus should be able to recognize, classify or predict spectral properties.

Spectral representation, as shown in Figure 4, is achieved by applying Fourier analysis to blocks of audio samples using a sliding-window that extracts short signal segments (also called frames) from the audio stream. It is often desirable to reduce the amount of information in spectral description so that it captures only essential spectral shapes. In doing so a balanced tradeoff between reducing the dimensionality of data and retaining the information contents can be achieved.

Estimation of IR is done as follows:

- Sound is analyzed and represented in terms of a sequence of \( N \) dimensional feature vectors \( v_t \) over time
- A block (macro-frame) of features vectors starting at time \( t \) and of duration \( L \) is selected \( v_{[t:t+L-1]} \). This macro frame represents a segment of the original sound in terms of features
- The sequence of feature vectors in a macro-frame is transformed into orthogonal representation using Singular Value Decomposition (SVD) (Cover and Thomas, 2006). This gives a set of approximately independent principal components \( u_{[1:L]} \) under Gaussian assumption. In other cases, independent component analysis can be
• Mutual information between past and present in macro-frame is estimated separately for every component $i$ of the feature vector $r_u[i]$. One method of estimation is the spectral flatness measure (Jayant and Noll, 1984; Dubnov, 2004) where each of the $N$ feature components is transformed from being a time sequence of $L$ elements in a macro-frame to a $K$ dimensional spectral magnitude vector $U_i$, estimated using Welch, Burg or other methods (Cover and Thomas, 2006). It can be shown that spectral flatness can be used as a measure of mutual information between a sample and its past, derived by comparison between signal variance and variance of the prediction error, equivalent also to so called “coding gain” of linear prediction (Jayant and Noll, 1984).

\[
e^{-2\rho[u_{1:L}(i)]} = \frac{\prod_{k=1}^{K} U_i(k)^{1/K}}{1/K \sum_{k=1}^{K} U_i(k)}
\]

• The values of mutual information of all elements are summed together to give the overall spectral anticipation (also called vector-IR) measure for that frame

\[
\rho(x_{[t:t+L-1]}) = \sum_{i=1}^{N} \rho[u_{1:L}(i)]
\]

• The process is repeated for the next macro-frame at time $t = t + \Delta$

This process is summarized in Figure 5.
Bayesian Surprise and Average Bayesian Surprise

In order to deal with long-term structures we introduce a model parameter space that is parameterized by some variable $\theta$. These parameters are themselves random variables that are distributed over a space of possible parameter values according to a probability function $P(\theta)$. Measuring mutual information between past and present in model space can be done by comparison between a prior distribution over the model parameters that is readily available to the listener before new data arrives (i.e. a model distribution based on the past), and a new model distribution that is inferred using new observations once the new musical data becomes available. Let us denote by $P(\theta)$ the prior (past) distribution and by $P(\theta | X)$ an a posteriori distribution that takes into account the data $X$. It should be noted the variable $X$ denotes a set of observations rather than a single present sample. Since the discussion here is rather general, we abandon the specific separation between past, present and future, and consider $X$ as a generalized notion of present in terms of a block of samples in an analysis window.

The difference between distribution over the space of model parameter after and before specific data was observed can be measured in terms of Kullback-Libler distance $KL[P(\theta | X) || P(\theta)]$ and is termed Bayesian Surprise (Itti and Balidi, 2005).

The difference between KL and IR is that IR measures the information transfer from past to the present in terms of mutual information between present data and prior model, which requires taking into account all possible combinations of model and data. As is shown in the next equation, IR between past model and present data is in fact an average KL over all possible values of the present, (see Box 6) where $P(\theta)$ and $P(X)$ are the marginal distributions of the “old” model parameters and the distribution of “new” data, respectively, and where without loss of generality we assumed a continuous probability distribution function of these variables.

