

Unsupervised Learning for Music Perception

Samer Abdallah* and Mark Plumbley†
Department of Electronic Engineering,
King's College London.

September 20, 1999

Abstract

Perception and cognition can be considered to be processes aimed at discovering the *independent causes* behind sensory input. Thus, the goal of perception might be to achieve a *factorial coding*.

Unsupervised learning with neural networks offers a way to implement this using techniques such as *sparse coding*. This would result in a representation in terms of an optimal set of features, rather than the heuristically guided selection often used now. The *Wigner Distribution* is a good choice of input to these algorithms for a number of reasons.

The principle of factorial coding, applied consistently, could result in natural representations for musical constructs such as melodic phrases or rhythmic motives; something which has obvious applications in music processing.

1 Introduction

The problem of music cognition is currently being attacked from a number of directions. One approach grows out of work being done on auditory scene analysis [6, 23], starting with an audio signal and modelling, at some level, the human auditory system. This has led to various biologically inspired representations of sound (such as cochleagrams and correlograms), and rules for interpreting these based on psychoacoustic and psychological evidence [20, 24, 10].

Another approach begins with the music already transcribed in some form, and attempts to model the higher levels of music cognition. This in-

cludes activities such as segmentation, detection of melodic or rhythmic similarities between phrases, analysis of harmony and tonality, and induction of metrical structure [7].

Adaptive neural techniques have been applied to various representations, dealing with percepts such as pitch, timbre, tonality, and genre [17, 15, 26].

What all these methods have in common is that the earliest stages of processing encode the input in some predetermined way. The different representations all embody choices about which features are to be important in later on. Whilst many of these choices may seem reasonable, we must recognise that they are not trivial and have a marked influence on the success or failure of any further processing.

Anagnostopoulou *et al* [1] investigated this in the case of melodic representations for pattern matching. There is a wide variety, from absolute semitone scales, to hierarchical elaboration trees. The decisions between them are often made heuristically, and contain a lot of 'predigested' musical knowledge, that is, they are based in part on intuitions of what makes sense musically in a given context. The ever-present risk in these situations is that something has been missed; perhaps a less obvious feature has not been given the prominence it deserves.

One suspects that what tends to happen is that various representations are tried until the algorithm or neural network performs as expected. To then claim that the network has 'discovered' this or that structure is to deny the importance of the representation: most of the work has been done already! The question is, can we build a system powerful enough to do that work by itself?

Perhaps we can gain something by looking at perception in general. Attempts to model perceptual

*e-mail: samer.abdallah@kcl.ac.uk

†e-mail: Mark.Plumbley@kcl.ac.uk

systems have given birth to the concept of *unsupervised learning* [5], and has resulted in a promising approach to perception with a sound basis in information theory. This has already produced elegant justifications for many features of early vision [3, 12], but has yet to be fully explored in the context of music perception.

2 First Principles

Perceptual systems in animals exist for one reason: to improve their owners' chances of survival. This is a fair assumption from an evolutionary viewpoint, but was suspected even before the advent of Darwinism. For example, in 1690 John Locke ¹ wrote,

Perception, I believe, is, in some degree, in all sorts of animals; though in some possibly the avenues provided by nature for the reception of sensations are so few, and the perceptions they are received with so obscure and dull, that it comes extremely short of the quickness and variety of sensation which is in other animals; but yet it is sufficient for, and wisely adapted to, the state and condition of that sort of animals who are thus made . . .

The need for an *ecological* theory of perception, that is, one that recognises the importance of and influence of the natural environment, has also been recognised more recently [2, 22], and also specifically in relation to music [17].

It is reasonable, therefore, to suppose that the auditory systems of each species are adapted to the kind of aural environment present during its evolution; furthermore, only those aspects of the environment that are *relevant* to the animal, such as the sounds made by predators and prey, or vocalisations of members of the same species. For example, these may occupy a certain range of frequencies and intensities; hence the sensitivity of the ear to those ranges and not others.

Information Theory Shannon's Information Theory [25] provides a natural mathematical framework for understanding perception, allowing us to quantify the information present in a signal about processes of interest.

