PROBABILITY AS METADATA: EVENT DETECTION IN MUSIC USING ICA AS A CONDITIONAL DENSITY MODEL

Samer A. Abdallah and Mark D. Plumbley

Department of Electronic Engineering,
Queen Mary, University of London.
samer.abdallah@elec.qmul.ac.uk
mark.plumbley@elec.qmul.ac.uk

ABSTRACT

We consider the problem of detecting note onsets in music under the hypothesis that the onsets, and events in general, are essentially *surprising moments*, and that event detection should therefore be based on an explicit probability model of the sensory input, which generates a moment-by-moment trace of the probability of each observation as it is made. Relatively unexpected events should thus appear as clear spikes. In this way, several well known methods of onset detection can be understood in terms of an implicit probability model. We apply ICA to the problem as an adaptive non-Gaussian model, and investigate the use of ICA as a *conditional* probability model. The results obtained using several methods on two extracts of piano music are presented and compared. Finally, we tentatively suggest an information theoretic interpretation of the approach.

1. INTRODUCTION

A wide variety of methods of onset detection have been proposed in the literature, (e.g. [1, 2, 3]) many of which perform adequately in their intended domains. What seems to be missing is any sense of an underlying design principle; the algorithms appear to the be result of an heuristic process relying on the insight and inventiveness of the engineer. Some systems [2] do make use of psychoacoustic data to guide the design process, but such imitation does not explain *why* those processing strategies should be used, or, indeed, why the human auditory system is as it is. Neither are the resulting algorithms applicable in other domains, such as detecting events in EEG traces, where there is no biological system to serve as a guide. Instead, the heuristic design process must be begun anew, as insights gained in one domain may not be relevant in another.

There is, however, a common motif in many of the algorithms, where a two-stage approach is adopted. First, the acoustic signal is processed to produce one or more new signals, which ideally are non-oscillatory, of lower bandwidth, and manifest a clear peak or other easily detected feature for each onset in the original signal. Subsequently, these reduced signals are analysed to locate the peaks and assign a time to each event thus detected—this is the stage where categorical decisions are made. The success of the second stage depends on the degree to which the reduced signals record a consistent, easily categorised response to any onset whilst rejecting other aspects of the sound, which may be considered "noise" in this application. This paper is concerned only with the first stage: we propose that the reduced signal should derive from the statisitical structure of the data, which in turn should be learned by an adaptive probability model.

1.1. A Probabilistic Approach

The underlying premise is that onsets in music, or more generally, significant events, are perceived as such because they are relatively *surprising* moments, during which the signal behaves in an unexpected or unpredictable way. This judgement is to be made by an observer relative to a statistical model of the signal, which generates a moment-by-moment trace of the probability of the signal under the model. When an event occurs, there will be a sudden dip in probability, but if the event is a stereotypical one familiar to the model, the initial surprise will be followed by a largely predictable consequent, so the dip in probability should be localised to the onset. Thus, for signals which are strongly event-based, we expect this "probability signal" to have some sparse temporal structure, from which event times can reliably be extracted.

Whether the surprises reflect genuine events or merely the observer's inability to make accurate predictions depends on the goodness of fit between observer's model and the data; it is therefore important to use an appro-