To summarize, the Bayesian approach takes into account the updates of the model distribution space that happen as a result of the arrival of new data without averaging over the chances that this data will appear. This approach requires an online learning of the model parameters, continuously tracking the changes in the “beliefs” about the models.

Spectral Recurrence

A different, though closely related anticipation considers the changes in distribution of the observations that occur due to changes of the underlying models. The idea of a Spectral Recurrence (SR) profile is to be able to summarize the repetition structure into a function of time that indicates how likely an empirical distribution of features in a current frame is in comparison to an overall
Musical Information Dynamics as Models of Auditory Anticipation

In order to do so, we start with a pair-wise comparison of feature distributions over some larger segments in time that we call macro-frames. Ideally, we would like to measure these changes in terms of KL distance between empirical distributions, but doing so is impractical since this requires many observations. Instead, one might consider only partial statistics, such as macro-frame centroids or correlation, and compute KL distance between their Gaussian distributions with these statistics (see Box 7).

If covariances are neglected then KL becomes Euclidian distance between the macro-frame centroids. To turn the distance matrix into a similarity matrix we use an exponential function

\[ S_{ij} = e^{-d_{ij}} \]

where \( d \) is a distance function and \( \beta \) is a scaling factor that controls how “fast” increasing distance translates to decreasing similarity, similar to the way it was done for audio texture modeling. Then, after some preprocessing, we use eigenvector analysis of the processed similarity matrix \( L \), known as spectral clustering (von Luxburg, 2007), to reveal data clusters. This clustering method does an embedding of the matrix \( L \) in a low dimensional space spanned by its first few eigenvectors. From this embedding cluster centers and their variances can be deduced, so that a new empirical sample can be compared to them.

In practice, it appears that only very few eigenvectors of processed musical self-similarity matrix are non-zero. A particularly simple case is one where the preprocessing consists of normalization of the similarity columns by their sums, notated below as an inverse of a diagonal matrix of these sums, called the rank matrix

\[ L = D^{-1}S, \quad D_{ii} = \sum_{j=1}^{N} S_{ij} \]

This effectively turns the similarity matrix into a stochastic or Markov matrix, where the columns describe the probability of transition from a state that is the column index, to next states that are the row indices. This method has great resemblance to the Markov modeling approach of sound texture with one slight difference that a transition between state \( i \) and \( j \) is based on similarity between actual macro-frames \( i \) and \( j \) rather than between the next frame \( i+1 \) and frame \( j \). In our experience, this distinction makes little difference in the outcomes.

One of the advantages of this simple case is that the eigenvector derived from the analysis of the stochastic matrix already contains the probabilities of appearance of the macro-frames relative to complete musical piece. In other words, the clusters themselves do not need be estimated, and the probability of a frame can be looked up directly from the value of the eigenvector. The disadvantage of this method is that relying on the first eigenvector alone is considered to give poor performance in spectral clustering. Moreover, conceptually, what we have is essentially a fully observed model without any inference of parameters or hidden variables in the usual sense. This would render much of the machinery developed so far unnecessary - a fully observed Markov chain model could be fitted, using the procedure described above to set the transition matrix, and

\[
KL[G(\mu_i, \Sigma_i), G(\mu_j, \Sigma_j)] = \frac{1}{2} \log \left( \frac{|\Sigma_j|}{|\Sigma_i|} \right) + \frac{1}{2} \text{Tr}(\Sigma_j^{-1}\Sigma_i) + \frac{1}{2} (\mu_j - \mu_i)^T \Sigma_j^{-1}(\mu_j - \mu_i) - \frac{N}{2}
\]
the information rate for the Markov chain can be computed in a straightforward manner, as will be explained in the following.

