IF THE INDEPENDENT COMPONENTS OF NATURAL IMAGES ARE EDGES, WHAT ARE THE INDEPENDENT COMPONENTS OF NATURAL SOUNDS?

Samer A. Abdallah[*] *and Mark D. Plumbley*[†]

Department of Electronic Engineering
King's College London

ABSTRACT

Previous work has shown that various flavours of Independent Component Analysis, when applied to natural images, all result in broadly similar localised, oriented band-pass feature detectors, which have been likened to wavelets or edge detectors.

In this paper, we present a similar analysis of 'natural' sounds drawn from two radio stations: one broadcasting mainly speech; the other mainly classical music. Many of the resulting basis vectors are quite wavelet-like, and can easily be characterised in terms of their position and spread in the time-frequency plane. Some of them, however, particularly from the set trained on music, do not fit that interpretation very well. The Wigner-Ville Distribution can be used to gain a clearer picture of time-frequency localisation of these basis vectors.

We conclude by suggesting that these results be compared with other widely used auditory representations such as short-term Fourier transforms, wavelet transforms, and physiologically derived models based on the auditory filterbank.

1. REDUNDANCY REDUCTION AS A GOAL OF PERCEPTION

It has been suggested [2, 1, 7] that the processing of sensory data in biological perceptual systems is best understood in the language of information theory. The wealth of structure present in natural phenomena means that sensory signals are highly redundant; characterising this structure in order to develop efficient, non-redundant representations might be an effective processing strategy. In a distributed code, a major source of redundancy is statistical dependency between units; independent or factorial coding will be an important tool in dealing with this.

ICA is Reduncany Reduction via Linear Transformation. If we restrict ourselves to instantaneous linear methods, then the best we can do is aim for a matrix operation that results in a vector whose elements are as independent as possible—that is, precisely the ICA problem. If the observed data is represented as an N-element vector $\mathbf{x}$, then we wish to find the linear transformation

$$\mathbf{y} = \mathbf{W}\mathbf{x}, \tag{1}$$

that minimises the mutual dependency between the elements of $\mathbf{y}$. In this paper, we further restrict ourselves to the case in which the weight matrix $\mathbf{W}$ is square.

After learning, the structure of the input is usually best revealed by examining the linear basis which will reconstruct the input $\mathbf{x}$ from the coded version $\mathbf{y}$. These basis vectors are given by the columns of $\mathbf{A} = \mathbf{W}^{-1}$.

2. ICA OF NATURAL SCENES

Field and Olshausen [8] showed that sparse coding (a technique closely related to ICA) of natural images results in a decomposition of the image in to a set features localised both spatially and in spatial frequency (that is, oriented and band-pass). Other similar experiments with ICA [4, 9] have produced comparable results.

These features can be interpreted as wavelets or edge detectors, providing some justification for the use of those methods as image processing tools, and suggesting that the learned features may actually be more appropriate in certain situations. They also show some correspondence with receptive fields of simple cells in visual cortex (V1), providing a possible explanation for operation of these cells.

There has been (to the authors' knowledge) no comparable study of natural sounds. Bell and Sejnowski [3] used ICA on sound, but this was limited to a very particular toothtapped rendition of Mozart's *Für Elise*. Casey [5] also used ICA, but again, learning was restricted to one auditory event at a time, not a long exposure to a representative selection of sounds, what one might call an auditory *environment*.

[*] e-mail: samer.abdallah@kcl.ac.uk
[†] e-mail: mark.plumbley@kcl.ac.uk

3. EXPERIMENTS WITH SPEECH AND MUSIC

We ran an ICA algorithm on several days' worth of largely unbroken radio output from two stations: BBC Radio 3, broadcasting mainly classical music but with some speech and other music; and BBC Radio 4, which outputs mainly speech. The input was presented as 512-sample blocks of waveform data; the analysis produced two sets of 512 basis vectors, one for Radio 3 and one for Radio 4, which we examine in the next section.