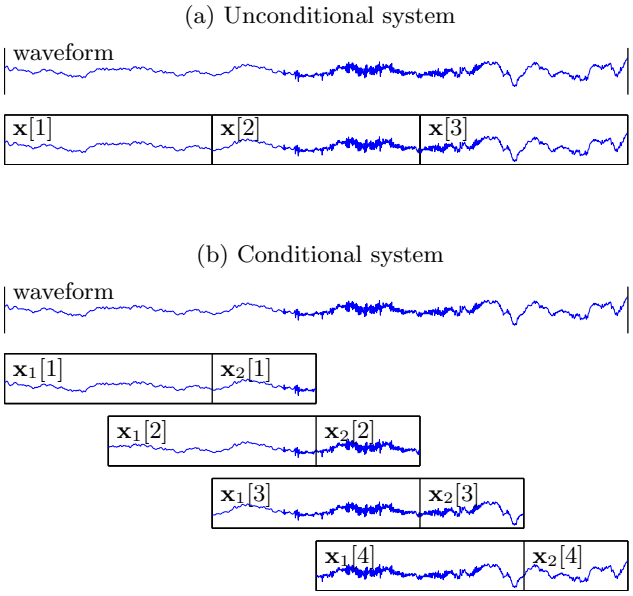

Figure 1: In the memoryless system (a) the signal is split into blocks $\mathbf{x}[k]$. The probability density $P(\mathbf{x})$ is then modelled as if the blocks were statistically independent. In the conditional system, (see § 3) each block is partitioned into two parts, $\mathbf{x}_1$ and $\mathbf{x}_2$, with the intention of modelling the conditional density $P(\mathbf{x}_2|\mathbf{x}_1)$.

priate class of model, to fit the parameters to the data adaptively, and, for signals with temporal dependencies, to use previous obervations to make *conditional* probability assessments.

The approach fits well with a wider perceptual theory (see [4] for a fuller discussion) which maintains that the goal of a perceptual system is to produce an efficient representation of its sensory input by learning about its statistical structure. One of the most direct responses such a system can make to any sensory scene is an estimate of the probability of that scene under the model embodied by the system.

In the rest of the paper, we will describe how this methodology translates into specific algorithms under different assumptions about the signal's statistical structure. In particular, when the unexpectedness of an observation is measured as a *negative log-probability* (a quantity which Attneave [5] called the "surprisal") some known methods of onset detection result directly. These can therefore be understood in terms of an implicit probability model, and hence judged objectively and quantitatively on the accuracy of that model.

2. MEMORYLESS MODELS

We will first consider models which do not condition their probability assessments on previous observations. The audio data is broken up into a sequence of n-tuples $\mathbf{x}[k] \in \mathbb{R}^n$, with $k \in \mathbb{Z}$, as shown in fig. 1(a). These are then treated as if they were independent and identically distributed, so the model amounts to an expression for $P(\mathbf{x})$, where the time index k has been dropped as it is no longer relevant. In this case, the "surprisal" is just a function of $\mathbf{x}$:

$$S(\mathbf{x}) = -\log P(\mathbf{x}). \tag{1}$$

A number of simple models fit this framework and translate directly into some well known methods of onset detection. (In all the following expressions, $P(\cdot)$ implicitly denotes a probability relative to the model under consideration, not the "true" probability, which is presumed unknown.)

2.1. Gaussian and non-Gaussian IID models

If the individual elements of $\mathbf{x}$ (that is, the original audio samples) are assumed to be independent and Gaussian with zero mean and variance σ^2, then the following probability density function is obtained:

$$P(\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{x_i^2}{2\sigma^2}, \tag{2}$$

which yields a measure of surprise which is essentially the signal energy (to within an additive constant):

$$S(\mathbf{x}) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \text{const.} \tag{3}$$

Hence, the energy is a measure of the unexpectedness of the observation if the signal is assumed to be Gaussian white noise. Alternatively, the samples may be assumed to be independent but non-Gaussian, with, for example, a Laplacian (or double-sided exponential) distribution, $P(x_i) = \frac{1}{2}e^{-\lambda|x_i|}$. This yields

$$S(\mathbf{x}) = \lambda \sum_{i=1}^{n} |x_i| + \text{const.}, \tag{4}$$

which is equivalent to a rectification and smoothing of the original signal, a common method of computing the amplitude envelope of the signal.

It has been observed [2] that neither the energy nor the amplitude envelope provides a good basis for onset detection, except for very percussive instruments. We suggest that this is because they both imply an underlying statistical model which is a poor fit to the data: for most real sounds, the audio samples are far from being independent.

