Structural and Affective Aspects of Music from Audio, to appear in Journal of the American Society for Information Science and Technology, Special Issue on Style.

Structural and Affective Aspects of Music from Statistical Audio Signal Analysis

S. Dubnov

University of California, San Diego, Department of Music, 9500 Gilman Dr. MC 0326, La Jolla, CA 92093-0326, E-mail sdubnov@ucsd.edu

S. McAdams

STMS-IRCAM-CNRS, 1 Place Igor Stravinsky, F-75004 Paris, France, and Département d'Etudes Cognitives, Ecole Normale Supérieure, 45 rue d'Ulm, F-75230 Paris, France, E-mail: smc@ircam.fr

R. Reynolds

University of California, San Diego, Department of Music, 9500 Gilman Dr. MC 0326, La Jolla, CA 92093-0326, E-mail reynolds@rogerreynolds.com

Abstract

Understanding and modeling human experience and emotional response when listening to music are important for better understanding of the stylistic choices in musical composition. In this work we explore the relation of audio signal structure to human perceptual and emotional reactions. Memory, repetition, and anticipatory structure have been suggested as some of the major factors in music that might influence and possibly shape these responses. The audio analysis was conducted on two recordings of an extended contemporary musical composition by one of the authors. Signal properties were analyzed using statistical analyses of signal similarities over time and information theoretic measures of signal redundancy. They were then compared to Familiarity Rating and Emotional Force profiles, as recorded continually by listeners hearing the two versions of the piece in a live concert setting. The analysis shows strong evidence that signal properties and human reactions are related, suggesting applications of these techniques to music understanding and music information retrieval systems.

Structural and Affective Aspects of Music from Statistical Audio Signal Analysis

Introduction

The question of style in music is commonly related, both qualitatively and quantitatively, to the presence of various factors that shape human experience of a musical work in a manner that is mostly unrelated to musical rules or other learned factors that might be specific to a particular musical “language”. For instance, music of different styles can be composed using very similar musical rules, with the difference being in the way the compositional planning and design are made and on the choice of music materials. The perception of musical materials might be influenced by a multitude of factors, which might include memorization, anticipation, perception of sound color or orchestration qualities (to be referred to as “timbre”), and many more. Determining these properties from musical recordings seems a formidable problem, still largely unsolved. In this paper we consider the goal of quantifying signal properties in relation to the perception of musical affect. Accordingly, it is hoped that the methods developed here will contribute to understanding and modeling of specific styles and stylistics in general.

The current work is based on a project that attempts to explore structural and affective aspects of human experience over time when listening to a musical work. The experiments were carried out on a contemporary musical piece, *The Angel of Death* by Roger Reynolds for piano, chamber orchestra and computer-processed sound, in a live concert setting. The experiment consisted of collecting continuous ratings on two scales: Familiarity and Emotional Force. For the Familiarity Rating (FR) scale, listeners were to continually estimate how familiar what they were currently hearing was to anything they had heard from the beginning of the piece on a scale from "Completely New" to "Very Familiar". For the Emotional Force (EF) scale, they were to

continually rate the force of their emotional reaction to the piece at each moment on a scale from "Very Weak" to "Very Strong". As a result, the obtained audio recordings and listener responses were aligned in time. This allowed us, among other things, to test various signal information processing methods in relation to human reactions. A preliminary report of the project was presented in McAdams et al. (2002).

Relatively few empirical studies of complete musical works have addressed the reaction of listeners across time. These works mostly relate the experience of musical emotions to psycho-physiological responses when listening to tonal music (e.g., Krumhansl, 1997). To the best of our knowledge, this is the first attempt to relate human experience to statistical properties measured directly on the acoustic signal. The two questions investigated in the work are whether signal similarity grouping and the predictability structure of signal features could be related to familiarity and emotional content of an audio signal, respectively. The basic assumptions were that global spectral similarity should be related to human familiarity judgments, while the local anticipation structure (i.e. the predictability of signal features on a short time scale) might be related to the emotional affect.

The signal similarity was evaluated in terms of groupings within a spectral similarity matrix across time (also called a signal recurrence matrix) using matrix-partitioning methods. As appropriate features, we used spectral envelopes that were estimated from short signal segments in a time-varying manner, represented by low-order cepstral coefficients. The similarity was obtained using a Euclidian distance or dot product between normalized cepstral feature vectors. Partitioning of the similarity matrix by singular value decomposition results in a vector that represents plausible similarity grouping structures. This vector was compared to mean FR profiles produced by the listeners.

The signal predictability was evaluated using the same cepstral feature vector sequences. The predictability was measured in terms of Information Rate (IR), a measure that represents the reduction of uncertainty that an information-processing system achieves when predicting future values of a stochastic process based on its past. Using a decorrelation procedure, the sequence of feature vectors is transformed into an alternative representation in which it can be regarded as a sum of approximately independent, time-varying expansion coefficients in an appropriate feature basis. The IR of a vector process may be computed then from the sum of the IRs of the individual components, as will be described below.

An additional signal feature that was employed for the estimation of EF was signal Energy (E). Both IR and E were compared separately to mean listener EF profiles. Moreover, a combined estimate of the two features was obtained using non-negative least squares regression over one-minute-long time segments. The weights of the regression, being positive values, might be considered as an indication for the relative importance of IR and E for EF estimation.

The structure of the paper is as follows: after a brief review of psychological research on human emotional experience when listening to music, we describe the structure of the musical piece that was especially composed for and used in the experiments. Next, we present the main methods of signal analysis, with some mathematical details deferred until the appendix, in order not to obscure the main focus of the paper. The amount of fit between FR, EF, and the various signal analysis methods is presented. Possible applications and future research directions are presented in the final discussion section.

Psychological Research

In the realm of tonal music, several approaches to the evolution of emotional experience have been used. Krumhansl (1997) related the experience of musical emotions to psychophysiological responses. Sloboda & Lehmann (2001) studied listeners' perceptions of emotionality in reaction to different interpretations of a Chopin Prelude. Schubert (1996) has developed techniques for two-dimensional, continuous response to emotional aspects of music. Fredericskon (1995) used a continuous response method to track the online evolution of perceived tension.

The study on which the present analysis is based (McAdams et al., 2002) recorded continuous responses by listeners in a live concert as they heard *The Angel of Death* for piano, chamber orchestra and computer-processed sound by Roger Reynolds. Two response scales were used: familiarity and emotional force. The first one concerned perceptual and cognitive aspects of musical structure processing, and the second one concerned emotional response to the music. The main findings of the analysis were that, although the piece had never been heard before and the style was unfamiliar to many of the listeners, the temporal shapes of the emotional experience and of the sense of familiarity were clearly related to the formal structure of the piece. Moreover, the piece elicits an emotional experience that changes over time, passing through different emotional states of varying force, and without having overlearned the stylistic conventions of the particular work or style.