The signal was pre-processed to compensate for any DC offset, and to (very roughly) normalise its amplitude. This was *not* done on each block independently, which would result in every input vector being of zero mean and unit variance, but rather by maintaining a slowly varying average of block means and variances, and using these to normalise each block. The idea was to tame the worst excesses of loudness variation over medium to long time scales, not to remove all dynamics. Letting $w_i[k]$ be the i'th sample in the k'th block of raw waveform data, the following steps were used to compute the elements $x_i[k]$ of the k'th input vector $\mathbf{x}[k]$:

$$\hat{\mu}[k] = \frac{1}{N}\sum_{i=1}^{N} w_i[k], \tag{2}$$

$$\hat{\sigma}[k] = \left\{ \frac{1}{N}\sum_{i=1}^{N} (w_i[k])^2 \right\}^{1/2}, \tag{3}$$

$$\mu[k] = \alpha_\mu \hat{\mu}[k] + (1-\alpha_\mu)\mu[k-1], \tag{4}$$

$$\sigma[k] = \alpha_\sigma \hat{\sigma}[k] + (1-\alpha_\mu)\sigma[k-1], \tag{5}$$

$$x_i[k] = (w_i[k] - \mu[k])/\sigma[k]. \tag{6}$$

The adaptation rates α_μ and α_σ were set to around 0.01, which, given a sampling rate of 11.025 kHz and blocks of 512 samples, implies a time-constant of about 5 seconds.

The version of ICA used was MacKay's covariant maximum-likelihood algorithm [10], which assumes a known prior distribution $p(y)$ for the independent components (*i.e.* the elements of $\mathbf{y}$). The weight update rule is

$$\Delta\mathbf{W} = \eta[\mathbf{I} - \mathbf{g}\mathbf{y}^T]\mathbf{W}, \tag{7}$$

where η is a learning rate parameter, and $\mathbf{g}$ is an element-wise nonlinear function of $\mathbf{y}$ given by

$$g_i = -\frac{d}{dy}\log p(y_i), \tag{8}$$

We experimented with both a Laplacian prior—$p(y) = e^{-|y|}$—and a Cauchy prior—$p(y) = (1+y^2)^{-1}$—producing broadly similar results; due to lack of space, we present only the Cauchy-derived results here.

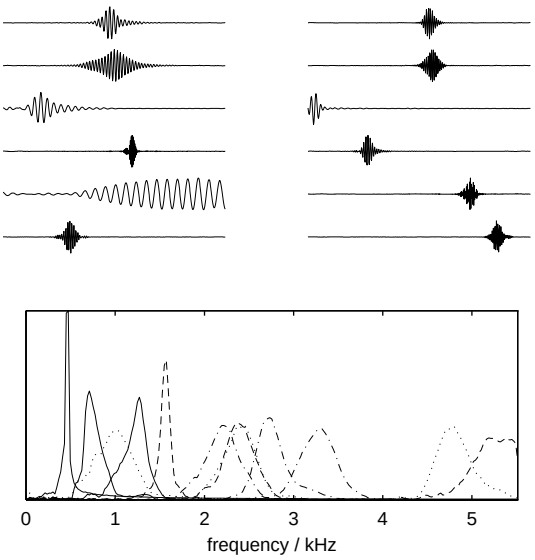

Fig. 1. A few of the Radio 4 basis vectors. The lower plot shows Fourier magnitude spectra of the (time domain) vectors in the upper plot.