2.2. Multivariate Gaussian models

One approach to modelling the dependencies between the x_i is to assume that $\mathbf{x}$ is Gaussian with a non-diagonal covariance matrix $\mathbf{C} = \mathrm{E}\ \mathbf{x}\mathbf{x}^T$ (where E denotes the expectation operator.) From the standard multivariate Gaussian density function:

$$P(\mathbf{x}) = \frac{\exp -\frac{1}{2}\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}}{\sqrt{(2\pi)^n \det \mathbf{C}}}, \tag{5}$$

we obtain

$$S(\mathbf{x}) = -\tfrac{1}{2}\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x} + \text{const.} \tag{6}$$

This is a *weighted* energy measured in a frame of reference defined by the eigenvectors of the covariance matrix: if $\mathbf{u}_i$ denotes the ith eigenvector of $\mathbf{C}$, with eigenvalue σ_i^2, then $\mathbf{C} = \sum_{i=1}^n \sigma_i^2 \mathbf{u}_i\mathbf{u}_i^T$, and

$$S(\mathbf{x}) = \sum_{i=1}^{n} \frac{(\mathbf{u}_i^T\mathbf{x})^2}{2\sigma_i^2} + \text{const.} \tag{7}$$

The $\mathbf{u}_i^T\mathbf{x}$ are the coordinates of $\mathbf{x}$ relative to the orthonormal basis formed by the $\mathbf{u}_i$. In this frame of reference, the "surprisal" is simply a sum of component-wise functions, similar to those in (3) and (4).

For audio data, we may reasonably assume that the covariance of two elements of $\mathbf{x}$ depends only on the time lag between them, so the covariance matrix will be Toeplitz, and when n is large, the eigenvectors will form an approximate Fourier basis [6]. Environmental sounds tend to have more energy at low frequencies than at high, so $S(\mathbf{x})$ is essentially a spectral energy measure weighted preferentially towards high frequencies, echoing Masri's [7] HFC (*high frequency content*) based onset detection system. In the present framework, greater weight is given to energy at high fequencies precisely because it is less expected, and the weights are assigned in a principled way.

2.3. A non-Gaussian model using ICA

A number of studies (e.g. [4, 6, 8]) have shown that natural and musical sounds have very non-Gaussian statistics. It should therefore be possible to improve on the Gaussian system described above. However, the Gaussian system demonstrated the utility of transforming the data to a representation whose elements are assumed to be independent, since the log-probability becomes a sum over those elements. These considerations motivate the use of ICA as a non-Gaussian probability model whose specific objective is to find a factorial representation.

The application of ICA to natural sounds has been described elsewhere [8, 9]; briefly, the data vectors $\mathbf{x} \in \mathbb{R}^n$ are assumed to be generated by a linear transformation $\mathbf{x} = \mathbf{A}\mathbf{s}$, where the n components of $\mathbf{s}$ are non-Gaussian and independent, and the square $(n \times n)$ matrix $\mathbf{A}$ is fixed, initially unknown, but may be estimated from the data using a maximum-likelihood algorithm [10]. If the independent components s_i have marginal densities $f_i(s_i)$, then the resulting probability model for $\mathbf{x}$ is

$$P(\mathbf{x}) = \det \mathbf{A}^{-1} f_{\mathbf{s}}(\mathbf{A}^{-1}\mathbf{x}), \tag{8}$$

$$\text{where} \quad f_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{n} f_i(s_i). \tag{9}$$

The "surprisal" is then a sum over the components:

$$S(\mathbf{x}) = \log\det \mathbf{A} - \sum_{i=1}^{n} \log f_i(s_i). \tag{10}$$

Previous work [4] has shown that a *generalised exponential*, $f(s) \propto \exp -|s|^\alpha$, with $\alpha \approx 0.3$, is a reasonably good approximation to the observed marginals, giving

$$S(\mathbf{x}) = \sum_{i=1}^{n} |s_i|^\alpha + \text{const.} \tag{11}$$

This is related to a non-Euclidean norm of $\mathbf{s}$, which, compared with the Euclidean norm, "expects" sparse activity in $\mathbf{s}$ and is "surprised" by non-sparse activations. An alternative is to use measured histograms to estimate the $f(s_i)$. The ICA algorithm requires derivatives of the $f_i(s_i)$, so the noisy histograms are not suitable for that purpose, but in practice, they do seem to be adequate for the computation of $S(\mathbf{x})$ even with very little data.