¹in *An Essay Concerning Human Understanding*, Book II, Ch. IX, §2

Attneave [4] suggested that, because sensory signals are highly correlated with each other, one possible strategy for a perceptual system is that of *redundancy reduction*. Barlow [5] proposed an information theoretic measure of redundancy that compares the amount of useful information in the code with the theoretical channel capacity.

Linsker's principle of information-maximisation [18, 19] says that we should try to maximise the amount of useful information in the code under various physical and biological constraints.

Atick and Redlich [3] highlight the difference between two approaches to redundancy reduction: we can either increase the information present in a channel (*i.e.* Linsker's method) *or* reduce the channel capacity, thus eliminating the 'slack'. They take the view that any excess capacity has an immediate biological cost, (not least in terms of energy) so that the evolutionary pressure will be to minimise this while maintaining a certain minimum of information transmission consistent with the needs of the animal.

Objects and Features Coding efficiency is not the only motive for redundancy reduction: perception is often framed in terms of object or feature recognition, which entities tend to manifest themselves as collections of highly correlated properties, whilst they themselves tend to remain relatively independent.

This principle could account for many of the features or percepts we attribute to sounds. For example, a musical note has a certain kind of redundancy due to it's being roughly periodic. A frequency analysis would reveal a set of harmonically related components, which tend to begin and end together and also suffer the same amplitude and frequency modulations (see figure 4). These features, being highly correlated, are thus bound together into a single percept and we are not usually aware of their separate existences. Of course, we already know about this and many other grouping rules from auditory scene analysis, but this gives us both a justification for these and a general principle for finding more, and perhaps more effective ones.

Factorial Coding The preceding discussion hints at the importance of *factorial coding*, that is, a description in terms of statistically independent

features: not only is this efficient, it may also reflect something of the structure of the world. In addition, Barlow suggested that factorial coding might help associative learning: neural models of associative memory are known to perform better with uncorrelated patterns. This raises the issue of biological viability: factorial representations may allow the brain to perform more complex, effective, and *intelligent* mappings between sensory input and motor activity.

Causality and Understanding Why is it that objects defined by correlated feature bundles seem to reflect useful structures in the world? The answer may be to do with *causation*. Perception and cognition might have as their goal the inference of the causal structure behind sense data: the objects and entities that we ultimately perceive are in some sense an *explanation* for the signals that eventually reach our sense organs.

These distinct causes can be assumed to be independent, because any dependent phenomena—‘suspicious coincidences’—invite further explanation in terms of underlying causes. If we equate objects with causes, then searching for objects by redundancy reduction automatically leads us to a description in terms of independent entities, which are good candidates for independent causes.

To give a flavour of how this might work in music, figure 1 shows a possible set of links between musical ‘causes’ (though in this case, the terminology seems a little strained; we might choose to call them ‘factors’ instead).

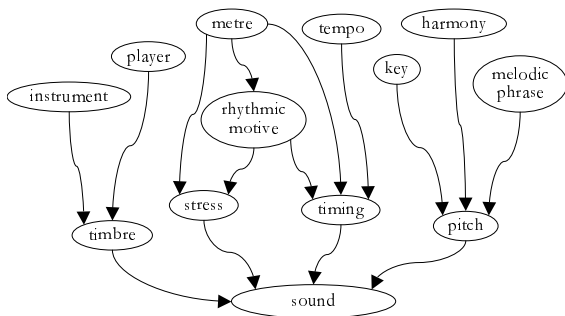


Figure 1: Causality diagram for music. To see how a rhythmic or melodic phrase can be said to ‘cause’ a section of music, one has only to think of the recurring four note pattern in Beethoven’s 5th symphony.

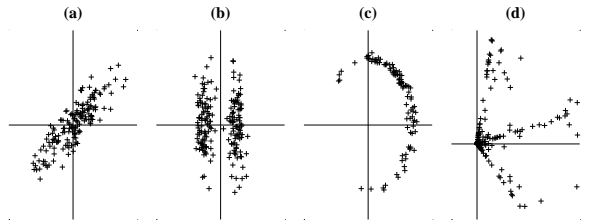


Figure 2: Some data distributions showing various kinds of structure.

3 Unsupervised Learning

Broadly speaking, unsupervised learning [5] addresses the following task: given a set of data, what can usefully be said about its distribution? What are its ‘interesting’ features? The learning process generally involves fitting some explicit or implicit stochastic model to the observed data.