The music

The structure of the piece (Reynolds, 2002) was conceived to allow experimental exploration of the way in which musical materials and formal structure interact. The piece is conceived in two main parts, one sectional (S) and the other a more diffusely organized domain (D) structure.

Certain musical materials occur at the same place in time and in nearly identical form (sometimes changing between piano and orchestral versions, sometimes between instrumental and computer-processed versions) in the two parts. Further the two parts can be played in either order (S-D or D-S), but the computer-processed part (evoking the angel) always starts at the end of the first part and continues throughout the second. This structure allowed for the study of the perception of certain materials under different formal settings (embedded in the sectional or the domain part, played alone or in the presence of the computer part, heard first in the sectional version or in the domain version, etc.). *The Angel of Death* thus provides a unique opportunity for music psychologists to study the relations between materials, form, and emotional response, and for signal analysts to explore the relations between signal properties and psychological responses.

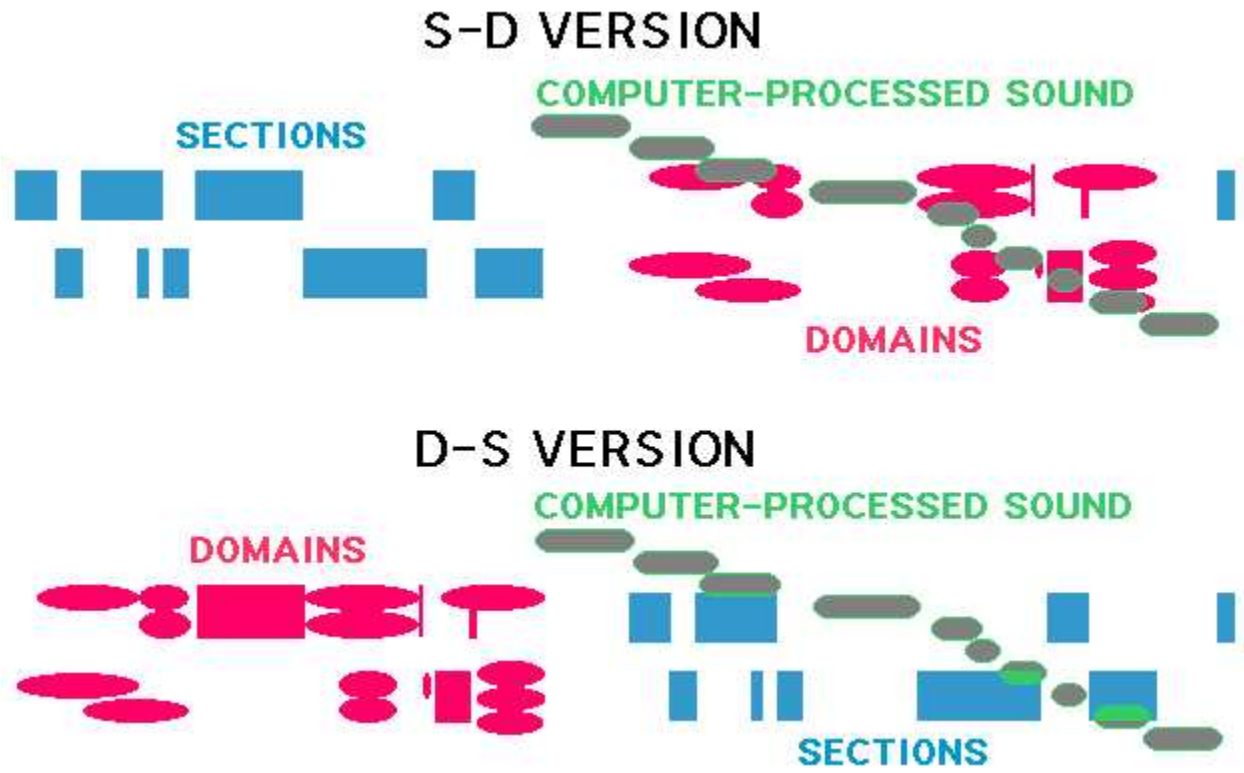
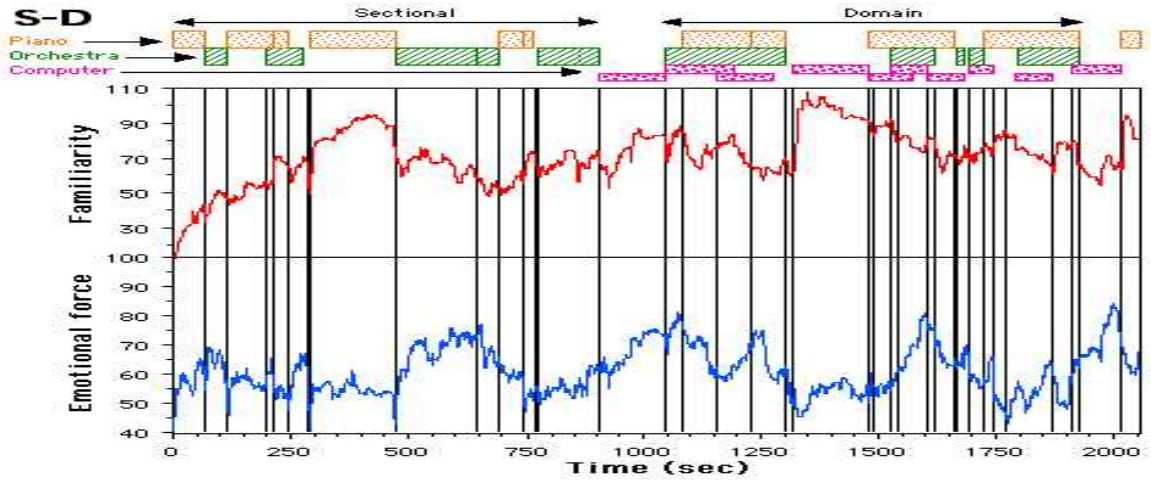


Figure 1

Graphical representation of the formal plan of the musical composition for the S-D and D-S versions.

Experimental Results

In this paper we will conduct several comparisons between statistical analyses of the audio features and profiles of continuous listener responses when listening to the S-D and D-S versions of the piece at their world premier in Paris in June 2001. Figure 2 shows the mean profiles of the listener Familiarity Rating (FR) and the Emotional Force (EF) responses, aligned with the formal structural scheme of the composition.



