Unsupervised onset detection: A probabilistic approach using ICA and a hidden Markov classifier

Samer Abdallah and Mark Plumbley,
Queen Mary, University of London.

March 11, 2003

Abstract

We describe an onset detection system that takes a two-stage approach, both of which are based on unsupervised learning in a probabilistic model.

The first stage uses independent component analysis (ICA) to fit a short-term non-Gaussian model to frames of audio data. This model is used to generate a *reduced signal* to be interpreted as the 'surprisingness' of the original audio signal. Our hypothesis is that onsets and events generally are perceived as so because they are temporally localised surprises.

The second stage uses a hidden Markov model (HMM) with Gaussian state-conditional densities to do unsupervised clustering of the 'surprise' signal as represented in a multidimensional embedding space. The clusters which emerge in this space can be associated the presence or absence of an onset, and so a trivial decision based on the current HMM state can be used to drive an onset detector.

1 Introduction

We take as our premise that perceptual systems can thought of as probabilistic machines that learn about the statistical structure of the data to which they are exposed. This enables such benefits as the efficient representation and transmission of information, the ability to make optimal probabilistic inferences about missing or uncertain data, and the ability to predict the range of likely outcomes in a given situation. (See [1, Chapter 2] for a fuller discussion.) A system that makes predictions has the capacity to be *surprised*; moreover, a system that models a time-conditional probability distribution can report exactly *how* surprised it is. A running report of this degree of surprise can be treated as a new signal and used as a basis for further processing. We conjecture that onsets, or events in generally, are essentially *surprising* moments, and can therefore be characterised as a 'clumping' or 'burstiness' in the 'surprise' signal. The fact that event perception is categorical—that is, either an event is present or it is not—suggests that we should look for some clustering in the space of surprise signals.

The first half of the paper describes how independent component analysis or ICA can be used as a statistical model of acoustic signals, and thus to generate a surprise signal, which will sometimes be referred to as a 'reduced signal,' since since it is meant to retain important information about the original signal but at a much reduced sampling rate.

The second half describes an unsupervised clustering method for classifying patterns in these reduced signals as either events or non-events. The reduced signal is projected into an n-dimensional embedding space by taking the n previous samples as coordinates. As the n sample window passes along the signal, the points form a clustered distribution, which we model using a hidden Markov model.

2 Generating a 'surprise' signal using ICA

The aim of the first stage of processing is to build a short-term probability model for the signal so that we can generate a new signal which is the conditional *negative log probability* of a short segment. Refering to fig. 2, the surprise signal is defined as

$$S[k] = -\log P(\mathbf{x}_2[k]|\mathbf{x}_1[k]) \quad (1)$$
$$= \log P(\mathbf{x}_1[k]) - \log P(\mathbf{x}[k]), \quad (2)$$

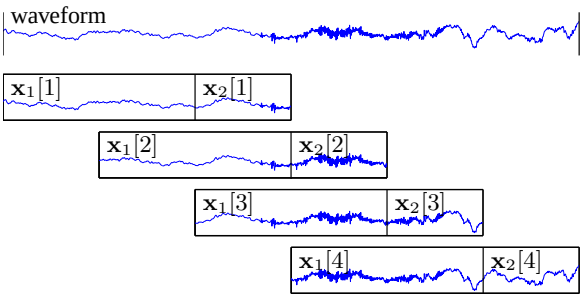


Figure 1: Each waveform block $\mathbf{x}$ is partitioned into two parts, $\mathbf{x}_1$ and $\mathbf{x}_2$ as shown above.

or, dropping the time index for simplicity, just $S(\mathbf{x}) = \log P(\mathbf{x}_1[k]) - \log P(\mathbf{x}[k])$. This quantity measures how surprising that segment appears to be according to the model. The negative log probability is related to information theoretic measures, and was called the 'surprisal' by Attneave [2].

The negative log-probability of a Gaussian density function is a quadratic form, and hence energy based methods of onset detection can be derived from Gaussian signal models. For example, a short term energy measure emerges if we model the signal as spherical (i.e. white) Gaussian noise, whereas a spectrally-weighted energy measure results if we use a more general non-spherical Gaussian model.

Audio signals tend to be extremely non-Gaussian [1] and hence, we used ICA as an initial attempt to model that non-Gaussianity. ICA systems trained (see [3] for algorithm) separately on $\mathbf{x}$ and $\mathbf{x}_1$ allow computation of the the conditional density $P(\mathbf{x}_2|\mathbf{x}_1)$. The results are shown in fig. 1, along with a comparison with other ways of generating a 'surprise' signal using other models.

3 'Peak picking' by unsupervised classification

Our approach to onset detection rests on the notion that in order to make reliable categorical decisions, there should be some *clustering* of the underlying data. This, in turn, is potentially learnable by an adaptive density model applied to the data presented in an appropriate form. The reduced signals described above begin to show such structure when a number of consecutive samples are taken as coordinates in a multidimensional *embedding* space.