3. A CONDITIONAL ICA MODEL

Consider the system illustrated in fig. 1(b), in which the audio data is arranged into overlapping blocks $\mathbf{x}$ of length n, each of which is partitioned into two pieces, $\mathbf{x}_1$ and $\mathbf{x}_2$, of lengths m and $n-m$ respectively, so if $\mathbf{x} \equiv (x_1, \ldots, x_n)$, then $\mathbf{x}_1 \equiv (x_1, \ldots, x_m)$ and $\mathbf{x}_2 \equiv (x_{m+1}, \ldots, x_n)$. If we have a model of $P(\mathbf{x})$, then we automatically have a model of the joint density $P(\mathbf{x}_1, \mathbf{x}_2)$, from which we may compute the conditional density

$$P(\mathbf{x}_2|\mathbf{x}_1) = \frac{P(\mathbf{x}_1, \mathbf{x}_2)}{P(\mathbf{x}_1)} = \frac{P(\mathbf{x})}{P(\mathbf{x}_1)}, \tag{12}$$

$$\text{where} \quad P(\mathbf{x}_1) = \int P(\mathbf{x}_1, \mathbf{x}_2)\ \mathrm{d}\mathbf{x}_2. \tag{13}$$

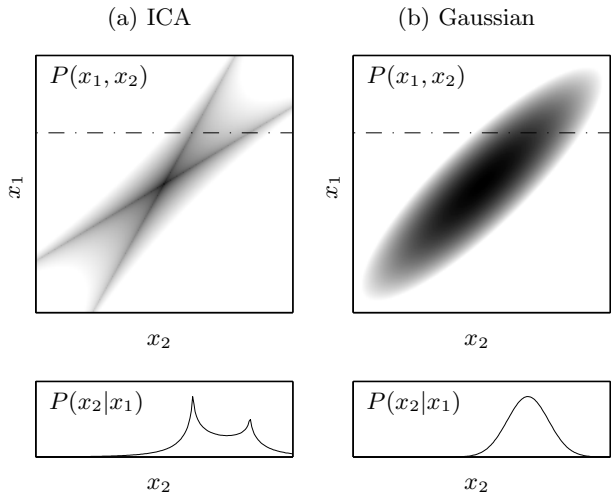

Figure 2: Joint probability densities and the conditional densities obtained by taking a horizontal slice (dashed line.) ICA can model a multimodal conditional density, whereas a Gaussian model can only ever produce a unimodal Gaussian conditional density.

For example, if $\mathbf{x}$ is Gaussian, then $\mathbf{x}_2|\mathbf{x}_1$ is also Gaussian with an expectation linearly related to $\mathbf{x}_1$. By comparison, an ICA model of the full density $P(\mathbf{x})$ is capable of modelling a multimodal conditional density, as shown in fig. 2.

Generally, then, we have

$$S(\mathbf{x}) = \log P(\mathbf{x}_1) - \log P(\mathbf{x}), \tag{14}$$

where the form of $P(\mathbf{x}_1)$ is implicit in $P(\mathbf{x})$. The probability associated with an observation $\mathbf{x}$ (or strictly, with the segment $\mathbf{x}_2$) is conditioned on the expectations set up by the previous observations contained in $\mathbf{x}_1$. If ICA is used to model $P(\mathbf{x})$, then the associated generative model can be written as

$$\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{pmatrix} = \begin{pmatrix}\mathbf{A}_1\\ \mathbf{A}_2\end{pmatrix}\mathbf{s}, \tag{15}$$

where the matrix $\mathbf{A}$ has been partitioned into the $m\times n$ matrix $\mathbf{A}_1$ and the $(n-m)\times n$ matrix $\mathbf{A}_2$, as described in a previous application of ICA to regression problems [11]. However, in regression, the aim is to predict a value of $\mathbf{x}_2$ given $\mathbf{x}_1$, using either a conditional expectation or the maximum of the conditional density. In contrast, we wish to compute the conditional probability of $\mathbf{x}_2$ after it has been observed, and thus cannot avoid computing $P(\mathbf{x}_1)$. By inspection of (15), $\mathbf{x}_1 = \mathbf{A}_1\mathbf{s}$; since $n > m$, this is equivalent to *overcomplete* ICA (see fig. 3), which is known to be a difficult problem [12]. There is no requirement to train the overcomplete system, since that is handled by the larger $n \times n$ ICA model, but the integral in (13) is generally intractible and some approximation is needed. Furthermore, the saddle-point approximation used in [12] is not applicable with the very strongly super-Gaussian priors needed to model the sort of multimodal conditional densities illustrated in fig. 2(a).