Let us assume that the listener recognizes and summarizes the types of musical materials that he hears in terms of some parameter $\theta$. This parameter can be as simple as enumeration of the different data types, or as sophisticated as estimation of parameter value in some model parameter space. If the transitions between different parameters obey Markov dynamics, a stationary probability $P^*(\theta)$ of the model parameters can be derived directly from the Markov model. If we assume that at a particular instance in time the listener identifies the musical materials as belonging to some type $\alpha$, then using the same reasoning as in the case of Bayesian surprise we may consider the KL distance between two probability distributions, one when the parameter is known $P(\theta | \alpha) = \delta(\theta - \alpha)$ and a stationary distribution $P^*(\theta)$, as

$$D[\delta(\theta - \alpha) \mid P^*(\theta)] \approx -\log P^*(\alpha)$$

The KL expression then becomes the negative log-likelihood of drawing parameter $\alpha$ from the stationary probability distribution, and it can be interpreted as information dynamics related to listener current choice of $\alpha$ from a set of possible parameters whose dynamics are generated by the Markov process.

Figure 6 shows a stationary distribution vector $P^*(\theta)$, which we call Spectral Recurrence, plotted together with Spectral Anticipation for the Bach Prelude whose recurrence matrix was shown in Figure 3.

Both profiles were derived using cepstrum coefficients as audio features, estimated over a macro-frame of 6 seconds in duration with advance in time of 1.5 seconds (overlap factor 4). The cepstrum coefficients where submitted to IR analysis for each macro frame, resulting in spectral anticipation values for every time step. The recurrence anticipation was obtained from a recurrence matrix, which in turn was estimated from distances between mean cepstral vectors in each macro-frame. The profiles were filtered in order to smooth fluctuations shorter than 5 seconds. The smoothing was done using a linear-
Musical Information Dynamics as Models of Auditory Anticipation

phase low pass filter with frequency cutoff at 0.3 of the window advance rate. It is evident from the graphs that both profiles have different behavior as they capture different aspects of information dynamics present in the recording.

INFORMATION GAP

Information gap is defined as mutual information between past and future when a gap of length L exists between the two. This can be considered as a generalization of Predictive Information with a missing segment of length L (present) between past of size T and future of size T’, as shown in Figure 7.

From definition of information gap the following information dynamic measure can be derived.

- Predictive information (PI) is equal to information gap with zero gap, i.e. L=0, where the future is taken to infinity.

\[ PI = \lim_{T' \to \infty} I_{\text{gap}}(T, T', 0) \]

- Information Rate (IR) is equal to information gap with zero gap, i.e. L=0, where the future is one time step and the past is taken to infinity.

\[ IR = \lim_{T' \to \infty} I_{\text{gap}}(T, T', 0) \]

- Predictive information rate (PIR) is the difference between information gap with L=0 (zero gap) minus information gap with gap of size one (or whatever the size of the present is).

\[ PIR = I(X,Y \mid Z) = H(Y \mid Z) - H(Y \mid X,Z) = H(Y \mid Z) - H(Y) + H(Y) - H(Y \mid X,Z) = -I(Y,Z) + I(Y,\{X,Z\}) = I_{\text{gap}}(L = 0) - I_{\text{gap}}(L = 1) \]

To demonstrate the usage of information gap we will apply it to a simple Markov case.

Let us assume a Markov chain with a finite state space \( S = \{1, \ldots, N\} \) and transition matrix \( P(S_{t+1} = i \mid S_t = j) = a_{ij} \). The stationary distribution of this process is \( \pi^a, \pi^a = a \pi^a \). Calculating the entropies of the stationary distribution and the entropy rate that takes into account the Markov structure gives

\[ H(S_t) = H(\pi^a) = \sum_{i=1}^{N} \pi_i^a \log(\pi_i^a) \]

\[ H(S_{t+1} \mid S_t) = H_r(a) = -\sum_{i=1}^{N} \pi_i^a \sum_{j=1}^{N} a_{ji} \log(a_{ji}) \]