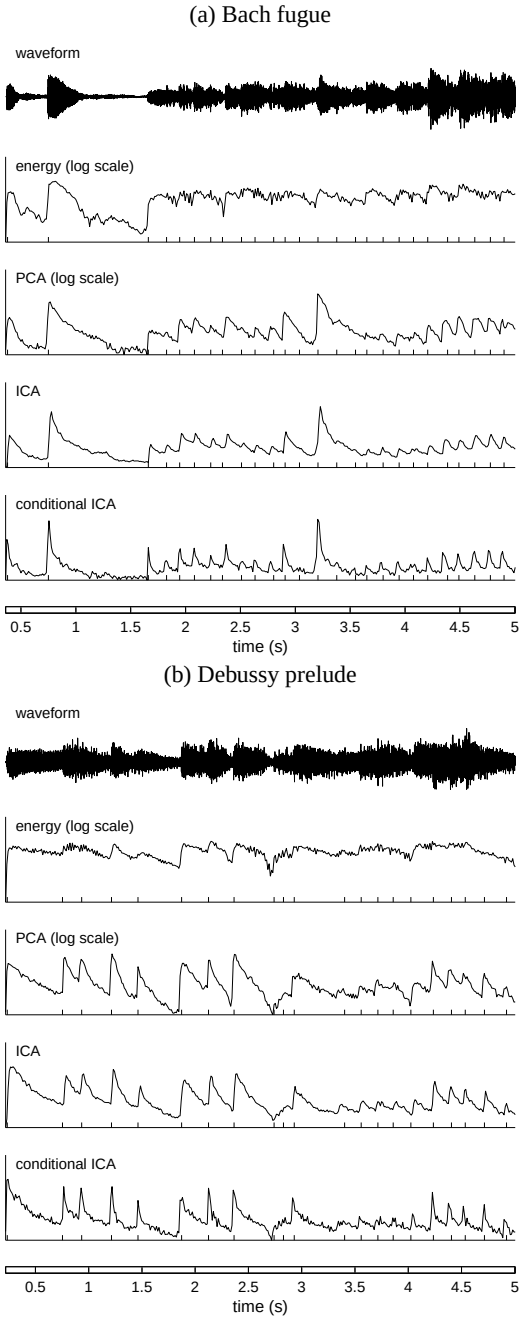
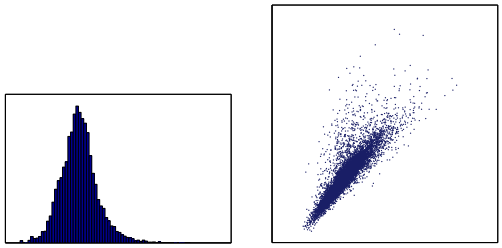


Figure 2: Comparative results for several methods applied to short extracts from two piano pieces. The tick marks on the axes indicate the actual onset times. The trace labelled 'energy' was computed as $S[k] = \frac{1}{2}\|\mathbf{x}_2[k]\|^2$ and is plotted on a logarithmic scale. The trace labelled 'PCA' (for principal component analysis) was generated by fi tting a mutivariate Gaussian model to the vectors $\mathbf{x}_2[k]$ and is also on a logarithmic scale.



(a) (b)

Figure 3: Distribution of surprise signal in 1 and 2-dimensional embedding spaces.

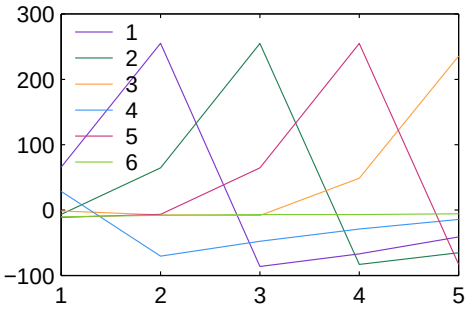


Figure 4: These are the mean signal shapes emitted by each state of the hidden Markov model.

A histogram showing the distribtion of points in a 1-dimensional embedding space, illustrated in fig. 2(a), shows no obvious clustering and therefore no reliable way to detect onsets. The 2-dimensional distribution shown as a scatter plot in fig. 2(b) begins to show some structure, with onsets tending to lie in the diffuse area above the diagonal. In higher dimensions, the clustering becomes more distinct.

Additional structure appears in higher dimensions: since the coordinates are taken from a sliding window on the reduced signal, the trajectory of points in the embedding space is highly constrained. If one imagines a separate cluster for each position within the window an onset may occur, then as an onset passes through the window, its associated point will jump from cluster to cluster in a regular way.

We have initially experimented with hidden Markov models (HMMs) with Gaussian states to capture both the clustering and the sequencing we expect to see in the embedding space. The model illustrated here has 6 states and was trained on a 5-dimensional embedding space. The means of the 6 resulting clusters are shown in fig. 3, and clearly represent an onset in one of 5 positions, or a non-onset. The associated cluster variances are shown in fig. 4 as the set of eigenvectors of each covariance matrix.