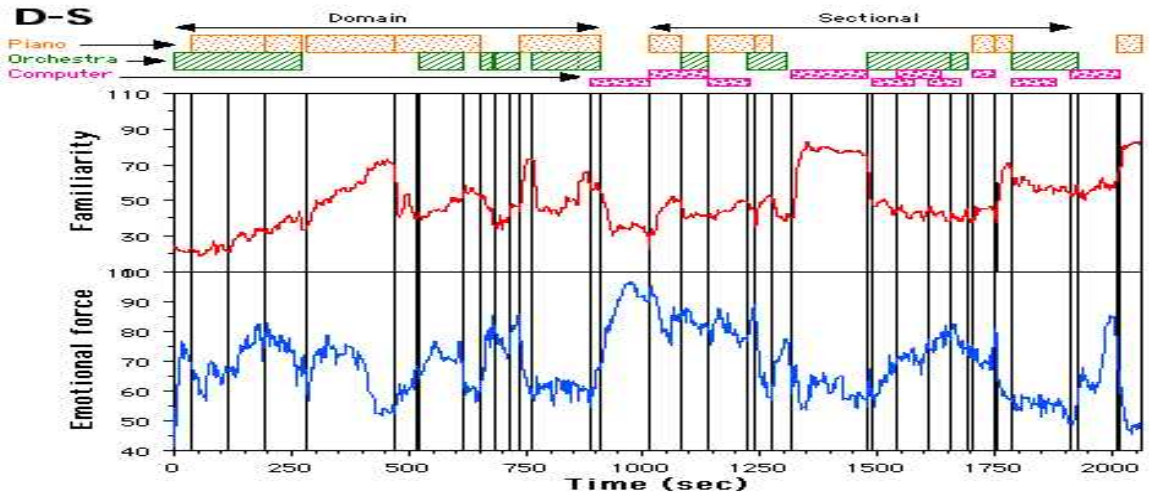


Figure 2
Average Familiarity Rating and the Emotional Force responses, aligned with the formal structural scheme of the composition of the S-D and D-S versions.

The Audio Features

The features considered for analysis of the audio signal were derived by cepstral analysis (Oppenheim and Schafer, 1989). In order to explain the method, we need to consider a so-called source-filter model for the audio signal. The source-filter model decomposes an acoustic signal into an input signal, usually called excitation, and a linear filter. Statistically speaking, the excitation usually carries the long-term correlation properties, such as periodic structure due to pitch or zero correlation between distant noise signal frames. In the frequency domain this corresponds to the finer details of the spectrum. The filter usually represents short signal correlations and corresponds to the smooth overall shape of the signal specturm.

Cepstral analysis provides a method for separating out the filter information from the excitation information. Using only the few first coefficients of the cepstrum, the cepstral components related to the filter part are retained. A reverse transformation can be carried out to

provide a smoothed spectrum of the filter part from an otherwise very detailed spectrum of the original signal. This smoothed spectrum is also called the “spectral envelope”.

Loosely defined, the real cepstrum $\hat{x}[n]$ of a signal $x[n]$ is defined in terms of its Z transform, which in turn is defined as the logarithm of the absolute value of the Z transform of the sequence $x[n]$. Alternately, we can write the definition for the cepstrum $\hat{x}[n]$ directly

$$\hat{x}[n] = Z^{-1}\{\log(|Z\{x[n]\}|)\}.$$

One of the more important properties of the cepstrum is that it is a homomorphic transformation. A homomorphic system is one in which the output is a superposition of the input signals, i.e., the input signals are combined by an operation that has the algebraic characteristics of addition. Under a cepstral transformation, the convolution of two signals $x_1[n] * x_2[n]$ becomes equivalent to the sum of the cepstra of the signals $\hat{x}_1[n] + \hat{x}_2[n]$. When the two signals correspond to excitation and filter components, and assuming that each one of them occupies a separate non-overlapping region in the cepstral domain (the filter having nonzero values at the low cepstral components and the excitation having nonzero values at the high cepstral components), separation of the signal into filter and excitation is possible using the so-called cepstral filtering or “liftering” operation, i.e. separately retaining and inverting[1] the cepstra of $\hat{x}_1[n], \hat{x}_2[n]$. We shall call the filter part "spectral envelope" and the excitation "spectral detail". These properties are described in full detail in Oppenheim and Schafer (1989).

In our analysis, we applied cepstral analysis to signal segments (to be called frames) of 200 milliseconds in duration. The analysis was repeated over successive frames in time, with advance

[1] One should note that exact inversion is possible only for the case of complex spectra, which is more complicated due to phase problems in the definition of a complex logarithm. In case of the real cepstrum, one still obtains the spectral amplitudes of the components, and a minimum phase version of the separate signals may be obtained.

(hop size) of 100 milliseconds, or an overlap of 50%. Only the first 32 real cepstral coefficients were retained. Assuming that $\hat{x}_1^i[n]$ represents the spectral envelope component of a signal at frame number *i*, we shall denote by $X_i = \left[\hat{x}_1^i[1], \hat{x}_1^i[1], ..., \hat{x}_1^i[32]\right]^T$ the cepstral feature vector.

The reason for this choice of signal features is that we were interested in a gross spectral envelope description of the sound, which captures sound properties that might be described as overall sound color or texture, without considering the more detailed features due to effects such as pitch, or notes and timbres of specific instruments. This choice was justified in part by the type of musical material that put a significant emphasis on orchestration aspects, while being less traditional in terms of melodic or harmonic or rhythmic patterns. Another practical reason for the choice of cepstral features was the ease and simplicity of their estimation.

Similarity Structure and Similarity Matrix Grouping

The first question considered was the relation between signal similarity and perception of musical familiarity. Using cepstral feature vectors, a distance between two signal frames can be estimated by calculating the Euclidian distance between the cepstral feature vectors. One can show that this is equal to a Euclidian distance between the logarithms of the spectral envelopes of the signal in the corresponding frames, i.e. between the two time instances. Using normalized versions of the cepstral vectors, a simplified distance matrix can be obtained directly from the dot product of the cepstral features of every pair of signal frames.

Let d be the distance between the feature vectors X_i and X_j at frames i and j,

$$d(i, j) = \frac{\left\langle X_i, X_j \right\rangle}{\left\| X_i \right\| \left\| X_j \right\|}$$

where $\left\langle X_i, X_j \right\rangle$ is the dot product defined as $|X_i||X_j|\cos(\theta)$, where θ is the angle between the vectors, and $|X_i|$ is the norm of X_i. This distance measure is large when the vectors are of high magnitude and similar, and because of the normalization, low magnitude and similar vectors also produce a large value.

For each time segment, these distance magnitude values are plotted on a similarity matrix. Figure 3 shows a similarity matrix of an example sound. We will be using this similarity matrix graph as a basis for partitioning the sound into perceptually similar groups.

This matrix is sometimes called a recurrence matrix or similarity matrix. It represents the spectral similarity between different time instants of the audio signal. An example of a recurrence matrix of the S-D recording (audio signal) is show in Figure 3. As can be seen from the figure, the signal at different times resembles or differs from the signal at other times. The goal of the similarity grouping procedure is to provide a function whose values correspond to plausible grouping structures based on the similarity matrix. Good criteria for grouping can be derived from considering the few first eigenvectors of the similarity matrix.