Using these expressions, IR can be simply obtained from the difference between the entropy of the stationary distribution and the entropy rate. Also information gap of size one can be obtained by considering the difference between the stationary entropy and a Markov prediction of two steps into the future, given by the matrix \( a^2 \)

\[ IR = I_{\text{gap}}(0) = I(S_{t+1}, S_t) = H(\pi^a) - H_r(a) \]

\[ I_{\text{gap}}(1) = I(S_{t+1}, S_{t-1}) = H(\pi^a) - H_r(a^2) \]

\[ PIR = I(S_{t+1}, S_t \mid S_{t-1}) = I(S_{t+1}, \{S_t, S_{t-1}\}) - I(S_{t+1}, S_{t-1}) = I_{\text{gap}}(0) - I_{\text{gap}}(1) = H_r(a^2) - H_r(a) \]

Example: Let us consider a simple Markov case of a circle of values with small probability for deviation from an exact sequence, as described in the graph below. The entropy of this model equals to log(5), while entropy rate for single or
two step prediction are close to zero. This results in a process that has maximal IR, since it has big difference between entropy and entropy rate, but close to zero PIR, since one and two step entropy rates are practically zero (Figure 8).

\[ H(\pi^e) \approx \log_2(5) \]
\[ H_r(a) \approx H_r(a^2) \approx 0 \]
\[ IR \approx Maximum, \ PIR \approx 0 \]

This example demonstrates the different meaning of IR and PIR, which is related to the difference between two listening scenarios: in the case of IR the listener uses the past to predict the present, and she does it maximally well by capturing the Markov structure that is very predictable. In the PIR scenario the listener is “absent minded” for one instance after having already memorized the past. In such a case a lapse in memory does not have a big effect since the sequence is approximately deterministic and can be reliably extrapolated into the future without considering the implications of “now”.

**RELATIONS BETWEEN DIFFERENT INFORMATION DYNAMICS MEASURES**

The characterizations of IR presented above, namely spectral, Bayesian and recurrence anticipations, can be unified under a single formalism that considers anticipation as a sum of different musical information processing functions. In the derivation of the “unified” theory (Dubnov 2008) an algebraic error appears in the sign of one of the information factors, as noted in (Abdallah 2008). The following discussion corrects this error and explains the specific assumptions behind some of the approximations in the unified formalism. The full derivation of this formalism is given in the appendix.

Let us assume that the information processing tasks facing the listener are as follows:

1. Constructing a model from a short window (macro-frame) of observations
2. Predicting new observations using that model
3. Changing the “belief” in the current model in view of newly arrived data
4. Comparing the current model to all other models that were constructed during listening to the same piece

Formal modeling of this listening act can be done as follows: considering the space of models in terms of parameterization \( \theta \) over the distribution of observations, we factor the probability of observations \( x_{[1:n]} \) into conditional probability \( P(x_{[1:n]} | \theta) \). Using expressions for approximations of entropy in the case of parametric distribution (Bialek et al, 2001), it is shown in the appendix that information rate can be approximated as a sum of several factors:

\[ \rho(x_{[1:n]}) \approx E[\rho(x_{[1:n]} | \theta)]_{\rho(\theta)} + I(x_n, \theta) - E[D(\theta || \theta^\prime)]_{\rho(\theta)} \]

This formulation accounts for three types of information dynamics:
• IR of the observations given a model $\rho(x_{[1:n]} \mid \theta)$, averaged over all models

• Information between individual observations and the model $I(x_n, \theta)$, also known as averaged belief

• Penalty factor in the form of a KL distance $D(\theta || \theta^*)$ between the distribution of observations in a model and a set of most likely distributions $\theta^*$, averaged over all models

In the second expression an averaging over model parameters occurs by definition of mutual information. It should be noted that in practice such an averaging over the set of all possible models is not performed and the most likely model at each macro-frame is used. This results in time varying functions or profiles that represent different aspects of information dynamics over time. In such a case the predictions needed for IR are performed using the model for the current macro-frame, and KL distance is estimated between the current model and the set of most likely models over the whole piece. Moreover, in practice the different profiles cannot be added together since they are estimated using different methods and their units of measurement are different. As will be explained below, the second factor of mutual information between individual observations and the model is often neglected. The remaining two factors can be estimated using spectral anticipation and negative of spectral recurrence.