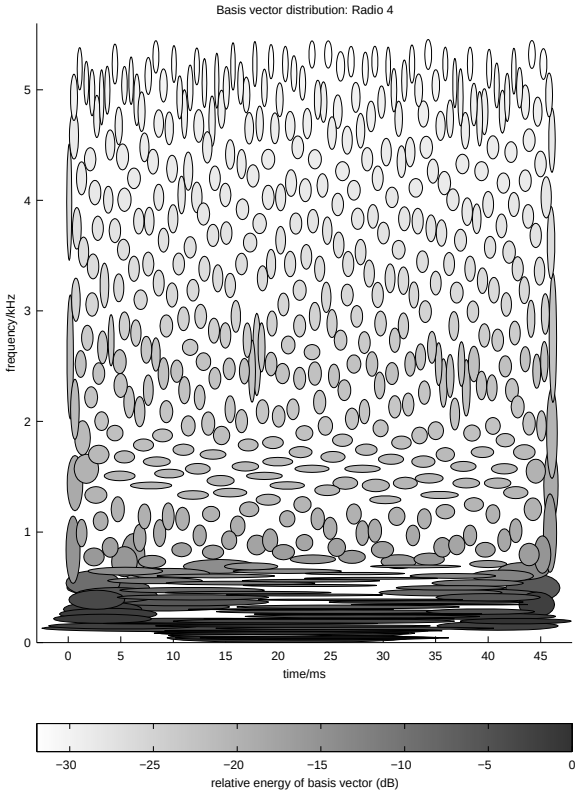

Fig. 2. Postion and spread in time and frequency of all 512 Radio 4 basis vectors. The grey scale encodes the overall energy of each basis vector relative to the one with the highest energy.

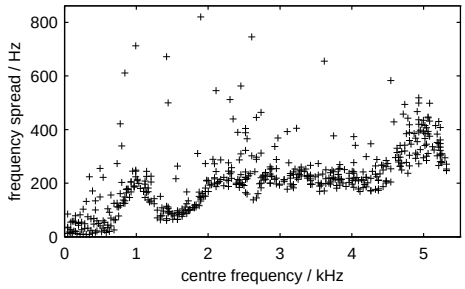

Fig. 3. Frequency domain bandwidth v. centre frequency for Radio 4 basis vectors.

4. PRELIMINARY ANALYSIS OF RESULTING BASES

We start by looking at the (speech dominated) Radio 4 basis as the results are easier to interpret than the Radio 3 basis.

4.1. Radio 4 Basis

The basis can be visualised in various ways: in the time domain, in the frequency domain, or in the time-frequency plane. Figure 1 illustrates time domain plots of some of the Radio 4 basis vectors, and their corresponding magnitude spectra. They are, on the whole, quite well localised in both domains, suggesting that it might be useful to characterise them by their position and spread in the time and frequency plane. This we did by squaring each element of a basis vector or its fourier transform and treating the resulting vector like a probability distribution, measuring the median and mean absolute deviation from that median [1].

Figure 2 is a combined plot of all 512 basis vectors—each one is represented by an ellipse indicating its position and spread in time and frequency. Figure 3 illustrates more clearly the relationship between centre frequency and bandwidth.

Several observations can be made from the plots:

- The basis vectors are fairly evenly distributed in time and frequency.
- The spectral widths are not exactly proportional to the centre frequencies, but there is a general increase bandwidths at higher frequnecies. The very lowest frequencies are not localised in time at all.
- There are edge effects, with short-time, wide-band features at the beginning and end of the block.

[1] This produced better results than the more obvious procedure of measuring means and variances, because the 'distributions' did not fit a Gaussian model especially well. In particular, the variance consistently overestimated the spread of the distributions as judged by eye.

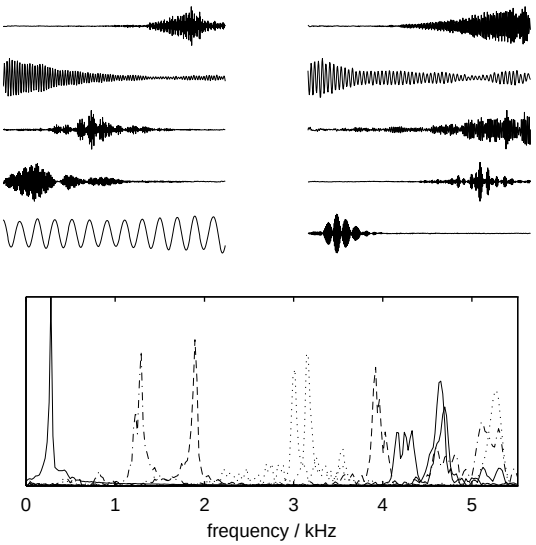

Fig. 4. Some of the Radio 3 basis vectors and their magnitude spectra.