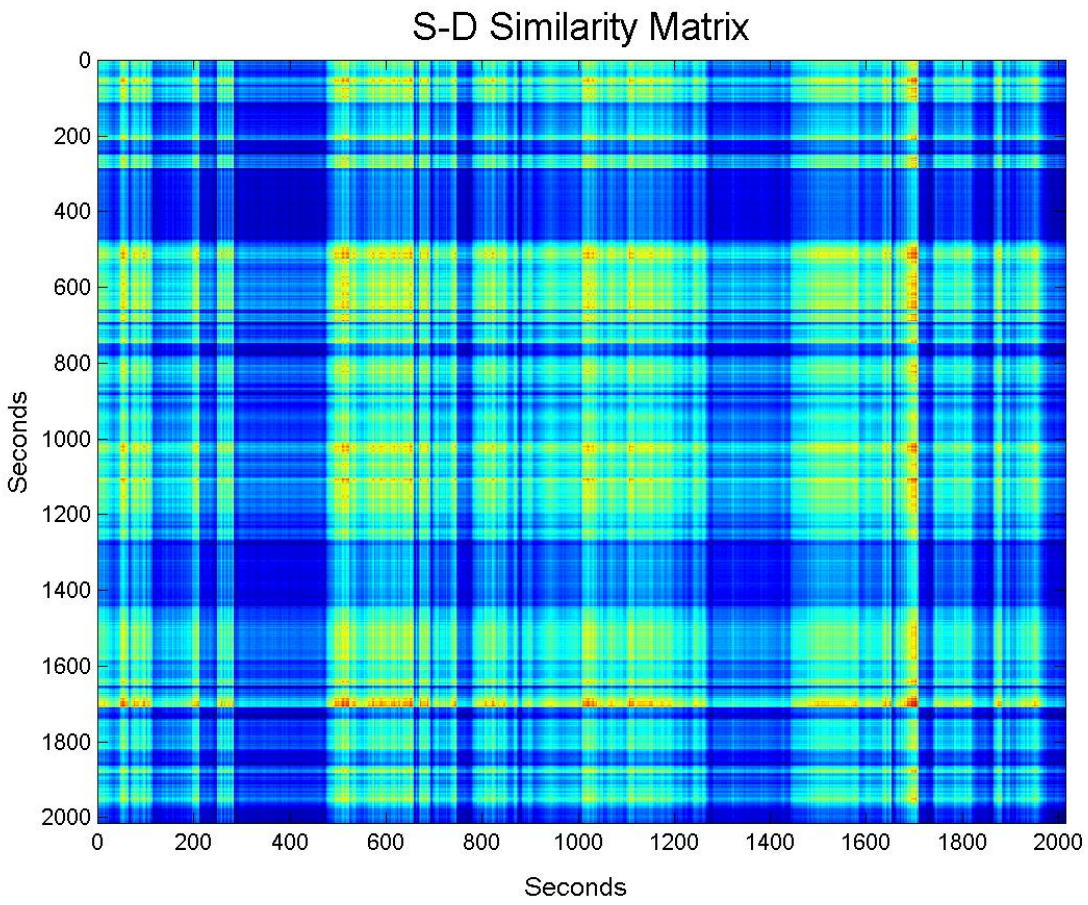


Figure 3

Similarity Matrix representing the distances between the music materials at different times in the S-D audio recording. The similarity is based on the dot product of cepstral feature vectors. Bright or red areas correspond to high similarity and dark or blue areas are different.

This method of grouping analysis, sometimes called spectral matrix clustering[2], or in general Spectral Clustering (Ng et al, 2002), recently emerged as an effective method for data clustering, image segmentation, Web ranking analysis, and dimension reduction. At the core of spectral clustering method is a graph that represents relations between different data points in terms of pairwise distances or similarities. The segmentation methods use the Laplacian of the

[2] The use of the term "spectral" has nothing to do with the actual audio signal spectrum and it comes from the usage of eigenvectors as a basis for clustering. The relation between spectrum and eigenvectors results in this terminology.

graph adjacency (pairwise similarity) matrix, using mathematical methods that evolved from spectral graph partitioning. It is beyond the scope of this paper to discuss these methods in detail. We shall say only that the eigenvector represents the main “direction” or pattern of behavior in time, according to which the similarity matrix is oriented.

Segmenting and Grouping

One method of approaching the perceptual grouping is to partition the data (image, text or audio in our case) into two maximally dissimilar groups. If necessary, these groups can then be sub-partitioned using the same procedure iteratively. That is, instead of searching for consistent features to be grouped in part of a graph, the spectral clustering methods attempt to separate regions in a top-down manner with the most dissimilar areas being separated first. One method, introduced by (Shi and Malik 2000), for creating these image segmentations uses the “normalized cut” to partition the graph.

Normalized Cut

A graph G = (V,E) in graph-theoretic language is a set of vertices V and a set of edges E. The graph can be segmented into two groups A and B by finding the “minimum cut” cut(A,B) of the graph. The minimum cut separates regions A and B by finding the regions that minimize the sum of the total weight of the edges cut. This criterion tends to select regions that are uneven in size, so the criterion is modified to create the normalized cut.

$$\mathrm{Ncut}(A,B) = \frac{\mathrm{cut}(A,B)}{\mathrm{assoc}(A, V)} + \frac{\mathrm{cut}(A,B)}{\mathrm{assoc}(B, V)},$$

where assoc(A, V) is the weight of all the connections between the nodes in A and all of the vertices. This normalization more nearly equalizes the sizes of the segmented groups. In our

case, the normalized cut criterion is used to segment our distance matrix. In this way the most dissimilar sound segments will be segmented by the first Ncut bipartition.

Eigenvector Method

It can be shown that the normalized cut can be calculated using methods based on eigenvectors of an affinity matrix. Using eigendecomposion of our recurrence matrix, the normalized cut can be calculated. We begin by performing an eigenvector decomposition of our recurrence matrix.

$$(D-W)v = \lambda Dv$$

Where $D_{ij} = d(i, j)$ is the recurrence matrix, $W_{ii} = \sum_j D_{ij}$ is the diagonal affinity matrix, λ are the eigenvalues of the system, and v are the eigenvectors of the system.

For clustering purposes the first eigenvector is usually used and each value of the eigenvector is assigned to one of two signal groups by setting up appropriate threshold or decision boundaries. One could consider this eigenvector as a data reduction or projection of the similarity matrix onto one salient dimension. The values of this eigenvector should fluctuate according to the most significant changes that occur in the similarity matrix. Accordingly, pairwise segmentation or grouping is possible by associating different values of the eigenvector to different groups. When more than pair-wise grouping is required, more eigenvectors might be used. Since in this work we are not interested in doing actual clustering, we compared the eigenvectors of the S-D and D-S versions of the piece to their corresponding FR profiles, as presented graphically in Figure 4. The y axis corresponds to normalized (zero mean and unit variance) values of the Familiarity Rating profile and the Similarity eigenvector.