The usefulness of an unsupervised learning algorithm in a given problem is largely determined by the probabilistic model it attempts to fit to the data and the assumptions that it involves. We can illustrate this and the limitations thereby imposed in two of the most commonly used forms of unsupervised learning: Principal Component Analysis and the Self-Organising Map, though we hasten to add that there are other methods that address some of these issues and are not limited in the same way.

PCA [21] works by finding those directions in the input space which account for most of the variance, and is often used for dimensionality reduction. The underlying model is essentially this: a gaussian data distribution (figure 2(a) with added spherical gaussian noise (that is, of the same variance in all directions)). PCA is not invariant to rescaling of the axes. If the axes are measured in different units, it may not be obvious what scaling to use. The ambiguity is resolved by the requirement that the noise have the same variance along each axis. It is this that ensures that we have the most to gain from high-variance components, because they have the highest signal-to-noise ratio.

The SOM [16] attempts to fit a low-dimensional lattice to the data distribution, assuming it to occupy some sub-manifold of the input space, as in figure 2(c). Thus, it is capable of non-linear dimensionality reduction. However, the use of a mean-square error measure in the derivation implies once again the assumption of spherical gaussian noise,

and a lack of invariance to scaling. Hence, the map that is eventually produced is heavily dependent on the representation chosen for the input.

The preceding discussion goes a long way to explaining why pre-processing is required in order to get sensible answers out of these methods: it is needed to squeeze the data distribution into a form more closely approximating that expected by the model implicit in the method. By generalising the models, we relax our assumptions about the data and obviate the need for preprocessing.

Sparse Coding A sparse code [11] is a one in which only a ‘few’ elements are significantly active at a time. In searching for a *sparse* factorial code for natural phenomena, we are saying something about the structure of the world: loosely speaking, of the very many possible (and largely independent) ‘things’ in the world, only a few of them are sensible at a time. This applies to all the senses because of their limited range: in audition, distant sources produce softer sounds, because of both obstructions and distance; one might call the effect ‘auditory perspective’.

Figure 2(d) illustrates a sparse distribution. By assuming this sort of model, we get a form of unsupervised learning known as sparse coding. The technique has been applied to vision, with some interesting results: Field and Olhausen [12] trained their neural network on a set of natural images. The result closely reproduced the receptive fields of the so-called simple cells in the visual cortex. Sparse coding also facilitates the rejection of certain kinds of noise, and Hyvärinen [14] developed an effective method for image de-noising based on this.

4 Time-Frequency Representations

There are a number of time-frequency representations that might serve as input to a neural network, the most familiar of these probably being the spectrogram. The Wigner-Ville Distribution (WD) [8], provides a unifying framework for considering many of them. The distribution $W(t, \omega)$ is defined in terms of an instantaneous auto-correlation func-

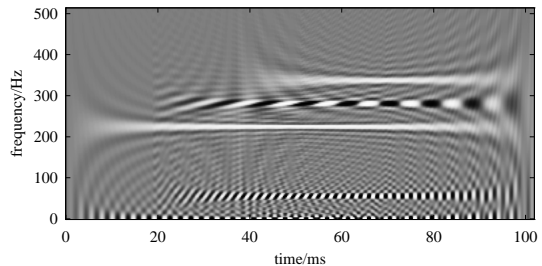


Figure 3: Wigner Distribution of two sinusoids illustrating the *interference* between them, in which the WD can take on negative values. This are generally held to be undesirable, but it does represent a perfectly valid interpretation of the signal: that of an amplitude modulated sinusoid of intermediate frequency (*i.e.* beats). The colour scale goes from black (-ve), to gray (0), to white (+ve).

tion $R(t, \tau)$:

$$R(t, \tau) = f(t + \frac{1}{2}\tau)f^*(t - \frac{1}{2}\tau), \quad (1)$$

$$W(t, \omega) = \int_{-\infty}^{\infty} R(t, \tau)e^{-i\omega\tau} d\tau. \quad (2)$$

The WD is *bi-linear* like a spectrogram, rather than linear like a Short Term Fourier Transform or Wavelet Transform. It is, in principle, invertible, unlike a spectrogram, where the loss of phase information is irreversible. Neither does it force a trade-off between time and frequency resolutions.