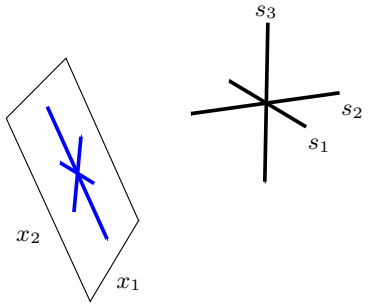

Figure 3: A square ICA model for $P(\mathbf{x}) \equiv P(\mathbf{x}_1, \mathbf{x}_2)$ implies an *overcomplete* ICA model for $P(\mathbf{x}_1)$, since there are more independent components than there are data components. For example, if the full system is three-dimensional, the distribution $P(s_1, s_2, s_3)$ has a six-pronged shape. The two-dimensional distribution $P(x_1, x_2)$, obtained by projecting onto a plane, also has six prongs, and hence cannot be modelled by a 2×2 ICA system.

The approach we have initially taken is to avoid using the implicit model of $P(\mathbf{x}_1)$, and instead to fit a separate $m \times m$ ICA system using $\mathbf{x}_1$ as training data. We do not expect this to fit the data as well as the overcomplete system, but the results illustrated in fig. 4 suggest that it does at least partially achieve the desired objective. Alternative methods of modelling $P(\mathbf{x}_1)$ are discussed in § 5.2.

4. RESULTS

Fig. 4 illustrates the results obtained with two short extracts of piano music. In both cases, the energy profile (and the amplitude profile, not illustrated) are sufficient to detect only the most intense and percussive onsets. A weighted energy was computed as in (7), by measuring second order statistics using a longer extract of music, and computing principal components via an eigenvalue decomposition. The principal components were indeed approximately sinusoidal, and thus the PCA-derived traces are essentially measures of spectral energy weighted by the reciprocal of the power spectrum. Note that both energy measures are plotted on a logarithmic scale; on a linear scale, the dynamic range is very high. In contrast, the ICA-derived traces have a built-in compressive nonlinearity due to the non-Gaussian marginal densities $f_i(s_i)$.