One of the important implications of the above equation is in the realization that several different factors combine together to create overall information dynamic. We will show that important information about musical structure may indeed be derived from various relations between the Spectral Anticipation and Spectral Recurrence factors, as two measures of information dynamics.

Special Cases of the “Unified” IR

Depending upon what aspect of the overall musical information are relevant for the listener different approximations to the unified IR equation can be derived. For instance, if the listener assumes a fixed data model without trying to consider model dynamics or probabilities of model change, only the first factor of the equation remains. Accordingly we call this factor data-IR and estimate it using spectral anticipation.

Alternatively, if a listener disregards prediction aspects of the observations and considers only model aspects of parameter dynamics, we are left with the right side of the equation that contains the KL distance between model distribution functions prior to and after observing the data. We call this term model-IR. It should be also noted that a model parameter is estimated using all available observations and not only the last (present) sample.

Considering the second factor of mutual information between sample and model, and of course depending on the type of musical data, it is likely that a single observation carries little information about the model. In such case the mean Bayesian anticipation factor is approximately zero. On the other extreme, it could be that a single observation carries a lot of information about the model. In such a case mutual information approximately equals the entropy of the model distribution

$$I(x_n, \theta) = H(\theta) - H(\theta \mid x_n) \approx H(\theta)$$

This can also be shown by plugging into the expression of Bayesian anticipation the function $P(\theta \mid x_n) = \delta(\theta - \alpha)$ when an individual sample implies a model parameter $\alpha$ and using the equivalence between distribution of model parameters and probability of observations in this case $\text{Prob}(x_n) \approx P(\alpha)$ (see Box 8).
Musical Information Dynamics as Models of Auditory Anticipation

Box 8.

\[
I(x_{\theta}, \theta) = E\{D[P(\theta \mid x_{n}) \| P(\theta)]\}_{P(x_{\theta})} = E\{D[\delta(\theta - \alpha) \| P(\theta)]\}_{P(x_{\theta})} \\
= E\{-\log P(\alpha)\}_{P(\alpha)} = H(\alpha)
\]

Table 1. Relation between cognitive and signal measures was tested for a large orchestral piece. See text for more details

<table>
<thead>
<tr>
<th>Cognitive Measure</th>
<th>Signal Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Familiarity (Recognition, Categorization)</td>
<td>Long-term Similarity Structure, Repetition Spectral Recurrence</td>
</tr>
<tr>
<td>Emotional Force (Anticipation, Implication-Realization)</td>
<td>Short-term Predictability Structure (Spectral Anticipation)</td>
</tr>
</tbody>
</table>

CONCLUSION: APPLICATIONS OF INFORMATION DYNAMICS

The question of relation between properties of musical works and their perception has attracted many researchers, mostly focusing on direct relationship between acoustical and perceptual features, such as intensity and loudness, relation between frequency contents and perceptions of pitch and tonality, and etc. Researches on emotional response try to relate overall properties such as mode, harmony, tempo, and etc. to primary emotions. These approaches largely disregard aspects of perception that are determined by temporal organization of music. Accordingly, potentially important applications of musical information dynamics arise from its ability to discover temporal information structure of musical data and relate it to dynamically to the act of listening to this data.

A first proof of concept for the use of information dynamics as a predictor of human listening responses was conducted in a large-scale experiment (Dubnov et al., 2006). Spectral Anticipation and Spectral Recurrence profiles were compared to continuous human judgments of Emotional Force and Familiarity when listening to two versions of a large orchestral piece. It was found that the two information profiles explain significant portions (between 20–40 percent) of the corresponding human behavioral responses. The approximate correspondence between signal derived measures and human judgment can be summarized in Table 1.