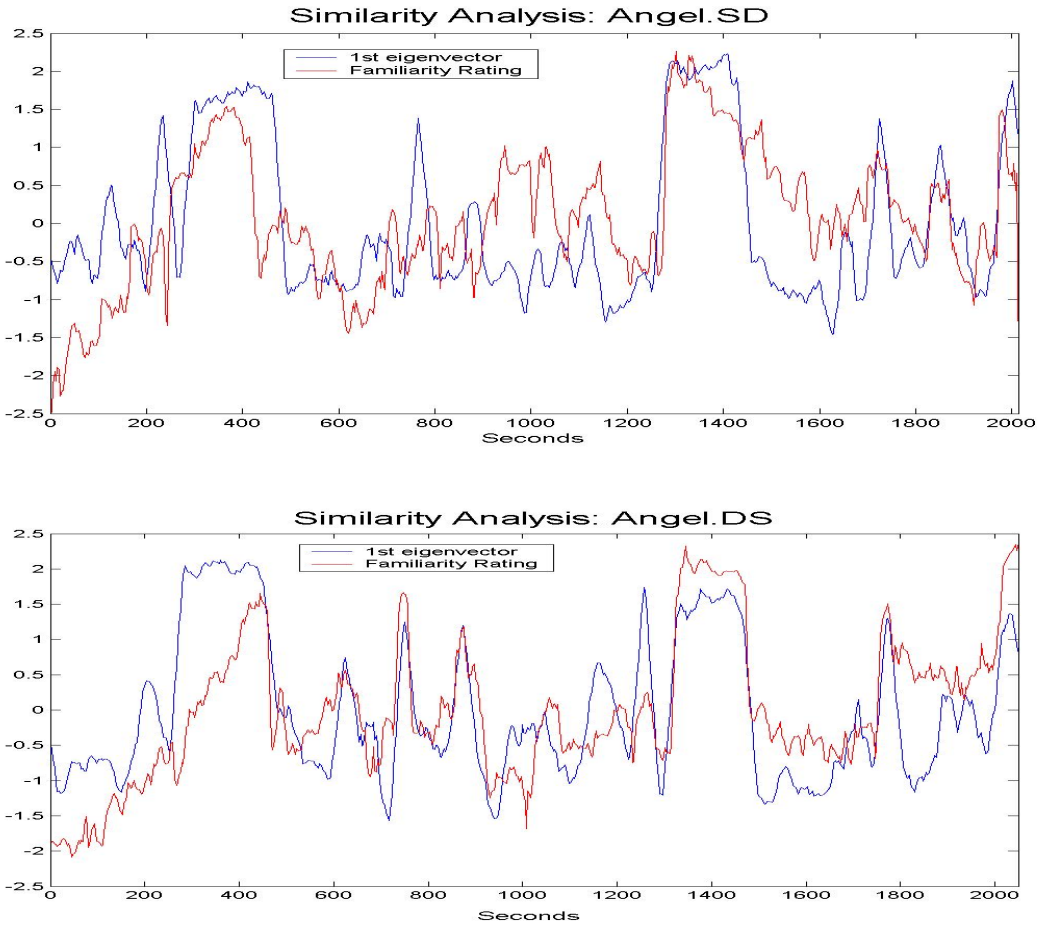


Figure 4

Familiarity Rating profiles of the human responses versus estimated similarity profile based on the first eigenvector of the similarity matrix. The figures show results for the S-D (top) and D-S (bottom) versions of the piece.

The correlations between the similarity eigenvectors and FR for the S-D and D-S versions of the piece are summarized in Table 1.

Table 1: Correlation between Similarity Matrix Eigenvector (normalized version) and experimental Familiarity Ratings by human listeners (df=682, p<.0001 in both cases).

Music Familiarity	Similarity Eigenvector correlation
S-D	0.54
D-S	0.65

The similarity eigenvector explains[3] 29% (R^2) of the variance in the mean FR profile for the S-D version and 42% in the D-S version. These correlations are highly significant, but also demonstrate that additional factors besides spectral similarity are required in order to better explain the familiarity ratings. Discovering these additional factors will be the subject of future research.

Predictability of Features and Information Rate

The second question that we investigated was whether predictability of signal features could be related to the emotional content of a signal. The predictability was evaluated in terms of Information Rate (IR) (Dubnov 2003), a novel feature that measures the relative growth of information in a

[3] The correlation coefficient (r) indicates the quality of linear fit between the predictor parameter and the subjective data. The number of degrees of freedom (df) in the test is equal to the number of points minus 2. The p value, determined with the Fisher r-to-z transform, indicates the probability that the two variables are completely independent, i.e. p values less than .05 indicate that the correlation is unlikely to be zero and the correlation is considered to be statistically significant. The coefficient of determination (R^2) is the most interesting measure since it indicates the amount of variance in the subjective data that is explained by the predictor.

random process for every additional sample that is observed over time. This also corresponds to the reduction in number of bits required to code a random process as a result of predicting this process from its past. The larger the difference between the number of bits needed to code the process "as is" as compared to the number of bits required when the next sample is predicted from its past, the higher is the information rate in the process. This also means that the knowledge of the next sample adds a significant amount of information to the total information needed to describe the process up to this point. For example, if on one extreme the process is a perfect noise, i.e. it is completely unpredictable, then the number of bits needed to code the process with or without prediction is the same and the resulting IR is zero. A similar situation (an inverted U relation) occurs if the process is almost constant or has very little variation. In such a case it might be that the next sample is almost precisely predicted, thus requiring only a few bits to code. Since the variation in the process is small, then the number of bits required to code it without prediction is also small. Eventually, the difference in bits between coding the process "as is" or coding the predictions is small and IR is low.

Information Rate (IR) is defined as the difference between the information contained in the variables $x_1, x_2, ..., x_n$ and $x_1, x_2, ..., x_{n-1}$, i.e. the additional amount of information that is added when one more sample of the process is observed

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

It can be shown that for large n, IR equals the difference between the marginal entropy $H(x)$ and entropy rate $H_r(x) = \lim_{n\to\infty} \frac{1}{n} H(x_1, ..., x_n)$ of the signal *x(t)*,

$$\rho(x) = \lim_{n\to\infty} \rho(x_1, ..., x_n) = H(x) - H_r(x).$$

In our experiments we have applied the IR analysis to a sequence of cepstral vectors that describe the evolution of the spectral envelope over time. It is shown in the Appendix that assuming

independence of the coefficients *s* after an appropriate transformation, one can generalize the IR definition to be the difference in information between sequences of vectors,

$$\rho_{IC}^{n}(X_1, X_2, ..., X_L) \triangleq I(X_1, X_2, ..., X_L) - \{I(X_1, X_2, ..., X_{L-1}) + I(X_L)\}$$
$$= \sum_{i=1}^{n} \rho(s_i(i), ..., s_i(L)$$

This block-wise information redundancy measure calculates the difference between the multi-information over L consecutive vectors versus the sum of multi-information of the first L-1 vectors and the multi-information in the last vector X_L. The convenient property of this measure is that it can be calculated from the marginal entropies of the n independent components, using the IR estimates of the single components.