The conditional ICA results were computed as the

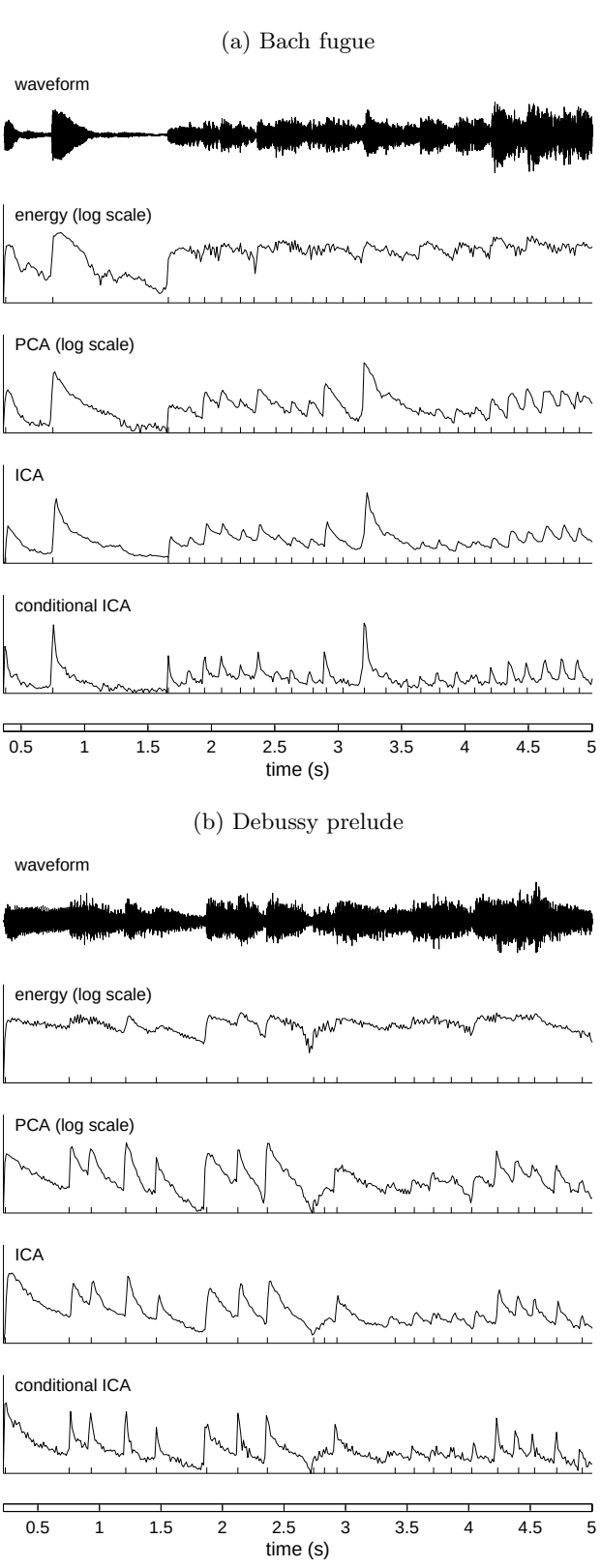

Figure 4: Comparative results for several methods applied to short extracts from two piano pieces. The tick marks on the axes indicate the actual onset times.

difference of the traces obtained from two separate ICA models: a 512 component model for $\mathbf{x}_1$ and a 768 component model for $\mathbf{x}$. The onset peaks are visibly sharper, though this seems to be at the expense of an increased "noisiness" in the trace. Whether this is intrinsic to the conditional ICA framework or due to the approximation used for $P(\mathbf{x}_1)$ is unknown at present.

5. DISCUSSION AND CONCLUSIONS

The material presented here was originally motivated by a particular application—that of onset detection in music—but is intended to illustrate two concepts that should be more widely applicable in problems of artificial perception and data analysis. The first is the use of independent component analysis (ICA) as a conditional probability model, which was initially investigated by Hyvärinen [11] in the context of regression. The second is the idea that, when a statistical model of the data is available, the probability assigned by the model to each observation is itself a useful form of data, with its own structure amenable to further analysis. When the observations are made in a temporal sequence, the probability assessments can be made as the data arrive, and hence can be imagined as a new signal emanating from the model, summarising the flow of expectation and surprise as time passes.

5.1. Relationships with other methods

The approach bears some similarity with the concept of a "novelty filter" [13]. However, we have not used that term since the events we are interested are not *novel* in the sense that they are unfamiliar. This would imply that they do not fit the model, or that the model needs changing. Rather, they are familiar but *relatively unlikely*. After the initial onset, the rest of the event should be well desribed by the conditional model, so that the probability signal records only a spike marking the onset time.

We have already noted how energy based methods correspond to Gaussian signal models. Other methods can be interpreted in this way too: for example, algorithms based on spectral difference [7] imply a conditional model for the short-term power spectrum in which the expected spectrum is equal to the previously observed one. Putting these methods on an equal footing in terms of an implied probability model allows objective comparisons to be made, by measuring the Kullback-Leibler divergence between the model and the observed data. The best model is the one with the lowest expectation E $S(\mathbf{x})$, something which can only be determined empirically.

5.2. Further Work

It should be clear from the discussion so far that the way to improve the performance of the system is to improve the fit of the models used for $P(\mathbf{x})$ and $P(\mathbf{x}_1)$. In particular, it has been shown [4] that ICA of audio data does *not* produce independent components, and that residual dependencies remain. These could be modelled using independent subspaces or topographic ICA [14].