Due to the difficulty of obtaining large amounts of user judgments, most researches rely on human musical experts in evaluating the results of information dynamics analysis. We already discussed the studies of Philip Glass’s Gradus and Two Pages using STM-LTM and Markov models of MIDI or symbolic representations of music (Potter et al 2007, Abdallah and Plumbley, 2009). Those studies indicate that significant correspondence exists between points indicated by an expert as significant or most surprising musical moments and peaks in model-information measures.

Combination of signal and model information measures as a way to detect musical interest points was suggested in (Dubnov 2008). It was shown that musical climax points could be detected consistently for different performances of the same piece by finding a moment that had both high novelty (small probability on the structural level measured in terms of negative quantized SR value) and high surface predictability (predictably structured musical passages that capture high values in quantized SA profile). These results are shown in Figure 9.
Future Work

Information Dynamics is a promising direction that includes aspects of temporal structure and prior listener knowledge in modeling of the listening experience. It differs from other works on musical perception that are concerned with direct mapping of sound features or their overall statistics to listening experience. The features of information dynamics allow the development of a formal mathematical framework for the manipulation of audio streams by providing alternative structures of manipulation that respect the temporal and probabilistic natures of music more than the usual structures used in audio content analysis applications do.

This formal framework leads to several applicable fields. Applications of information dynamics to automatic structure discovery include segmentation, recognition of auditory scenes in terms of higher-level features, detection of auditory surprise and so on. Information dynamics could be considered as a meta-feature that analysis of audio on a higher semantic level related to musical interest. This will require additional research for introducing more features into the analysis, as well as finding ways to combine these features into meaningful musical categories. Such high level labels, performed in temporal manner, will
allow variable length content-based matching of audio and will lead to better concatenation type of synthesis that are important for texture synthesis or recombinant music generation. In the context of sound transformations the ability to characterize salient perceptual moments in music may lead to new types of processing, such as content driven analysis-synthesis, or computer-aided human interaction with machine improvisation. Other potential applications of information dynamics include characterization of preferred structures and automated quality assessment, extending the scope of music recommendation systems today that rely on overall listening preference profiles and social filtering.

A promising direction for improving the estimation of information contents in audio comes from new work on music information geometry (Cont, 2008). Information geometry is a recent field of mathematics in particular of statistical inference that studies the notions of probability and information by the way of differential geometry (Amari and Nagaoka, 2000). The notion of Audio Oracle is generalized to include Bregman mutual information between observations and oracle states, thus incorporating data information rate in the process oracle construction. Deriving model information from AO-derived similarity structure is another aspect of this ongoing research.

REFERENCES


Cont, A. (2008). Modeling Musical Anticipation: From the time of music to the music of time. (PhD thesis University of Paris 6 (UPMC) and University of California San Diego (UCSD)). See also http://imtr.ircam.fr/index.php/Music_Information_Geometry


In (Dubnov 2008, eq. 16) there was an error in the sign of the right hand factor (KL distance). Also, as explained in the text, the mutual information between a single sample and the model parameters was neglected, assuming that for most practical cases the model cannot be represented by one sample. Another point worth considering is the relation between the different factors in this equation and the SR and SA estimators presented in this chapter. SA estimator uses spectral flatness measure to estimate IR of the decorrelated (approximately independent) features. This estimation can be also performed by using linear prediction, in which case it is based on the compression gain computed from the ratio of the signal variance and the residual error variance for every feature component. The important point here is that the prediction and the error estimation, or the spectral flatness estimation, are all done over macro-frames, which means that they estimate IR relative to the current empirical distribution. This indeed corresponds to the first factor of the unified IR expression. The second factor can be estimated using average Bayesian surprise that was described in this chapter. To be able to use this method, the definition of IR has to be extended to include more then one sample as the signal present. Doing so requires extending SA to future predictions. The most problematic factor in this equation is the last factor that computes KL distance between the empirical distribution and some other distribution denoted by θ*. In case when θ* contains a single peak in model space that is close to the empirical distribution, KL measures model perturbation and is proportional to Fisher information that the observed data carries about a parameter. In case when the distribution of the model space comprises of several peaks, KL distance measures the distance of the empirical distribution to all other most likely distributions. This could be done using spectral clustering method and estimating the cluster parameters by models such as GMM. Currently we use SR as a crude approximation to this process, with reservations that were discussed in the section on Spectral Recurrence.
APPENDIX A