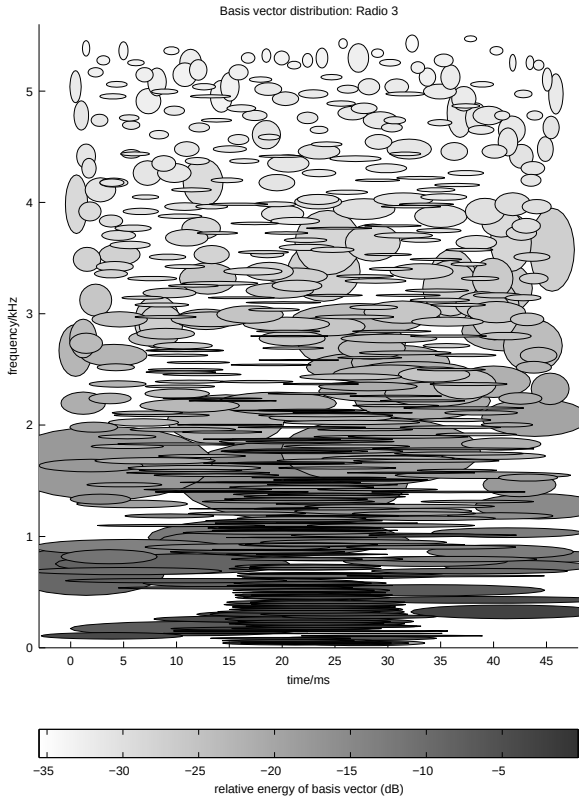

Fig. 5. Postion and spread in time and frequency of all 512 Radio 3 basis vectors.

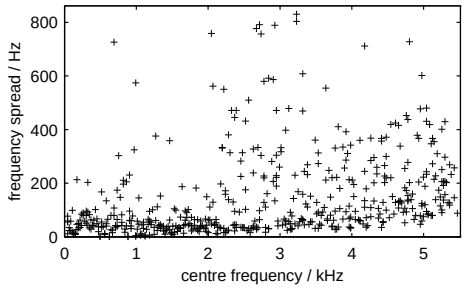

Fig. 6. Frequency domain bandwidth v. centre frequency for Radio 3 basis vectors.

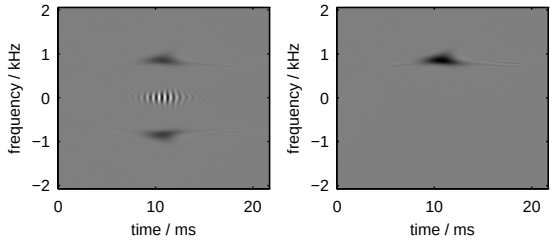

Fig. 7. Wigner Distribution of one of the Radio 4 basis vectors (left). The right-hand image is the WD of the Hilbert transform of the basis vector. The Hilbert transform filters out the negative frequency components and with them the oscilliatory cross term in the middle. (The middle grey represents zero, going to black for more positive values.)

- There is a reversal of the bandwidth trend between 1 and 1.5kHz. There are also a few anomalous features at around 2.5kHz near 17ms and 37ms.

It is not clear what significance the bandwidth behaviour between 1 and 2kHz has, though it may be an adaptation to the formant structure of speech.

4.2. Radio 3 Basis

When we examine the Radio 3 (music derived) basis in the same way, the interpretation is not so clear. In figure 5, there appear to be many poorly localised basis vectors. Referring to figure 4, we can see that some of the vectors have multimodal spectra, and a 'pulsating' envelope in the time domain. The centre/spread model is not appropriate for the description of these forms.

5. FURTHER ANALYSIS USING THE WIGNER DISTRIBUTION

In an effort to gain a clearer picture of the time-frequency behaviour of the basis vectors, especially the poorly localised ones, we analysed them with the Wigner-Ville distribution (WD) [6], which is a type of time-frequency representation,