As will be shown in the Appendix, using expressions for the entropy and entropy rate of a Gaussian process, one has the following relation between Spectral Flatness Measure (SFM) (Jayant and Noll, 1984) and the IR of a signal, $SFM(x) = \exp(-2\rho(x))$, or equivalently $\rho(x) = -\frac{1}{2}\log(SFM(x))$. For the purpose of IR estimation, the separate components are treated as separate signals and their IRs are estimated from their SFMs.

As described in the section on feature vectors, the cepstral analysis was performed over signal frames of 200 milliseconds with time advance of 100 milliseconds (50% overlap). Among all cepstral coefficients, the first 32 coefficients were retained for cepstral envelope characterization. Moreover, the first component that contains the signal energy in the frame was removed from the cepstral data that was submitted for IR analysis and was considered separately as an Energy (E) feature. As described in the Appendix, IR estimation of sequences of vectors requires a decorrelation step to be performed prior to estimation of the IR of the individual components. Applying Singular Value Decomposition (SVD) (Hayes 1996) to the cepstral

feature vector matrix effectively achieves the desired decorrelation. Summing up the IR's of the individual components then performs the IR analysis.

Vector IR estimation algorithm:

The IR analysis procedure consisted of the following steps:

i) Preprocessing:

The cepstral coefficients were calculated over time frames of 200 milliseconds, with no overlap. Since the first cepstral coefficient contains the energy of the signal, it was not considered as part of the IR analysis and was later used as a separate feature.

ii) Decorrelation:

The cepstral feature matrix was submitted to Singular Value Decomposition (SVD) analysis. This resulted in a set of independent (in the case of a Gaussian process) expansion coefficients over time.

iii) Information Redundancy:

The IR was calculated separately from IRs of the coefficients. The method of calculation is based on an estimation of spectral flatness of each individual component, considered as a scalar time signal, as explained in the Appendix.

As will be shown below, it was found that both E and IR have high correspondence to Emotional Force (EF), with IR being superior to E by approximately 10%. Below we present several graphs that summarize our experiments and principal findings.

Energy and EF

Figure 5 shows the relation between EF (red) and signal energy (blue), estimated as the first cepstral coefficient, using analysis frame size (Ta) of 200 milliseconds and averaged over macro-segments of 3 seconds (segment size) with no overlap between the macro-segments (segment step is 3 seconds as well). One should note the distinction between signal frames and macro-segments (sometimes called segments): signal frames are on the order of hundreds of milliseconds and are used for extracting the short-time features, in our case cepstral coefficients. The sequence of cepstral coefficients is further divided into macro-frames within which E or IR are evaluated. Thus, the E and IR estimations are done using multiple cepstral vectors over a 3-second period. The E feature is obtained by averaging the first cepstral coefficient over the duration of the whole macro-frame.

Applying a 10-segment-long moving-average filter additionally smoothed the E and IR results. The experimental EF profile was subsampled and interpolated accordingly, so as to provide interpolated values of EF every 3 seconds in correspondence to the E and IR results (the original EF was recorded with a 2-Hz sampling rate). As can be seen from Figure 5, certain portions of the Energy curve fit closely to the EF data, while other portions differ significantly. The correlation coefficients between Energy and the S-D and D-S EF profiles are 0.51 and 0.36, respectively.

Figure 5
Estimated Energy feature versus human Emotional Force profile response. The figures show results for the S-D (top) and D-S (bottom) versions of the piece.

IR Analysis

Figure 6 presents IR analyses of the recordings (audio signals) of the D-S and S-D versions. The IR property is evaluated independently for every macro-segment. As explained previously for the case of the Energy feature, IR is evaluated over time using sequences of cepstral features that are organized into macro-frames of 3 seconds duration. IR results are additionally smoothed using a moving average 10 segments long. The correlations with EF were 0.63 and 0.46 for the S-D and D-S

versions, respectively. The results of IR and Energy correlation to the listeners' mean EF responses are summarized in Table 2. The explained variances for IR and Energy in the S-D case are 41% and 25%, respectively, and explained variances for IR and E in the D-S case are 22% and 10%, respectively.