Derivation of the Unified IR Measure

Starting with parameterization of the probability function

\[ P(x_{[1:n]}) = \int P(x_{[1:n]} \mid \theta)P(\theta)d\theta \]

we use the approximation (Bialek et al. 2001) to express the probability in terms of a conditional probability relative to an empirical distribution and an expression that measures the distance between this empirical distribution and all other distributions generated by this model space

\[ P(x_{[1:n]}) = P(x_{[1:n]} \mid \theta) \int \frac{P(x_{[1:n]} \mid \alpha)}{P(x_{[1:n]} \mid \theta)} P(\alpha)d\alpha = P(x_{[1:n]} \mid \theta) \int e^{-\int \theta(\alpha)}P(\alpha)d\alpha \]

\[ \approx P(x_{[1:n]} \mid \theta) \int e^{-nD(\theta(\alpha))}P(\alpha)d\alpha \]

This expression allows writing the entropy of a macro-frame in terms of conditional entropy in a model and log-partition function

\[ Z_n(\theta) = \int e^{-nD(\theta(\alpha))}P(\alpha)d\alpha \]

\[ H(x_{[1:n]}) = -\int P(\theta) \int P(x_{[1:n]} \mid \theta) \log P(x_{[1:n]} \mid \theta)dx_{[1:n]}d\theta \]

\[ -\int P(\theta) \int P(x_{[1:n]} \mid \theta) \cdot \log \int e^{-\int \theta(\alpha)}P(\alpha)d\alpha dx_{[1:n]}d\theta = H(x_{[1:n]} \mid \theta) - E[\log Z_n(\theta)]_{P(\theta)} \]

The entropy of a single observation cannot be approximated in this way, so we express it in terms of conditional entropy and mutual information between the sample and the empirical model

\[ H(x_n) = H(x_n \mid \theta) + I(x_n, \theta) \]

Combining these expression into the definition of information rate gives

\[ IR(x_{[1:n]}) = H(x_n) + H(x_{[1:n-1]} \mid x_n) - H(x_{[1:n]}) = \]

\[ = H(x_n \mid \theta) + I(x_n, \theta) + H(x_{[1:n-1]} \mid \theta) - E[\log Z_{n-1}(\theta)]_{P(\theta)} - H(x_{[1:n]} \mid \theta) + E[\log Z_n(\theta)]_{P(\theta)} = \]

\[ = IR(x_{[1:n]} \mid \theta) + I(x_n, \theta) + E[\log \frac{Z_n(\theta)}{Z_{n-1}(\theta)}]_{P(\theta)} \]
This expression can be further simplified if we assume that the space of models comprises of several peaks centered around distinct parameter values. In such case the partition function $Z_n(\theta)$ can be written through Laplace’s method of saddle point approximation in terms of a function proportional to its arguments at extreme values of $\theta = \theta^*$. This allows writing the right hand of previous equation as

$$\log \frac{Z_n(\theta)}{Z_{n-1}(\theta)} \approx -D(\theta, \theta^*)$$

resulting in an expression of information rate

$$IR(x_{[1:n]}) = IR(x_{[1:n]} \mid \theta) + I(x_n \mid \theta) - E[D(\theta \mid \theta^*)]_{p(\theta)}$$