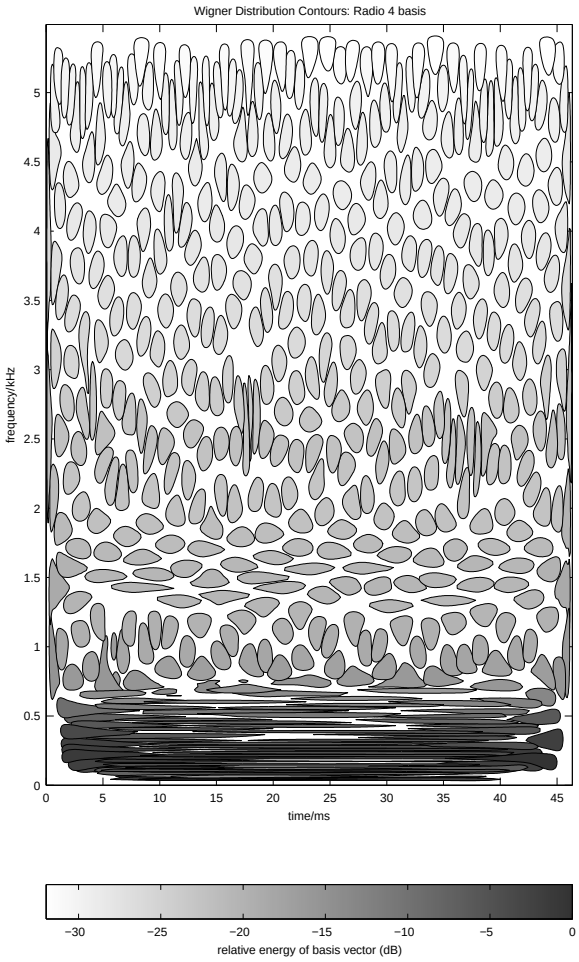

Fig. 8. Combined contour plots of all the Radio 4 basis vector Wigner distributions. Each one is represented by a contour at 0.6 times its peak value. The grey scale represents the total energy of each basis vector—not the value of the Wigner distribution.

similar to a spectrogram, but without the time and frequency resolution limits of a spectrogram. It is defined for a continuous time complex-valued signal $x(t)$ as

$$W(t,\omega) = \int_{-\infty}^{+\infty} x(t+\tfrac{1}{2}\tau)x^*(t-\tfrac{1}{2}\tau)e^{-2i\omega\tau}dt \qquad (9)$$

There are some subtleties involved in defining the discrete time analogue of this; suffice it to say that what we actually used was a Type II quasi-Wigner distribution as defined in [11].

Being essentially a quadratic function of the signal, the WD of the sum of two signals is *not* equal to the sum of the individual WDs. So-called *cross-terms* appear halfway between the individual WDs (also called *auto-terms*). These cross-terms are often larger then the auto-terms, and generally a distraction as far as visualisation goes—this is the

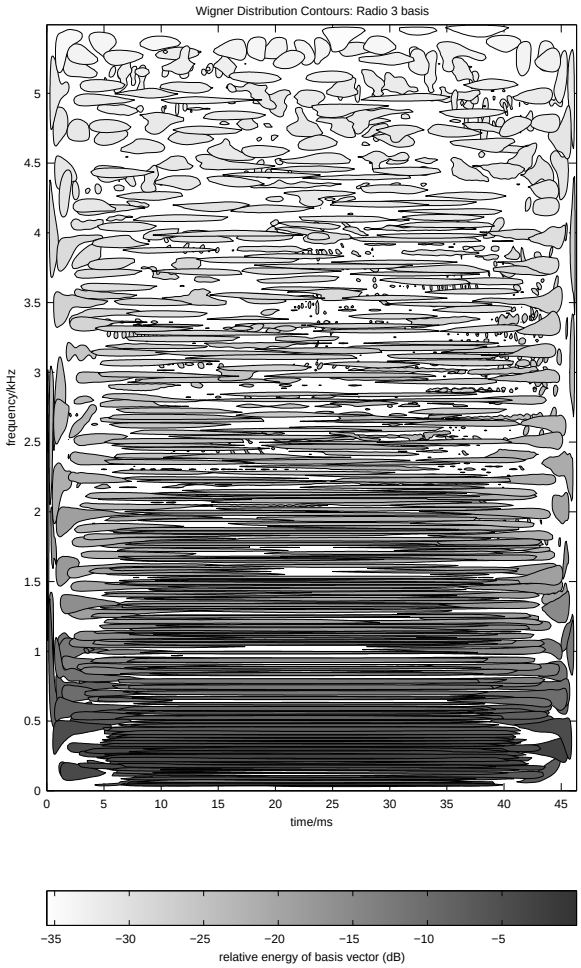

Fig. 9. Contour plots of the Radio 3 basis vector Wigner distributions, produced the same way as in figure 8.