The approximation used for $P(\mathbf{x}_1)$ remains to be validated by computing the integral in (13), which may be possible using Monte Carlo methods. An alternative approach is to use a more tractable overcomplete generalisation of ICA [15] which would be capable of modelling the multiple "spikes" of the distribution shown in fig. 3. The causal interpretation afforded by the additive ICA model would be absent, but this is unimportant since all that is required is a function for $P(\mathbf{x}_1)$.

It may be objected that, for continuously distributed variables, $P(\mathbf{x})$ and hence $S(\mathbf{x})$ are not invariant to invertible transformations of $\mathbf{x}$, so if, for example, $\mathbf{x}$ is transformed into $\mathbf{y}$ and $P(\mathbf{y})$ modelled instead, a different probability signal will result. This discrepancy can be resolved by acknowledging that, in a physical system, continuous variables cannot be measured to infinite precision, and there will always be some noise. This means that an observation is characterised not by a precise value of $\mathbf{x}$, but a posterior distribution, $P(\mathbf{x}|\mathcal{D})$, where $\mathcal{D}$ denotes the observational data. If we let $P(\mathbf{x}|\mathcal{M})$ denote the distribution defined by the current the current state of observer's model, we may define a generalised "surprisal" as

$$S(\mathcal{D}) = \int P(\mathbf{x}|\mathcal{D}) \log \frac{P(\mathbf{x}|\mathcal{D})}{P(\mathbf{x}|\mathcal{M})} \, \mathrm{d}\mathbf{x}, \qquad (16)$$

which is the Kullback-Leibler divergence between the two distributions. This *is* invariant to invertible transformations of representation, and has a satisfying interpretation as the *information gained* during the observation, leading directly to the hypothesis that sensory data is perceived as event-based precisely to the extent that information arrives in bursts, and that events are essentially concentrated "packets" of information.

6. REFERENCES

[1] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 588–601, Jan 1998.

[2] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1999.

[3] Masataka Goto and Yoichi Muraoka, "Music understanding at the beat level — real-time beat tracking for audio signals," in *Working Notes of the IJCAI-95 Workshop on Computational Auditory Scene Analysis*, August 1995, pp. 68–75.

[4] Samer A. Abdallah, *Towards Music Perception by Redundancy Reduction and Unsupervised Learning in Probabilistic Models*, Ph.D. thesis, Department of Electronic Engineering, King's College London, 2002.

[5] Fred Attneave, *Applications of Information Theory to Psychology*, Holt, New York, 1959.

[6] Liubomire G. Iordanov and Penio S. Penev, "The principal component structure of natural sound," in *Advances in Neural Information Processing Systems*, Michael S. Kearns, Sara A. Solla, and David A. Cohn, Eds. 1999, vol. 11, MIT Press.

[7] P. Masri, *Computer Modeling of Sound for Transformation and Synthesis of Musical Signal*, Ph.D. thesis, University of Bristol, 1996.

[8] Michael S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[9] Samer A. Abdallah and Mark D. Plumbley, "If edges are the independent components of natural scenes, what are the independent components of natural sounds?," in *3rd Intl. Conf. on Independent Component Analysis and Signal Separation, ICA2001*, San Diego, 2001, pp. 534–539.

[10] Jean-François Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. on Signal Processing*, vol. 44, no. 12, pp. 3017–30, Dec. 1996.

[11] Aapo Hyvärinen, "Regression using independent component analysis, and its connection to multi-layer perceptrons," in *Proc. Intl. Conf. on Artificial Neural Networks ICANN'99*, Edinburgh, 1999, pp. 491–496.

[12] Michael S. Lewicki and Terrence J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, pp. 337–365, 2000.

[13] Teuvo Kohonen, *Self-organization and Associative Memory*, Springer-Verlag, Berlin, 1984.

[14] Aapo Hyvärinen, Patrik Hoyer, and Mika Inki, "Topographic independent component analysis," *Neural Computation*, vol. 13, no. 7, pp. 1527–1558, 2001.

[15] Geoffrey E. Hinton, Max Welling, Yee Whye Teh, and Simon K. Osindero, "A new view of ICA," in *3rd Intl. Conf. on Independent Component Analysis and Signal Separation, ICA2001*, San Diego, 2001, pp. 746–751.