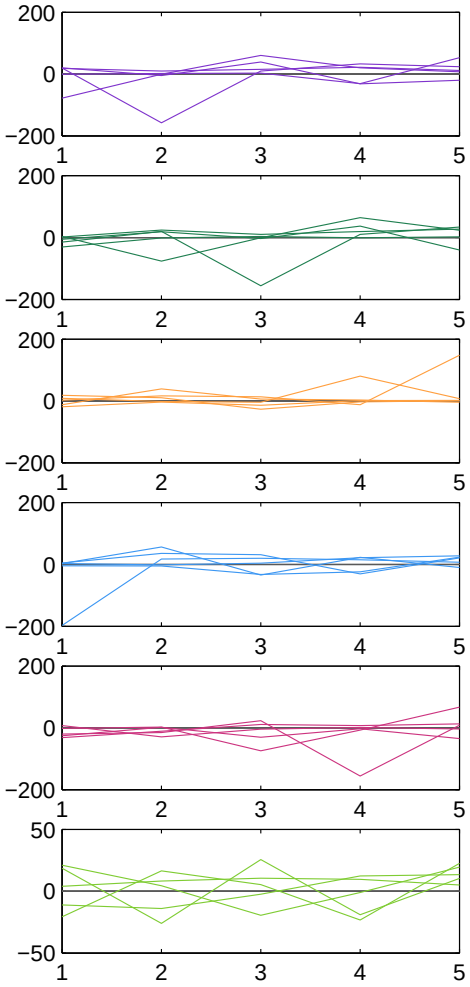


Figure 5: Eigenvectors of state covariance matrices. These model the variability in the observations produced from each state in the hidden Markov model.

The state-to-state transition probabilitiy matrix is shown in fig. 5 and confirms our earlier intuitions about the typical sequence of states the model passes through as an onset passes through the signal window. At each detected onset, the hidden state variables goes through the states 3,5,2,1 and 4 before returning to state 6. Some of the transitions, however, are less reliable than others, with states 3 and 4 having a significant probability of lingering for more than one time step. Hence, it would be best to use one of states 2, 1, 4 as an onset trigger.

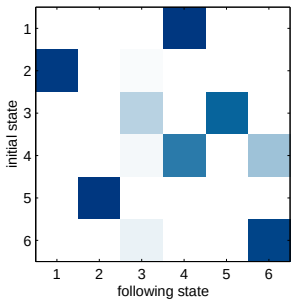


Figure 6: Markov Transition matrix The typical sequence of states for each onset is $6 \rightarrow 3 \rightarrow 5 \rightarrow 2 \rightarrow 1 \rightarrow 4 \rightarrow 6$, with states 3 and 4 occasionally persisting for more than one time step.

The sequence of states (obtained by Viterbi decoding) for the Debussy extract used in fig. 1(b) is shown in fig. 6. In this example, 42 out of 43 onsets are correctly detected, but there are 3 spurious onsets.

4 Conclusions

These results show that it is feasible to build a an onset detection system in an almost completely unsupervised way, with only the final classification of clusters (as either onset or not onset) requiring any external input. The formulation in terms of probabilistic models in both stages of the process means that different models can objectively be compared in terms of how well they fit the data. Improving the performance is then a question of finding and training the appropriate models.

One possibility that we have been investigating is to use a HMM with more states, to improve the fit of the HMM as a density model, and add a further stage of unsupervised analysis to group the states into event and non-event clusters.

References

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

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

[3] J.-F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 44(12):3017–30, Dec. 1996.

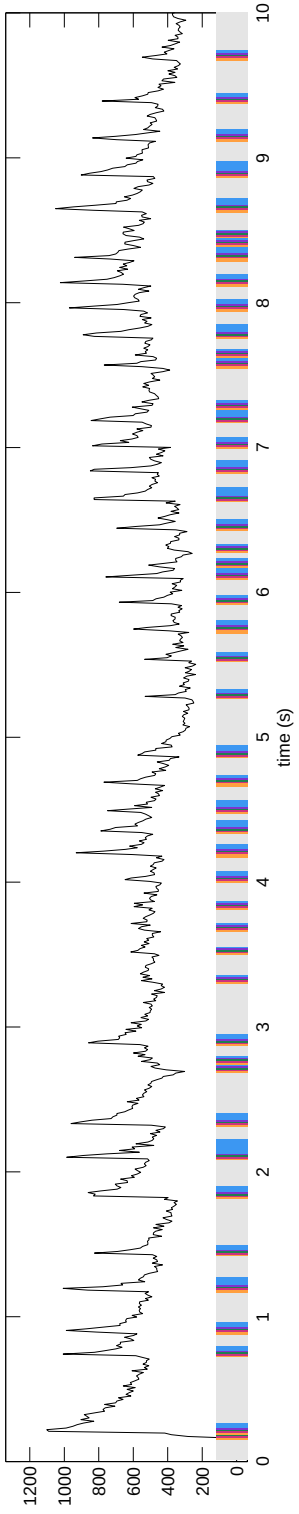


Figure 7: The original reduced signal and the inferred most likely sequence of hidden states.