price we pay for the improved resolution of the WD. The usual approach to removing them is some sort of smoothing in the time-frequency plane, which inevitably leads to some loss of resolution. However, in the case of real-valued, wavelet-like signals, the main source of cross-terms is interference between the positive and negative frequency components. Thus, we removed the negative frequency parts using a discrete Hilbert transform. This proved to be quite effective for many of the basis vectors. (See figure 7.)

The Radio 4 basis vectors all produce well-localised Wigner Distributions, as illustrated in figure 8, which is a compilation of all 512 distributions, each represented by one contour. The Radio 3 vectors however, seem to fall in to three groups: many narrow-band features covering most of the width of the window (these are basically pure sinusoids); a number of compact wide-band features towards the top of the spectrum; and a few features with very fragmented Wigner distributions. These are the basis vectors which appeared as large ellipses in figure 5.

Looking at their spectra jointly with their Wigner Distributions (see figure 10), it is possible to discern that in many cases, the fragmentation is due to large cross-terms between the components of a multi-harmonic basis vector— the spectra of these features have two or more clear peaks. Some of these have components whos frequencies are in small integer ratios, which is what one would expect from the spectrum a low musical note with multiple harmonics or overtones. Since the cross-terms oscillate at a frequency equal to the *difference* between the frequencies of the components, they actually illustrate the frequency of the implied low note.

This still leaves a few basis vectors that defy explanation: some have multiple components which are not in small integer ratios, and some are so irregular that it leads us to suspect that the algorithm has either not converged properly, or has not learned an optimal solution, but fallen into a local minumum.

6. DISCUSSION AND CONCLUSIONS

This experiment has shown that ICA can learn interesting representations of audio signals that, under certain circumstances, correspond closely with a wavelet basis. In particular, the basis trained on speech shows a very clear and regular time-frequency structure. The features are well localised in time and frequency, with bandwidths that generally increase with the centre frequency.

Similar experiments with visual scenes have produced results the compare favourably with what is thought to occur in the early stages of the human visual system— the comparison ought to be made between these results and the human auditory system, in particular, the auditory filter-bank [12]. This will be the subect of further work, but a cursory investigation suggests that the bandwidths are a little too narrow. One possible explanation for this is that the continuous speech on which our system was trained may not be the sort of auditory environment for which the human auditory system is best adapted. We might get a closer match if we train on an environment including more non-speech sounds, such as mechanical noises, animal calls, rustling bushes etc. In this respect, the television may a better source of training data than radio!

Another interesting avenue of investigation into the speech derived basis is to discover how well adapted (if at all) it is to representing speech. Is the shape of the bandwidth *vs.* centre frequency plot (see figure 3) significant? Does it yeild a more efficient coding of speech than a wavelet basis, or other methods?

The Radio 3, music derived results are less conclusive. The basis did include many narrow-band sinusoids covering the whole width of the analysis window, suggesting that a