A spectrogram can be obtained from the WD by a two-dimensional convolution, *i.e.* by *blurring* it with a particular kernel. This has the effect of removing much of the interference (see figure 3) and ensuring that the result is entirely non-negative, but at the expense of a loss of resolution. A wavelet-like constant-Q transform can be obtained by smoothing with a position dependent kernel, which is tall and narrow at high frequencies, and broad and short at low; this is called *affine smoothing* [13]. A broad class of time-frequency distributions, known as *Cohen’s Class* [8], can be obtained by using a general convolution.

As an interesting aside, it can be shown that something very much like the auditory correlogram [9] can be obtained from the WD by means of a linear transformation equivalent to a pre-multiplication before affine smoothing. It has the effect of preserving information that would otherwise be lost by the smoothing process.

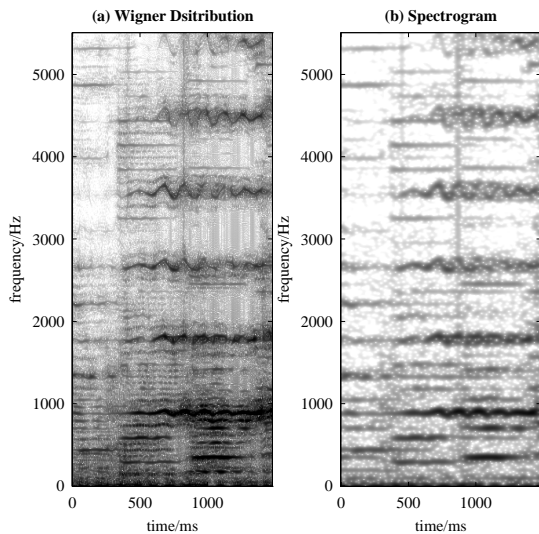


Figure 4: Comparison of two time-frequency distributions: a snippet of Bach’s 2nd *Brandenburg Concerto*, containing harpsichord, flute and violin.

5 Proposal

The simpler forms of unsupervised learning such as PCA or the SOM are not capable of extracting high level musical information unaided. This does not mean, however, that more sophisticated forms cannot succeed. We propose to bring these methods to bear on the problem of music perception in searching for the independent causes behind musical signals. This is justified by the hypothesis that our own perceptual systems operate along the lines outlined in section 2: if this is indeed true, then we should be able to achieve the same result by whatever means are available to us. We should be able to infer the presence of the same auditory ‘objects’ or ‘causes’ that a human observer would detect, even though the intermediate steps may be different.

Part of the motivation for this approach is a desire to remove any significant pre-processing and the free parameters that it entails, such as filter bandwidths or adaptation rates. The Wigner Distribution is a good candidate starting point: it has essentially no free parameters, and out of all the bilinear time-frequency distributions, it has the best resolution and can generate the others via a linear transformation; hence, anything a neural network could do with a spectrogram or a correlogram, it

can also do with the WD.

Sparse coding is an attractive method for initial experimentation: it has a simple neural implementation, and as had good results in vision. However, more powerful techniques will eventually be required, relaxing, for example, the assumption of spherical gaussian noise.

6 Discussion

A popular approach has been to take as a starting point a more or less faithful model of the early stages of the human auditory system. Why should we do otherwise? From a computational point of view, we would like to reproduce the effect of this processing in the most efficient way possible. For example, simulating the firing of individual neurons is a laborious process that we would like to avoid if possible. In the words of Lyon and Mead [20], “Nerve impulses are by their very nature a horrible medium into which to translate sound. They are noisy and erratic, and can work over only a limited dynamic range of firing rates.” The question is, how important are the particular details of auditory transduction to later processing, and do we need to replicate the mechanisms of the human ear in order to model human-like hearing?

If auditory percepts *can* be derived from redundancy reduction, or a search for independent causes, then we should not need to start from an auditory model: nature has had to *evolve* those particular mechanisms because of the limitations of the hardware available to it, be we do not. We are free to optimise the whole lot. Using our technology, the optimal configuration for the early stages need not match the human auditory system.