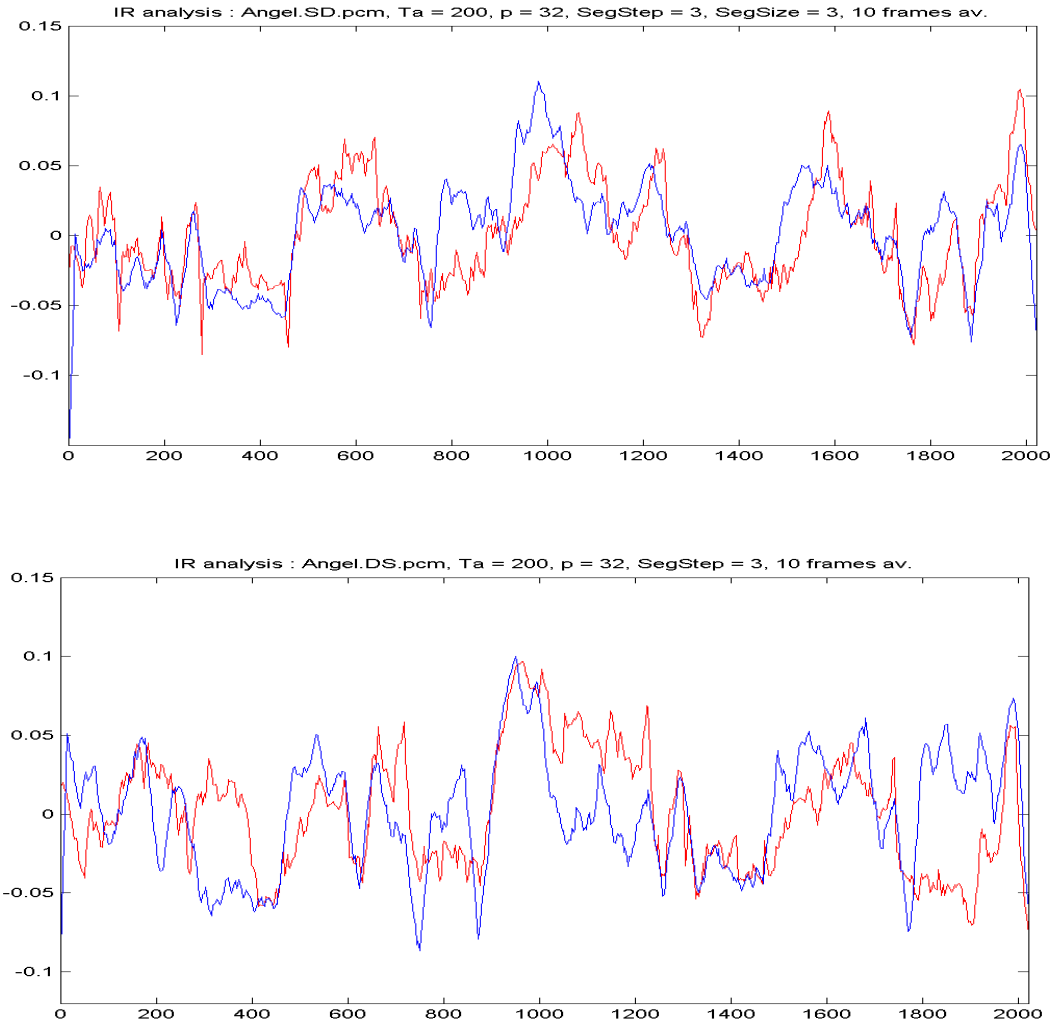


Figure 6
Estimated IR feature and human Emotional Force profile response. The figures show results for the S-D (top) and D-S (bottom) versions of the piece.

Table 2: IR and Energy correlation to the human responses of EF (df=682, p<.0001 in all cases).

Emotional Force	IR correlation	Energy correlation
SD	0.63	0.51
DS	0.47	0.33

EF estimation using combined Energy and IR

In order to better approximate the EF from signal analyses, we have performed a least-squares fit of E and IR curves to the EF profile. In the following, we shall denote E and IR as predictors, in order not to confuse them with the cepstral *features* that are derived directly from the audio signal. E and IR predictors could be considered as higher-order features needed for the higher-level processing involved with emotional responses.

Using a combination of predictors for estimation of EF, the predictor weights might change slowly over time, depending on various factors related possibly both to the nature of the signal or to the listening process. A tradeoff exists when considering a time-varying regression: one should note that in principle a perfect fit is possible if the weight coefficients vary every sample. On the other hand, we cannot expect to have a single constant set of weights over the whole duration of the signal. As a reasonable compromise we have chosen a block of 1-minute duration as the regression period over which the weight coefficients would be estimated.

Additionally, we should require a non-negative contribution of the predictors to the total EF response. This decision is justified by the claim that the various factors can contribute positively to

the emotional response but they cannot cancel each other out or inhibit the total EF response. Accordingly, we employed a Non-Negative Least Squares (NNLS) regression for estimation of the EF match from E and IR. The NNLS algorithm was first introduced by (Lawson and Hanson 1974). NNLS solves the algebraic equations of the Least Squares problem subject to the added constraint that the fitting parameters contain no negative elements.

Figure 7 shows the results of NNLS regression of E and IR so as to match EF in a time-varying manner with regression weights varying every minute. The correlations between the NNLS fit of IR and E and EF are summarized in Table 3. Note that the gain in predictability obtained with the multiple correlation is quite large, resulting in 82% and 83% of the explained variance for the two versions.

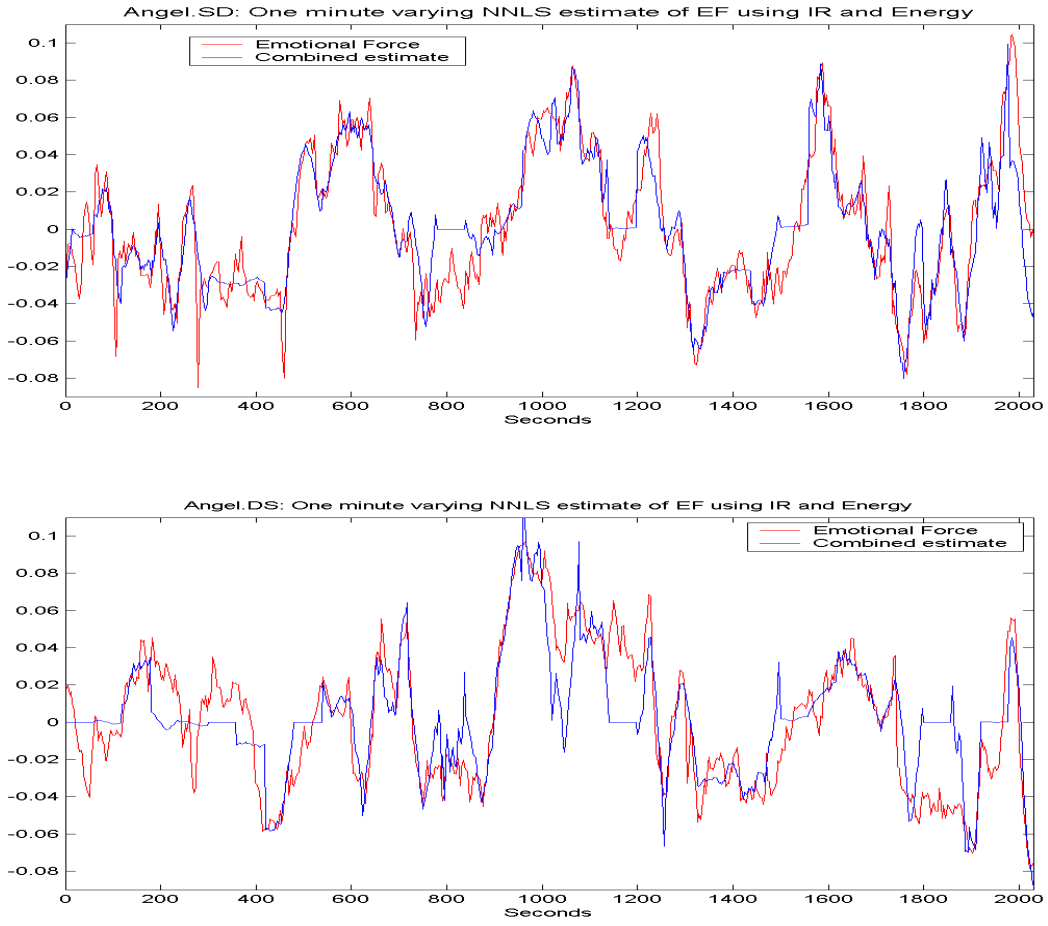


Figure 7
Non-Negative Least Squares regression of the IR and E feature matching to human Emotional Force profile response. The figures show results for the S-D (top) and D-S (bottom) versions of the piece.

Table 3:
Correlations between the NNLS fit of IR and E and EF for the S-D and D-S versions (df=682, p<.0001 in both cases).