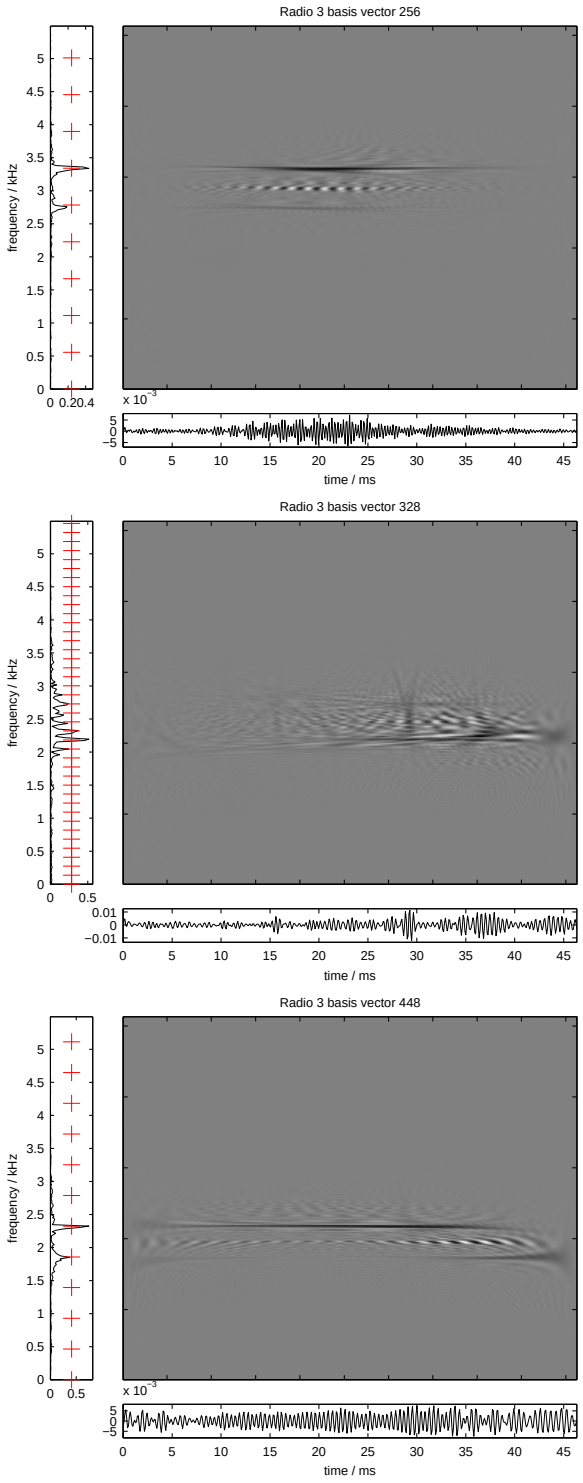

Fig. 10. Some of the multi-component Radio 3 basis vectors in time, frequency, and time-frequency. The crosses on the spectral plots are regularly spaced, showing how some of the components are in small integer ratios. Note that in bottom Wigner plot, the two components are consecutive rather than simultaneous, corresponding, in musical terms, to a drop of a Major Third.

Fourier basis is not wholey inappropriate for analysis music, at least at a time scale of 50ms. There were also some wavelet-like features at higher centre frequencies. However there are a significant number of features that are difficult to characterise. Though some consist of harmonically related components, others seem to be very irregular. We suspect that this may be the result of poor learning on the part of the ICA algorithm, especially when we consider the large amount of rather varied training data.

On a more encouraging note, preliminary tests involving listening to the basis vectors has revealed some interesting pitch structure, and we are currently investigating just how much musical structure has been encoded within them.

7. REFERENCES

[1] J. J. Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, 3(2):213–251, 1992.

[2] H. B. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.

[3] A. J. Bell and T. J. Sejnowski. Learning the higher-order structure of a natural sound. *Network*, 7(2), 1996.

[4] A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[5] M. A. Casey. *Auditory Group Theory with Applications to Statistical Basis Methods for Structured Audio*. PhD thesis, M.I.T., 1998.

[6] L. Cohen. Time-frequency distributions—A review. *Proc. IEEE*, 77(7):941–981, 1989.

[7] D. J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

[8] D. J. Field and B. A. Olshausen. Emergence of simple-cell receptive field properties by learing a sparse code for natural images. *Nature*, 381:607–609, 1996.

[9] A. Hyvärinen, E. Oja, P. Hoyer, and J. Hurri. Image feature extraction by sparse coding and Independent Component Analysis. In *Proc. Intl. Conf. on Pattern Recognition (ICPR'98)*, pages 1268–1273, Brisbane, Aus., 1998.

[10] D. J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis, 1996.

[11] J. C. O'Neill and W. J. Williams. Shift covariant time-frequency distributions of discrete signals. *IEEE Transactions on Signal Processing*, 47(1), 1999.

[12] R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice. SVOS final report: The auditory filterbank. Technical Report Apu 2341, Cambridge Electronic Design, 1988.