The optimisation could conceivably reproduce some features of our own auditory system, in the same way that similar techniques have reproduced properties of the visual system. This would provide a theoretical explanation for those features, showing that they are genuine adaptations and not just evolutionary accidents. For example, the non-linear compression and dynamic adaptation in the ear could be an informationally-optimal response to the statistics of loudness in natural sounds.

To conclude, the ideal of redundancy reduction provides a constant and elegant organising principle, which promises to provide a unifying descrip-

tion of all levels of music cognition from the raw audio signal to abstract musical concepts. If this promise is indeed fulfilled, we will be able to design artificial auditory systems that meet the same ends as biological ones, but are better adapted to the technology used to implement them.

References

- [1] C. Anagnostopoulou, D. Hörnel, and K. Höthker. Investigating the influence of representations and algorithms in music classification. In *Proceedings of the AISM'99 Symposium on Music Creativity*, pages 35–41, 1999.
- [2] J. J. Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, 3(2):213–251, 1992.
- [3] J. J. Atick and A. N. Redlich. Towards a theory of early visual processing. *Neural Computation*, 2(3):308–320, 1990.
- [4] F. Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193, 1954.
- [5] H. B. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.
- [6] A. S. Bregman. *Auditory Scene Analysis*. MIT Press, 1990.
- [7] E. Cambouropoulos, T. Crawford, and C. S. Iliopoulos. Pattern processing in melodic sequences: Challenges, caveats and prospects. In *Proceedings of the AISM'99 Symposium on Music Creativity*, pages 42–47, 1999.
- [8] L. Cohen. Time-frequency distributions—A review. *Proc. IEEE*, 77(7):941–981, 1989.
- [9] D. P. W. Ellis and D. F. Rosenthal. Mid-level representations for computational auditory scene analysis. In *Proc. Intl. Joint Conf. on Artif. Intell. Workshop on Computational Auditory Scene Analysis, Montreal*, 1995.
- [10] D. P. W. Ellis and B. L. Vercoe. A perceptual representation of sound for auditory signal separation. Presented at the 123rd meeting of the Acoustical Society of America, Salt Lake City, May 1992.
- [11] D. J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.
- [12] D. J. Field and B. A. Olhausen. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.
- [13] P. Flandrin and O. Rioul. Affine smoothing of the wigner distribution. In *ICASSP'90*, pages 2455–2458, 1990.
- [14] A. Hyvärinen. Denoising of sensory data by maximum likelihood estimation of sparse components. *Neural Computation*, 11(7):1739–1768, 1999.
- [15] T. Järvinen, P. Toivainen, and J. Louthivuori. Classification and categorization of musical styles with statistical analysis and self-organizing maps. In *Proceedings of the AISM'99 Symposium on Music Creativity*, pages 54–57, 1999.
- [16] T. Kohonen. *Self-Organizing Maps*. Springer, 1995.
- [17] M. Leman and F. Carreras. The self-organization of stable perceptual maps in a realistic musical environment. In *Proc. Journée d'Informatique Musicale*, Caen, May 1996.
- [18] R. Linsker. Toward an organising principle for a layered perceptual network. In Anderson, editor, *Neural Information Processing Systems*, pages 485–494. American Institute of Physics, 1988.
- [19] R. Linsker. An application of the principle of maximum information preservation to linear systems. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 1, pages 186–194. Morgan-Kaufman, 1989.
- [20] R. F. Lyon and C. Mead. An analog electronic cochlea. *IEEE Transactions on Acoustics Speech and Signal Processing*, 36(7):1119–1144, 1988.
- [21] E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5:927–935, 1992.
- [22] B. A. Olhausen and D. J. Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(2):333–339, 1996.
- [23] E. D. Scheirer. Bregman's chimerae: Music perception as auditory scene analysis. In *Proc. Int. Conf. on Music Perception and Cognition*, Montreal, 1996.
- [24] S. Seneff. A computational model for the peripheral auditory system: Application to speech recognition research. In *Intl. Conf. on Acoustics, Speech and Signal Processing*, pages 37.8.1–4, 1986.
- [25] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–623–, 1948.
- [26] J.-P. Thouard, P. Depalle, and X. Rodet. Pitch classification of musical notes using Kohonen's self organising feature map. In *Proc Intl. Neural Network Conf.*, volume 1, 1990.