Emotional Force	NNLS fit of IR and E
SD	0.91
DS	0.91

Discussion and Conclusions

These results indicate that structural and affective analyses from statistical properties of the audio signal are plausible. In our analysis we employed a simplified signal representation by means of a sequence of spectral envelopes, represented by 32 real cepstral coefficients that were calculated over signal segments of 200 milliseconds. These features were used for estimation of the similarity matrix, which serves as the basis for structural analysis of the signal repetition structure. We found that 29-41% of the variance in the Familiarity property is explained by the grouping profile as represented by the first eigenvector of the similarity matrix. Using the same features over macro-frames of 3 seconds, local time-varying parameters of Energy and IR were derived. We found that Energy alone explained 11-26% of the variance in EF. Using IR of the cepstral coefficients explained 22-40% of the variance in EF, if used alone, and up to 79% of the variance when both features are optimally combined in non-negative manner over segments of 1 minute.

Although our model is grossly over-simplified and cannot capture the true complexity of music, we found a significant correspondence for FR and a high correspondence for EF between our very

simple statistical audio signal analysis and experimental human responses. In this paper we provided a principled formalization of the concepts of signal recurrence and signal information rate, as new features for signal characterization. We examined the relation of these features to human judgments of familiarity and emotional force.

Considering additional musical parameters, such as melody, harmony and rhythm might improve the results and will probably be required in order to deal with additional musical styles. One should note that these musical features are complicated for estimation from a raw audio signal. A further subject for future research is to find a method for machine learning of regression coefficients. Although it is plausible to assume that human listening criteria might vary for different types of music and for different listeners, we would like to determine a priori the rules of regression for different types of signals. Applications of the method to signal classification, music summarization and automatic music appreciation are the subject of future work. Another interesting future direction might be high temporal resolution fMRI studies correlating brain states to these statistics.

Potential significance of the above results for musical and cognitive research seems to be quite exciting. Music can be viewed as successions of sonic events that unfold over time to create temporal patterns and expectancies. Leonard Meyer's *Emotion and Meaning in Music* (1956) relied heavily on psychological insights and psychologically based arguments in describing music, suggesting strong dependencies between expectation, emotion and meaning. It has since become a commonly accepted view that music operates on perceptual and cognitive aspects of our listening experience by the forming and violation of anticipations in order to create tension and resolution over time. The work presented here suggests that principled research into these questions is possible using our analytic features and statistical methods.

References

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*, John Wiley & Sons, New-York.

Dubnov, S. (2003). Non-Gaussian Source-Filter and Independent Components Generalizations of Spectral Flatness Measure, *Proceedings of International Conference on Independent Components Analysis (ICA2003)*, Nara, Japan.

Frederickson, W. E. (1995). A comparison of perceived Musical tension and æsthetic response, *Psychology of Music*, 23:81-87.

Hayes, M. (1996) *Statistical Signal Processing and Modeling*, Wiley.

Jayant, N.S. and Noll, P. (1984). *Digital Coding of Waveforms*, Prentice-Hall Signal.

Krumhansl, C. L. (1997). An exploratory study of musical emotions and psychophysiology, *Canadian Journal of Experimental Psychology*, 51:336-352.

Lawson, C. L. & Hanson, B. J. (1974), *Solving Least Squares Problems*, Prentice-Hall (Englewood Cliffs, NJ).

McAdams, S., Smith, B. K., Vieillard, S., Bigand, E., and Reynolds, R., (2002) Real-time perception of a contemporary musical work in a live concert setting, *Proceedings of the 7th International Conference on Music Perception and Cognition*, Sydney, edited by C. Stevens, D. Burnham, G. McPherson et al. (Causal Productions, Adelaide [CD-ROM], Sydney).

Meyer, Leonard B. (1956). *Emotion and Meaning in Music*, Chicago: Chicago University Press.

Ng, A., Jordan M.I., and Weiss, Y., (2002) On spectral clustering: Analysis and algorithm , In Dietterich, T.G., Becker S., and Ghahramani Z., editors, *Advances in Neural Information Processing Systems*, volume 14, The MIT Press

Oppenheim, A.V. and Schafer, R.W. (1989). *Discrete-Time Signal Processing.* Prentice-Hall. Englewood Cliffs.

Reynolds, R., (2002). Compositional strategies in The Angel of Death for piano, chamber orchestra and computer processed sound, *Proceedings of the 7th International Conference on Music Perception and Cognition*, Sydney, edited by C. Stevens, D. Burnham, G. McPherson et al. (Causal Productions, Adelaide [CD-ROM], Sydney).

Schubert, E. (1996). Continuous response to music using a two-dimensional emotion space, *Proceedings of the 4th International Conference on Music Perception and Cognition, Montreal*, pp. 263 268.

Shi, J. and Malik J. (2000). Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.

Sloboda, J. A. & Lehmann, A. C. (2001). Tracking performance correlates of changes in perceived intensity of emotion during different interpretations of a Chopin piano prelude, *Music Perception*, 19:87-120,.

Appendix

A.1 Spectral Flatness Measure

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

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

Rewriting it as a discrete sum gives

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

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

A.2 Information Redundancy

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

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

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

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

This function measures the average amount of common information contained in variables $x_1, x_2, ..., x_n$. Using the mutual information, we define marginal information redundancy (sometimes simply called Information Redundancy or IR) (Dubnov 2003) to be the difference between the common information contained in the variables $x_1, x_2, ..., x_n$ and the set $x_1, x_2, ..., x_{n-1}$, i.e. the additional amount of information that is added when one more variable is observed.

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

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

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

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

A.3 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 a signal spectrum and its entropy. The entropy of a "white" Gaussian random variable is given by

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

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

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

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

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

One can also see that IR is equal to half the logarithm of SFM.

A.4 Vector IR as the sum of independent component scalar IRs

In order to consider a sequence of random vectors, we generalized the idea of IR by representing the vector process as independent linear combinations of n-dimensional basis vectors. We denote by X the original vectors, **A** the basis and s the coordinates or coefficients of *X* in the basis **A**.

$$\left[X_1 X_2 ...\right] = \mathbf{A}\begin{bmatrix} s_1(1) & s_1(2) & ... \\ s_2(1) & s_1(2) & ... \\ \vdots & \vdots & ... \\ s_n(1) & s_1(2) & ... \end{bmatrix}$$

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 relation 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 $x_1 ... x_{nL}$,

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

we write the vector IR as

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

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 $\rho_L^n(X_1, ..., X_L) = H(S_L) - H(S_L \mid S_1, ..., S_{L-1})$ (note that the dependence upon the determinant of *A* is cancelled by subtraction). If there are no dependencies across different coefficients and the only dependencies are within each coefficient sequence as a function of time

(i.e. the trajectory of each coefficient is time dependent but the coefficients among 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 gives the desired result

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